

## Association entre l'exposition à la pollution atmosphérique et la santé : utilisation des séries chronologiques

Daniel Eilstein<sup>1</sup>, Sophie Larrieu<sup>2</sup>, Véréne Wagner<sup>3</sup>, Abdelkrim Zeghnoun<sup>4</sup> et  
Agnès Lefranc<sup>5</sup>

### Title

Relationships between air pollution and health : a time series study

### Résumé

Au cours des vingt dernières années, l'analyse des liens à court terme entre la pollution atmosphérique et la santé a fait l'objet de nombreux travaux. La méthode la plus souvent utilisée est l'analyse de séries chronologiques. Celle-ci vise à quantifier les liens pouvant exister entre les variations quotidiennes du niveau d'un indicateur d'exposition à la pollution atmosphérique et celles du nombre d'occurrences d'un événement sanitaire (décès, hospitalisations, etc.). Afin d'obtenir une estimation non biaisée de l'association à court terme entre la pollution atmosphérique et la santé, il est nécessaire de prendre en compte dans l'analyse tous les facteurs susceptibles de modifier cette association (ces facteurs doivent présenter des liens avec les niveaux de l'indicateur de pollution atmosphérique et avec l'indicateur sanitaire). L'approche actuelle est une régression de Poisson basée sur un modèle additif généralisé (GAM). Ce dernier recourt à des fonctions de lissage qui permettent d'ajuster le modèle au plus près de la forme des relations entre la variable sanitaire et les variables explicatives. Aussi, sont intégrés au modèle la variable d'intérêt (teneurs atmosphériques d'un polluant) ainsi que la tendance à long terme et la saisonnalité, le jour de la semaine, les jours fériés, les périodes de vacances, les conditions météorologiques (température), les épisodes de grippe, les périodes de pollinisation, etc. Les données sanitaires présentent souvent une surdispersion qui est prise en compte par l'hypothèse d'une distribution quasi-poissonnienne de la variable sanitaire. Les paramètres de la fonction de lissage (*spline* pénalisée) utilisée pour prendre en compte la tendance à long terme et la saisonnalité sont sélectionnés de façon à minimiser l'autocorrélation partielle des résidus. Le modèle de régression permet d'estimer le coefficient associé à l'indicateur de pollution et de calculer un risque relatif. La répétition de telles études de séries chronologiques sur des périodes successives est très utile car elle permet d'assurer une surveillance épidémiologique des liens à court terme entre pollution atmosphérique et santé grâce à des données enregistrées en routine.

**Mots-clés** : Pollution atmosphérique, épidémiologie, séries chronologiques, modèle additif généralisé.

<sup>1</sup>Institut de veille sanitaire, d.eilstein@invs.sante.fr

<sup>2</sup>Institut de veille sanitaire, sophie.larrieu@invs.sante.fr

<sup>3</sup>Institut de veille sanitaire, v.wagner@invs.sante.fr

<sup>4</sup>Institut de veille sanitaire, a.zeghnoun@invs.sante.fr

<sup>5</sup>Institut de veille sanitaire, a.lefranc@invs.sante.fr

### Abstract

During the past twenty years, short-term relationships between air pollution and health have been largely investigated, mainly through time series studies. Their aim is to estimate the association between daily levels of an air pollution indicator and daily numbers of a health event (death, hospital admission, etc.). In order to get a non-biased estimation of this short term association, all factors which can modify this relationship have to be taken into account (these factors are related to air pollution and health events). The current approach consists of using a Poisson regression based on a Generalized Additive Model (GAM). This model fits nonparametric functions to allow for non-linear effects and provide a better estimation of the relationships between the different variables. The following factors are included in the model : long-term trend and seasonality, day of the week, holidays, vacations, temperature, influenza epidemics, pollen counts, and the indicator of air pollution. A quasi-Poisson distribution allows taking into account the frequently observed over dispersion of the health data series. The parameters of the smoothing function (penalized spline) used for trend and seasonality are chosen in order to minimize the partial autocorrelation of the model residuals. This model allows to estimate the parameter of the pollutant in this multivariate model and to estimate a relative risk. Repeating such a study on successive periods is very useful in order to provide a real epidemiologic surveillance on the health risks related to air pollution, based on routinely registered data.

**Keywords :** *Air pollution, Epidemiology, Time series, Generalized Additive Model.*

**Mathematics Subject Classification:** (62M10, 62G08, 62P10)

## 1 Introduction

L'analyse des relations entre l'exposition des populations à la pollution atmosphérique et la santé fait l'objet d'un très grand nombre de travaux. On distingue, aujourd'hui, deux grands types d'effets sanitaires : les effets dits « à court terme », faisant suite immédiatement à l'exposition (conventionnellement : survenant quelques jours au plus après celle-ci) et les effets « à long terme », liés à une exposition chronique à la pollution atmosphérique. Ces derniers sont approchés, en général, à l'aide de méthodes relevant de l'analyse de survie, fondée sur le suivi de cohortes de sujets exposés à différents niveaux de polluants et nécessitent le suivi sanitaire individuel de ces sujets [7, 23, 37, 36]. Pour les effets « à court terme », l'approche la plus répandue actuellement est de type écologique (l'analyse porte sur des données agrégées et non sur des données individuelles) et utilise l'analyse de séries chronologiques [42, 21, 44, 35, 19, 5, 51, 10, 18, 26, 45, 2, 25, 12, 24, 17]. Les séries étudiées sont sanitaires et environnementales. Les variables sanitaires sont, par exemple, le nombre de décès par maladies cardiovasculaires, par maladies pulmonaires ou toutes causes confondues (hors les accidents et les suicides), le nombre d'admissions hospitalières pour les mêmes causes, le nombre de cas d'infarctus du myocarde, de crises d'asthme, le nombre de délivrances de médicaments pour certaines pathologies, etc. [42, 44, 5, 35, 45, 51, 10, 26, 25, 12, 24, 17]. Les variables environnementales sont, classiquement, les niveaux de concentration de polluants dans l'atmosphère, la température (minimale et maximale journalières), l'humidité ou la pression atmosphérique, les niveaux de pollens dans l'air, etc. Les indicateurs de pollution atmosphérique les plus souvent étudiés sont le dioxyde de soufre, les particules, le dioxyde d'azote ( $\text{NO}_2$ ) et l'ozone, notamment, car ils sont mesurés en routine dans la plupart des grandes agglomérations. Les effets propres de la plupart d'entre eux sur la santé ont été bien documentés dans le cadre d'études expérimentales. Chacun de ces

polluants est par ailleurs un « traceur » d'une source ou d'un type de pollution atmosphérique. L'analyse aboutit à l'estimation de risques relatifs (RR) qui quantifient la force de l'association statistique entre les variations de concentrations du polluant et celles de l'indicateur sanitaire.

En 1997, l'Institut de veille sanitaire (InVS) a mis en place un dispositif de surveillance des effets de la pollution de l'air sur la santé de la population française dans neuf villes de la métropole (Bordeaux, Le Havre, Lille, Lyon, Marseille, Paris, Rouen, Strasbourg et Toulouse) suite à la Loi sur l'air et l'utilisation rationnelle de l'énergie de 1996. L'objectif de ce programme est d'étudier les risques sanitaires associés à l'exposition à la pollution atmosphérique et de suivre leurs évolutions dans le temps. Ainsi, depuis sa mise en place, ce programme a permis d'estimer, sur plusieurs périodes consécutives, les relations à court terme entre pollution atmosphérique et santé dans ces 9 agglomérations [19, 18]. Ce programme a pour objectif de fournir des outils méthodologiques pour l'application des recommandations de la surveillance qui, contrairement à la recherche, nécessite le recours à des méthodes « réactives » et à des données facilement accessibles en routine. Ceci explique que l'InVS, pour assurer sa mission de veille sanitaire, ait opté pour une approche écologique par analyse de séries chronologiques. Les données de mortalité (nombres de décès journaliers pour une pathologie donnée dans chacune des villes du dispositif) peuvent être obtenues auprès du Centre d'épidémiologie sur les causes médicales de décès (Cépi-DC), les données d'admissions hospitalières (nombre journalier d'hospitalisations pour une maladie donnée) sont extraites des fichiers du Programme de médicalisation des systèmes d'information (PMSI) généralisé aux hôpitaux publics et privés, géré par les Départements d'information médicale (DIM) de ces établissements. Ces données sont collectées en routine, à des fins épidémiologiques (décès) ou médico-économiques (PMSI) et donc disponibles avec un décalage correspondant à la consolidation des données et à leur mise à disposition. Notons (mais ça n'est pas l'objet de cet article) qu'à partir des RR, il est possible, lorsqu'un ensemble de critères montre une forte présomption de causalité [16], de calculer un nombre de cas (décès, admissions) attribuable à une augmentation donnée du polluant.

Nous nous proposons, dans cet article, de présenter la méthode utilisée dans le cadre de ce programme pour étudier des relations à court terme entre l'exposition à un polluant et l'occurrence d'événements sanitaires. Dans un premier temps, nous présenterons le principe de cette méthode ; puis nous détaillerons le modèle utilisé et la nature des données nécessaires ; enfin, nous illustrerons ces notions par un exemple issu de l'analyse des liens à court terme entre les concentrations en  $\text{NO}_2$  et la mortalité dans l'agglomération de Bordeaux sur la période 2000-2004.

## 2 Principe

L'approche retenue pour étudier l'association à court terme entre les niveaux de la pollution atmosphérique et les comptes journaliers d'un indicateur sanitaire est de type écologique temporel. L'étude repose sur des données agrégées à l'échelle de l'agglomération et non sur des données individuelles. L'unité d'observation est le jour. Par conséquent, les facteurs individuels qui sont constants dans le temps, comme le sexe, ou les facteurs dont la distribution dans la population étudiée ne varie pas d'un jour à l'autre, comme le tabagisme, n'influencent pas la relation étudiée et n'ont pas besoin d'être pris en compte dans l'analyse. En revanche, tous les

facteurs pouvant varier au cours du temps avec les niveaux de l'indicateur d'exposition à la pollution atmosphérique et l'indicateur sanitaire peuvent conduire à un biais dans l'estimation de l'association existant à court terme entre ces deux indicateurs. Le principe de base de la modélisation des liens à court terme entre pollution atmosphérique et santé repose donc sur le contrôle adéquat de ces facteurs dits de confusion.

Les **indicateurs sanitaires** (variables expliquées) tels que la mortalité, les admissions hospitalières, les occurrences de cas pathologiques ou la délivrance des médicaments, généralement enregistrés sur un pas de temps journalier, sont caractérisés par un faible nombre d'événements quotidiens survenant au sein d'une population d'effectif relativement important. Le nombre journalier d'événements sanitaires peut être modélisé par une variable aléatoire dont la distribution suit une loi de Poisson. Ces indicateurs sanitaires présentent aussi des variations temporelles à long terme (tendance), à moyen et court termes (variations saisonnières, hebdomadaires, journalières, . . .), une autocorrélation des comptes journaliers (d'événements sanitaires) et une surdispersion de ces derniers (la variance des données observées est souvent supérieure à la variance théorique et donc à la moyenne de la variable *compte journalier*). Cette autocorrélation et cette surdispersion des données sont habituellement liées à des phénomènes externes à la variable expliquée, identifiés (conditions météorologiques, épidémies de grippe, pollens) mais aussi à des variations saisonnières représentatives de déterminants non mesurés.

Les **indicateurs de pollution** (concentrations journalières) sont aussi soumis à des variations à court, moyen et long termes, dues essentiellement aux variations d'émission de ces polluants et à des facteurs météorologiques intervenant à moyen (saisons) et à court termes. Ils présentent aussi des autocorrélations.

Par ailleurs d'autres facteurs, comme la température, l'humidité relative ou les épidémies de grippe interviennent comme facteurs de confusion potentiels puisque leurs variations temporelles à court terme sont liées à la fois à celles des polluants et à celles des données sanitaires [21, 27, 19, 18]. La prise en compte de ces facteurs environnementaux et temporels permet généralement de stationnariser la série, de réduire l'autocorrélation et la surdispersion.

L'**étude de la relation** entre les indicateurs sanitaires et les indicateurs de pollution doit donc tenir compte :

1. de la tendance à long terme et des variations saisonnières des différentes variables ;
2. des facteurs de confusion potentiels ;
3. de l'autocorrélation des données au sein des séries d'indicateurs sanitaires et environnementaux intervenant à court terme ;
4. de la surdispersion de l'indicateur sanitaire (la mortalité, dans l'exemple qui sera présenté ici).

Dans les premiers travaux du programme [21, 44], et jusqu'à la fin des années 90 l'analyse était basée sur une régression de Poisson et un modèle linéaire généralisé (GLM) [32, 21, 27]. Elle exprimait le logarithme népérien de la moyenne journalière de l'indicateur sanitaire comme une combinaison linéaire des variables explicatives (tendance, variations saisonnières, épidémies de

grippe, température, polluant, etc.) :

$$Y_t \sim \mathbb{P}(\mu_t) \quad \text{et} \quad \mu_t = E(Y_t),$$

$$\ln(\mu_t) = \eta_t = \sum_{j=1}^p \beta_j x_{tj}$$

où  $Y_t$  est la variable à expliquer au temps  $t$  ( $t = 1, 2, \dots, T$ ) et représente l'indicateur sanitaire,  $\mu_t$  est l'espérance de  $Y_t$ , au temps  $t$ ,  $\eta_t$  est le prédicteur linéaire au temps  $t$ . Les  $\beta_j$  sont les paramètres à estimer,  $x_{tj}$  est la valeur du  $j^{\text{ème}}$  facteur explicatif (température, humidité, etc.) au temps  $t$ . Pour des raisons liées au calcul de la vraisemblance, il est d'usage d'attribuer à la fonction de lien la forme d'une fonction logarithme népérien (l'expression formelle de la loi de Poisson s'accorde bien à la forme logarithmique du lien en simplifiant le calcul de la vraisemblance).

Pour prendre en compte une éventuelle non-linéarité entre la variable à expliquer et les variables explicatives, un modèle additif généralisé (GAM) a été utilisé par la suite [11]. Ce dernier modèle n'est en fait qu'une extension du GLM : le prédicteur linéaire du GLM est remplacé par un prédicteur dit additif [14] :

$$Y_t \sim \mathbb{P}(\mu_t) \quad \text{et} \quad \mu_t = E(Y_t),$$

$$\ln(\mu_t) = \eta_t = \sum_{j=1}^p f_j(x_{tj}), \quad (1)$$

$$\text{Var}(Y_t) = \tau \mu_t$$

La relation entre la variance et l'espérance de  $Y_t$  représente l'extra-variabilité poissonnienne (modèle quasi-Poisson). Les **fonctions**  $f_j$  sont des fonctions quelconques d'une ou plusieurs variables. Elles peuvent être paramétriques (polynomiales, trigonométriques, *splines* de régression, etc.) ou non paramétriques (*splines* non paramétriques ou de lissage, fonctions de régression locale pondérée [42, 43]). L'utilisation de telles fonctions résume en grande partie l'intérêt des GAM puisqu'elle permet d'appréhender avec peu d'*a priori* la forme de la relation unissant les variables explicatives et la variable à expliquer.

Actuellement, dans la thématique « air santé », ce sont les fonctions *splines* qui sont les plus utilisées. Rappelons que, si l'on considère une suite donnée de  $k$  nœuds  $(\xi_j)_{j=1, \dots, k}$ , une fonction *spline*  $s(x)$  d'ordre  $m$  est, par définition, une fonction polynomiale par morceaux, de degré  $(m - 1)$  entre deux nœuds successifs. Le niveau de régularité de  $s(x)$  est assuré en imposant un certain nombre de contraintes de continuité ou de nullité sur les dérivées successives de  $s(x)$  aux points-nœuds  $(\xi_j)_{j=1, \dots, k}$ . Ces contraintes varient selon le type de fonctions *splines* construites (puissances tronquées, *B-splines*, *natural-splines*, etc.). Les fonctions *splines* se déclinent en deux groupes : les *splines* de régression et les *splines* de lissage. Les ***splines* de régression** présentent des qualités mathématiques de par la simplicité des calculs et une expression explicite :

$$s(x) = \sum_{j=1}^m \theta_j x^{j-1} + \sum_{j=1}^k \delta_j (x - \xi_j)_+^{m-1} \quad (2)$$

où  $(x - \xi_j)_+ = \max(0, x - \xi_j)$  et  $\varphi = (\theta_1, \dots, \theta_m, \delta_1, \dots, \delta_k)$  est le vecteur des coefficients à estimer selon une méthode de régression linéaire classique. De façon à simplifier la notation,

on définit la base de fonctions de puissances tronquée

$$\begin{aligned} B_j(x) &= x^{j-1}, \quad j = 1, \dots, m \\ B_{m+j}(x) &= (x - \xi_j)_+^{m-1}, \quad j = 1, \dots, k \end{aligned}$$

L'expression (2) peut alors s'écrire comme :

$$s(x) = \sum_{j=1}^{m+k} \varphi_j B_j(x)$$

L'utilisation des bases de fonctions de puissances tronquées peut entraîner certains problèmes de stabilité numérique, rendant parfois délicate la mise en œuvre sur le plan pratique. Une autre base beaucoup plus utilisée que celle des puissances tronquées est celle des *B-splines*. Celles-ci sont souvent préférées pour leurs propriétés numériques et pour leur facilité de mise en œuvre algorithmique [6, 9].

Les *natural-splines* sont des fonctions *splines* avec une contrainte de linéarité imposée aux extrémités, via une dérivée seconde nulle sur les premier et dernier nœuds, ce qui laisse moins de liberté à la fonction mais stabilise la variance des estimations aux extrémités. Les fonctions *splines* de régression souffrent d'inconvénients majeurs. Elles sont sensibles au nombre et à la position des nœuds. Ajouter un nœud supplémentaire ou le déplacer modifie la forme de la relation. De plus, la position des nœuds est souvent distribuée uniformément sur les données. L'ajustement dans les régions avec peu de fluctuations dans les données se fait alors avec le même nombre de paramètres que dans les régions présentant beaucoup de variabilité.

L'autre groupe de *splines* est composé de fonctions non paramétriques : les fonctions *splines* de lissage. Contrairement aux *splines* de régression qui utilisent souvent des méthodes plutôt intuitives ou par essais et erreurs pour déterminer l'ensemble des nœuds, les *splines* de lissage déterminent les nœuds en minimisant un critère bien précis. Celui-ci combine la mesure classique de la qualité de l'ajustement, la somme des carrés des résidus et une mesure du degré de lissage (définie ici par une intégrale) :

$$\frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2 + \lambda \int [f^{(m)}(t)]^2 dt$$

où  $\lambda$  est le paramètre de lissage et  $m$  est fixé et sert à définir le degré des polynômes ajustés. Le paramètre de lissage détermine la flexibilité de l'estimateur. Plus  $\lambda$  est proche de 0, plus l'estimateur est flexible. Plus on augmente la valeur de  $\lambda$ , plus on donne de l'importance à la deuxième partie du critère et, de ce fait, plus l'estimateur est lisse. Il a été montré que l'unique fonction qui minimise ce critère est une *natural-spline* de degré  $2m - 1$  avec des nœuds à chacune des valeurs distinctes de la variable explicative.

Les *splines* de lissage ont l'avantage de se baser sur un critère précis. Cependant, utilisées dans le cadre des GAM, leur emploi peut induire une sous-estimation de la variance quand plusieurs fonctions non paramétriques sont incluses dans le modèle et lorsque les variables sont corrélées [39]. Une alternative combinant à la fois la flexibilité des fonctions *splines* non paramétriques et le cadre bien défini des GLM repose sur l'utilisation de **fonctions *splines* pénalisées**. Ces dernières peuvent être vues comme des *splines* de lissage mais avec plus de

flexibilité dans le choix des nœuds : le nombre et la position des nœuds ne sont pas fixés. L'idée développée par Eilers et Marx [9, 31] consiste à positionner plus de nœuds que nécessaire puis à retirer ceux n'apportant pas suffisamment d'information en pénalisant l'expression des moindres carrés par des différences finies portant sur les coefficients des *B-splines* adjacentes. Le degré étant choisi *a priori*, les *splines* pénalisées ainsi définies, dépendent d'un seul paramètre (de pénalisation).

Le modèle (1) met en relation l'indicateur sanitaire et le polluant en tenant compte de l'ensemble des variables ayant un rôle présumé dans les variations de l'indicateur sanitaire et du polluant. L'indicateur de pollution atmosphérique est introduit dans le modèle sous la forme linéaire en son coefficient :  $\beta X$  (on tient compte de la probabilité associée au test de significativité du paramètre  $\beta$ ). Les autres variables explicatives apparaissent comme fonction *splines* ou sont non transformées, selon le cas (voir tableau 1).

Notons que ce modèle ne modélise pas l'autocorrélation résiduelle. Si de l'autocorrélation persiste dans les résidus, des termes autorégressifs peuvent être introduits dans le modèle.

Ce type de modèle peut être estimé en utilisant la fonction GAM fournie dans le package *mgcv* du logiciel R [38]. Le lecteur pourra se référer à l'ouvrage de Wood [48] pour un exposé détaillé des modèles GAM et des fonctions *splines* pénalisées ainsi que pour leur utilisation à l'aide du logiciel R.

L'estimation du coefficient relatif au polluant permet de calculer le RR. Ce dernier est, par définition, un rapport de probabilité :

$$RR = \frac{\mathbb{P}(M|E)}{\mathbb{P}(M|\bar{E})}$$

Au numérateur figure la probabilité de présenter un événement sanitaire lorsqu'on est exposé au facteur de risque. Au dénominateur, la probabilité de présenter l'événement si l'on n'est pas exposé. Si le RR est significativement supérieur à 1, du point de vue statistique, le facteur correspondant est dit « facteur de risque ». S'il est inférieur à 1, il est dit « facteur protecteur ». Dans le cas d'une exposition à la pollution atmosphérique, dans la mesure où l'absence totale d'exposition ne correspond à aucune situation réelle, le RR se présentera plutôt de la façon suivante :

$$RR = \frac{\mathbb{P}(M|E_2)}{\mathbb{P}(M|E_1)}$$

Avec, au numérateur, la probabilité de présenter un événement sanitaire (M), lorsqu'on est exposé à un certain niveau de polluant (la condition  $E_2$  correspond au niveau de concentration du polluant,  $C_2$ ) et, au dénominateur, la probabilité de présenter l'événement si l'on est exposé au niveau de concentration de référence (la condition  $E_1$  correspond au niveau de concentration de référence,  $C_1$ ).

L'estimation centrale du RR est reliée à celle du coefficient  $\beta$  du polluant par la relation :

$$RR = e^{\beta(C_2 - C_1)}$$

Le RR est exprimé par son estimateur central affecté d'un intervalle de confiance (généralement à 95 %), dont les bornes sont elles-mêmes déterminées par un calcul similaire au précédent à partir des bornes de l'intervalle de confiance de l'estimation du coefficient  $\beta$ .

## 3 Le modèle

### 3.1 Nature des variables introduites dans le modèle

Le modèle exprime le logarithme népérien de l'espérance de la variable sanitaire comme une somme de fonctions des variables *temps*, *jour de la semaine*, *jours fériés*, *épidémies de grippe*, *pollens* (si nécessaire), *température minimale journalière*, *température maximale journalière* et un indicateur de pollution.

La variable *sanitaire* est le nombre d'événements journalier (décès, hospitalisations, etc.). Elle figure, dans le modèle, sous forme de logarithme.

La variable *temps* apparaît dans le modèle comme la série des indices des jours de l'étude. Elle est transformée par une fonction *spline* pénalisée. La modélisation du temps par une telle fonction va permettre de contrôler à la fois la tendance et les variations saisonnières. D'autre part, la durée de la période d'observation étant courte, la population exposée subit peu de variations quantitatives ou qualitatives au cours de cette période. Aussi, fait-on l'hypothèse que la population étudiée varie peu sur le court terme. En fait, la variable *temps* prend en compte la variation (faible) éventuelle de la population, en taille et en structure. Le paramètre de lissage de la fonction *spline* du temps devra permettre de modéliser suffisamment l'impact de ces variables tout en évitant d'intégrer l'effet de la pollution. Dans cette optique, la minimisation de la somme des autocorrélations partielles des résidus guide le choix de ce paramètre de lissage [47].

Le jour de la semaine est également un facteur de confusion : le nombre de décès (plus encore les hospitalisations) présente une variation périodique hebdomadaire. Il en est de même pour les polluants issus du trafic routier. Le *jour de la semaine* figure dans le modèle comme une variable qualitative. Cette variable n'est pas transformée.

Les *jours fériés* sont pris en compte sous forme de variable binaire. Cette variable n'est pas transformée.

Dans certains cas, l'effet des vacances doit également être pris en compte dans le modèle (par exemple, lorsque la variable sanitaire étudiée est le nombre d'hospitalisations car l'activité des hôpitaux est influencée par ce facteur). La variable qualitative *type vacances* sera alors introduite sans transformation dans le modèle (les types sont : Noël, février, Pâques, été, tous-saint).

Les épidémies de grippe ont un impact sur les données de santé, que l'on qualifie de « limitée » à « majeur » selon le sérotype du virus circulant. Aussi, pour chaque période d'épidémie, une variable *grippe* est créée, permettant ainsi d'individualiser des périodes épidémiques caractérisées par des virus de types différents. Ces variables prennent la valeur 0, en dehors de la période d'épidémie et les valeurs de 1 à  $n$ , pendant la période d'épidémie avec  $n$ , durée de l'épisode grippal exprimé en jours. Les variables *grippe* apparaissent sous forme de *B-splines* à 3 degrés de liberté, permettant ainsi de prendre en compte la variabilité inter-épidémie dans l'intensité et la dynamique des effets sanitaires dus à la grippe.

Les variables météorologiques prises en compte dans l'analyse sont les *températures minimales* et *maximales*. Ce sont des variables continues exprimées en degré Celsius. Ces variables

apparaissent sous forme de *natural-splines* à 3 degrés de liberté. La température maximale journalière est affectée d'un retard de un jour, tandis que la température minimale n'est pas décalée. En effet, un certain nombre d'études ont montré que les fortes températures avaient un effet immédiat sur la santé, alors que les faibles températures avaient un effet décalé dans le temps [3]. La température minimale, observée en général la nuit, traduit mieux l'effet des fortes chaleurs estivales car, pendant l'été, les personnes se protègent le jour des plus fortes températures mais ouvrent leur fenêtre la nuit. La température maximale traduit quant à elle mieux l'impact des jours froids car, en période de grands froids, les gens restent confinés la nuit et ne sont exposés aux conditions météorologiques qu'au cours de leurs déplacements diurnes. L'humidité et la pression atmosphériques ne sont habituellement pas intégrées dans le modèle car leur prise en compte n'apporte pas d'information supplémentaire par rapport à celle déjà intégrée au modèle au travers des variables températures minimales et maximales.

Parfois, les périodes de pollinisation peuvent également être prises en compte, en particulier lorsque l'événement sanitaire étudié est un indicateur de pathologie respiratoire. Dans ce cas, les comptes journaliers des grains de pollens pour les taxons pertinents (possédant des propriétés allergisantes) sont sélectionnés selon la zone étudiée et inclus dans le modèle sous forme d'une variable sans transformation.

Le *polluant* est une variable continue exprimée en  $\mu\text{g}/\text{m}^3$  car son effet est supposé linéaire et sans seuil. Il apparaît sous forme d'un terme linéaire égal à la moyenne des niveaux de polluant du jour même et de la veille. Ce décalage est sélectionné *a priori*. En effet, l'attitude classique, abandonnée à présent, consistait à tester un ensemble de décalages (de 0 à 5 jours) mais ceci augmentait l'erreur de première espèce associée à la réalisation de tests multiples. Le choix de ce décalage, supposant des effets similaires des niveaux de polluants le jour et la veille du décès sur le risque de décès, est par ailleurs biologiquement vraisemblable, et le plus fréquemment étudié dans la littérature.

Ci-dessous (Tableau 1) figurent les variables du modèle et leur transformation.

### 3.2 Source des données et élaboration des indicateurs

Les données concernant les niveaux de polluants sont fournies par les Associations agréées de surveillance de la qualité de l'air (Aasqa). Pour reconstruire l'exposition de la population à la pollution atmosphérique ambiante, l'hypothèse de départ est que les mesures enregistrées par l'Aasqa sont une estimation non biaisée de la moyenne des valeurs individuelles de l'exposition des individus composant la population d'étude [34]. Dans la pratique, cette hypothèse se traduit par l'utilisation de la moyenne arithmétique des valeurs journalières enregistrées par différentes stations de mesure situées sur la zone géographique étudiée (les valeurs journalières d'une station étant, elles-mêmes, les moyennes des 24 mesures horaires enregistrées par la station). Cette moyenne constitue l'indicateur d'exposition ambiante qui, chaque jour, est attribué à l'ensemble de la population. Le calcul de cette moyenne est d'autant plus licite que les niveaux de pollution enregistrés par l'ensemble des stations de mesure sont homogènes dans le temps et dans l'espace [19, 18]. Ainsi, lorsqu'on considère deux séries de mesure du polluant, le percentile 75 de l'une doit être supérieur au percentile 25 de l'autre et la différence entre les moyennes ne doit pas excéder  $15\mu\text{g}/\text{m}^3$ ; de plus, le coefficient de corrélation, pour deux séries données, doit être supérieur à 0,6 [15, 4, 50].

TABLEAU 1 – Variables du modèle initial, nature, transformation et décalage.

Variable	Nature	Transformation	Décalage(jours)
<i>variable sanitaire</i> <sup>(a)</sup>	discrète	logarithme népérien	-
<i>temps</i>	discrète	<i>p-spline</i>	-
<i>jour de la semaine</i>	qualitative	aucune	0
<i>jours férié</i>	binaire	aucune	0
<i>vacances scolaires</i>	qualitative	aucune	0
<i>grippe</i>	discrète	<i>B-spline</i>	0
<i>température minimale</i>	continue	<i>n-spline</i>	0
<i>température maximale</i>	continue	<i>n-spline</i>	1
<i>compte pollinique</i>	continue	aucune	0
<i>polluant</i>	continue	aucune	0-1

(a) En fait c'est l'espérance de la variable nombre de cas incidents qui apparaît dans le modèle.

Le nombre de décès est fourni par le Centre d'épidémiologie sur les causes médicales de décès (CépiDC (Inserm)).

Les périodes d'épidémie grippale sont fournies par la coordination nationale des Groupes régionaux d'observation de la grippe (Grog). Ces périodes d'épidémies sont définies au moyen de données d'activité en médecine de ville (consultations pour syndrome grippal) et de données virologiques (validation de la présence du virus de la grippe dans les prélèvements effectués sur les patients par test rapide ou analyse dans un centre de référence).

Les données météorologiques proviennent de Météo France.

Enfin, les comptes polliniques sont fournis par le Réseau national de surveillance aérobiologique (RNSA).

### 3.3 Les outils de l'analyse

À chaque étape de l'élaboration du modèle, un ensemble d'outils (analyse des résidus, comparaison des séries observée et prédite par le modèle) est utilisé afin de déterminer la structure du modèle la mieux adaptée aux données.

La série initiale (nombre journalier de cas incidents ou de décès) est plus ou moins fortement autocorrélée. La prise en compte des variations temporelles et des facteurs de confusion potentiels permet de réduire significativement cette autocorrélation. Au fur et à mesure de la progression de l'analyse, l'autocorrélation doit donc diminuer dans les résidus du modèle. La somme des autocorrélations doit s'approcher au maximum de 0 et, idéalement, en fin d'analyse, la fonction d'**autocorrélation partielle des résidus** doit être celle d'un bruit blanc. L'utilisation des autocorrélations partielles est basée sur le fait qu'elles permettent de déterminer l'ordre  $p$  d'un processus autorégressif,  $AR(p)$  : l'autocorrélation partielle au retard  $k$  doit être nulle pour  $k > p$ . Notons encore que l'examen de la fonction d'autocorrélation partielle des résidus permet de déterminer si le modèle n'a pas été sur-paramétré ce qui se traduirait par de l'autocorrélation négative dans les résidus. D'autre part, le critère d'Akaike (AIC) permet, en comparant plusieurs modèles, de dire si le nombre de paramètres rajoutés dans un modèle réduit « de façon importante » la déviance ou non.

L'observation du **graphe des résidus** permet de vérifier qu'il ne persiste pas, après modélisation, une structure particulière (tendance, saisonnalité). Le graphe des résidus *versus* les valeurs prédites peut aussi permettre d'observer la structure des résidus.

La **comparaison des graphes des valeurs observées et prédites** permet visuellement de juger de la qualité du modèle et de compléter l'information apportée par le **Chi-2 d'adéquation**. Cela permet ainsi de déterminer si un brusque changement de la variable sanitaire, comme la fermeture de services pendant les mois d'été ou les vacances scolaires, a bien été pris en compte. Si tel n'était pas le cas, une variable binaire pour cette période particulière peut être utilisée en relation avec une fonction de lissage spécifique.

L'**effet partiel** de chacun des facteurs sur la variable sanitaire peut être mis en évidence de façon graphique. Les graphes des effets partiels permettent de vérifier la cohérence des effets, notamment au regard des connaissances épidémiologiques et biologiques accumulées par ailleurs.

### 3.4 Choix du modèle

Les critères de sélection du modèle sont basés sur un compromis entre ce que l'on sait, d'après la littérature [42, 44, 36, 19, 45, 18] des effets de la pollution atmosphérique sur la mortalité (choix des variables à inclure dans le modèle, des retards affectés à ces variables, . . .), les critères de comparaisons classiques (AIC, ANOVA, . . .) et l'analyse des résidus. Ces derniers doivent être assimilés à un bruit blanc gaussien, *i.e.*, de moyenne zéro, de variance constante, et sans autocorrélations résiduelles significatives. L'autocorrélation résiduelle est particulièrement influencée par le choix du paramètre de lissage de la fonction *spline* associée à la variable « temps ». Celui-ci est choisi en utilisant un algorithme permettant de minimiser la valeur absolue de la somme des autocorrélations partielles des résidus. La normalité des résidus est vérifiée graphiquement (histogramme, QQPLOT) et en utilisant le test de Shapiro-Wilk. La significativité statistique des autocorrélations partielles est vérifiée, en pratique, graphiquement, sur le corrélogramme partiel, en utilisant l'intervalle de confiance associé à chaque retard de l'autocorrélation ce qui est équivalent au test de significativité de chaque autocorrélation.

Notons que le modèle adopté ici est le fruit d'un ensemble de réflexions antérieures menées

par d'autres et par nous-mêmes, après avoir testé d'autres formes de relations entre les variables : régression en composantes saisonnières reposant sur l'utilisation d'une somme de fonctions trigonométriques [27], ajout d'une variable humidité supplémentaire dans le modèle, terme d'interaction température-humidité, ajustement de la variable grippe et des variables températures à différents retards, transformation de la variable jour de la semaine à l'aide d'une fonction *spline*, etc. [19, 18].

## 4 Application

Nous illustrons cette approche par l'analyse des relations à court terme entre le dioxyde d'azote ( $\text{NO}_2$ ) et la mortalité « toutes causes non accidentelles et hors suicide » sur la période 2000-2004 au sein de la population habitant dans l'agglomération de Bordeaux, constituée de 23 communes.

### 4.1 Statistique descriptive

L'effectif de la population étudiée est de 604 238 habitants (recensement 1999), dont 15,7% de personnes âgées de 65 ans et plus. La période d'étude s'étend du 1er janvier 2000 au 31 décembre 2004, en excluant toutefois l'épisode de canicule de 2003 (1<sup>er</sup> au 20 août) du fait du caractère exceptionnel de cette période à la fois en termes de pollution, de climat et de mortalité. La moyenne, la médiane, les 25<sup>ème</sup> et 75<sup>ème</sup> centiles de la distribution des concentrations du  $\text{NO}_2$  sont respectivement égaux à  $22,2 \mu\text{g}/\text{m}^3$ ,  $20,8 \mu\text{g}/\text{m}^3$ ,  $14,5 \mu\text{g}/\text{m}^3$  et  $28,3 \mu\text{g}/\text{m}^3$  sur l'ensemble de la période. Le nombre journalier moyen de décès toutes causes hors accidents et suicides pour toute la population est de 12,4 (écart type : 3,8) et s'étend de 3 à 29 au cours de la période étudiée.

Les figures 1 à 3 illustrent les caractéristiques de l'indicateur sanitaire étudié, à savoir la mortalité toutes causes dans l'agglomération de Bordeaux. La figure 1 représente le nombre journalier de décès pour l'ensemble de la période d'étude et sa moyenne mobile sur 7 jours ; on constate une composante saisonnière importante, avec un nombre de décès plus élevé en hiver qu'en été. La figure 2 représente le corrélogramme partiel de cet indicateur, et montre une forte autocorrélation en particulier sur les 3 premiers retards due à des facteurs extrinsèques. Enfin, la figure 3 montre que la mortalité journalière a une distribution assez proche de la densité de probabilité d'une loi de Poisson de paramètre  $\lambda = 12,4$  (correspondant au nombre moyen de décès journaliers observés).

La figure 4 montre que l'indicateur de pollution est soumis à des variations saisonnières encore plus marquées que pour la mortalité, avec des concentrations journalières en  $\text{NO}_2$  beaucoup plus importantes en hiver qu'en été, ce qui illustre bien l'importance d'ajuster correctement sur la saisonnalité.

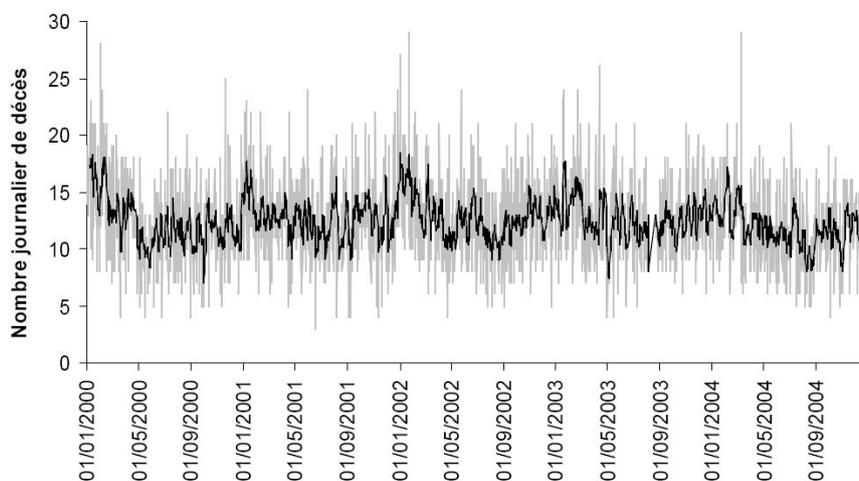


FIGURE 1 – Nombre journalier de décès toutes causes hors accidentelles dans l'agglomération de bordeaux et moyenne mobile sur 7 jours, 2000-04

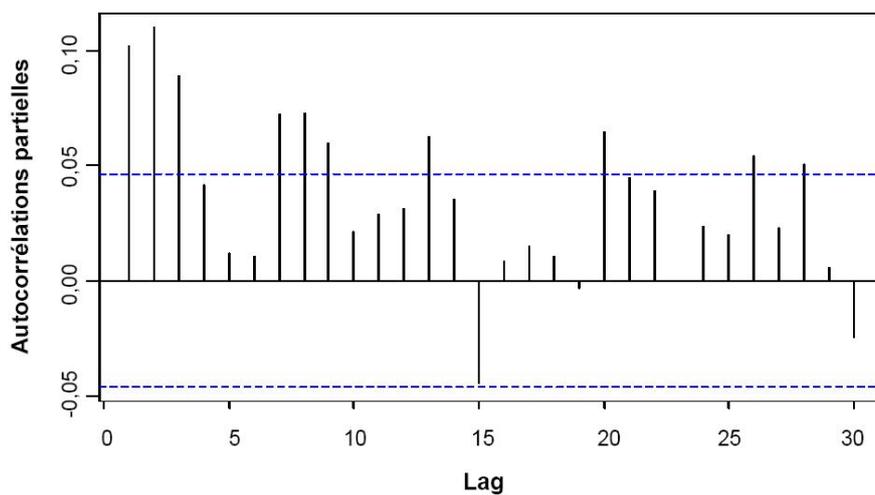


FIGURE 2 – Auto-corrélogramme partiel de la mortalité journalière

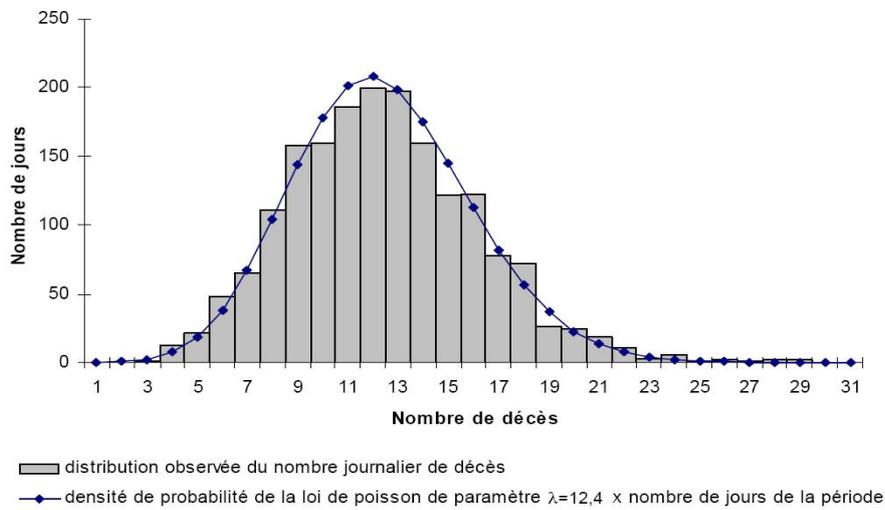


FIGURE 3 – *Distribution des jours de l'étude selon le nombre de décès journalier dans l'agglomération de Bordeaux, 2000-24*

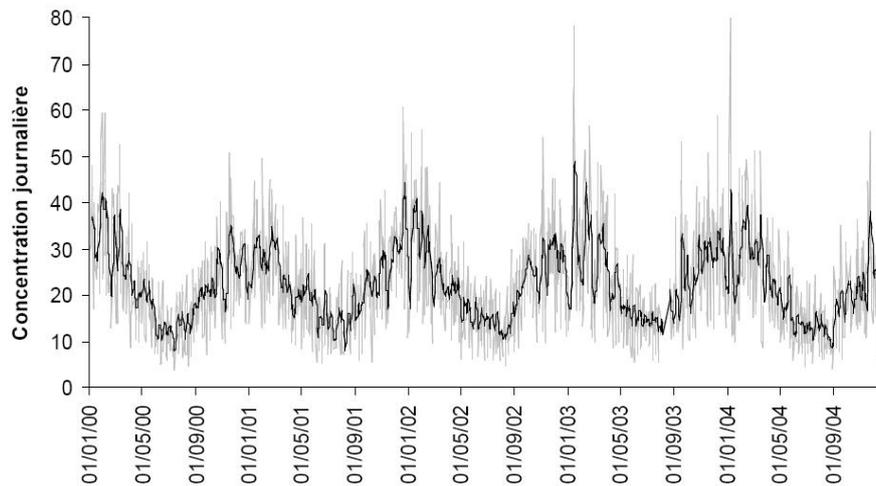
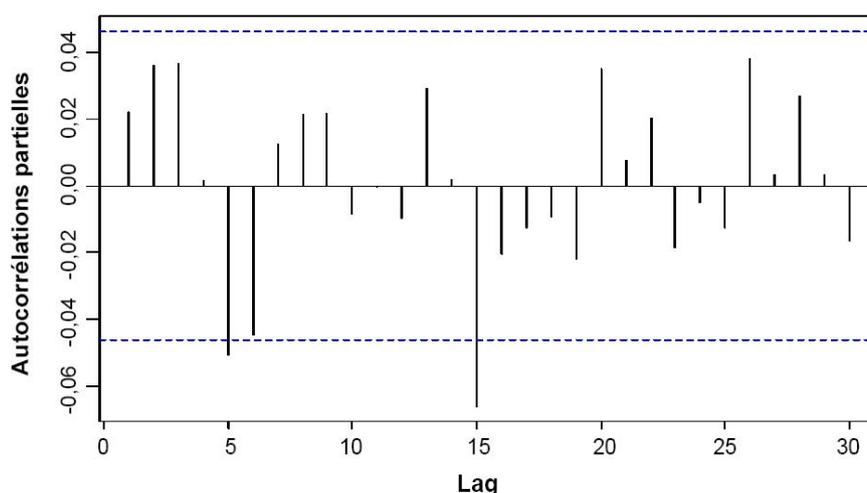


FIGURE 4 – *Concentration journalière en NO<sub>2</sub> (en  $\mu\text{g}/\text{m}^3$ ) mesurée dans l'agglomération de Bordeaux, 2000-04.*

FIGURE 5 – *Corrélogramme partiel des résidus du modèle.*

## 4.2 Analyse

Les analyses ont été réalisées au moyen de la version 2.5.0 du logiciel R, en utilisant le *package* *mgcv*. Rappelons que l'indicateur de pollution ainsi que les différentes variables d'ajustement sont prises en compte de différentes façons selon leur effet supposé sur la mortalité (tableau 1).

Néanmoins, si la qualité du modèle n'avait pas été satisfaisante (du fait notamment de la persistance d'autocorrélations des résidus élevées), des modèles alternatifs auraient pu être envisagés.

Une fois le modèle construit, plusieurs éléments permettent de vérifier son adéquation. Tout d'abord, le Chi-deux ou le rapport du Chi-deux au nombre de degrés de liberté (paramètre de dispersion), ce dernier devant être proche de 1. Dans le cas présent, il était de 1,0409. D'autre part, la somme des autocorrélations partielles des résidus doit s'approcher de zéro (avec le modèle construit pour l'exemple présenté ici, celle-ci était égale à 0,02). C'est une condition nécessaire mais pas suffisante. Idéalement, la fonction d'autocorrélation partielle des résidus doit tendre, en fin de modélisation, vers celle d'un bruit blanc. Celle-ci est visualisée à l'aide d'un corrélogramme partiel représenté dans la figure 5. En dehors d'une faible autocorrélation au retard 15, toutes les PACF, ici, sont inférieures à 0,1 et non significatives. L'observation du graphe des résidus et de leur distribution permet également de vérifier l'adéquation du modèle. La figure 6 présente les résidus obtenus pour chaque jour de la période d'étude, ainsi que leur distribution. La représentation graphique des résidus ne semble révéler aucune tendance ou saisonnalité particulière et leur distribution semble conforme à une distribution normale centrée réduite (test de Shapiro-Wilk :  $p = 0,42$ ).

L'analyse des résidus suggère donc que l'on est en présence d'un bruit blanc et que, *a priori*, l'ensemble des facteurs ayant un effet sur la mortalité a bien été pris en compte dans la modélisation. La comparaison des graphes des valeurs observées et prédites peut également aider à juger de la qualité de la modélisation. Cette comparaison ne se fait pas de façon absolue

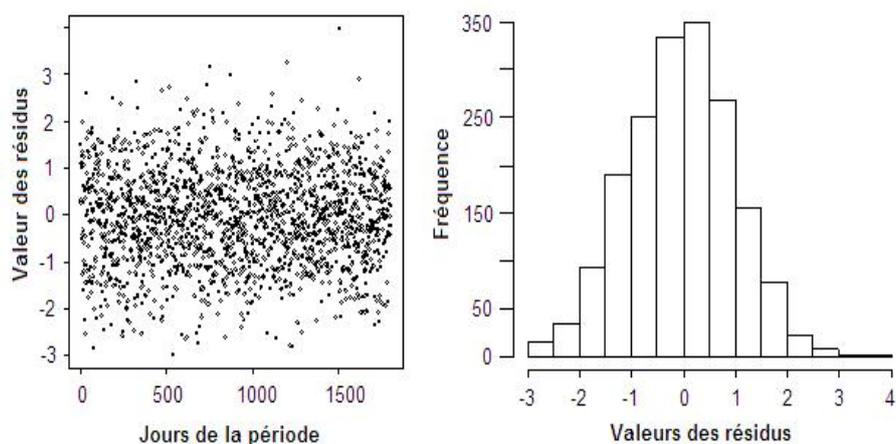


FIGURE 6 – Représentation graphique (à gauche) et distribution (à droite) des résidus du modèle.

car les données observées sont beaucoup plus dispersées que les valeurs prédites, mais elle peut permettre de vérifier que les grandes tendances temporelles, et notamment saisonnières, ont bien été prises en compte. Enfin, l'effet partiel de chaque variable, dont le polluant, peut être visualisé graphiquement, comme le montre la figure 7 (« grippe 1 », « grippe 2 », ... « grippe 5 ») sont des variables représentant des épisodes de grippe distincts). Ainsi, le coefficient obtenu pour chacun de ces facteurs ainsi que la probabilité associée permettent de mesurer l'allure et la force de leur association avec la mortalité et d'éliminer éventuellement ceux qu'il ne serait pas pertinent de conserver dans le modèle. Ainsi, le nombre de ddl et la statistique F pour chacun de ces facteurs, ainsi que la probabilité associée, permettent de mesurer la significativité de leur association avec la mortalité (tableau 2).

Lorsque le modèle ne semble pas correctement ajusté (ce qui n'est pas le cas ici), et notamment s'il reste de l'autocorrélation dans les résidus, plusieurs modifications peuvent être effectuées pour diminuer celle-ci. Parfois, il est nécessaire d'élargir les périodes d'épidémies de grippe, la survenue d'une infection grippale pouvant entraîner à terme le décès avec un certain décalage. Ce phénomène est modélisé par l'extension des périodes de grippe quelques semaines au-delà des périodes épidémiques correspondant à la circulation du virus proprement dit. Les décalages choisis *a priori*, notamment pour la température, peuvent également être modifiés. L'ajout de facteurs supplémentaires tels que des variables climatiques (humidité, pression) peut également être testé et discuté. Enfin, si des pics d'autocorrélation persistent sur les premiers retards en dépit de tous ces ajustements, il est également possible d'intégrer dans le modèle un terme autorégressif afin de les réduire.

Une fois que tous les paramètres ont été correctement pris en compte, le risque relatif (RR) associé à une augmentation de polluant peut être estimé à partir du coefficient obtenu dans le modèle. Par convention et dans un souci de comparaison avec les autres études, ce RR est généralement estimé pour une augmentation de  $10 \mu\text{g}/\text{m}^3$  des niveaux du polluant le jour et la veille du décès. Dans cet exemple, le coefficient, relatif au polluant, est trouvé égal à 0,004 ( $\sigma = 0,001$ ), et le RR estimé est donc de 1,041 avec un intervalle de confiance (IC) à 95 % de

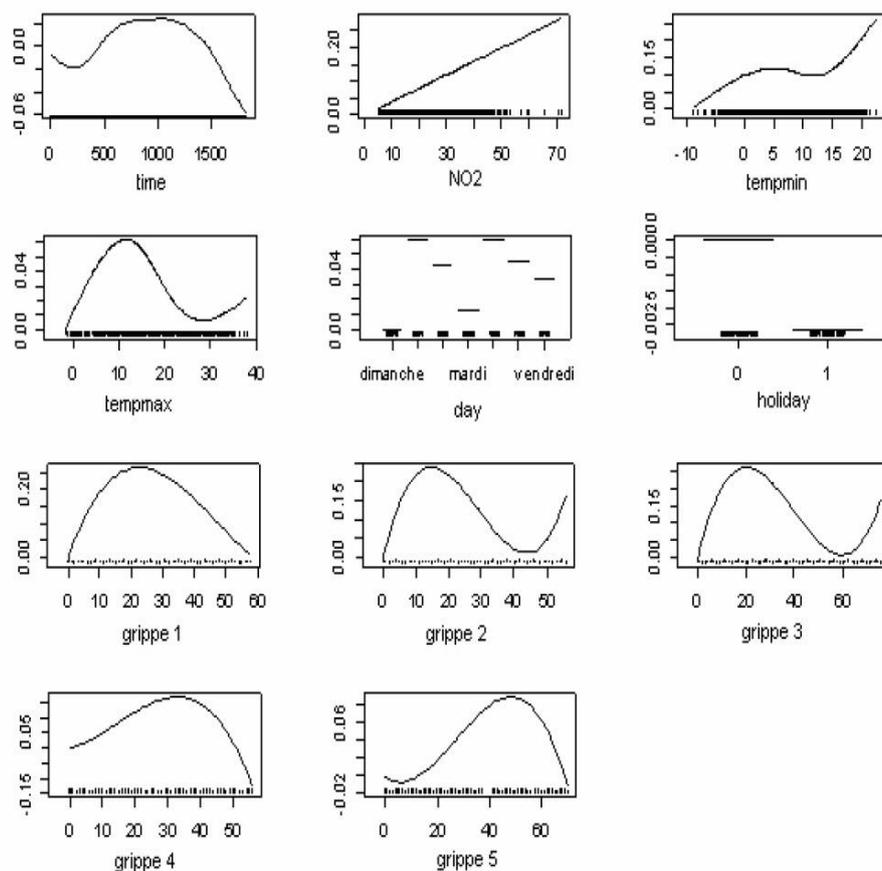


FIGURE 7 – Effet partiel des différentes variables du modèle : tendance,  $\text{NO}_2$ , température minimale et maximale, jour de la semaine, jours fériés et épidémies de grippe.

TABLEAU 2 – Variables du modèle : significativité.

Variabes	ddl <sup>(e)</sup>	valeur de F	p
NO <sub>2</sub>	1	15,3749	9,15 e-05
Température minimale (ns) <sup>(a)</sup>	3	1,749	0,155088
Température maximale (ns)	3	0,843	0,470207
Jour de la semaine (fact) <sup>(b)</sup>	6	1,449	0,192187
Jours fériés (fact)	1	0,004	0,948228
Grippe 1 (bs) <sup>(c)</sup>	3	6,417	2,54 e-04
Grippe 2 (bs)	3	5,044	1,76 e-03
Grippe 3 (bs)	3	7,911	3,05 e-05
Grippe 4 (bs)	3	2,715	0,043409
Grippe 5 (bs)	3	0,750	0,522223
Tendance (s) <sup>(d)</sup>		2,723	0,00554

(a) Transformation *natural spline* ; (b) variable factorielle ; (c) transformation *B-spline* ; (d) transformation *spline* ; (e) degrés de liberté.

[1, 020 ; 1, 062]. Ainsi, lorsque les concentrations ambiantes en NO<sub>2</sub> augmentent de 10 µg/m<sup>3</sup> dans l'agglomération de Bordeaux, le risque de décès augmente de 4,1 % le jour même et le lendemain parmi les personnes qui résident dans cette zone.

### 4.3 Comparaison aux résultats de la littérature

Le RR estimé ici est plus élevé que ceux obtenus dans les grandes études multicentriques utilisant la même méthodologie. Ce résultat, correspondant à un risque régional (susceptibilité de la population, mode de vie, composition des mélanges de polluants), ne doit pas être interprété isolément mais combiné avec les RR obtenus dans les 8 autres villes du programme Psas. Le RR combiné (métarisque) Psas est de 1,013 (IC 95 % = [1, 006; 1, 019]) et est représentatif de la situation française en général. En effet, tout l'intérêt des études multicentriques est d'obtenir suffisamment de puissance et d'être témoins d'une diversité géographique et climatique telle que les résultats puissent être considérés comme fiables, robustes et globalement représentatifs de la situation sur l'ensemble du territoire couvert par les centres participant à l'étude [28]. Ainsi, le résultat obtenu sur l'ensemble des 9 villes est tout à fait cohérent avec ceux de la

littérature et plus particulièrement ceux des études multicentriques et des méta-analyses publiées sur le sujet, même s'il est légèrement supérieur. En Europe, deux études multicentriques similaires à celle du Psas et développées à l'échelle d'un pays ont produit des résultats similaires avec une méthodologie comparable : l'étude MISA, en Italie, avec un RR de 1,009 (IC 95 % = [1,006 ; 1,013]) [2] et l'étude EMECAM, en Espagne avec un RR de 1,007 (IC 95 % = [1,004 ; 1,009]) [40]. Par ailleurs, le programme européen APHEA, incluant 26 grandes métropoles, a également mis en évidence une association significative entre les niveaux de NO<sub>2</sub> et la mortalité toutes causes avec un RR égal à 1,003 (IC 95 % = [1,002 ; 1,004]) [41]. Les grandes études nord américaines, quand à elles, ne se sont pas penchées sur l'indicateur NO<sub>2</sub> mais ont montré des associations du même ordre de grandeur avec des indicateurs de la pollution atmosphérique particulaire [8].

## 5 Conclusion

Rappelons l'intérêt de l'analyse des séries chronologiques dans le cadre de la surveillance épidémiologique, en général et de la surveillance des effets de la pollution atmosphérique sur la santé, en particulier. Les données utilisées sont produites en routine et donc facilement et rapidement accessibles. La méthode présente une puissance souvent élevée du fait du grand nombre d'observations habituellement analysées. Cette puissance permet de mettre en évidence des risques dits faibles qui caractérisent l'épidémiologie environnementale. Elle permet aussi de prendre en compte le décalage entre l'exposition au facteur de risque et la survenue de l'événement sanitaire et d'étudier l'impact d'une exposition cumulée sur plusieurs jours. Elle permet, enfin, de s'affranchir d'un ensemble de paramètres intrinsèques à la population puisque celle-ci est comparée à elle-même. Le recours à une régression par GAM avec splines pénalisées [49] autorise une flexibilité suffisante pour contrôler une possible non-linéarité des facteurs de confusion. Une autre approche utilisée pour mesurer les liens pollution-santé repose sur les études cas (témoins)-croisés. Ces dernières comparent les expositions aux polluants les jours « avec événement sanitaire » aux expositions les jours « sans événement » [29, 30, 20]. Elles sont intéressantes quand on veut prendre en compte un certain nombre de facteurs individuels (notamment, une estimation individualisée de l'exposition à la pollution atmosphérique ou de certains autres facteurs de risque) mais n'apportent pas de réel avantage par rapport aux séries chronologiques dans une stratégie de surveillance fondée sur un dispositif écologique (les modèles mis en œuvre sont, par ailleurs dans ce cas, mathématiquement équivalents dans le cas des études de séries chronologiques et de cas-croisés).

En ce qui concerne l'exposition de la population étudiée, nombre de discussions portent sur les biais écologiques potentiels pouvant mener à une mauvaise estimation du risque sanitaire lié à la pollution atmosphérique. Les analyses et les exemples évoqués montrent que les biais évoqués auraient plutôt tendance à minimiser le RR [1, 46, 33, 13].

Notons encore, puisqu'il s'agit de surveillance, que la relation mise en évidence entre les niveaux de polluants et la mortalité, même quand l'association est statistiquement significative, ne préjuge pas de la nature causale du lien. La causalité doit s'appuyer sur un faisceau d'arguments [16, 22], qui sont présents dans le cas de la pollution atmosphérique. Cette notion de causalité est nécessaire lorsque l'on veut passer à des mesures de préventions, pour lesquelles il

est nécessaire d'agir sur les causes identifiées (ici l'exposition à la pollution atmosphérique). Cependant, la surveillance épidémiologique s'appuyant sur des méthodes telles que celle présentée ici est indispensable à la quantification des relations entre pollution atmosphérique et santé et permet, dans un deuxième temps, de quantifier les impacts sanitaires (nombre de décès anticipés par exemple) de l'exposition à ce facteur de risque ou, même, les bénéfices sanitaires attendus d'une mesure de réduction des niveaux de polluants.

## Remerciements :

Les auteurs remercient chaleureusement les relecteurs pour leurs remarques pertinentes et l'aide précieuse qu'ils ont apporté à la rédaction de cet article tant sur le plan de la rigueur méthodologique que sur la clarté des messages.

Les auteurs tiennent aussi à remercier le coordinateur de ce numéro thématique ainsi que les éditeurs pour l'appui à la mise en forme de l'article.

## Références

- [1] B. ARMSTRONG – « Effect of measurement error on epidemiological studies of environmental and occupational exposures », *Occupational and Environmental Medicine* (1998), no. 55, p. 651–656.
- [2] A. BIGGERI, M. BACCINI, P. BELLINI & P. TERRACINI – « Meta-analysis of the italian studies of short-term effects of air pollution 1990-1999 (misa) », *Int J Occup Environ Health* (2005), no. 11, p. 107–22.
- [3] A. BRAGA, A. ZANOBETTI & J. SCHWARTZ – « The time course of weather-related deaths », *Epidemiol* (2001), no. 12, p. 662–7.
- [4] B. BRUNEKREEF – « The measurement of exposure to air pollution : concepts, methods and stakes », in *40e anniversaire de l'Association pour la Prévention de la Pollution Atmosphérique*, 1998.
- [5] M. DANIELS, F. DOMINICI, J. SAMET & S. ZEGER – « Estimating particulate matter-mortality dose-response curves and threshold levels : an analysis of daily time-series for the largest us cities. », *Am J Epidemiol* (2000), no. 152, p. 397–406.
- [6] C. DE BOOR – *A practical guide to splines*, Springer-Verlag, New York, 1978.
- [7] D. DOCKERY, C. POPE, X. XU, J. SPENGLER, J. WARE, M. FAY, B. FERRIS & F. SPEIZER – « An association between air pollution and mortality in six u.s. cities. », *N. Engl. J. Med.* (1993), no. 329, p. 753–9.
- [8] F. DOMINICA, A. MC DERMOT & M. DANIELS – « Revised analyses of the national morbidity mortality and air pollution study : mortality among residents of 90 cities », *J of Toxicol Environ Health* (2005), no. 68, p. 1071–92.
- [9] P. EILERS & B. MARX – « Flexible smoothing with b-splines and penalties », *Statistic Science* (1996), no. 11, p. 89–121.

- [10] D. EILSTEIN, P. QUÉNEL, G. HEDELIN, J. KLEINPETER, D. ARVEILER & P. SCHAFER – « Pollution atmosphérique et infarctus du myocarde. strasbourg, 1984-1989. », *Rev. Epidemiol. Sante Publique* (2001), no. 49, p. 13–25.
- [11] D. EILSTEIN, A. ZEGHNOUN, A. LE TERTRE, S. CASSADOU, C. DECLERCQ, L. FILLEUL, A. LEFRANC, S. MÉDINA, C. NUNES, L. PASCAL, H. PROUVOST, P. SAVIUC, D. CAMPAGNA & P. QUÉNEL – « Modélisation des liens à court terme entre la pollution atmosphérique et la santé : une actualisation des méthodes d’analyse des séries temporelles. », *Rev. Epidemiol. Sante Publique* (2004), no. 52, p. 583–9.
- [12] L. FILLEUL, S. CASSADOU, S. MÉDINA, P. FABRES, A. LEFRANC, D. EILSTEIN, A. LE TERTRE, L. PASCAL, B. CHARDON, M. BLANCHARD, C. DECLERCQ, J. JUSOT, H. PROUVOST & M. LEDRANS – « The relation between temperature, ozone and mortality in 9 french cities during the heat wave 2003 », *Environ Health Perspect* (2006), no. 114, p. 1344–7.
- [13] S. GREENLAND & J. ROBINS – « Invited commentary : Ecological studies - biases, misconceptions, and counterexamples », *American Journal of Epidemiology* (1994), no. 139, p. 747–60.
- [14] T. HASTIE & R. TIBSHIRANI – *Generalized additive models*, 1st ed., Chapman and Hall, London, 1990.
- [15] M. HATCH & D. THOMAS – « Measurement issues in environmental epidemiology », *Environmental Health Perspectives* (1993), no. 101, p. S49–S57.
- [16] A. HILL – « The environment and disease : association or causation ? », *Proceedings Royal Soc Med* (1965), no. 58, p. 295–300.
- [17] S. HOST, S. LARRIEU, L. PASCAL, M. BLANCHARD, C. DECLERCQ, P. FABRE, J. JUSOT, B. CHARDON, A. LE TERTRE, V. WAGNER, H. PROUVOST & A. LEFRANC – « Short-term associations between fine and coarse particles and cardiorespiratory hospitalizations in six french cities », *Occup. Environ. Med.* (2008), no. 65(8), p. 544–51.
- [18] Institut de Veille Sanitaire – Saint-Maurice, *Programme de surveillance air et santé 9 villes ; surveillance des effets sur la santé liés à la pollution atmosphérique en milieu urbain - phase ii : rapport de l’étude*, 2002, 184 pages.
- [19] Institut de Veille Sanitaire – Saint-Maurice, *Surveillance des effets sur la santé liés à la pollution atmosphérique en milieu urbain : rapport de l’étude. institut de veille sanitaire*, mars 1999, 148 pages.
- [20] H. JANES, L. SHEPPARD & T. LUMLEY – « Case-crossover analyses of air pollution exposure data : referent selection strategies and their implications for bias », *Epidemiology* (2005), no. 16, p. 717–26.
- [21] K. KATSOUYANNI, J. SCHWARTZ, C. SPIX, G. TOULOUMI, D. ZMIROU, A. ZANOBETTI, B. WOJTYNIAK, J. VONK, A. TOBIAS, A. PÖNKÄ, S. MÉDINA, L. BACHÁROVÁ & H. ANDERSON – « Short-term effects of air pollution on health : a european approach using epidemiologic time series data : the apha protocol », *Journal Epidemiol Community Health* (1995), no. 50, p. S12–S18.
- [22] M. KUNDI – « Causality and the interpretation of epidemiologic evidence », *Environ Health Perspect* (2006), no. 114, p. 969–74.

- [23] F. LADEN, J. SCHWARTZ, F. SPEIZER & D. DOCKERY – « Reduction in fine particulate air pollution and mortality : extended follow-up of the harvard six cities study », *Am. J. Respir. Crit. Care Med.* (2006), no. 173, p. 667–72.
- [24] S. LARRIEU, J. JUSOT, M. BLANCHARD, H. PROUVOST, C. DECLERCQ, P. FABRE, L. PASCAL, A. LE TERTRE, V. WAGNER, S. RIVIÈRE, B. CHARDON, D. BORRELLI, S. CASSADOU, D. EILSTEIN & A. LEFRANC – « Short term effects of air pollution on hospitalizations for cardiovascular diseases in eight french cities : The psas program », *Sci. Total Environ.* (2007), no. 387, p. 105–12.
- [25] A. LE TERTRE, A. LEFRANC, D. EILSTEIN, C. DECLERCQ, S. MÉDINA, B. BLANCHARD, B. CHARDON, P. FABRE, L. FILLEUL, J. JUSOT, L. PASCAL, H. PROUVOST, S. CASSADOU & M. LEDRANS – « Impact of the 2003 heatwave on all-cause mortality in 9 french cities », *Epidemiology* (2006), no. 17, p. 75–9.
- [26] A. LE TERTRE, P. QUÉNEL, D. EILSTEIN, S. MÉDINA, H. PROUVOST, L. PASCAL, A. BOUMGHAR, P. SAVIUC, A. ZEGHNOUN, L. FILLEUL, C. DECLERCQ, S. CASSADOU & C. LEGOASTER – « Short term effects of air pollution on mortality in nine french cities : a quantitative summary », *Arch Environm Health* (2002), no. 57, p. 311–25.
- [27] A. LE TERTRE, P. QUÉNEL, S. MÉDINA, Y. LE MOULLEC, B. FESTY, R. FERRY & W. DAB – « Modélisation des liens à court terme entre la pollution atmosphérique et la santé. un exemple : SO<sub>2</sub> et mortalité totale, paris, 1987-1990. », *Rev Epidemiol Sante Publique* (1998), no. 46, p. 316–28.
- [28] A. LE TERTRE, J. SCHWARTZ & G. TOULOUMI – « Empirical Bayes and adjusted estimates approach to estimating the relation of mortality to exposure of pm(10) », *Risk Anal* (2005), no. 25, p. 711–8.
- [29] M. MACLURE – « The case-crossover design : a method for studying transient effects on the risk of acute events », *Am J Epidemiol* (1991), no. 133, p. 144–53.
- [30] M. MACLURE & M. MITTLEMAN – « Should we use a case-crossover design ? », *Annu Rev Public Health* (2000), no. 21, p. 193–221.
- [31] B. MARX & P. EILERS – « Generalized linear regression on sampled signals and curves : a p-spline approach », *Technometrics* (1999), no. 41, p. 1–13.
- [32] P. MCCULLAGH & J. NELDER – *Generalized linear models*, 2nd ed., Chapman and Hall, London, 1989.
- [33] H. MORGENSTERN & D. THOMAS – « Principles of study design in environmental epidemiology », *Environmental Health Perspectives* (1993), no. 101, p. S23–S38.
- [34] P. PIRARD, P. QUÉNEL & P. LAMELOISE – « Étude de l'utilisation d'une moyenne arithmétique des mesures d'un réseau de surveillance comme indicateur de niveau d'exposition de pollution atmosphérique en milieu urbain. indice des fumées noires, région parisienne, 1991-1992 », *Pollution Atmosphérique* (1992), no. 146, p. 59–66.
- [35] C. POPE & J. SCHWARTZ – « Time series for the analysis of pulmonary health data », *Am J Respir Crit Care Med* (1996), no. 154, p. S229–S233.
- [36] C. I. POPE, R. BURNETT, M. THUN, E. CALLE, D. KREWSKI, K. ITO & G. THURSTON – « Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution », *J. Am. Med. Assoc.* (2002), no. 287, p. 1132–41.

- [37] C. I. POPE, M. THUN, M. NAMBOODIRI, D. DOCKERY, J. EVANS, F. SPEIZER & J. HEATH – « Particulate air pollution as a predictor of mortality in a prospective study of us adults », *American J. Respir. Critical Care Med.* (1995), no. 5, p. 669–74.
- [38] R DEVELOPMENT CORE TEAM – *R : A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2008, ISBN 3-900051-07-0.
- [39] T. RAMSAY, R. BURNETT & D. KREWSKI – « The effect of concavity in generalized additive models linking mortality to ambient particulate matter », *Epidemiol* (2003), no. 14, p. 18–23.
- [40] M. SAEZ, F. BALLESTER & M. BARCELO – « A combined analysis of the short-term effects of photochemical air pollutants within the emecam project », *Environ Health Perspect* (2002), no. 110, p. 221–8.
- [41] E. SAMOLI, E. AGA & G. TOULOMI – « Short-term effects of nitrogen dioxide on mortality », *Eur. Respir. J.* (2006), no. 27, p. 1129–38.
- [42] J. SCHWARTZ – « Air pollution and hospital admissions for the elderly in birmingham, alabama », *Am J Epidemiol* (1994), no. 139, p. 589–98.
- [43] — , « Non parametric smoothing in the analysis of air pollution and respiratory illness », *Can. J. Statist.* (1994), no. 22(4), p. 471–87.
- [44] J. SCHWARTZ, C. SPIX, G. TOULOUMI, L. BACHAROVA, T. BARUMAMDZADEH, A. LE TERTRE, T. PIEKARKSI, A. PONCE DE LEON, A. PÖNKÄ, G. ROSSI, M. SAEZ & J. SCHOUTEN – « Methodological issues in studies of air pollution and daily counts of deaths or hospital admissions », *J Epidemiol Community Health* (1996), no. 50(Suppl 1), p. S3–S11.
- [45] J. SUNYER, X. BASAGANA, J. BELMONTE & J. ANTO – « Effect of nitrogen dioxide and ozone on the risk of dying in patients with severe asthma », *Thorax* (2002), no. 57(8), p. 687–93.
- [46] D. THOMAS, D. STRAM & J. DWYER – « Exposure measurement error : influence on exposure-disease relationships and methods of correction », *Annual Review of Public Health* (1993), no. 14, p. 69–93.
- [47] G. TOULOUMI, R. ATKINSON, A. LE TERTRE, E. SAMOLI, J. SCHWARTZ, C. SCHINDLER, J. VONK, G. ROSSI, M. SAEZ, D. RABSZENCKO & K. KATSOUYANNI – « Analysis of health outcome time series data in epidemiological studies », *Environmetrics* (2004), no. 15, p. 101–17.
- [48] S. WOOD – *Generalized additive models. An introduction with R*, Chapman and Hall, London, 2006.
- [49] F. YAO & T. LEE – « Penalized spline models for functional principal component analysis », *J. Royal Statistical Soc., B* (2006), no. 68-1, p. 3–25.
- [50] A. ZEGHNOUN, P. CZERNICHOW & P. QUÉNEL – « Estimation de l'exposition dans les études écologiques temporelles des effets sanitaires de la pollution atmosphérique en milieu urbain : agglomération rouennaise, 1994-96 », *Pollution Atmosphérique* (1999), no. 160, p. 135–45.

- [51] A. ZEGHNOUN, D. EILSTEIN, P. SAVIUC, L. FILLEUL, C. LE GOASTER, S. CASSADOU, A. BOUMGHAR, L. PASCAL, S. MÉDINA, H. PROUVOST, A. LE TERTRE, C. DECLERCQ & P. QUÉNEL – « Surveillance des effets à court terme de la pollution atmosphérique sur la mortalité en milieu urbain. résultats d'une étude de faisabilité dans 9 villes françaises. », *Rev Epidemiol Sante Publique* (2001), no. 49, p. 3–12.