Bilan de l'existant en matière de prévision statistique des pics de pollution

Serge Guillas,
Ecole des Mines de Douai et Université Paris VI,

Noureddine Rhomari,
Université Paris VI,

Jiantong Zhang,
Université Paris VI

Décembre 2000

Table des matières

Résumé

Dans cette étude, une description des différentes techniques statistiques utilisées jusqu'à présent pour prévoir les pics de pollution est réalisée. Trois types d'approche sont explorées.

Tout d'abord, nous présentons les méthodes des séries temporelles : la suite des observations est indicée par le temps. Les données du temps présent dépendent des données du passé de manière plus ou moins linéaire. On peut citer les modèles ARMA à coefficients variables ou à seuils, ARIMA, GARCH ou des variantes non-paramétriques (c'est à dire avec des paramètres fonctionnels). Le fait de considérer l'influence de facteurs météorologiques conduit dans cette approche à la régression (linéaire ou non) ; et l'exemple des modèles additifs non-linéaires est tout particulièrement détaillé compte tenu de nombreuses applications. Notons également une technique originale pour des prévisions à très court terme : l'utilisation de la théorie du chaos.

Ensuite, nous montrons l'apport des réseaux de neurones à la prévision des niveaux de pollution. Ce type de modèles s'inspire du fonctionnement du cerveau humain. Le modèle classique de régression par le perceptron multi-couches est détaillé car les prévisions ont donné de bons résultats. Puis l'application de modèles plus complexes (BAM et HAM) est illustrée.

Dans le dernier type de modèles, nous avons regroupé les méthodes d'analyses discriminantes et la méthode CART (arbres de régression). Il s'agit de méthodes dont le principal intérêt réside dans la compréhension des intéractions entre les différentes variables entrant en ligne de compte. Les méthodes de construction sont particulièrement détaillées. De plus, l'estimation de densités par directions révélatrices est une technique participant de la même stratégie et est par conséquent expliquée à la fin de cette partie pour son application de prévision des pics de SO_2.

Chacune de ces parties est constituée d'une introduction au domaine considéré et des graphiques viennent illustrer la pertinence des prévisions (il s'agit le plus souvent de comparaisons entre les données réelles et les prévisions).

Une annexe renseigne sur les critères d'efficacité de ces méthodes. Des comparaisons déjà effectuées dans la littérature permettent de se forger une opinion sur la qualité relative des modélisations.

1 Introduction

Le souci d'information de la population et le désir de la part des autorités de répondre suffisamment tôt et de manière adaptée aux problèmes posés par la pollution atmosphérique ont conduit les scientifiques à construire des modèles de prévision des niveaux de pollution (en particulier de dépassement de seuils). Il est possible de consulter Fromage (1996) [26] pour une vision d'ensemble y compris dans son aspect politique et décisionnel. Nous nous intéressons ici aux méthodes fondées sur la statistique mathématique. L'appréhension d'un phénomène incertain par un modèle aléatoire est à l'origine de cette démarche, adoptée depuis longtemps dans d'autres domaines comme en économétrie par exemple.

Le premier atout d'une méthode statistique de prévision réside dans le fait qu'il n'est pas nécessaire de connaître les transformations physico-chimiques permettant l'apparition de grandes quantités de certains polluants. Il s'agit donc de méthodes "ignorantes" : les données pourraient provenir d'un tout autre phénomène que celui de la pollution atmosphérique et la méthode serait identique. La consultation d'un expert est ainsi moins nécessaire. Le second atout se situe au niveau de la rapidité : il ne faut bien souvent qu'environ une heure à un algorithme évolué pour rendre son verdict. C'est pourquoi les réseaux de surveillance de la qualité de l'air montrent un grand intérêt pour ces méthodes.

La diversité des approches est néanmoins très grande. Notre travail consiste ici à présenter l'état de l'art en cette matière. Nous nous intéresserons aux méthodes fondées sur les séries temporelles, les méthodes discriminantes et les réseaux de neurones. Les méthodes fondées sur la statistique des extrêmes n'ont malheureusement pas permis de prévoir les pics à une échéance courte, mais seulement de décrire un peu mieux le phénomène, cf. Bellanger (1999) [2], Killam et Bhattacharyya (1996) [38], Leadbetter (1995) [42], Nakamura et al (1993) [51].

Cette étude est réalisée à partir d'articles, thèses, ouvrages et rapports de recherche publiés dans le monde entier. Malheureusement, les rapports de recherche ne sont pas toujours publiés ou disponibles : il s'agit de la part des laboratoires de recherche d'une activité contractuelle avec les réseaux. Cependant, l'étude "Retour d'expérience en matière de prévision" réalisée pour le Laboratoire Central cette année par Sahli (2000) [60] au sein de l'INERIS fera le point sur l'utilisation actuelle dans les ASSQA de ces méthodes. De plus, un rapport [25] très complet concernant la mise en oeuvre concrète d'une procédure de prévision de l'ozone a été édité par l'agence américaine de l'environnement à destination des réseaux américains. Il est disponible sur le site *www.epa.gov*.

Lorsque les modèles sont construits à l'aide de variables explicatives - ce qui n'est pas toujours le cas- et que la méthode est suffisamment explicite, nous détaillerons les types de variables utilisées, les méthodes de sélection des variables utilisées pour la prévision, la prise en compte des phénomènes locaux particuliers et les indices de confiance dans la prévision.

Dans l'annexe A, un certain nombre de critères de comparaison des différentes

méthodes seront détaillés. Un article récent Bel et al (1999) [3] se penche sur cette question.

Enfin, le lecteur non spécialiste pourra, s'il le souhaite, éviter les passages les plus techniques vu que plusieurs parties du texte sont indépendantes.

2 Séries temporelles et régression

2.1 Introduction

Lorsque des données évoluent au cours du temps de manière continue -comme c'est le cas pour le taux d'un polluant dans l'atmosphère noté X_t (au temps t) - une première approche est de faire des relevés à intervalles réguliers : on travaille alors en temps discret.

L'idée de base des séries temporelles (on dit aussi séries chronologiques) est de prendre en compte le rapport entre la donnée à un instant et le passé au travers d'une relation perturbée par un bruit. Le bruit s'explique par l'inadéquation du modèle à la réalité. Les résultats de mesure ne sont alors qu'une réalisation d'un processus aléatoire. Par exemple le modèle autorégressif d'ordre 1 noté $AR(1)$ s'écrit

$$X_t = \phi X_{t-1} + \varepsilon_t, t \in \mathbb{Z}$$

où $\phi \in \mathbb{R}$ et (ε_t) est un bruit blanc, c'est à dire une suite de variables aléatoires de moyenne nulle et de variance σ^2 et non corrélées, par exemple identiquement distribuées (souvent gaussiennes dans ce cas), voir figure 3. On assimile à ce modèle les processus (X_t) tels que pour tout $t \in \mathbb{Z}$, $EX_t = \mu$ vérifiant[1]

$$X_t - \mu = \phi(X_{t-1} - \mu) + \varepsilon_t, t \in \mathbb{Z}.$$

Ce modèle est assez bien adapté aux statistiques environnementales cf. Wilks (1995) [66] : ϕ sera positif car cela correspond à une persistance et ce phénomène est généralement observé. Une estimation de ϕ, notée $\hat{\phi}$ est donnée par le coefficient d'autocorrélation empirique

$$\hat{\phi} = \frac{\sum_{i=1}^{n-1}(x_i - \bar{x}_-)(x_{i+1} - \bar{x}_+)}{\left[\sum_{i=1}^{n-1}(x_i - \bar{x}_-)^2 \sum_{i=2}^{n}(x_{i+1} - \bar{x}_+)^2\right]^{1/2}}$$

où lorsqu'on considère les données observées $(x_1, ..., x_n)$, $\bar{x}_-$ est la moyenne des $x_1, ..., x_{n-1}$ et $\bar{x}_+$ est la moyenne des $x_2, ..., x_n$.

[1] E désigne l'espérance mathématique (ou moyenne) d'une variable aléatoire, c'est à dire le résultat moyen attendu.

On peut généraliser le modèle $AR(1)$ par le modèle autorégressif d'ordre p moyenne mobile d'ordre q (AutoRegressive Moving Average) $ARMA(p,q)$ en rajoutant une combinaison linéaire des bruits précédents :

$$X_t = \sum_{i=1}^{p} \phi_i X_{t-i} + \varepsilon_t + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} \tag{1}$$

où les (ϕ_i) et les (θ_j) sont des coefficients réels généralement compris entre -1 et 1 et (ε_t) est un bruit blanc. Les processus $ARMA(p,q)$ sont stationnaires, c'est à dire que leur propriétés statistiques ne changent pas au cours du temps : (X_t) est dit stationnaire si

i) $E\left|X_t\right|^2 < \infty$ pour tout $t \in \mathbb{Z}$

ii) $EX_t = m$ pour tout $t \in \mathbb{Z}$

iii) $Cov(X_r, X_s) = Cov(X_{r+t}, X_{s+t})$ pour tous $r, s, t \in \mathbb{Z}$.

On définit alors pour un processus stationnaire (X_t) la fonction d'autocovariance

$$\gamma_X(h) = Cov(X_{t+h}, X_t) \text{ pour tous } t, h \in \mathbb{Z}$$

et la fonction d'autocorrélation

$$\rho_X(h) = \gamma_X(h)/\gamma_X(0) = Corr(X_{t+h}, X_t) \text{ pour tous } t, h \in \mathbb{Z}.$$

Un cas particulier des processus stationnaires est celui des processus strictement stationnaires : les lois conjointes de $(X_{t_1}, ..., X_{t_k})$ et $(X_{t_1+h}, ..., X_{t_k+h})$ sont les mêmes pour tous $k \in \mathbb{N}, t_1, ..., t_k, h \in \mathbb{Z}$. Et ces deux notions coïncident dans le cas des processus gaussiens (si tous les X_t suivent une loi normale).

Néanmoins, de nombreux phénomènes ne sont pas stationnaires. On peut penser à la température maximale mesurée chaque jour pendant plusieurs années pour fixer les idées. La moyenne m de X_t n'est pas forcément constante : elle peut évoluer au cours des années à cause du réchauffement du climat (et on parle de tendance), elle peut évoluer au cours de l'année du fait des saisons (on parle de saisonnalité avec une certaine période). On peut aussi observer une variance $\sigma_t^2 = Var(X_t)$ qui n'est pas constante -ce qui nie la condition *iii*)- et cela arrive lorsque la dispersion des valeurs de X_t change au cours du temps.

Avant d'éliminer la non stationnarité due à la tendance et à la saisonnalité, il peut être utile de modifier les données de départ par une transformation qui stabilise la variance : la transformée de Box-Cox (pour des données positives), utilisée dans Graf-Jacottet (1993) [34] par exemple au choix :

$$U_t = \begin{cases} \dfrac{X_t^\lambda - 1}{\lambda}, X_t \geq 0, \lambda > 0 \\ \ln X_t, X_t > 0, \lambda = 0. \end{cases} . \tag{2}$$

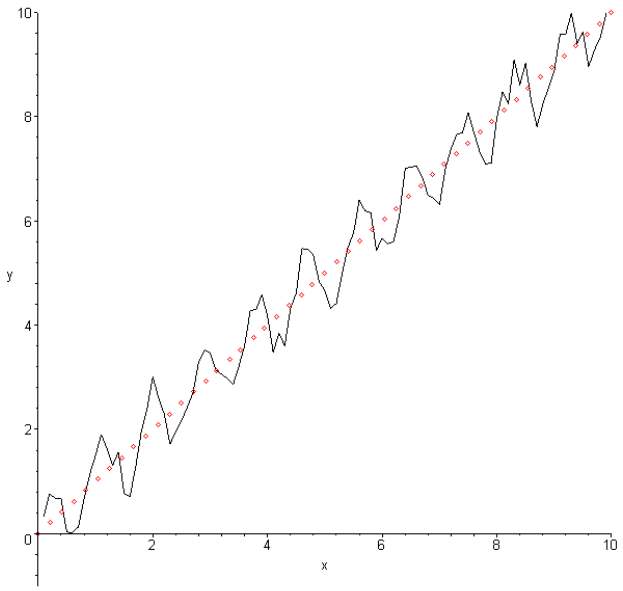


FIG. 1 – Présence d'une tendance et d'une saisonnalité : X_t en fonction de t.

où λ est un paramètre.

Pour se ramener ensuite à un processus stationnaire il est alors possible de considérer deux méthodes.

Premièrement, on peut souvent présenter des processus non stationnaires sous la forme

$$X_t = m_t + s_t + Y_t$$

où m_t est la tendance, s_t la saisonnalité et Y_t est stationnaire. Plusieurs techniques permettent d'estimer et donc d'éliminer m_t et s_t cf. Brockwell et Davis (1991) [9]. Les figures 1, 2, 3 illustrent ces techniques.

En l'absence de saisonnalité on peut faire une estimation de m_t au sens des moindres carrés à l'aide d'une famille de fonctions ou lisser grâce à une moyenne mobile. En présence d'une tendance peu importante et d'une saisonnalité, il est possible de considérer que la tendance est constante lors d'une période et d'obtenir des estimateurs de m_t et s_t. En présence d'une tendance importante et d'une saisonnalité une moyenne mobile peut être utilisée pour estimer m_t.

Deuxièmement, on peut utiliser la méthode dite de Box-Jenkins (1976) [5] qui consiste à différencier un certain nombre de fois afin d'obtenir un processus stationnaire. Plus précisément, on considère le processus

$$Y_{t,1} = (X_t - X_{t-1})$$

puis éventuellement

$$Y_{t,2} = (Y_{t,1} - Y_{t-1,1})$$

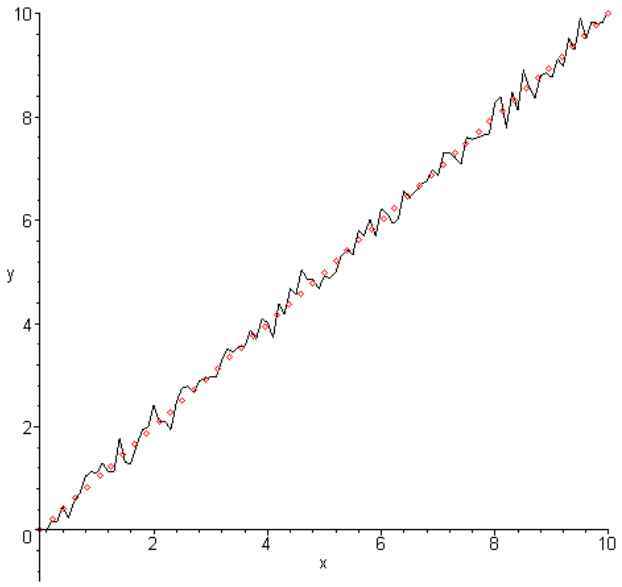


FIG. 2 – Elimination de la saisonnalité

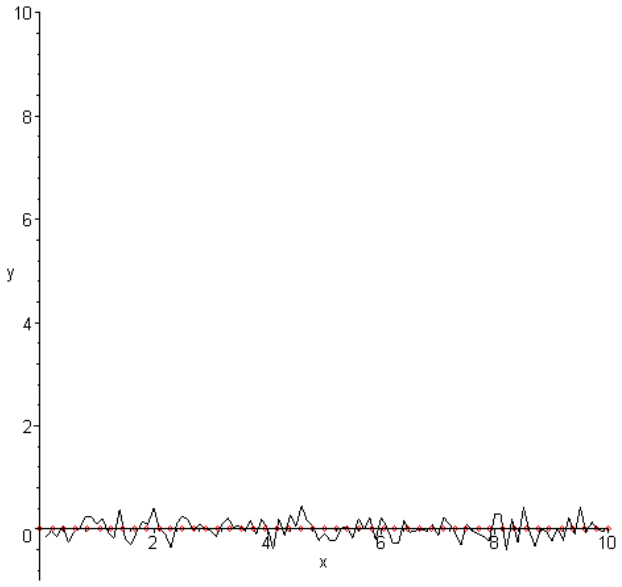


FIG. 3 – Elimination de la tendance et de la saisonnalité : bruit blanc

et ainsi de suite jusqu'à obtenir après d étapes un processus stationnaire que l'on modélise par un $ARMA(p,q)$. De ce fait le modèle d'origine est appelé $ARIMA(p,q,d)$ (pour Autoregressive Integrated Moving Average). Enfin, les séries temporelles (X_t) saisonnières (c'est à dire présentant une forte corrélation à des intervalles réguliers) peuvent être appréhendées par le modèle $SARIMA$ (pour Seasonal $ARIMA$) dont l'expression est relativement compliquée mais qui signifie que d'une période sur l'autre les X_t sont liés entre eux par une relation de type $ARMA$ et que le bruit résultant de cette approche est lui même $ARMA$ mais de proche en proche et non de période en période.

Expliquons dans le cadre du modèle $AR(p)$ (et c'est en fait souvent celui qui fonctionne assez correctement sur des données environnementales) la manière de choisir p et d'estimer les coefficients $\phi_1, ..., \phi_p$. Pour déteminer $\hat{\phi}_1, ..., \hat{\phi}_p$ estimateurs de $\phi_1, ..., \phi_p$, il faut utiliser les équations dite de Yule-Walker

$$\begin{aligned}
r_1 &= \hat{\phi}_1 + \hat{\phi}_2 r_1 + \hat{\phi}_3 r_2 + ... + \hat{\phi}_p r_{p-1} \\
r_2 &= \hat{\phi}_1 r_1 + \hat{\phi}_2 + \hat{\phi}_3 r_1 + ... + \hat{\phi}_p r_{p-2} \\
r_3 &= \hat{\phi}_1 r_2 + \hat{\phi}_2 r_1 + \hat{\phi}_3 + ... + \hat{\phi}_p r_{p-3} \\
&\vdots \\
r_p &= \hat{\phi}_1 r_{p-1} + \hat{\phi}_2 r_{p-2} + \hat{\phi}_3 r_{p-3} + ... + \hat{\phi}_p
\end{aligned}$$

où les r_p sont les fonctions d'autocorrélation empiriques. On peut alors en déduire une estimation $s^2_\varepsilon(p)$ de la variance du bruit.

Pour éviter de choisir p trop grand [2] on utilise un des deux critères de sélection suivants à minimiser :

$$\begin{aligned}
BIC(p) &= n \ln \left[\frac{n}{n-p-1} s^2_\varepsilon(p) \right] + (p+1) \ln n \\
AIC(p) &= n \ln \left[\frac{n}{n-p-1} s^2_\varepsilon(p) \right] + 2(p+1)
\end{aligned}$$

où n est la taille de l'échantillon.

Une vérification de la qualité de la modélisation peut être menée en vérifiant que les résidus (c'est à dire les restes) sont assimilables à un bruit blanc. Il existe des tests ayant cette fonction.

Ainsi les prévisions peuvent se faire dans le cadre $AR(p)$ d'une manière simple en annulant le bruit et en posant donc la valeur au temps $t+1$ égale à la combinaison linéaire des valeurs passées. Des méthodes algorithmiques de moindres carrés plus poussées existent dans un cadre général cf.[9].

En tout cas il est à ce stade très utile de fournir un **intervalle de confiance**, c'est à dire un ensemble de valeurs possibles pour la prévision dont on sait que

[2] Plus p est grand plus on améliore l'adéquation du modèle, mais en faisant de mauvaises prévisions : on parle d' "overfitting".

la vraie valeur sera dedans avec un certain niveau de confiance (en général 90 ou 95%). Cela s'obtient avec une estimation de la variance si l'on suppose le bruit gaussien.

Des programmes informatiques performants réalisent la plupart de ces opérations.

Cependant, pour décrire l'évolution de la pollution, ce modèle linéaire est souvent insuffisant : il a donc été étudié des modèles ARMA dont les coefficients varient suivant l'heure de la journée cf. Barrat et al (1990) [1], des séries chronologiques avec seuils cf. Mélard et Roy (1988) [49], des modèles GARCH [35], des améliorations non paramétriques cf. Gonzalez-Manteiga (1993) [31], ainsi qu'une approche régressive plutôt qu'autorégressive : on explique les concentrations (Y_n) d'un polluant en fonction d'autres variables $(X_n^1, ..., X_n^d)$ (des données météorologiques, la concentration du polluant à un instant passé) :

$$Y_n = f(X_n^1, ..., X_n^d) + \varepsilon_n,$$

et le cas particulier du modèle additif non linéaire a été considéré par Chèze-Payaud et al (1998) [12].

2.2 Modèle ARMA avec coefficients variables

Cette approche a été adoptée pour la prévision de concentrations de NO par l'ESPAC (association pour l'Etude et la Surveillance de la Pollution Atmosphérique de Caen) cf. Barrat et al (1990) [1]. Les données utilisées sont celles de la concentration en NO heure par heure au centre ville de Caen durant les mois de Janvier, Février et Mars 1985. L'horizon de prévision est de deux heures, ce qui est peu. La figure 4 présente la réalisation d'une prévision par la méthode de Box et Jenkins classique. On remarque une sous-estimation des pics et un défaut un peu moins grave : un décalage temporel.

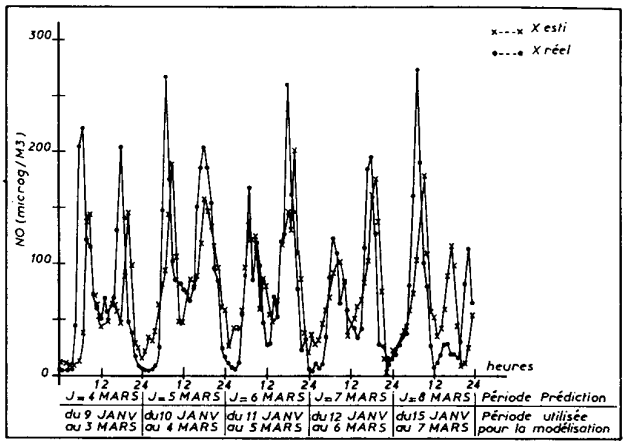
Pour éviter l'écueil de la non stationnarité, les séries temporelles $u_h(J)$ considérées sont les données prises à la même heure h, aux jours successifs J et centrées par la soustraction de la moyenne $m(h)$ à l'heure h calculée à partir des échantillons observés à l'heure h. Le modèle est alors le suivant :

$$u_h(J) = \sum_{i=1}^{p} a_{i,h} u_{h-i}(J) + e_h(J) \tag{3}$$

c'est à dire un modèle AR dont les coefficients $a_{i,h}$ dépendent de l'heure considérée. Les deux avantages sont la prise en compte de la variation diurne et saisonnière (par un renouvellement des $a_{i,h}$ effectué sur quelques jours).

Les estimations des $a_{1,h}, ..., a_{p,h}$ sont obtenues par la méthode des moindres carrés en écrivant (3) sur plusieurs jours consécutifs ; ceci étant fait pour chaque heure h.

On obtient alors assez facilement des prévisions $\hat{u}_{h+f}(J)$ f heures à l'avance sous forme de combinaison linéaires (dépendant de h) des $u_{h-i}(J)$ pour i entre 1 et p.



Comparaison des valeurs réelles et des valeurs prédites pour les concentrations de NO. Prédiction 2 h à l'avance. Modèle autorégressif d'ordre $p = 25$.

FIG. 4 – Coefficients fixes

La supériorité du modèle (3) par rapport au modèle AR classique ou au modèle $SARIMA$ ressort de l'étude, notament dans le décalage dans le temps du moment prévu des pics, cf. figure 5.

2.3 Modèles ARMA avec seuils

Cette méthode a été introduite par Tong (1983) [64], et exploitée dans le cadre de la prévision des pics de pollution en SO_2 par Mélard et Roy (1988) [49]. Elle consiste par exemple dans le cas d'un $AR(1)$ avec un seuil à modéliser la série temporelle (X_t) de la manière suivante :

$$X_t = \begin{cases} \phi_1' X_{t-1} + \varepsilon_t \text{ si } X_{t-1} < \alpha \\ \phi_1'' X_{t-1} + \varepsilon_t \text{ si } X_{t-1} \geq \alpha \end{cases}$$

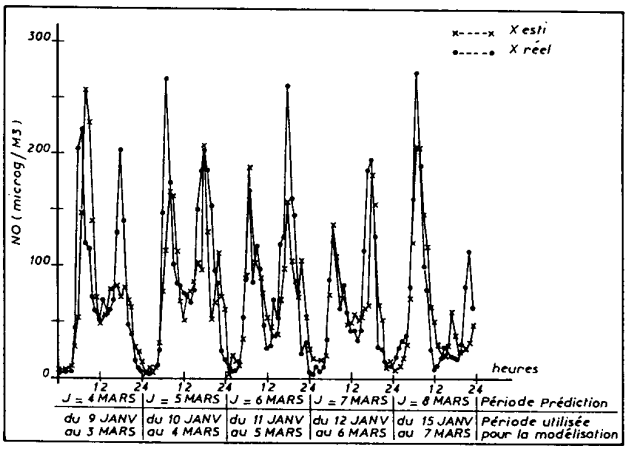
où α est le seuil, ϕ_1', ϕ_1'' des coefficients et ε_t est le bruit. On peut dire grosssièrement que l'on "change de régime" en fonction des valeurs passées.

Ce modèle se généralise en un modèle noté $TAR(l, p_1, ...p_l)$ (Threshold Auto Regressive) lorsqu'on est en présence d'une autre série temporelle (Y_t) - par exemple la température - qui va déterminer le comportement de (X_t) :

$$X_t = \theta_{0i} + \sum_{j=1}^{p_i} \phi_i(j) X_{t-j} + \varepsilon_t$$

pour tout t tel que $\alpha_{i-1} \leq Y_t < \alpha_i, i = 1, .., l$. Les seuils sont données par

$$\alpha_o = -\infty < \alpha_1 < ... < \alpha_{l-1} < \alpha_l = \infty$$



Comparaison des valeurs réelles et des valeurs prédites pour les concentrations de NO. Prédiction 2 h à l'avance. Modèle autorégressif à coefficients variables.

FIG. 5 – Coefficients variables

et ainsi dans le cas où $\alpha_{i-1} \leq Y_t < \alpha_i$, (X_t) est un $AR(p_i)$ de coefficients $(\phi_i(j))_{j=1,\ldots,p_i}$. Dans le cas où $Y_t = X_{t-b}$ pour un $b > 0$, le modèle est dit de type $SETAR$ (Self Exciting Threshold AutoRegressive).

Tous ces modèles ont pour but de corriger un défaut des modèles ARMA : leur linéarité. Comme de nombreux phénomènes naturelles n'évoluent pas de cette manière -on peut penser bien sûr à la pollution- ces modèles ont été appliqués avec succès à des données hydrologiques ou aux tâches du soleil cf. Tong (1983) [64].

Mélard et Roy (1988) [49] ont développé des méthodes pour estimer les paramètres et les seuils dans le modèle assez général $ARMA(p,q)$ avec seuils :

$$X_t = \mu_i + \sum_{j=1}^{p} \theta_i(j)(X_{t-j} - \mu_{I(t-j)}) + \varepsilon_t - \sum_{k=1}^{q} \theta_i(k)\varepsilon_{t-k}$$

si $\alpha_{i-1} \leq Y_t < \alpha_i$ $(i = 1, .., l)$ où la fonction $I(t)$ est définie par $I(t) = i$ si $\alpha_{i-1} \leq Y_t < \alpha_i$.

Ils ont étudié des estimateurs des coefficients par maximum de vraisemblance et estimé les seuils en utilisant les valeurs rangées par ordre croissant du processus (Y_t).

L'application à des moyennes journalières de concentration en SO_2 de la station de Warsage en Belgique a été réalisée sur les données de l'année 1977 (dont on a auparavant pris le logarithme). La méthode de Box et Jenkins fournit un modèle $AR(3)$. En prenant comme valeurs du processus qui détermine les seuils les données de vitesse du vent, on obtient plusieurs modèles à seuils, dont le meilleur est un $AR(1)$.

Mélard et Roy concluent à la supériorité des modèles $ARMA$ à seuils, grâce à des critères de type AIC ou BIC. Il semble assez clair que l'on peut attendre une certaine amélioration de la prévision avec ce genre de modèles.

2.4 Modèle GARCH

Graf-Jacottet et Jaunin [35] ont développé un modèle de séries chronologiques permettant de prévoir les concentrations d'ozone et de NO_2 à Payerne et Taenikon en Suisse. Ils n'ont pas intégré de variables explicatives car leur philosophie est la suivante : l'information nécessaire pour la prévision est contenue dans la série statistique et il vaut mieux se focaliser sur une meilleure modélisation mathématique que sur la recherche d'autres variables pertinentes.

Le modèle utilisé ici est le modèle GARCH (Generalized Autoregressive Conditional Heteroscedastic model). Il s'agit de modèles de séries chronologiques dont les résidus (ε_t) ne constituent pas un bruit blanc : on a

$$\varepsilon_t|\varepsilon_{t-1}, \varepsilon_{t-2}, ... \sim N(0, \sigma_t^2)$$
$$\sigma_t^2 = a + \sum_{i=1}^{p} a_i \varepsilon_{t-i}^2 + \sum_{i=1}^{q} b_i \sigma_{t-i}^2,$$

modèle noté $GARCH(p, q)$. Ces modèles sont massivement utilisés en finance cf [32] et [23] pour modéliser la volatilité, car la variabilité est reliée au risque.

Notons y_i la moyenne sur 24h de la concentration d'un certain polluant le jour $i, i = 1, 2, ..., N$ et x_{i+12} le nombre d'heures entre le lever et le coucher du soleil du jour i. Ici, la modélisation se fait en deux temps. Tout d'abord, une transformation est appliquée aux données : la TBS (tranform both sides) de Caroll et Ruppert [10]. Notons $g(y, \lambda)$ la transformée de Box-Cox (2) et le modèle s'écrit (avec le raffinement TBS-AR(p) de [34])

$$g(y_i, \lambda) = g(f(x_i, \beta_0, ..., \beta_r), \lambda) + \varepsilon_i$$

où ε_i suit un modèle autorégressif d'ordre p dont les erreurs (a_i) sont sensées être un bruit blanc. Seulement, ce dernier critère n'est pas vérifié, ce qui est courant pour des données environnementales. C'est pourquoi l'ajustement se fait sur un modèle $GARCH(1, 1)$ sur les $(|a_i|)$ dont le résultat est :

$$|a_i| = \omega + (\alpha + \beta)\,|a_{i-1}| - \beta\nu_{i-1} + \nu_i$$

où (ν_i) est un bruit blanc, et les paramètres ω, α, β sont positifs. Puis une procédure d'estimation-vérification est entreprise pour connaître les paramètres du modèle. Des intervalles de confiance sont calculés également. On peut observer les prévisions à un jour sur la figure 6.

Les résultats sont intellectuellement intéressant car une variabilité évolutive au cours du temps est prise en compte. Une comparaison avec les résultats de modèles

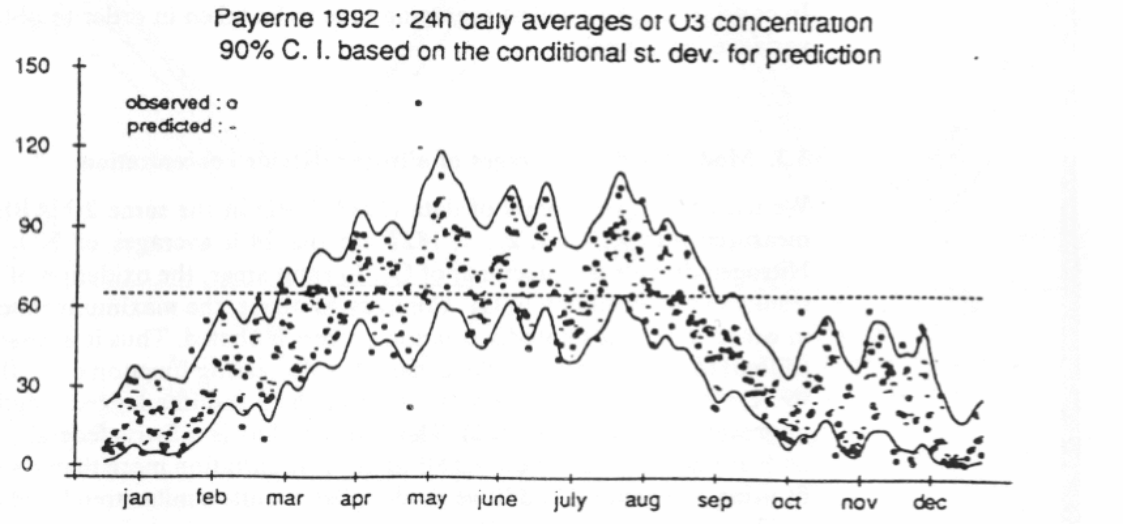


FIG. 6 – Intervalles de confiance construits par modèle GARCH. Unités : ppb.

comprenant un nombre important de variables explicatives tourne légèrement à l'avantage de ce dernier modèle. Cependant, l'intérêt de la modélisation GARCH réside aussi dans le fait qu'aucun effort n'est fait pour recueillir les données météorologiques et s'avère donc peu chère.

2.5 Méthodes non-paramétriques avec résidus ARIMA

Cette méthode a été utilisée par Gonzalez-Manteiga et al [31] pour la prévision de concentrations en SO_2 à côté d'une centrale électrique à As Pontes en Espagne pour une prévision à très court terme : une demi-heure (c'est le temps qu'il faut à un opérateur pour réduire la production). Les données arrivent toutes les 5 minutes, donc la prévision se fait à l'horizon $t+6$. Cette méthode pourrait semble-t-il s'adapter à une prévision à 6 heures voire plus avec des données fournies heure par heure. Les données utilisées pour modéliser et prévoir sont celles des 6 dernières heures. L'observation des valeurs indique des changements abruptes et une instabilité de la variance, phénomènes éminemment non linéaires.

La modélisation $ARIMA$ n'étant pas satisfaisante - cf. figure 7 - pour cause de paramètres trop grands, une méthode non paramétrique a été utilisée : en notant (X_t) le processus, $E(X_{t+6}|X_t, X_{t-1})$ a été estimée [3] par $\hat{E}(X_{t+6}|X_t, X_{t-1})$ en utilisant l'estimateur de régression de Nadaraya-Watson avec une fenêtre choisie par validation croisée et un noyau gaussien, cf. figure 8.

Mais pour prendre en compte l'historique et non seulement les dernières valeurs, une classification en neuf classes de 500 tableaux de la forme $(X_{\tau-1}, X_\tau, X_{\tau+6})$

[3] $E(.|.)$ désigne l'espérance conditionnelle, c'est à dire le résultat moyen attendu connaissant certains événements ou certaines variables aléatoires.

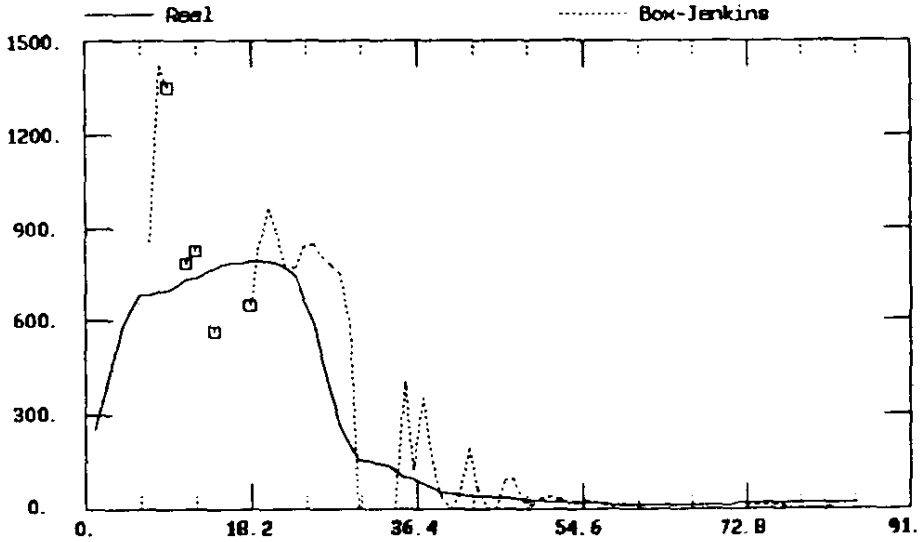


FIG. 7 – Méthode ARIMA. Prévision de SO_2 à 30 mn.

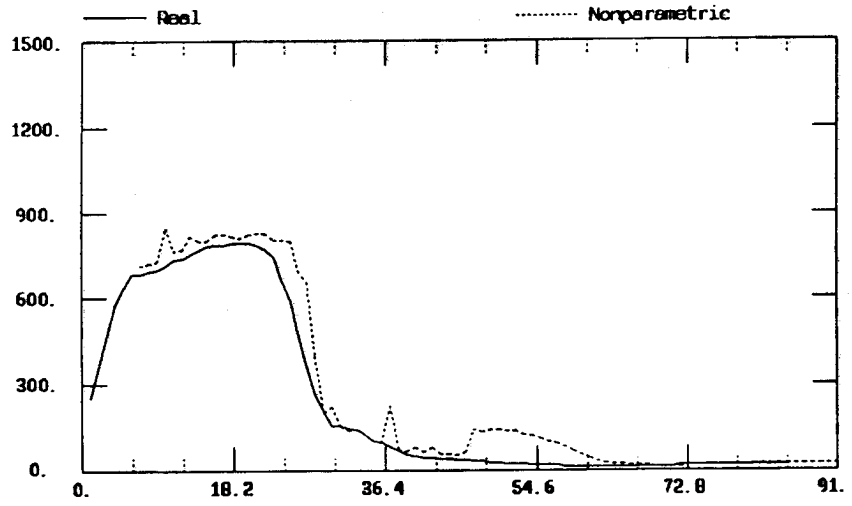


FIG. 8 – Utilisation de méthodes non paramétriques. Prévision de SO_2 à 30 mn.

en fonction de $X_{\tau+6}$ a été réalisée, ce qui améliore les choses.

Seulement, lorsqu'on calcule les résidus $\hat{Z}_i = X_i - \hat{E}(X_i|X_{i-6}, X_{i-7})$, on trouve que ceux-ci ne sont pas toujours assimilables à un bruit blanc. Ainsi, une modélisation paramétrique de type $ARIMA$ a été entreprise sur les $\hat{Z}_{t-64}, ..., \hat{Z}_t$ afin de prévoir $\hat{Z}_{t+6}$ et la prévision de X_{t+6} est alors

$$\hat{E}(X_{t+6}|X_t, X_{t-1}) + \hat{Z}_{t+6}.$$

Deux types d'intervalles de confiance ont été calculés : avec la méthode de Box et Jenkins sur la partie $ARIMA$ ou par une méthode de bootstrap sur cette même partie.

La conclusion est que l'approche semi-paramétrique (non paramétrique et $ARIMA$) est meilleure que la méthode non paramétrique simple et est de loin bien meilleure que la méthode $ARIMA$ sur les (X_t).

Prada-Sanchez et al [54] ont amélioré récemment ce modèle, compte tenu de l'obligation de prévoir une heure à l'avance et non une demi-heure. Pour cela ils ont rajouté une combinaison linéaire de variables explicatives exogènes. Ce choix d'une action linéaire est fait pour des raisons pratiques de calcul et pour des raisons théoriques : "la malédiction de la dimension" (le modèle additif non linéaire est en ce sens plus avancé, voir partie 2.6.2). Plusieurs variantes ont été étudiées, toutes de la forme générale

$$Y_n = V_n^t \beta + \varphi(Z_n) + \varepsilon_n$$

où Y_n est la réponse, Z_n est le vecteur aléatoire d'intérêt et V_n^t est le vecteur aléatoire constitué des variables explicatives. Le travail qui consiste à estimer β et φ se fait en utilisant différentes techniques statistiques : régression linéaire, séries temporelles et noyau. Pour l'application aux séries de SO_2, voici la démarche

Types de variables utilisées :

- vitesse et direction du vent
- radiation solaire
- température et différences entre les températures à 10 et 80 m., à 10 et 30 m.
- précipitations accumulées
- humidité

Méthodes de sélection des variables utilisées pour la prévision :

Examen de la pertinence, de la fiabilité, de la stabilité et de la redondance.

Prise en compte des phénomènes locaux particuliers

Non

Indices de confiance dans la prévision :

Les résultats ne sont pas donnés sous forme d'intervalles de confiance, mais les erreurs de type MSE, MAE, MRE, MRAE (cf. Annexe A) sont calculée pour diverses variantes du modèle. Pour une journée où il y a eu un pic la figure 9 illustre assez bien les résultats.

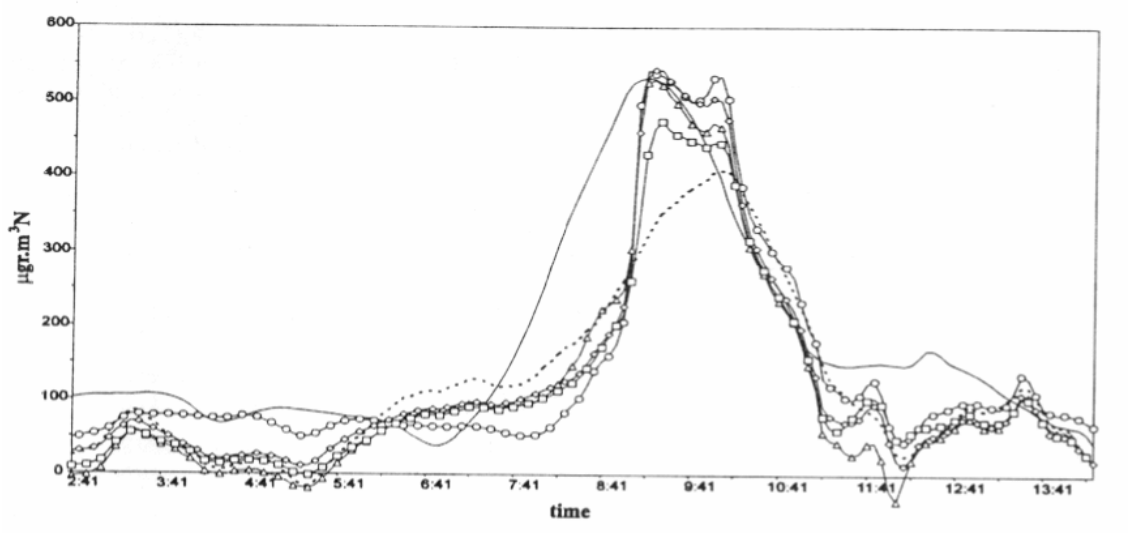


FIG. 9 – Différents essais de prévision du SO_2 une heure à l'avance utilisant une combinaison de variables exogènes (— : réel)

2.6 Modèles de régression

D'une manière générale, il est question de calculer les valeurs de la variable d'intérêt en fonction d'un certain nombre de variables dites explicatives. Par exemple, le modèle non paramétrique général

$$Y_n = f(X_n^1, ..., X_n^d) + \varepsilon_n, \tag{4}$$

où (ε_n) est un bruit et $X_n^1, ..., X_n^d$ sont les variables explicatives représente formellement cette relation. Cependant, la complexité de calcul a poussé les statisticiens à s'intéresser à un certain nombre de cas particuliers.

2.6.1 Modèle linéaire

Il s'agit d'une modèle de régression simple, qui peut tout à fait avoir un intérêt pour certains polluants. La fonction f dans (4) est alors une combinaison linéaire des $X_n^1, ..., X_n^d$. En particulier Comrie et Diem [16] ont effectué cette démarche pour prévoir le niveau de CO à Phoenix.

Types de variables utilisées :

14 variables explicatives ont été utilisées, par le biais des données suivantes à différentes heures et de prévisions météorologiques.

- vitesse du vent
- pression
- température et différences de température
- niveau de CO observé et index (décrivant le comportement suivant le jour)

– temps de formation d'une inversion

Méthodes de sélection des variables utilisées pour la prévision :

Examen des différents modèles possibles et utilisation de procedures standards pour le modèle linéaire.

Prise en compte des phénomènes locaux particuliers

A priori seulement des effets de jours de semaine.

Indices de confiance dans la prévision :

Les résultats indiquent que les erreurs sont entre

– 5 et 15% au quantile de 80%

– 25 et 45% au quantile de 90%.

Cela signifie que la méthode est somme toute assez bonne, mais bien sûr, les résultats se dégradent lorsqu'on cherche à évaluer les extrêmes compte tenu du modèle linéaire utilisé. Il est important de souligner que les variations du CO sont semble-t-il plus linéaires que celles de l'ozone par exemple.

2.6.2 Modèle additif non linéaire

Cette méthode a été introduite par Hastie et Tibshiriani [37] et utilisée dans le cadre de la prévision de l'ozone par Chèze-Payaud et al [12] et Davis et Speckman [18]. Elle a été utilisée durant l'été 1998 par AIPARIF afin de repérer les maximums de concentration d'ozone.

Le modèle est le suivant :

$$Y_n = \mu + \sum_{i=1}^{d} f^i(X_n^i) + \varepsilon_n \tag{5}$$

où μ est une constante, Y_n est la variable à expliquer - par exemple la concentration d'un certain polluant -, $X_n^1, ..., X_n^d$ sont les variables explicatives et ε_n est un bruit. Les fonctions f^i sont à estimer.

L'avantage de ce modèle est d'être plus souple que le modèle linéaire paramétrique

$$Y_n = \mu + \sum_{i=1}^{d} \theta_i X_n^i + \varepsilon_n$$

où les θ_i sont des coefficients multiplicatifs, et de posséder une vitesse de convergence en moyenne quadratique optimale meilleure (elle ne se dégrade pas avec la dimension d) que le modèle non paramétrique général

$$Y_n = f(X_n^1, ..., X_n^d) + \varepsilon_n.$$

L'estimation des f^i est faite sous l'hypothèse suivante d'identifiabilité du modèle : pour tout i

$$E\left[f^i(X_n^i)\right] = 0.$$

La méthode utilisée par Chèze-Payaud et al [12] s'intitule estimation par intégration marginale. L'estimateur de f^i obtenu est de la forme

$$\hat{f}^i(x^i) = \frac{1}{n}\sum_{j=1}^{n} \hat{m}(X_j^1, ..., X_j^{i-1}, x^i, X_j^{i+1}, ..., X_j^d)$$

avec $\hat{m}$ estimateur de la fonction de régression

$$m(x) = E(Y|(X^1, ..., X^d) = x).$$

L'estimateur $\hat{m}$ choisi est celui obtenu par la méthode du noyau.

Il est également possible d'interpréter $\hat{f}^i$ comme une somme pondérée par des fonctions en x^i des Y_k (cf.[12]).

Cette technique est justifiée par un théorème de normalité asymptotique, pour plus de détails voir Linton et Hardle (cf.[44]).

Des simulations détaillées dans [12] pemettent de choisir judicieusement les fenêtres de la méthode du noyau afin de réduire l'erreur de prévision.

Types de variables utilisées :

- le niveau d'ozone de la veille
- la température maximale de la veille
- le vent moyen de l'après-midi de la veille

Méthodes de sélection des variables utilisées pour la prévision :

Observation graphique de l'effet des variables sur l'ozone du lendemain.

Prise en compte des phénomènes locaux particuliers

La donnée campagne/ville concernant chaque station.

Indices de confiance dans la prévision :

La technique qui a été utilisée est la construction d'intervalle de confiance par la méthode du bootstrap (rééchantillonage). Les auteurs souhaitent approfondir cette méthode.

Les résultats de cette technique appliquée à la région parisienne sont comparés à d'autres méthodes et développés dans l'Annexe A.

Davis et Speckman [18] ont utilisé ce modèle pour prévoir le maximum et une moyenne durant 8 heures de la journée dans l'agglomération de Houston. Le défaut de cette étude réside dans l'utilisation a posteriori des variables météorologiques et non l'utilisation des prévisions. De ce fait, il ne s'agit pas à proprement parler de prévisions. Néanmoins cela représente un assez grand intérêt lorsque les prévisions météorologiques sont de bonne qualité.

Types de variables utilisées :

- composante est-ouest du vent (positive provenant de l'ouest) u
- composante sud-nord du vent (positive provenant du sud) v
- couverture nuageuse opaque $opcov$
- température maximale $t\,\mathrm{max}$
- hauteur de mélange du matin $mixam$

– niveau d'ozone du jour précédent $\max lag_1$.

Les trois premières variables ont été exploitées à travers la moyenne des valeurs horaires de 20h à 5h (indice 1), de 6h à 9h (indice 2), de 10h à 17h (indice 3). Le modèle obtenu grâce au module GAM (Generalized Additive Models) du logiciel S-PLUS est :

$$\begin{aligned}\log E(y) = {} & f_1(u_1, v_1) + f_2(u_2, v_2) + f_3(u_3, v_3) \\ & + f_4(opcov_3) + f_5(\max lag_1) + f_6(t\max) + f_7(mixam),\end{aligned} \tag{6}$$

où $E(y)$ est la prévision de la concentration maximale en ozone.

Méthodes de sélection des variables utilisées pour la prévision :

L' idée exploitée ici est de construire différents modèles de régression avec certaines variables explicatives par une méthode locale ("loess") implantée sur le logiciel S-PLUS et d'examiner les coefficients de corrélations. Ainsi les variables sus-citées ont été prises en compte et des modèles plus compliqués n'ont pas été utilisés car le gain n'était qu'infime. Ce choix a été confirmé par des critères AIC par exemple, ainsi que par l'examen des résidus à l'aune de l'homoscédasticité (stabilité de la variance) et de la symétrie.

Prise en compte des phénomènes locaux particuliers

La brise de mer n'a pas été prise en compte par les modélisateurs à leur regret : il s'agit d'un paramètre important. Ils proposent d'améliorer le modèle en exploitant cette variable, mais aussi le niveau de radiation solaire dont les données n'étaient pas correctement mesurées.

Indices de confiance dans la prévision :

L'erreur de prévision, mesurée par la racine carrée de l'erreur quadratique moyenne s'étale de 18,5 à 22,0 ppb suivant les stations pour le niveau maximum d'ozone - le niveau "seuil" qui intéresse les Etats-Unis est de 120 ppb -. En général, le modèle sous-estime le niveau d'ozone, mais pendant une petite période, le modèle a surestimé ce même niveau, phénomène que n'arrivent pas à expliquer les modélisateurs, voir figure 10.

Cobourn et Hubbard [13], [14] ont utilisé une démarche similaire mais moins aboutie sur le plan statisitique. Dans une première approche [13], le modèle de régression était le suivant :

$$\begin{aligned}\sqrt{Y} = {} & b_0 + b_1 X_1 + b_2 X_1^2 + b_3 X_1^4 + b_4 \exp(\theta X_2) \\ & + b_5 X_3 + b_6 X_4 + b_7 X_5 + b_8 X_6 + b_9 X_7 + b_{10} X_8\end{aligned} \tag{7}$$

où Y est le niveau maximum d'ozone prévu, et les variables explicatives $X_1, ..., X_8$ sont la température maximale, la vitesse du vent, l'intensité de la radiation solaire, la température minimale, la couverture nuageuse, les précipitations, le nombre de périodes de vent calme durant la nuit et le jour de la semaine. Les résultats sont assez médiocres car aucun des neufs pics (au dessus de 120 ppb) n'ont été prévus.

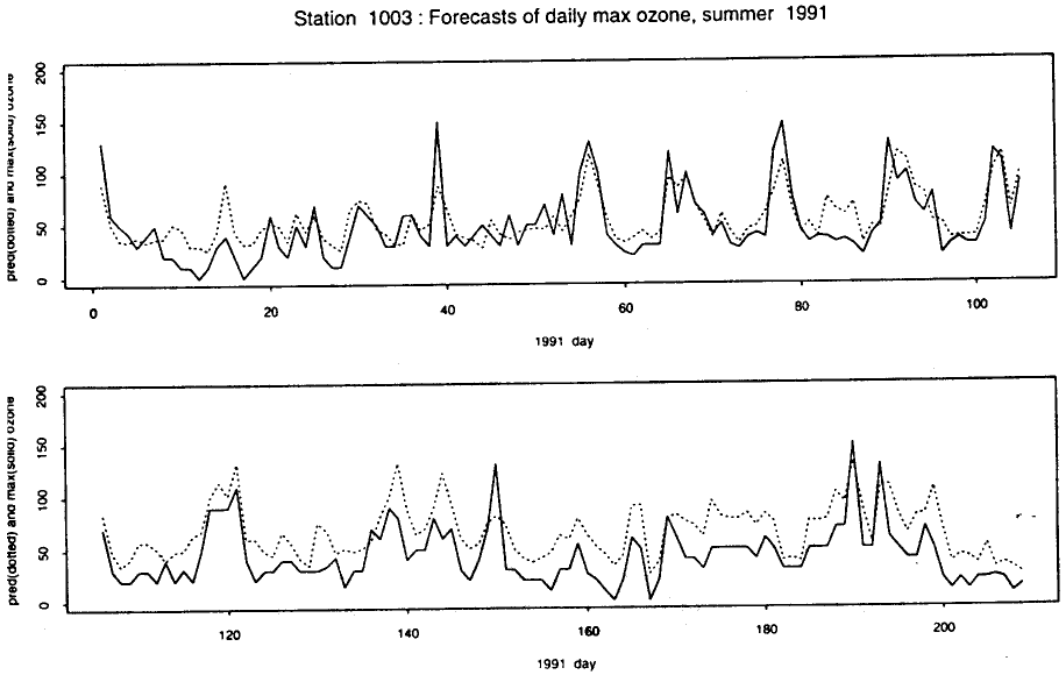


FIG. 10 – Maximums d'ozone mesurés (en gras) et prévisions (en pointillé). Modèle de régression (6).

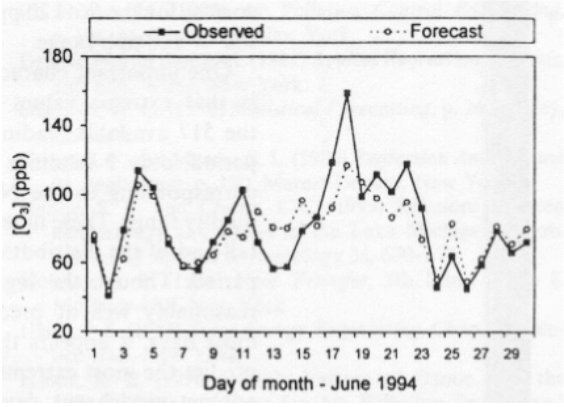


FIG. 11 – Prévision suivant le modèle de régression non linéaire (7). Louisville, KY.

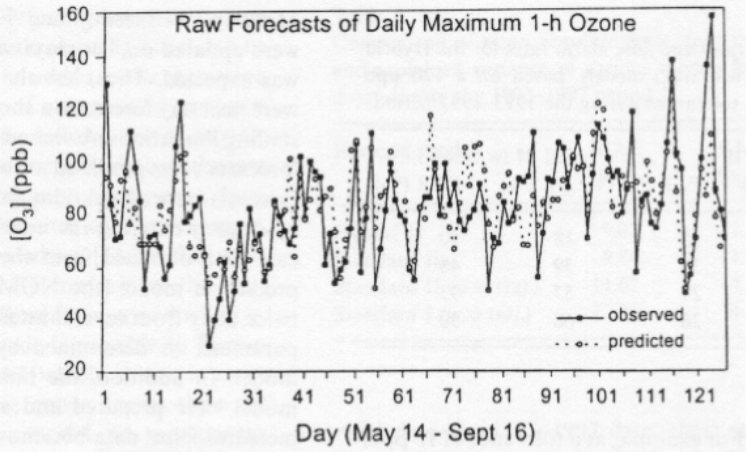


FIG. 12 – Prévision suivant le modèle de régression non linéaire (8)

Dans une deuxième approche, une variable supplémentaire a été introduite : la trajectoire de la masse d'air, sous la forme d'un paramètre égal à 0 ou 1 suivant la situation. Le modèle s'écrit

$$O_3 = b_0 + b_1 O_{3nl} + b_2 CC + b_3 DOW + b_4 LOD + b_5 NC + b_6 RF + b_7 TRAJ \tag{8}$$

L'illustration de ces résultats est visible sur les figures 11 et 12.

2.7 Utilisation de la théorie du chaos

Certains chercheurs - cf. [11], [43] et [55] - se sont apperçus que l'évolution de la concentration en ozone au cours du temps présentait un caractère chaotique. C'est pourquoi, dans le cadre d'une prévision à court terme, l'utilisation de cette théorie peut s'avérer utile. En revanche, sur le long terme, la sensibilité aux conditions initiales de ce genre de modèles ne permet pas de faire de bonnes prévisions.

Kocak et al. [41] ont réalisé une étude de ce type en matière d'ozone sur le site d'Istanbul. **L'horizon de leur prévision est d'une heure**, et la fréquence des mesures est d'une heure également, le pas de prédiction $p = 1$. La technique est relativement ardue sur le plan théorique. Il s'agit de reconstruire l'attracteur dans un espace de grande dimension. Cet attracteur représente géométriquement la part de stabilité de la série temporelle (x_i) en question. On cherche à trouver un entier m tel que par le biais de m décalages, les vecteurs

$$X_i = (x_i, x_{i-\tau}, ..., x_{i-(m-1)\tau}) \in \mathbb{R}^m$$

soient situés sur l'attracteur en question. τ est le temps de corrélation, pour lequel la corrélation entre les coordonnées s'annule. Ici l'observation de la fonction

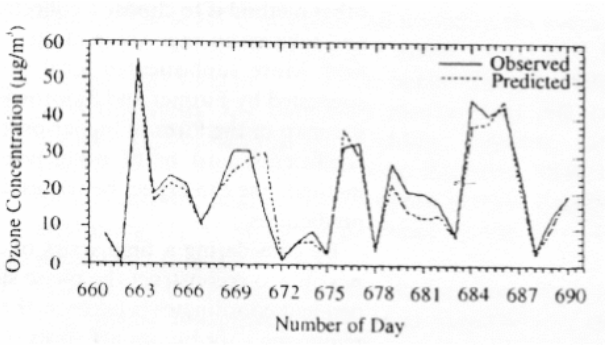


FIG. 13 – Prévision à une heure par modèle chaotique

d'autocorrélation a permis d'obtenir $\tau = 75$. Puis le choix de $m = 3$ a été effectué par une technique graphique.

La relation entre X_t et X_{t+p} sur l'attracteur est approchée par

$$X_{t+p} \simeq F(X_t).$$

La méthode de prévision locale permet de trouver une fonction F convenable, polynôme de plusieurs variables de degré 1, dont les coefficients dépendent du voisinage du point de départ.

L'erreur relative $RMSE/\sigma$, où σ est l'écart-type de la série temporelle observée est de 0.3. On peut comparer avec le modèle issu de [11] : 0.42 ; et pour des modèles $AR(2), AR(5), AR(10)$, on obtient 0.92, 0.94, 0.98.

Il n'y a ici que l'étude de la série temporelle brute et non l'utilisation de variables explicatives. Les résultats sont assez satisfaisants, cf figure 13, même si une prévision à une heure est assez aisée.

3 Réseaux de neurones

3.1 Introduction

Sous le terme de réseaux de neurones, on regroupe aujourd'hui des modèles dont l'intention est d'imiter quelques unes des fonctions du cerveau humain en reproduisant certaines de ses structures de base.

Dans le cerveau, il existe environ 10^{11} neurones, 10^{15} connexions (synapses) qui échangent des signaux électriques. On sait mesurer d'une manière ou d'une autre l'activité d'un petit nombre de neurones ou d'une grande population de neurones. Au niveau d'une cellule, le potentiel de membrane peut être representé par un processus de sauts, ou par une diffusion. Il y a émission d'un "spike" lorsque le potentiel dépasse un certain seuil. Dans le cerveau, on distingue des sous-ensembles constitués de couches ou de colonnes, avec des entrées, des sorties, des connexions internes et externes réalisant une certaine tâche.

Historiquement, les origines de cette discipline sont très diversifiées. En 1943, McCulloch et Pitts [48] étudièrent un ensemble de neurones formels interconnectés et montrèrent leurs capacités à calculer certaines fonctions logiques. C'est en 1957 que Rosenblatt [56] décrivit le premier modèle opérationnel de réseaux de neurones, mettant en oeuvre les idées de McCulloch et Pitts : le perceptron, inspiré du système visuel, capable d'apprendre à calculer certaines fonctions logiques en modifiant ses connexions synaptiques.

Des applications des réseaux de neurones ont été développées dans beaucoup de domaines, par exemple en reconnaissance de formes, en prévision, en modélisation, etc... On trouve beaucoup d'utilisations des réseaux de neurones dans l'industrie ainsi qu'en finance et en économie.

Récemment, quelques travaux ont été menés dans le domaine de la prévision de la pollution. On peut citer par exemple Comrie (1997) [15], Ruiz-Suarez et al (1995) [58], Yi et Prybutok (1996) [67].

3.2 Les modèles généraux de réseaux de neurones

D'une façon plus générale, on peut définir un neurone formel par les cinq éléments suivants :

- la nature de ses entrées ;
- la fonction d'entrée totale qui définit le prétraitememt effectué sur ses entrées ;
- la fonction d'activation du neurone qui définit son état interne en fonction de son entrée totale ;
- la fonction de sortie qui calcule la sortie du neurone en fonction de son état d'activation ;
- la nature de la sortie du neurone.

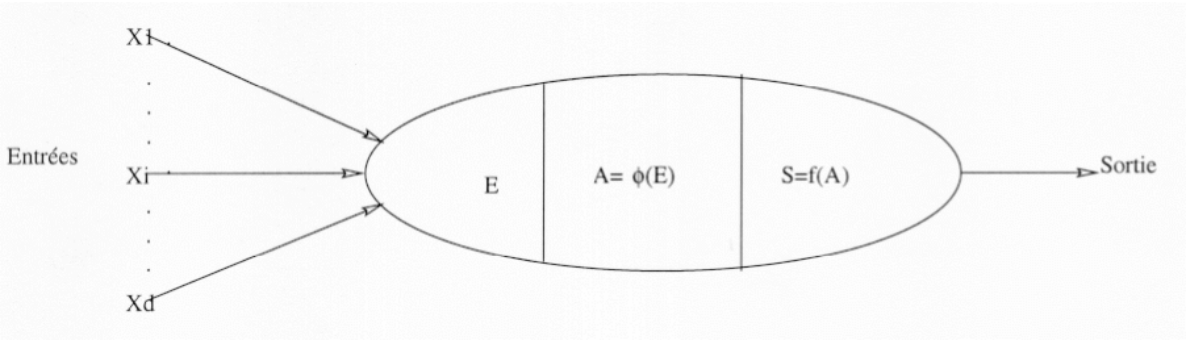


FIG. 14 – Neurone formel

La figure 14 représente un neurone formel. Nous adopterons par la suite les notations suivantes :

- $(x_i)_{i=1...d}$ seront les entrées ;
- h sera la fonction d'entrée totale ;
- f sera la fonction de sortie ;
- ϕ sera la fonction d'activation ;
- $E = h(x_1, x_2, \ldots, x_d)$;
- $A = \phi(E)$ comme état du neurone ;
- $S = f(A)$ comme sortie.

La fonction d'activation ϕ joue un rôle très important, elle peut être linéaire ou bien non linéaire. Voici quelques exemples de fonctions d'activation :

Exemple 1 *fonction linéaire* $\phi(x) = ax$.

Exemple 2 *fonction linéaire par morceaux.*

$$\phi(x) = \begin{cases} T & x > T \\ x & t \leq x \leq T \\ t & x < t \end{cases}$$

Exemple 3 *fonction de signe*

$$\phi(x) = \begin{cases} 1 & si\ x \geq 0 \\ 0 & sinon \end{cases}$$

Exemple 4 *fonction sigmoïde exponentielle*

$$\phi(x) = \frac{1}{1 + e^{-x}}$$

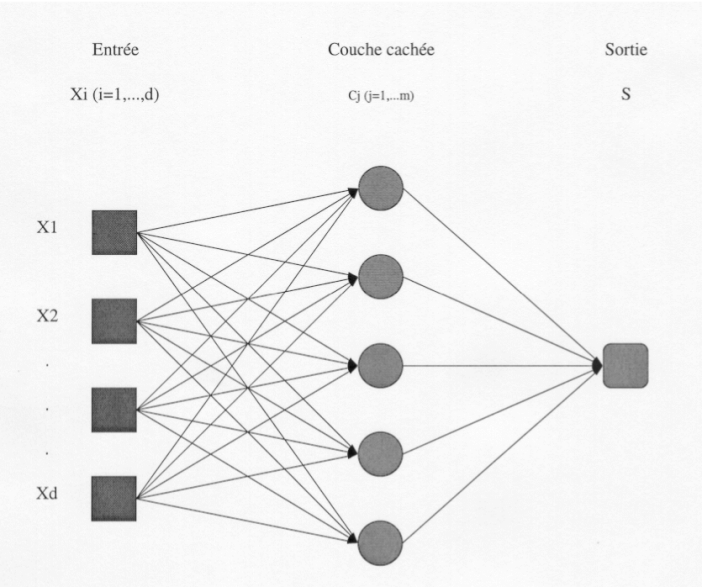


FIG. 15 – PMC

D'une manière générale, l'architecture des réseaux de neurones formels peut aller d'une connectivité totale (tous les neurones sont reliés les uns aux autres), à une connectivité locale où les neurones ne sont reliés qu'à leurs plus proches voisins.

Nous présentons ci-dessous deux modèles classiques.

1. Les réseaux à couches

   On utilise une structure de réseaux à couches telle que les neurones qui appartiennent à une même couche ne soient pas connectés entre eux, chacune des couches recevant des signaux de la couche précédente, et transmettant le résultat de ses traitements à la couche suivante. Les deux couches extrêmes correspondent à la couche qui reçoit ses entrées du millieu extérieur d'une part, et à la couche qui fournit le résultat des traitememts effectués d'autre part. Les couches intermédiaires sont appelées couches cachées, leur nombre est variable. Un exemple type des réseaux à couches est le Perceptron Multi-Couches dit PMC. Un PMC comportant une couche cachée avec un fonction d'activation non linéaire et une couche de sortie (voir figure 15) permet d'approcher toute fonction à support borné (Cybenko 1989) [17]. On utilise souvent ce modèle.

2. Les réseaux entièrement connectés

   Dans ces réseaux, chaque cellule est reliée à toutes les autres et possède même un retour sur elle-même (voir figure 16).

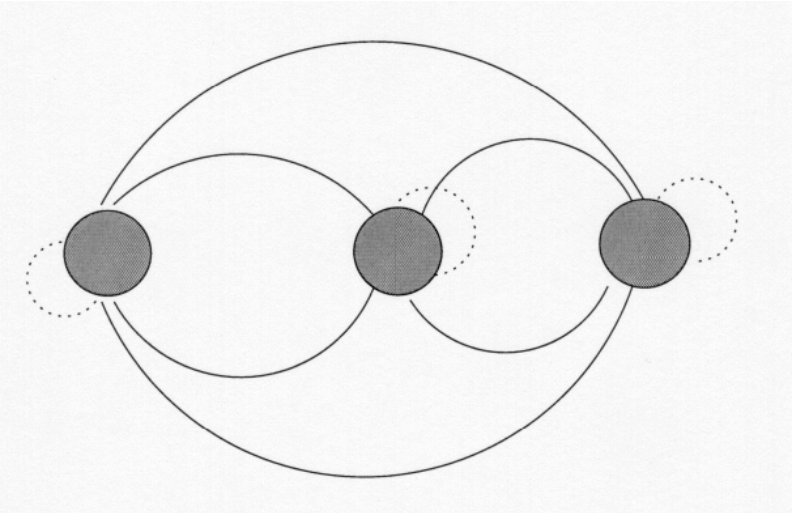


FIG. 16 – Réseaux entièrement connectés

3.3 Régression par le PMC

Dans la partie précédente, on a déja présenté un peu le PMC. Remarquons que le modèle est équivalent à un modèle linéaire, lorsque les fonctions d'activation sont linéaires. En général, la fonction d'activation est une sigmoïde afin que les modèles soient non linéaires et s'adaptent à la complexité de la réalité. Les applications sont nombreuses : prévision, reconnaissance de formes, classification.

Présentons maintenant le modèle statistique de la régression par le PMC. Considérons les données observées (l'ensemble d'apprentissage) $(x_t, y_t)_{t=1,\dots,n}$ comme la réalisation d'un échantillon, c'est à dire d'une suite de variables aléatoires $(X_t, Y_t)_{t=1,\dots,n}$ indépendantes et identiquement distribuées à valeur dans $\mathbb{R}^d \times \mathbb{R}$. On veut estimer la regression

$$f(x) = E(Y/X = x)$$

Nous utiliserons le modèle suivant :

$$Y_t = f(X_t) + \varepsilon_t,\ t \in \mathbb{Z}$$

où (ε_t) est un bruit blanc. Le problème consiste à estimer cette fonction $f(x)$ à partir des données observées. Pour réaliser une estimation fonctionnelle avec des PMC avec une couche cachée on peut utiliser par exemple l'ensemble $\mathcal{F}$ des perceptrons avec m unités dans la couche cachée. Cet ensemble est défini de la manière suivante :

$$\mathcal{F} = \left\{ f_m : f_m = \sum_{i=1}^{m} c_i \phi(a_i.x + b_i) + c_0, a_i \in \mathbb{R}^d; c_0, b_i, c_i \in \mathbb{R} \right\}$$

Pour estimer les paramètres, on utilise la méthode des moindres carrés, ce qui revient à minimiser la somme suivante :

$$E = \sum_t (Y_t - f_m(X_t))^2.$$

L'algorithme dit de "retro-propagation" est bien adapté à la résolution de ce problème. Les details peuvent être trouvés dans le livre de Bishop (1995). En outre, Maier et al (1998a et 1998b) ont fait une étude empirique sur cet algorithme. Cet algorithme est résumé ci-dessous

1. initialiser les poids du réseau
2. présenter premièrement les vecteurs d'entrée à partir des données d'apprentissage dans le réseau
3. envoyer les vecteurs d'entrée à travers le réseau pour obtenir une sortie
4. calculer un signal d'erreur entre la sortie réelle et la sortie désirée
5. envoyer le signal d'erreur en arrière à travers le réseau
6. corriger les poids pour minimiser l'erreur
7. répéter les étapes 2-6 avec le prochain vecteur d'entrée jusqu'à ce que l'erreur soit suffisament petite.

3.3.1 Prévision de l'ozone par le PMC

Ce travail a été effectué par Yi et Prybutok (1996) [67], ils ont utilisé le modèle PMC avec une couche cachée pour la prévision des maximums de concentration d'ozone dans une région urbaine industrielle.

Variables d'entrée :

- le niveau d'ozone à 9 h du matin
- la température actuelle maximale
- le niveau de dioxyde carbone
- le niveau d' oxydes d'azote (NO_x)
- le niveau de dioxyde d'azote(NO_2)
- le niveau de monoxyde d'azote (NO)
- la vitesse du vent
- la direction du vent

Méthodes de sélection du nombre de neurones H dans la couche cachée :

H=nombre de données d'apprentissage/$(5 \times (I + 0))$, où I est le nombre d'entrées et O est le nombre de sorties.

Prise en compte de phénomènes locaux particuliers

Ce modèle est spécialement adapté aux régions urbaines. Pour les régions rurales, on doit construire un autre modèle. Les données choisies sont extraites

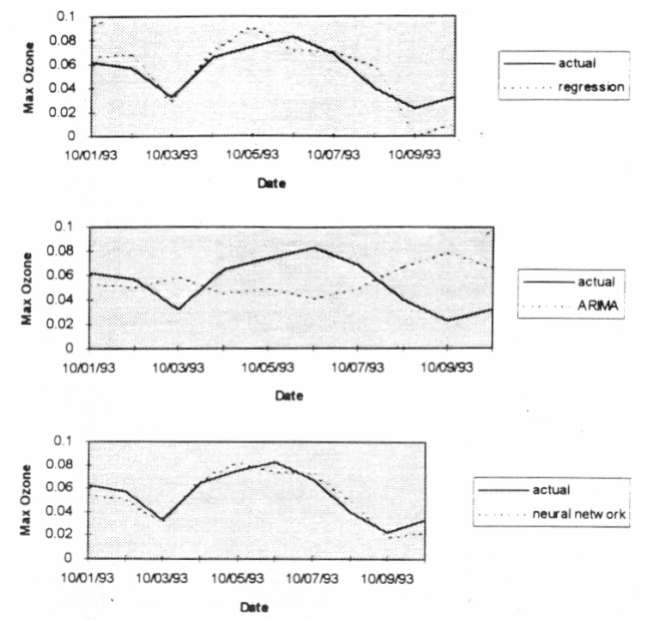


FIG. 17 – Comparaison avec des modèles de régression et des modèles ARIMA

du premier juin au 30 septembre, car la saison d'été représente le cas le plus mauvais, ce modèle étant déconseilleé pour les autres saisons.

Comparaison avec des modèle de régression et ARIMA.

Les résultats de cette technique sont comparés avec autres méthodes, modèle de régression et modèle ARIMA, voir figure 17. Pour comparer ces trois modèles empiriques, on a utilisé le test de Friedman (Siegel et Castellan 1998). Le test de Friedman a indiqué la supériorité significative du modèle de réseaux de neurones par rapport aux modèles de regression et ARIMA avec l'indice de confiance 5% (F=21.748, p-value<0.00). La figure 17 illustre ce résultat.

3.3.2 Prévision à court terme de concentrations de SO_2 par le PMC

Boznar et al (1996) ont présenté la méthode du PMC pour la prévision à court terme de concentrations de SO_2 dans des régions industrielles hautement polluées. Ils ont utilisé un PMC avec une couche cachée pour la station de Zavodnje.

Types de variables d'entrée

- vitesse et direction du vent
- température
- radiation solaire
- humidité
- heure
- concentration en SO_2 prises au moment présent et historique

Sélection des variables d'entrée

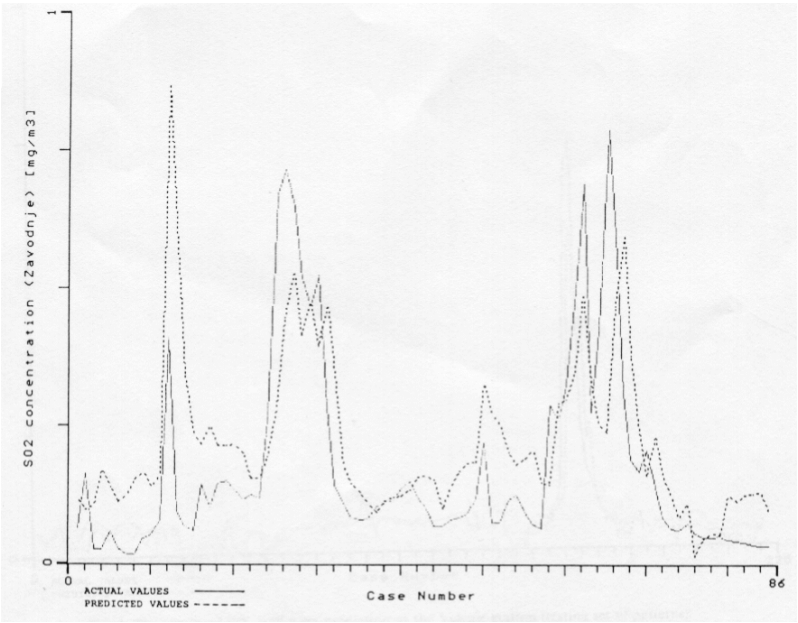


FIG. 18 – Les résultats de la prévision de SO2 à la station de Zavodnje

L'importance de certaines variables peut être déterminée par des études de topographie et la connaissanse des conditions météorologiques locales, ainsi que par comparaison de leurs facteurs de contribution avec les facteurs de contribution des autres variables.

Prise en compte des phénomènes locaux particuliers

L'inversion thermique a été prise en compte. Les résultats montent qu'un réseau de neurones reconnaît facilement les conditions d'inversion.

Indice de confiance dans la prévision

L'illustration de ces résultats est visible sur la figure 18.

3.3.3 Prévision à court terme de concentrations de NO_x et de NO_2 par le PMC

Gardner et Dorling (1999) [28] ont utilisé le PMC avec deux couches cachées pour prévoir les concentrations de NO_x et de NO_2 à Londres. Les oxydes d'azote (NO_x=NO+NO_2) sont émis dans l'atmosphère par l'échappement des véhicules.

Types de variables d'entrée :

- quantité de nuages bas (LOW)
- niveau de nuages le plus bas (BASE)
- Visibilité (VIS)
- température sèche (DRY)
- pression de la vapeur (VP)
- vitesse du vent

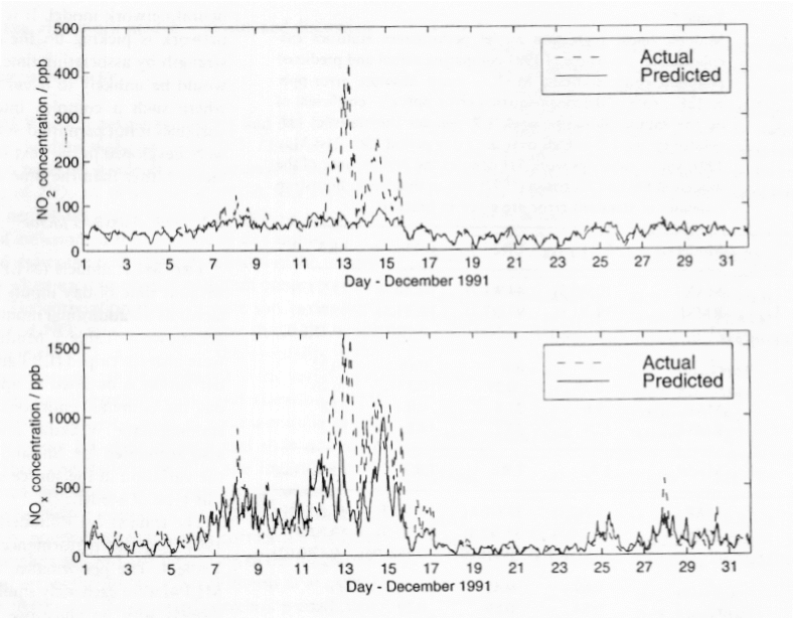


FIG. 19 – Prévision de NO_2 et de NO_x avec le PMC

Prise en compte des phénomènes locaux particuliers :

On a comparé les modèles de PMC avec facteur d'emissions et sans facteur d'emissions. En fait, les résultas des deux modèles avec et sans facteur d'emissions sont extrêmement similaires. On peut conclure que le facteur d'emission peut être remplacé par la donnée de l'heure. La figure 19 illustre ces résultats en utilisant le PMC sans facteur d'emission.

Indices de confiance dans la prévision :

Les erreurs en moyenne absolue (MAE) s'étalent de 9.8 à 38.5 ppb, et les racines carrées des erreurs en moyenne quadratique sont situées entre 18.2 et 78.6 ppb. La corrélation entre la prévision et le réel pour le NO_2 est de 0.62, et de 0.77 pour NO_x.

On voit que le modèle s'adapte bien dans le cas où le niveau de pollution n'est pas très élevé, mais pendant la période du 11 au 17 décembre 1991, les prévisions sont sous estimées, cf. figure 19.

Les résidus du modèle du PMC étant très correlés, Gardner et Dorling (1999) ont construit leurs modèles de PMC en ajoutant la pollution avec une heure de décalage. Ces modèles prédisent pafaitement les concentration NO_2et de NO_x une heure à l'avance, voir figure 21. Mais malheureusement, la prévision une heure à l'avance est peu utile. Par conséquent, des modèles de PMC avec prise en compte de la pollution avec 24 heures de décalage ont été développés. Il est évident que ces derniers modèles sont meilleurs que les modèles du PMC sans variable décalées et moins bons que les modèles du PMC en incluant la variable "pollution" décalée d'une heure. La figure 21 illustre ces résultats.

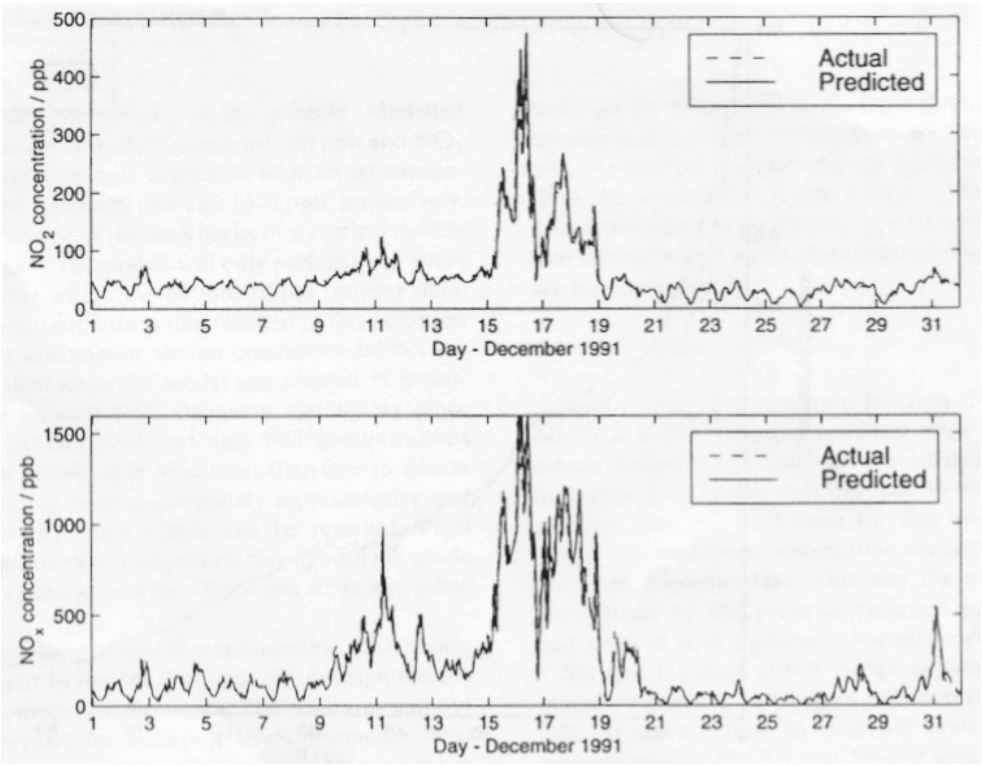


FIG. 20 – Prévision de NO_2 et NO_X en utilisant le PMC avec une heure de décalage

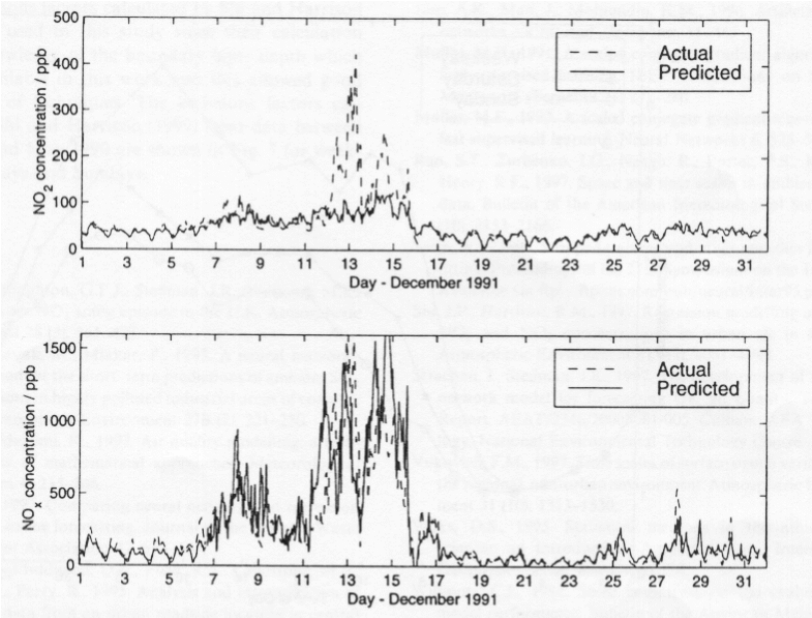


FIG. 21 – Prévision à 24 heures de NO_2 et NO_X en utilisant le PMC

3.4 Modèles BAM et HAM

Le BAM (Bidirectional Associative Memory) et Le HAM (Holographie Associative Memory) ont été utilisés dans le cadre de la prévision de l'ozone à Mexico par Ruiz-Suarez et al (1995) [58]. Nous présentons respectivement ces deux modèles et leurs résultats de la prévision.

3.4.1 Prévision de l'ozone à court terme par les modèles BAM

Le BAM est un système avec deux couches, les informations passent d'une couche à l'autre par une matrice de l'opération. Et puis les informations repassent en arrière par la matrice transposée.

Types de variables d'entrée :

- direction du vent (WD)
- vitesse du vent (WV)
- température (T)
- humidité relative (RH)
- SO_2
- CO
- O_3
- NO_2
- NO_x

Sélection des variables :

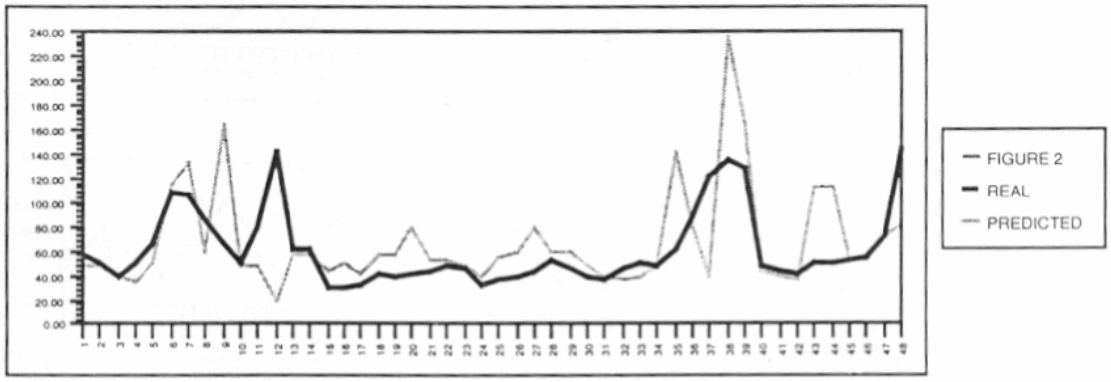


FIG. 22 – prévision de l'ozone par BAM

On a enlevé quelques variables d'entrée ci-dessus, mais on a trouvé que les meilleurs résultats avaient été obtenus en utilisant tous les variables ci-dessus.

Prise en compte des phénomènes locaux particuliers :

On a consideré la corrélation entre diverses stations à Mexico en entrant les données fournies par chaque station considérée.

La figure 22 montre la prévision de la concentration de l'ozone à la station de Pedregal.

3.4.2 Prévision de l'ozone à court terme par les modèles du HAM

Les modèles HAM superposent plusieurs associations stimulus-réponse dans le même neurone. Pour les détails de ces modèles HAM, on peut voir [63]. Notons que les variables d'entrée et la méthode de sélection des variables sont de même type que dans la partie précédente, c'est pourquoi on omet les détails. Par le HAM, on a construit 5 modèles dits HAM_i $(i = 0, 1, ..., 4)$. Les sorties sont les concentration de l'ozone à $t + i$. La figure 23 montre les résultats. La prévision à une heure est relativement bonne. Néanmoins, les résultats deviennent moins bons lorsque l'horizon de prévision est plus lointain.

3.4.3 Conclusion

Les deux réseaux de neurone BAM et HAM sont capables de prédire les concentrations de l'ozone. L'avantage des ces deux modèles est que nous avons une mémoire associative parfaite ainsi qu'une bonne robustesse. Comparativement, le HAM est plus rapide et a plus de capacité de mémoire que BAM. Nous trouvons aussi que le BAM et le HAM sont meilleurs que le PMC.

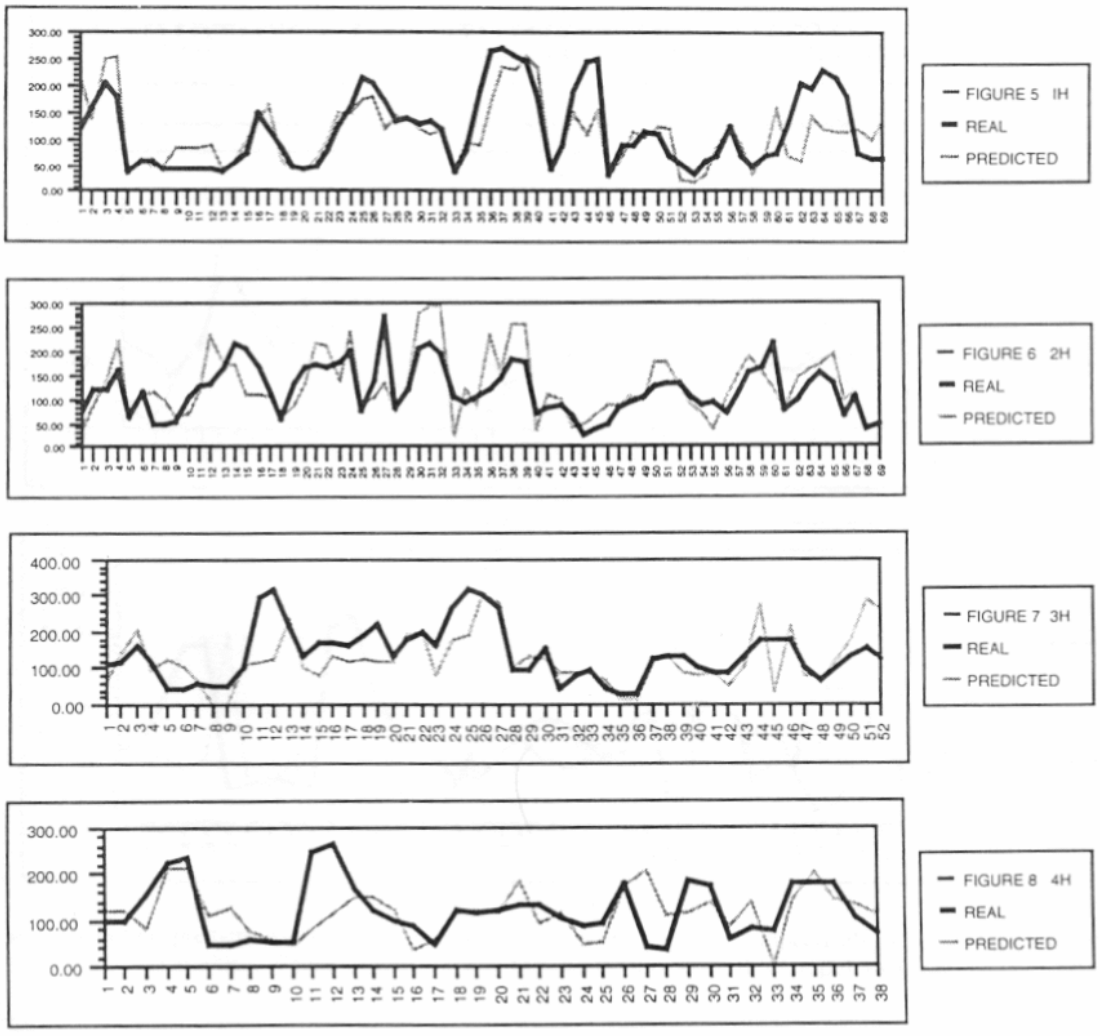


FIG. 23 – Prévision de l'ozone par HAM

4 Analyses discriminantes et méthode CART

4.1 Introduction

Nous abordons ici les méthodes, dites CART -Classification And Regression Trees- qui regroupent, entre autre, les méthodes de classification et de régression par arbre binaire, cette dernière ayant été utilisée dans le programme "Automatic Interaction Detection" développé par Morgan et Sonquist [50]. Une application à la prévision des pics d'ozone a été réalisée par Ghattas [30] et une étude comparative menée à AIRPARIF se trouve dans l'Annexe A.

Ces méthodes sont présentées dans Breiman et al [8]. La classification, descriptive et/ou prédictive, consiste à subdiviser un ensemble d'individus en J classes, chacune portant une étiquette (un label, par exemple l'état d'un phénomène), suivant des caractères ou variables (numériques ou qualitatifs) puis prédire la classe d'un nouvel individu. Par exemple, pour la pollution, classer les jours en 2 classes (ou plus), {non-alerte, alerte}, à l'aide de mesures météorologiques ainsi que celles de certains polluants ; puis prédire la classe du jour suivant en se basant sur les mesures du jour présent.

Donc le but est, d'une part, de produire une règle adéquate de classification (ou de prévision) et d'autre part de découvrir les structures prédictives du phénomène. Si on s'intéresse à ces dernières, c'est qu'on essaye de comprendre comment les variables ou leurs interactions induisent le phénomène en donnant de simples caractéristiques. Mais souvent le but est double : prévision et compréhension. La difficulté majeure est la détermination de cette partition en classes. On doit disposer d'un échantillon qu'on appelle d'apprentissage et, éventuellement, suivant le critère de validation de la partition, adopter un autre échantillon appelé témoin.

La deuxième méthode qui est la régression par arbre est une méthode prédictive relativement plus simple que la précédente. On explique un phénomène par des fonctions simples de variables explicatives.

Ces méthodes sont à la fois descriptives et décisionnelles.Elles présentent plusieurs atouts : elles sont souples par rapport aux dimensions des variables qui peuvent changer d'un individu à l'autre et les valeurs manquantes sont faciles à manipuler.

L'arbre est générée par un algorithme itératif de partition binaire. Chaque nœud est successivement partitionné jusqu'à l'obtention d'un grand arbre dont les nœuds terminaux sont "purs" ou contiennent "peu" d'individus. Cet arbre est ensuite élagué suivant des critères jusqu'à optimalité. Un critère important d'une bonne procédure de classification n'est pas seulement de produire une règle adéquate mais de produire aussi un aperçu et une compréhension de la structure prédictive des données. Que se soit pour la classification ou la prévision, un nouvel individu cheminera dans l'arbre final, suivant la règle obtenue, dans un nœud terminal, il portera alors le label ou la valeur prévue pour ce nœud.

L'idée de la classification et de la régression par arbre binaire est presque la même. On peut dire que la classification est utilisée lorsque la variable à expliquer est entière et la régression quand la variable à expliquer est continue mais le vocabulaire change.

On trouve des aperçus de ces deux méthodes dans [36] et [30] avec des applications dans le domaine médical, pour le premier, et la prévision des pics d'ozone pour le deuxième. Une discussion sur l'implantation de variantes de ces méthodes en S-plus, illustrée par des exemples, est faite dans [65, chapitre 14].

4.2 Classification par arbre binaire

Supposons que chaque individu est representé par un vecteur $x = (x^1, ..., x^k)$ de dimension k, où x^i est la mesure du $i^{ème}$ caractère ou variable et $x \in \mathcal{X}$ une partie de $\mathbb{R}^k$. Pour la simplicité, nous supposerons que tous les individus sont representés par des vecteurs de même dimension ; mais la méthode s'applique aussi bien au cas de dimensions différentes, c'est un des atouts de ces méthodes. Ces individus sont répartis dans J classes, $1, ..., J$. Soit $C = \{1, ..., J\}$ l'ensemble de ces classes.

On utilisera indifférement les termes règle de classification ou règle de prévision. Une règle de classification ou classificateur est une règle d'affectation d'une classe d'appartenance à chaque x de $\mathcal{X}$ i.e. une fonction $d : \mathcal{X} \longrightarrow C$ qui à chaque x de $\mathcal{X}$ associe la classe $d(x)$ de C. Cette définition est équivalente à la décomposition de $\mathcal{X}$ en une partition $(A_j, j = 1, ..., J)$, $\mathcal{X} = \cup_j A_j$ avec $A_j = \{x \in \mathcal{X} : d(x) = j\}$.

Etant donné un espace $\mathcal{X}$ et un ensemble de classes C, le problème principal est de déterminer le "meilleur" classificateur d. La construction de la règle est basée sur le "passé" observé. Cette règle est construite sur la base d'un échantillon, dit d'apprentissage $\mathcal{A} = \{(x_1, j_1), ..., (x_n, j_n)\}$ de n individus.

4.2.1 Mesure de la qualité d'un classificateur

Pour chaque règle d sur $\mathcal{X}$, basée sur $\mathcal{A}$, on définit la quantité $R^*(d)$ qui est le taux d'erreur de d -la proportion de mauvais classements- que l'on peut formuler, d'une manière générale, comme suit :
Soit P une probabilité sur $\mathcal{X} \times C$ et soit un échantillon (X, Y), indépendant de $\mathcal{A}$, tiré avec la loi P, on définit le taux d'erreur de d par

$$R^*(d) = P(d(X) \neq Y)$$

qui peut être estimé

– soit par reclassement des x_i de $\mathcal{A}$ avec d :

$$R(d) = (1/n)\#\{(x, j) \in \mathcal{A} : d(x) \neq j\},$$

où $\#A$ désigne le nombre d'éléments dans A. Mais cette estimation est trop optimiste puisque la plus part des règles tentent de la minimiser.

– soit à l'aide d'un échantillon témoin, $\mathcal{A}' = \{(x'_1, j'_1), ..., (x'_{n'}, j'_{n'})\}$ qui servira à tester la règle construite :

$$R^{tem}(d) = (1/n')\#\{(x', j') \in \mathcal{A}' : d(x') \neq j'\}.$$

– soit par validation croisée : donnons nous une partie $\mathcal{A}_v$ de $\mathcal{A}$ qu'on écartera lors de la construction de d et qui servira d'échantillon témoin pour cette construction, puis on répètera le même procédé pour un certain nombre de sous-échantillons de $\mathcal{A}$ de cardinaux proches de $\mathcal{A}_v$; la qualité sera la moyenne de celle de chacune des estimations, d^v, de d. Ceci supposera qu'on puisse appliquer la même procédure de construction de la règle à tout sous-échantillon. Plus précisément, partageons $\mathcal{A}$ en V sous-échantillons $\mathcal{A}_v$, $v = 1, ..., V$ de tailles comparables. Pour chaque sous-échantillon $\mathcal{A}_v$ on estimera d, par d^v, avec comme sous-échantillon d'apprentissage $\mathcal{A}^v = \mathcal{A} - \mathcal{A}_v$. On estimera la qualité de chaque d^v par

$$R^{tem}(d^v) = (1/n_v)\#\{(x, j) \in \mathcal{A}_v : d^v(x) \neq j\},$$

où $n_v = \#\mathcal{A}_v$ qui est proche de n/V.
Ainsi la qualité de la procédure de construction sera estimée par la moyenne de ces dernières quantités :

$$R^{vc}(d) = (1/V)\sum_{v=1}^{V} R^{tem}(d^v).$$

La méthode bootstrap peut être aussi utilisée pour estimer $R^*(d)$, mais elle n'est pas toujours satisfaisante (cf. [8]).

4.2.2 Règle de Bayes

Soit P une probabilité sur $\mathcal{X} \times C$ et soit un vecteur aléatoire (X, Y) tiré avec la même loi P : on dit que d_B est une règle de bayes si elle est meilleure que toute autre classificateur d i.e.

$$R_B := R^*(d_B) = P(d_B(X) \neq Y) \leq R^*(d) = P(d(X) \neq Y).$$

Dans le cas où $\mathcal{X}$ est un espace de dimension finie, si on note π la probabilité marginale sur C et $f_j(x)$ la densité de $P(X|Y = j)$ alors la règle de Bayes est donnée par (cf. [8, th.1.15, p.14])

$$d_B(x) = j \quad \text{sur} \quad A_j = \left\{x : f_j(x)\pi(j) = \max_i [f_i(x)\pi(i)]\right\},$$

et $R_B = 1 - \int \max_j [f_j(x)\pi(j)]\mathrm{d}\,x$.
Cette expression de d_B était à l'origine de plusieurs méthodes de discrimination basées sur les densités (voir aussi la partie 4.5).
Mais dans la pratique ni π ni f_j ne sont connues ; on les estime alors par des méthodes convenables. Pour plus de détail on peut voir [8], ainsi que ses références.

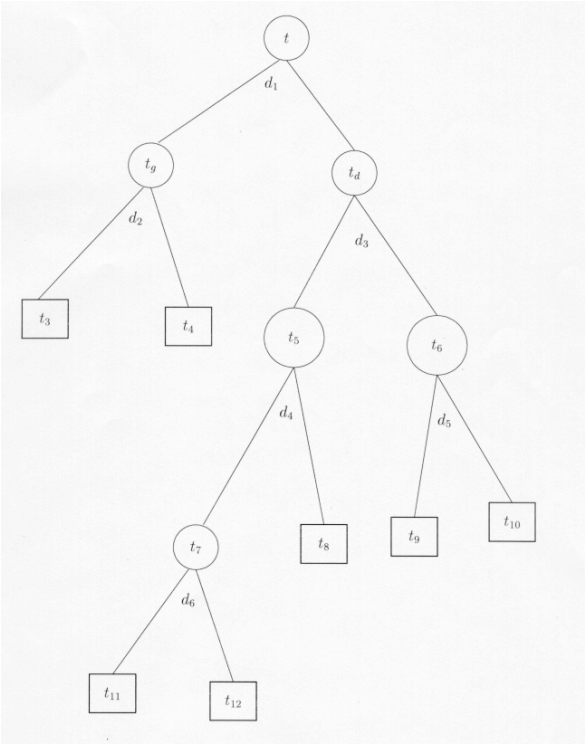


FIG. 24 – Arbre binaire

4.2.3 Classification par arbre : structure et construction

La classification est obtenue en trois étapes :

1. la construction d'un grand arbre binaire sans s'imposer de règles d'arrêt,
2. élaguer ce grand arbre, suivant un critère basé sur le taux de mauvais reclassements de l'arbre et de sa taille, pour obtenir une suite de sous-arbres,
3. utiliser une des mesures de qualité pour en extraire le meilleur sous-arbre.

Construction Un arbre binaire est obtenue par divisions successives de l'échantillons $\mathcal{A}$ en deux descendants (voir figure 24 où $t = \mathcal{A}$). A chaque étape la division est telle que chacun des descendants est plus "homogène". Elle est appliquée à chaque nœud, suivant une des variables (ou une combinaison), pour réduire l'impureté des deux sous-groupes.

On associe à chaque nœud t une mesure d'impureté $i(t)$ qui est une fonction des proportions de chaque classe dans t et vérifiant certaines conditions. La division d est choisie parmi celles qui maximisent la réduction de l'impureté définie par :

$$\Delta i(d,t) = i(t) - p_g i(t_g) - p_d i(t_d),$$

où p_g et p_d sont, respectivement, les proportions des individus de t dans t_g et t_d qui sont les descendants de t ($t = t_g \cup t_d$ et $t_g \cap t_d = \emptyset$). Puis on continue avec les nœuds descendants t_g et t_d jusqu'à ce que les nœuds soient terminaux. Un nœud est dit terminal s'il est pur, i.e. ne contient que des individus d'une même classe, ou "petit" ou ne contient que des individus ayant les mêmes mesures.

A chaque nœud est affecté la classe majoritaire de ses individus.

Exemples :
1) Plusieurs choix de $i(t)$ sont donnés dans la littérature, les plus utilisés sont :
- l'entropie : $i(t) = -\sum_{j=1}^{J} p(j/t) \log p(j/t)$,
- l'index de Gini : $i(t) = \sum_{l \neq j} p(l/t)p(j/t) = 1 - \sum_{j=1}^{J} p^2(j/t)$,

où $p(j/t)$ est la proportion de la classe j dans le nœud t.
2) Les divisions d qu'on considère souvent sont simples et sont de la forme $x^i < c$.

Elagage et choix de l'arbre La procédure d'élagage permet d'avoir des sous-arbres réduits. L'élagage est basé sur l'idée que si une division ou une suite de divisions n'améliorent pas le taux de reclassement par rapport à celui du nœud père elle sera ou elles seront supprimées. Parmi les sous-arbres restants on choisit celui de meilleure qualité suivant l'une des mesures citées plus haut.

Remarque : Contrairement aux autres méthodes de classification qui choisissent un critère d'arrêt, ces méthodes CART développent d'abord un arbre aussi grand que possible puis l'élaguent pour obtenir des sous-arbres qu'on compare à l'aide d'un échantillon test ou par la validation croisée.

4.3 Régression par arbre binaire

Elle s'apparente à la classification binaire dans le principe, sauf qu'au lieu des classes étiquetées on a une variable numérique continue Y, que l'on veut expliquer. Supposons que $\mathcal{A} = \{(x_1, y_1), ..., (x_n, y_n)\}$ où les x_i sont dans $\mathcal{X}$ et les y_i sont des valeurs réelles. De même qu'en classification, on peut définir un prédicteur comme une règle d qui associe à chaque x, de $\mathcal{X}$, un nombre réel $d(x) = y$. La question est de savoir comment construire d à partir de $\mathcal{A}$.
La qualité de l'ajustement est mesurée par la moyenne de l'écart de y à $d(x)$:

$$R^{\varphi}(d) = E\varphi(Y - d(X))$$

pour une certaine fonction positive φ.
Lorsque $\varphi(u) = |u|$, on parle de régression L^1 tandis que pour $\varphi(u) = u^2$ on parle d'erreur des moindres carrées, qu'on notera R^*, et le prédicteur, noté d_B, qui la minimise est la régression des moindres carrées, la plus souvent utilisée. Le

meilleur prédicteur, par rapport à R^*, est $d_B(x) = E(Y|X = x)$. On ne présentera que la régression correspondant à l'erreur R^*, les autres étant analogues. Les mêmes méthodes sont utilisées pour estimer R^*.

4.3.1 Mesure de la qualité d'un prédicteur

Étant donné un échantillon d'apprentissage $\mathcal{A}$, on veut construire le prédicteur d et estimer son erreur R^*. L'erreur peut être estimée :

– soit par comparaison des y_i de $\mathcal{A}$ avec $d(x_i)$:

$$R(d) = (1/n)\sum_{i=1}^{n}(y_i - d(x_i))^2.$$

– soit à l'aide d'un échantillon témoin, $\mathcal{A}' = \{(x'_1, y'_1), ..., (x'_{n'}, y'_{n'})\}$ qui servira à tester le prédicteur construit :

$$R^{tem}(d) = (1/n')\sum_{i=1}^{n'}(y'_i - d(x'_i))^2.$$

– soit par validation croisée : On supposera qu'on puisse appliquer la même procédure de construction du prédicteur à tout sous-échantillon. Soit $\mathcal{A}_v$, $v = 1, ..., V$, de tailles comparables, des sous-échantillons de $\mathcal{A}$. Pour chaque sous-échantillon $\mathcal{A}_v$ on estimera d, par d^v, avec comme sous-échantillon d'apprentissage $\mathcal{A}^v = \mathcal{A} - \mathcal{A}_v$. La qualité de d^v est estimée par

$$R^{tem}(d^v) = (1/n_v)\sum_{(x,y)\in\mathcal{A}_v}(y - d^v(x))^2,$$

où $n_v = \#\mathcal{A}_v$ qui devrait être proche de n/V.
Ainsi la qualité de la procédure de construction sera estimée par la moyenne de ces dernières quantités :

$$R^{vc}(d) = (1/V)\sum_{v=1}^{V} R^{tem}(d^v).$$

4.3.2 Construction de l'arbre binaire de régression

La méthode est similaire à la classification binaire en changeant la réduction de l'impureté dans la division d'un nœud par la réduction de la variation des deux descendants autour de leurs moyennes (si on considère le critère L^2). Soit

$$R(t) = \sum_{x_i\in t}(y_i - \bar{y}(t))^2,$$

avec $\bar{y}(t)$ désigne la moyenne du nœud t. Alors, la division d du nœud t en t_g et t_d est celle qui réduit le plus la variation de t_g et t_d i.e. qui maximise

$$\Delta R(d,t) = R(t) - R(t_g) - R(t_d).$$

La deuxième étape consiste à élaguer le grand arbre pour choisir le meilleur sous-arbre. Une branche est élaguée si elle n'améliore pas significativement la variation par rapport à son nœud père. Enfin, par un échantillon test ou la validation croisée on prilivégie un sous-arbre qu'on retient.

Comme pour la classification on affecte à chaque nœud terminal la moyenne de ses individus qui sera la prévision d'un nouvel individu atteignant ce nœud par la règle obtenue.

Remarque :1) On peut aussi pondérer ces différentes quantités par le nombre d'individus du nœud auquel sont associées, et au lieu de $R(t)$ choisir la variance $s^2(t)$ du nœud.
2) On peut remplacer les carrés par des valeurs absolues et les moyennes par les médianes.

L'interêt de ces méthodes est principalement leur simplicité, et leurs résultats sont faciles à interpréter. Elles traitent aussi bien les valeurs manquantes que les variables de dimensions différentes. Elles sont appliquées avec succés dans divers domaines : social, médical, météorologique, pollution, apprentissage automatique, ... On peut citer par exemple [36], [30], [29], [59], ... Plusieurs variantes ou généralisations sont apparues, ces dernières années, [39, 40], [45], [19], ...

4.4 Applications et généralisations de la méthode CART dans la littérature

La prévision et l'étude de l'évolution de l'ozone a été l'objet et l'application de plusieurs travaux :
- [30] étudie la prévision des pics d'ozone à l'aide de certaines variables de pollution telles que maximum d'ozone du jour précédent, l'ozone à certaines heures et les maximum du dioxyde de soufre et du dioxyde d'azote du même jour et des variables météorologiques. L'arbre est représenté sur la figure 25.

- Différentes méthodes ont été comparées pour l'étude de la concentration d'ozone dans [29] qui montre que la concentration horaire d'ozone nécessite une intéraction non-linéaire entre les variables explicatives pour bien résumer l'évolution de l'ozone. Bien que les méthodes CART soient facilement interprétables, les réseaux de neurones expliquent mieux la relation entre les variables météorologiques, les variables de pollution et la concentration horaire d'ozone (voir également l'Annexe A).

- Un choix bayésien des divisions des noeuds a été proposé par Denison et al [19], voir figure 26.

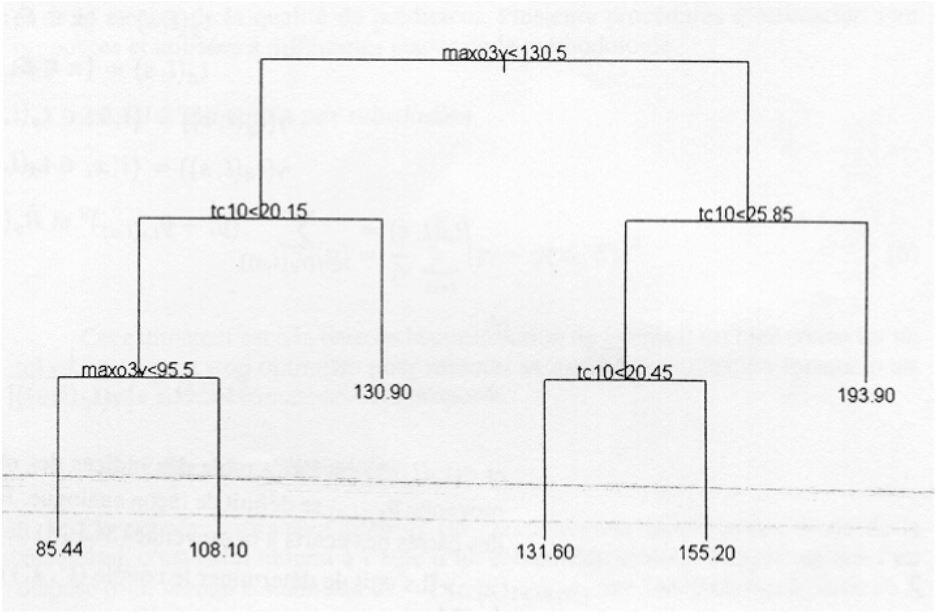


FIG. 25 – variable expliquée (maxo3) : max d'ozone du jour j, variables explicatives (maxo3v, tc10) : max d'ozone du jour $j-1$, température du jour j à 10 h.

4.5 Estimation de densité par directions révélatrices

Cette méthode est expliquée dans [22]. L'expérimentation a eu lieu sur les données de SO_2 fournies par le réseau industriel Alpolair de la région lyonnaise. Les six descripteurs de la situation météorologique sont les suivants :

- pression atmosphérique
- température sèche
- pluie
- direction du vent
- vitesse du vent
- rafales.

On ne cherche pas à prédire le niveau de pollution mais l'accroissement entre le moment présent et l'instant futur : celui-ci a été découpé en deux classes : accroissement supérieur à 10 $\mu g/m^3$ (20.84% des observations) et accroissement inférieur à 10 $\mu g/m^3$ (79.16% des observations).

On note (X, Y) la variable aléatoire observée, où X représente le vecteur des six paramètres météorologiques et Y la classe d'accroissement de la pollution. La fonction qui à une situation météorologique donnée permet de décider dans quelle classe est l'accroissement est notée $g : \mathbb{R}^6 \to \{1, 2\}$ et est obtenue par la règle bayésienne de minimisation de l'erreur

$$L = P(g(X) \neq Y).$$

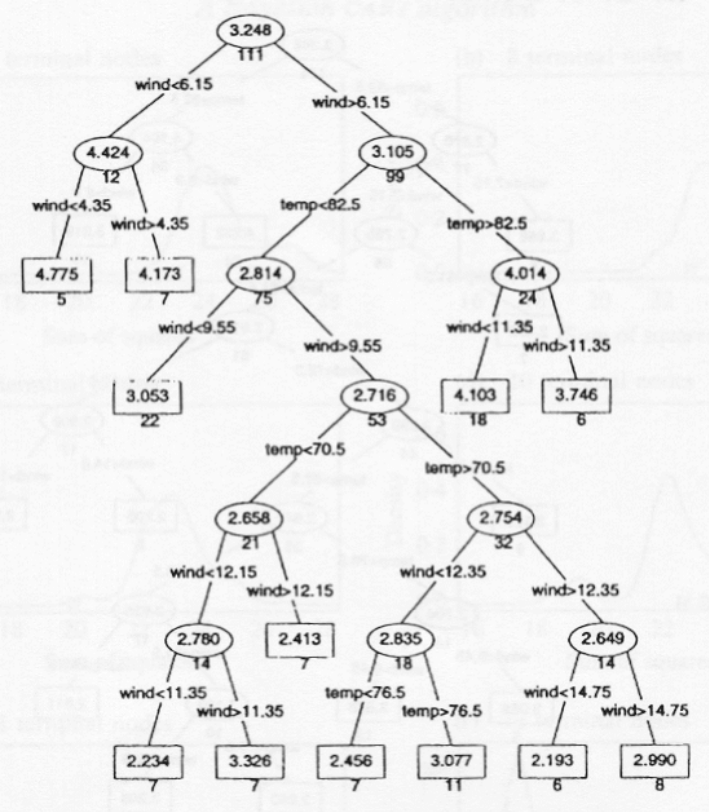


FIG. 26 – Structure d'arbre bayésien, données d'ozone. La moyenne est inscrite et le nombre d'occupants est donné en dessous du noeud terminal

On note f la densité marginale de X, f_i les densités conditionnelles de X sachant $Y = i$, et π_i la probabilité marginale $P(Y = i)$. La règle g minimisant L est celle pour laquelle

$$\pi_{g(X)} f_{g(X)}(x) = \max_{i=1,2} \pi_i f_i(x).$$

Le choix de $\pi_1 = 0.2$ et $\pi_2 = 0.8$ est fait en fonction des observations sur un grand échantillon. Il suffit donc de connaître une estimation des f_i pour pouvoir obtenir la règle g.

En dimension 6 (donc assez élevée), le problème de l'estimation d'une densité est délicat : les méthodes classiques (histogramme et noyau) sont touchées par le phénomène dit de "la malédiction de la dimension" qui dégrade très fortement les estimations avec l'accroissement de la dimension. La méthode adoptée ici est celle des directions révélatrices PPDE (Projection Pursuit Density Estimation). Soit f une densité sur $\mathbb{R}^d$. On approche f par

$$\tilde{f}(x) = g_0(x) \prod_{i \geq 1} g_i(a_i' x)$$

où les a_i sont des vecteurs normés, g_0 une densité sur $\mathbb{R}^d$ et les g_i des fonctions de $\mathbb{R}$ dans $\mathbb{R}$, appelées fonctions correctrices. En fait ce qui est construit, c'est plutôt une suite d'approximations

$$g^{(k)}(x) = g_0(x) \prod_{i=1}^{k} g_i(a_i' x).$$

Pour mener à bien cette recherche, on utilise l'information de Kullback qui mesure la proximité de deux densité f et g :

$$I(f, g) = E_f(\log \frac{f}{g}) = \int f \log \frac{f}{g},$$

où l'intégrale est prise sur le support de f toujours supposé inclus dans le support de g. On introduit également la marge de f suivant une direction quelconque a : si X est une variable aléatoire de densité f, f_a est par définition la densité de la projection $a'X$.

Dans le cas où f est une densité sur $\mathbb{R}^d$ admettant un moment d'ordre 2, on construit succesivement

- $g_0 = g^{(0)}$ la densité gaussienne de mêmes moyenne et variance que f.
- pour chaque $k \in \mathbb{N}^*$, $g^{(k)}(x) = g^{(k-1)}(x) g_k(a_k' x)$, où $g_k = \dfrac{f_{a_k}}{g_{a_k}^{(k-1)}}$ et a_k est tel que $I(f, g^{(k)})$ soit minimum.

Pour l'estimation pratique, on utilise les moments empiriques de f, et pour f_{a_k} il est possible d'utiliser la méthode du noyau puisqu'on est alors en dimension un.

Pour la marge $g_{a_k}^{(k-1)}$, il suffit d'intégrer $g^{(k-1)}$, qui est connue, par une méthode de Monte Carlo afin de gagner en rapidité de calcul.

Deux Théorèmes de consistance [22, p. 43 et p.49] permettent de justifier cette technique.

Les résultats numériques ont été comparés avec les méthodes d'analyse discriminantes linéaire et quadratique, d'estimations à noyau des densités conditionnelles, et des k plus proches voisins.

5 Conclusion

Le choix d'une méthode à appliquer à tel ou tel phénomène est souvent très délicat, et la prévision d'un pic de pollution n'échappe pas à ce principe. Certaines méthodes sont simples et facilement interprétables, comme les modèles linéaires et les méthodes CART mais supposent des intéractions linéaires. Des modèles compliqués, comme les modèles non linéaires ou les réseaux de neurones sont mieux adaptés lorsque des intéractions complexes sont en jeu.

Les méthodes statistiques évoquées dans cette étude sont maintenant assez largement utilisées. On peut observer une accélération de la dynamique de recherche sur ces questions, notamment à l'étranger. Assez fréquemment, plusieurs méthodes sont combinées ou utilisées en parallèle pour produire des prévisions.

Malgré ces progrès méthodologiques et l'augmentation de la puissance des ordinateurs, l'horizon de prévision raisonnable reste au maximum 24 heures actuellement, et la qualité des prévisions n'est pas encore très grande. Sur le long terme, des travaux ont été menés sur l'évolution de la tendance ou du nombre de dépassements de seuils. A cause de l'incertitude des prévisions météorologiques à plus de quelques jours, il semble illusoire d'imaginer faire des prévisions de pics de pollution à long terme.

L'utilisation de variables explicatives est quasi-systématique : quelques modélisateurs préfèrent s'en tenir à la série statistique en expliquant qu'elle contient assez d'information. Les modèles qui semblent donner le plus satisfaction (modèle additif non linéaire, réseaux de neurones, CART, classification) utilisent un certain nombre de variables météorologiques par exemple. Notons qu'une méthode assez générale de choix des variables explicatives est l'examen de la causalité. Pitard et Viel [53] ont utilisé - dans le contexte de la pollution justement - cet outil statistique. Le tableau ci-après résume les caractéristiques de chacun des modèles.

L'Annexe A donne un apperçu de comparaisons de méthodes effectuées dans la littérature. L'agence américaine de protection de l'environnement cf. [25] a également édité un rapport très complet à l'attention des réseaux de surveillance de la qualité de l'air en ce qui concerne la mise en place d'un programme de développement de prévision des pics d'ozone.

Des progrès importants restent à réaliser à la fois sur le plan technique et sur le plan méthodologique. En effet, certaines variables météorologiques sont peu ou

mal mesurées alors qu'elles semblent avoir une influence assez forte (on peut citer la radiation solaire dans le cas de l'ozone par exemple). D'un point de vue statistique, l'étude des modèles non linéaires est relativement récente. La modélisation en temps continu des polluants pourrait permettre de mieux prévoir, grâce à une meilleure représentation de la réalité. Les outils mathématiques correspondants sont actuellement en développement.

Enfin, si sur certains points la recherche française en ce domaine est bien avancée, des progrès ne peuvent venir que d'un renforcement des équipes en France et/ou d'une coopération internationale accrue.

Modèles	Séries temporelles	Régression	Réseaux de neurones	CART
Polluants d'intérêt	NO, SO_2,	ozone,	NO_2, CO	
Variables explicatives	aucune	données météorologiques	données météorologiques	données météorologiques
Limitations	intéractions linéaires, faiblesse pour la prévision des extrêmes	faiblesse pour la prévision des extrêmes	effort de modélisation important	intéractions linéaires
Commentaires	pas de données météorologiques	utilisation facile et classique	intéractions non-linéaires prises en compte	requiert de l'expertise

A Annexe : critères d'efficacité de méthodes

Commençons par citer les critères classsiques de pertinence de prévisions. Notons $(X_i)_{i=1,\dots,n}$ les observations et $\left(\hat{X}_i\right)_{i=1,\dots,n}$ les prévisions. On peut calculer :

– l'erreur en moyenne quadratique MSE (mean squared error)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{X}_i - X_i)^2$$

et la très utilisée racine carrée de cette erreur RMSE (root mean squared error)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{X}_i - X_i)^2}$$

– l'erreur en moyenne absolue MAE (mean absolute error)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{X}_i - X_i\right|$$

– l'erreur en moyenne relative MRE (mean relative error), en pourcentage

$$MRE = \frac{100}{n}\sum_{i=1}^{n}(\hat{X}_i - X_i)/X_i$$

– l'erreur en moyenne relative absolue MRAE (mean relative absolute error), en pourcentage

$$MRAE = \frac{100}{n}\sum_{i=1}^{n}\left|\hat{X}_i - X_i\right|/X_i$$

Récemment, Gardner et Dorling [29] ont comparé trois modèles : le perceptron multicouches, la régression par arbre et le modèle linéaire simple. Le terrain de comparaison est constitué de plusieurs régions britanniques et le polluant est l'ozone. Le choix de l'erreur $RMSE$ s'explique par sa plus grande sensibilité aux extrêmes que la MAE par exemple. Ces comparaisons sont visibles sur la figure 27. La conclusion est que le perceptron donne les meilleurs résultats que les deux autres modèles mais qu'il n'est pas facilement interprétable. De manière imagée, Gardner et Dorling affirment que pour effectuer une prévision il vaut mieux un bon modèle "boîte noire" qu'un modèle bien compris et ayant une assise physique mais relativement faible.

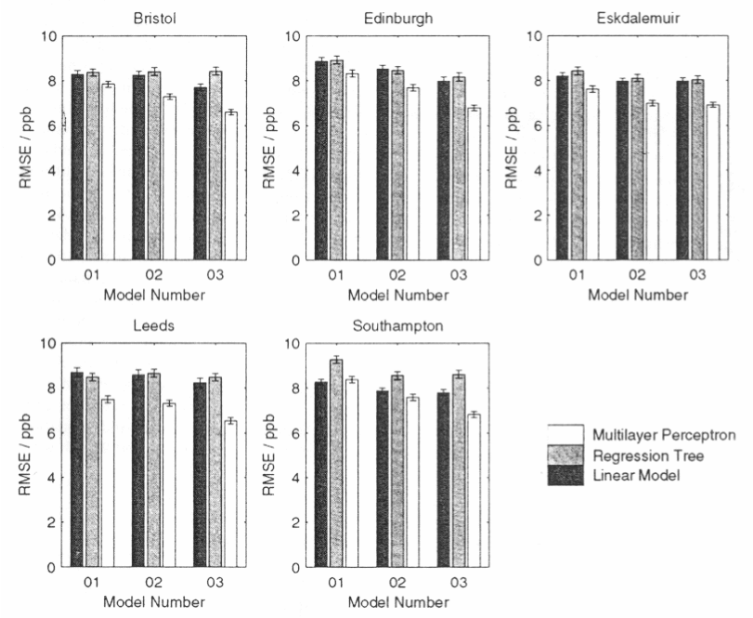


FIG. 27 – Performance relative (en RMSE) des trois modèles

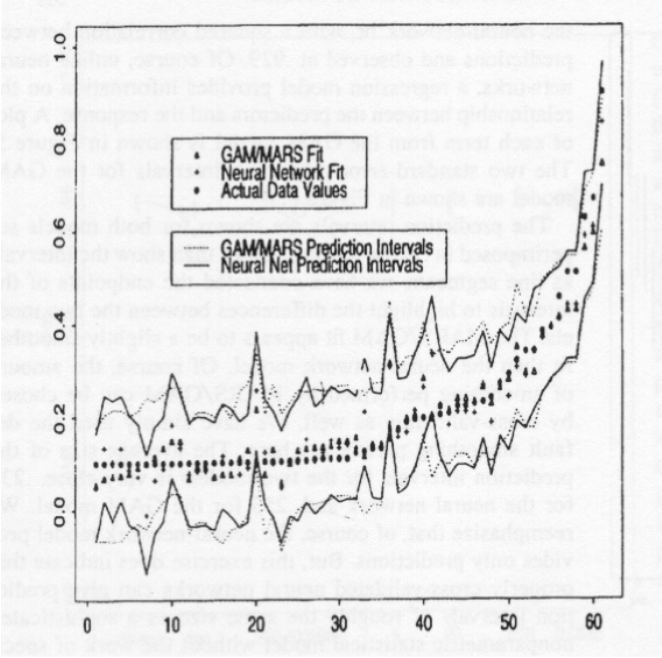


FIG. 28 – Prévision d'intervalles de confiance par réseaux de neurones et par régression non linéaire additive (GAM).

La comparaison avec le modèle linéaire simple n'est pas adéquate, vu que la non linéarité n'est pas prise en compte. Pour une comparaison avec le modèle additif non linéaire, mais avec des données issues de de la mécanique des polymères, voir [21] et la figure 28.

Dans le travail de Bel et al (1999) [3], quatre modèles de prévision des pics d'ozone ont été comparés. Il s'agit du modèle de régression général (4), du modèle non linéaire additif (5), du modèle CART et du modèle classification-discrimination-régression, noté CDR. Les prévisions sont faites à 06 h T.U. le jour j pour l'après-midi du même jour j. Après avoir réglé sur un ensemble de jours, appelé ensemble d'apprentissage, les différents modèles, la comparaison se fait eu égard aux prévisions sur un ensemble de jours identique -bien sûr disjoint de l'ensemble d'apprentissage- , appelé échantillon test.

L'ensemble d'apprentissage va du 01/01/92 au 31/12/96. Plusieurs échantillons test ont été utilisés : des étés entiers, des quinzaines, et un échantillon de 33 jours.

Les critères de comparaison sont de type score, vu la caractère décisionnel du problème. Le tableau suivant est utilisé

		Prévu		
		Niveau 0	Niveau 1	Niveau 2
	Niveau 0	Bonne alerte	FA	FA
Réalisé	Niveau 1	ANF	Bonne alerte	FA
	Niveau 2	ANF	ANF	Bonne alerte

où FA signifie Fausse alerte et ANF signifie Alerte non faite. Ainsi, plus les résultats sont sous la forme d'un tableau diagonal et plus les prévisions sont bonnes. Il est alors possible de calculer grâce à ce tableau plusieurs nombres qui synthétisent la qualité de la prévision en suivant la finalité : on multiplie les résultats de chacune des cases par un coefficient donné.

– un coût uniforme :

	Niveau 0	Niveau 1	Niveau 2
Niveau 0	0	1	1
Niveau 1	1	0	1
Niveau 2	1	1	0

– un coût "santé publique" qui pénalise les erreurs de prévision et avantage les alertes prévues lorsqu'elles concernent des niveaux élevés de pollution :

	Niveau 0	Niveau 1	Niveau 2
Niveau 0	0	1	3
Niveau 1	1	-2	2
Niveau 2	3	2	-3

– un coût "préfecture" qui ressemble au coût "santé publique" mais qui

n'avantage pas les alertes prévues :

	Niveau 0	Niveau 1	Niveau 2
Niveau 0	0	1	3
Niveau 1	1	0	2
Niveau 2	3	2	0

.

On peut également utiliser des taux de réussite. Plus précisément, on compte et on classe les résultats dans le tableau suivant

	Prévu 0	Prévu 1 ou 2
Réalisé 0	t_{00}	t_{01}
Réalisé 1 ou 2	t_{10}	t_{11}

.

Puis on calcule :

- le taux de non détection, c'est à dire le nombre d'alertes non faites sur le nombre total de dépassements de seuils

$$t_1 = \frac{t_{10}}{t_{11} + t_{10}}.$$

- le taux de fausses alertes, c'est à dire le nombre de fausses alertes sur le nombre total de prévisions de dépassements de seuils

$$t_2 = \frac{t_{01}}{t_{11} + t_{01}}$$

- le **threat score**, égal au nombre de prévisions de dépassements de seuils correctes sur le nombre total de dépassements prévus (bien ou non) et non prévus

$$t_3 = \frac{t_{11}}{t_{11} + t_{10} + t_{01}}$$

Plus une méthode est bonne, plus t_1 et t_2 sont petits et plus t_3 est grand.

Les résultats du travail de comparaison effectué dans [3] sont les suivants : Pour les critères de coûts, toutes les méthodes sont assez proches sauf la méthode CART qui est distancée. Pour les critères classiques de mesures d'erreurs et pour les critères t_1 et t_2, la méthode CDR est devant. La méthode additive obtient le meilleur threat score, et permet de faire plus de prévisions. Bien sûr, ces comparaisons dépendent des échantillons test.

Signalons que bien souvent les threat score sont de l'ordre de 30 à 40% et exceptionellement 66%. De plus, les dépassements de seuils les plus élevés (de niveau 2) sont rarement prévus . Ces résultats sont faibles mais comparables aux autres études dans le monde, ce qui démontre la difficulté de ce type de prévisions.

Références

[1] M. Barrat, Y. Lecluse, Y. Slamani (1990), Etude comparative de différents modèles mathématiques pour la prédiction des niveaux de pollution atmosphérique, analyse univariable, R.A.I.R.O. APII **24**, n°3, p. 283-298.

[2] L. Bellanger (1999), Statistique de la pollution de l'air. Méthodes mathématiques. Applications au cas de la région parisienne. Thèse de doctorat de l'Université Paris XI Orsay.

[3] L. Bel, L. Bellanger,V. Bonneau, G. Ciuperca, D. Dacunha-Castelle, C. Deniau, B. Ghattas, M. Misiti, Y. Misiti, G. Oppenheim, J.M. Poggi, R. Tomassone (1999), Eléments de comparaison de prévisions statistiques des pics d'ozone, Revue de Statistique Appliquée **XLVII**, n°3.

[4] C. M. Bishop (1995) Neural networks for pattern racognition. Clarendon Press, Oxford.

[5] G.E. Box, G.M. Jenkins (1976), Time series analysis. Forecasting and control. Holden Day, San Francisco.

[6] M. Boznar, M. Lesjak, P. Mlakar (1993), A neural network-based method for short-term predictions of ambient SO_2 concentrations in highly polluted industrial areas of complex terrain, Atmospheric Environnement **27B**, n°2, p. 221-230.

[7] L. Breiman (1996), Heuristic of instability and stabilization in model selection, The Annals of Statistics, Vol 24, N° 6, 2350-2383.

[8] L. Breiman, J. Friedman, R. Ohlsen, R. Stone (1984), Classification and regression trees. Belmont, CA : Wadsworth.

[9] P.J. Brockwell, R.A. Davis (1991), Times series : theory and methods, second edition Springer-Verlag.

[10] Caroll R.J., Ruppert D. (1988), Transformation and weighting in regression, Chapman&Hall, Londres.

[11] J.L. Chen, S. Islam, P. Biswas (1998), Nonlinear dynamics of hourly ozone concentrations : nonparametric short-term prediction, Atmospheric Environnement **32**, p. 1839-1848.

[12] N. Chèze-Payaud, J-M. Poggi, B. Portier (1998), Un modèle additif pour la prévision de l'ozone à trois échéances, Rapport de contrat de recherche IUT de Paris V-Equipe de statistique d'Orsay.

[13] W.G. Cobourn, M. C. Hubbard (1998), Development of a regression model to forecast ground-level ozone concentration in Louisville, KY, Atmospheric Environnement **32**, p. 2637-2647.

[14] W.G. Cobourn, M. C. Hubbard (1999), An enhanced ozone forecasting model using air mass trajectory analysis, Atmospheric Environnement **33**, p. 4663-4674.

[15] A.C. Comrie (1997), Comparing neural netwoks and regression models for ozone forecasting, Air & Waste Management Association **47**, p. 653-663.

[16] A.C. Comrie, J.E. Diem (1999), Climatology and forecast modeling of ambient carbon monoxide in Phoenix, Arizona, Atmospheric Environnement **33**, p. 5023-5036.

[17] G. Cybenko (1989), Approximation by Superposition of a sigmoide function, Mathematics and control, Signals, and Systems, 2, 303-314.

[18] Davis, J.M., Speeckman P. (1999), A model for predicting maximum and 8 h average ozone in Houston, Atmospheric Environment **33**, 2487-2500.

[19] D. Denison, B. Mallick, A. Smith (1998), A bayesian CART algorithm, Biometrika **85**, 2, p. 363-377.

[20] G. Der Megreditchian, P. Pilibossian (1983), Recherche d'une fonction de régression par le critère minimax, Publications de l'Institut de Statistique des Universités de Paris **XXVIII**, fasc. 3, p. 59-92.

[21] R.D. De Veaux, J. Schumi, J. Schweinsberg, L. H. Ungar,(1998) Prediction intervals for neural networks via nonlinrae regression. Technometrics, **40**, 273-282.

[22] E. Elguero (1988), Estimation de densité par directions révélatrices : une application aux alertes de pollution atmosphérique, Thèse de doctorat, Université des sciences et techniques du Languedoc.

[23] Engle R.F. (1982), Autoregressive conditional heteroscedasticity with estimates of the variance of the United Kingdom inflation, Econometrica, **50**, 987-1006.

[24] F. Engel, C. Viel, H. Chanut (1998), Développement d'un modèle statistique de prévision à 24 heures d'un dépassement du seuil d'information de la population pour l'ozone, Pollution atmosphérique juillet-septembre 1998, p. 59-63.

[25] Environmental Protection Agency (1999), Guidelines for developping an ozone forecasting program, USA, www.epa.gov.

[26] A. Fromage (1996), Prévision des pointes de pollution atmosphérique : état de l'art dans le monde et perpectives pour la région Ile-de-France, Thèse professionnelle, Institut Supérieur d'Ingéniérie et de Gestion de l'Environnement.

[27] M. W. Gardner , S.R. Dorling (1998), Artificial neural networks (the multilayer Perceptron)-a review of applications in the atmospheric sciences, Atmospheric Environnement 32 n°14/15.p 2627-2636.

[28] M. W. Gardner , S.R. Dorling (1999), Neural network modelling and prediction of hourly NO_x and NO_2 concetrations in urban air in London. Atmospheric Environnement **33**, 709-719.

[29] M. W. Gardner , S.R. Dorling (2000), Statistical surface ozone models : an improved methodology to account for non-linear behaviour. Atmospheric Environnement **34**, 21-34.

[30] B. Ghattas (1999), Prévisions des pics d'ozone par arbres de régression simples et agrégés par bootstrap, Revue de Statistique Appliquée **XLVII**, n°2.

[31] W. Gonzalez-Manteiga, J.M. Prada-Sanchez, R. Cao, I. Garcia-Jurado, M. Febrero-Bande, T. Lucas-Domingez (1993), Time series analysis for ambient concentrations, Atmospheric Environnement **27A**, n°2, p. 153-158.

[32] Gouriéroux C. (1997), ARCH models and financial applications. Springer Series in Statistics. Springer-Verlag, New York.

[33] C. Gourieroux, A. Monfort (1990), Séries temporelles et modèles dynamiques, Economica.

[34] M. Graf-Jacottet (1993), A flexible model for ground ozone concentration, Environmetrics **4**, n°1, p. 23-37.

[35] M. Graf-Jacottet, M-H. Jaunin (1998), Predictive models for ground ozone and nitrogen dioxide time series, Environmetrics **9**, p. 393-406.

[36] A. Gueguen et J.P. Nakache (1988), Méthode de discrimination basée sur la construction d'un arbre de décision binaire, Revue de Statistique Appliquée **XXXVI**, n°1, p. 19-38.

[37] Hastie, T.J., Tibshirani R.J. (1990), Generalized additive models, Chapman & Hall, New-York.

[38] B. Killam, B.B. Bhattacharyya (1996), Time series modelling of SO_2 pollution data and the distribution of yearly maximum, Journal of Applied Statistical Science **4**, p. 285-303.

[39] S.H. Kim (1994), A General property among nested, pruned subtrees of a decision-support tree, Commun. Statist.Theory Meth., 23(4), pp.1221–1238.

[40] S.H. Kim (1996), Model selection for tree-Structured Regression, J. Korean Statis. Soc., vol. 25, 1, pp.1–24.

[41] K. Kocak, L. Saylan, O. Sen (2000), Nonlinear time series prediction of O_3 concentration in Istanbul, Atmospheric Environnement **34**, p. 1267-1271.

[42] M.R. Leadbetter (1995), On high level exceedance modeling and tail inference, Journal of statistical planning and inference **45**, p. 247-260.

[43] I.F. Lee, P. Biswas, S. Islam (1994), Estimation of the dominant degrees of freedom for air pollutant concentration data : application to ozone measurement, Atmospheric Environnement **28**, p. 1707-1714.

[44] O.B. Linton, W. Härdle (1996), Estimation of additive regression models with known links, Biometrika **83**, p. 529-540.

[45] W-Y. Loh, N. Vanichsetakul (1988), Tree-structured classification via generalized discriminant analysis, Journal of the American Statistical Association **83**, n°403, p.715-725.

[46] H. R. Maier, G.C. Dandy (1998), Understanding the behaviour and optimising the performance of back-progation neural netwoks : an empirical study, Environmental Modeling &Software 13 179-191.

[47] H. R. Maier, G.C. Dandy (1998), THe effect of internal parameters and geometry on the performance of bac-progation neural networks : an empirical study, Environmental Modeling &Software 13 193-209.

[48] W.S. McColloch et W. Pitts (1943), A logical Calculus of the ideas immanent in nervous activity, Bull. Math. Biophysics 5, 115-133.

[49] G. Mélard, R. Roy (1988), Modèles de séries chronologiques avec seuils, Revue de Statistique appliquée **XXXVI**, n°4, p. 5-24.

[50] J.M. Morgan and J.N. Sonquist (1964), The detection of interaction effects. Ann Arbor : Institute for Social Research, University of Michigan.

[51] M. Nakamura, R. Perez Abreu, V. Perez Abreu (1993), Un modelo estaditico para excedentes de episodios altos de ozono en la ciudad de Mexico, Cienca **44**, p. 397-407.

[52] P. Pérez, A. Trier, J. Reyes (2000), Prediction of $PM_{2.5}$ concentrations several hours in advance using neural networks in Santiago, Chile, Atmospheric Environnement **34**, 1189-1196.

[53] A. Pitard, J.F. Viel (1999), A model selection tool in multi-pollutant time series : the Granger-causality diagnosis, Environmetrics **10**, 53-65.

[54] J.M. Prada-Sanchez, M. Febrero-Bande,T. Cotos-Yanez, W. Gonzalez-Manteiga, L. Bermudez-Cela, T. Lucas-Domingez (2000), Prediction of SO_2 pollution incident near a power station using partially linear models and an historical matrix of predictor-response vectors, Environmetrics **11**, p. 209-225.

[55] G.B. Raga, L. Le Moyne (1986), On the nature of air pollution dynamics in Mexico City-I. Nonlinear analysis, Atmospheric Environnement **30**, p. 3987-3993.

[56] F. Rosenblatt (1957), The perceptron, A perceiving and recognizing automation, Cornell Aeronautical Laboratory Report 85-460-1.

[57] J.C. Ruiz-Suarez, O.Mayora, R.Smith-Perez, L.G.Ruiz-Suarez (1994), A neural network-based predition model model of ozone for México City. In Air Pollution 94, Vol.1. Computational Mechanics Publications, Southampton.

[58] J.C. Ruiz-Suarez, O.A. Mayora-Ibarra, J. Torres-Jimenez, L.G. Ruiz-Suarez (1995), Short-term ozone forecasting by artificial neural networks, Advances in Engineering Software **23**, p. 143-149.

[59] W. F. Ryan (1995), Forecasting ozone episodes in the Baltimore metropolitan area, Atmospheric Environnement **29**, n°17, p. 2387-2398.

[60] A. Sahli (2000), Retour d'expérience en matière de prévision, rapport INERIS pour le LCSQA .

[61] S. Siegel, N.J.Jr. Castellan (1998). Nonparametric statistics, 2nd Ed.. McGraw-HILL. 87-94.

[62] G. Soja, A-M, Soja (1999). Ozone indices based on simple meteorological parameters : potentials and limitations of regression and neural network models. Atmospheric Environnement 4299-4307.

[63] J. G. Sutherland (1992). The holographic neural method. In Fuzzy, Holographic and parallel intelligence, ed. B. Soucek. John Wiley and Sons, New York.

[64] H. Tong (1983), Threshold models in non-linear time series analysis, Lecture Notes in Statistics 21, Springer-Verlag, New-York.

[65] W.N. Venables and B.D. Ripley (1997), Modern Applied Statistics with S-plus, 2^{nd} ed., Springer.

[66] D.S. Wilks (1995), Statistical methods in the atmospheric sciences, Academic Press.

[67] J. Yi, V.R. Prybutok (1996), A neural network model forecasting for prediction of daily maximum ozone concentration in an industrialized urban area, Environnemental Pollution **92**, n°3, p. 349-357.