REVUE DE STATISTIQUE APPLIQUÉE

L. BELLANGER
R. TOMASSONE

La pollution de l'air dans la région parisienne: étude de la tendance dans les hautes valeurs d'ozone

Revue de statistique appliquée, tome 48, nº 1 (2000), p. 5-28
<http://www.numdam.org/item?id=RSA_2000__48_1_5_0>

© Société française de statistique, 2000, tous droits réservés.

L'accès aux archives de la revue « Revue de statistique appliquée » (http://www.sfds.asso.fr/publicat/rsa.htm) implique l'accord avec les conditions générales d'utilisation (http://www.numdam.org/conditions). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM
Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
http://www.numdam.org/

Rev. Statistique Appliquée, 2000, **XLVIII** (1), 5-28

LA POLLUTION DE L'AIR DANS LA RÉGION PARISIENNE : ÉTUDE DE LA TENDANCE DANS LES HAUTES VALEURS D'OZONE

L. Bellanger* et R. Tomassone**

** CNRS URA D0743, Modélisation Stochastique et Statistique,*
Université Paris-Sud, 91405 Orsay cedex
email : lise.bellanger@math.u-psud.fr
*** Département Mathématique & Informatique,*
Institut National Agronomique, 75231 Paris cedex 05
email : rr.tomassone@wanadoo.fr

RÉSUMÉ

Le but de cette étude est de déterminer si les tendances observées sur les données de la région parisienne sont réelles ou si elles sont la conséquence des changements météorologiques affectant les conditions de formation de l'ozone. Nous avons choisi une approche par processus ponctuels, fondée sur les dépassements d'un seuil élevé, pour modéliser l'ozone troposphérique. Ainsi, nous utilisons un processus de Poisson non-homogène pour tenir compte de la relation entre les très hautes valeurs d'ozone et les conditions météorologiques, et donc pour détecter une tendance à moyen terme.

Mots-clés : *ozone, dépassement de seuil, processus de Poisson non-homogène, régression logistique, tendance.*

ABSTRACT

The aim of this study is to establish whether observed trends for Paris area data are real or whether they are the consequences of meteorological changes affecting the conditions under which ozone is generated. A point process approach is chosen to modelling trend in tropospheric ozone based on exceedances of a high threshold. In this way, a non-homogeneous Poisson process is used to account for the relationship between very high values of ozone and the meteorological conditions, and so detect a general trend.

Keywords : *ozone, level exceedance, non-homogeneous Poisson process, logistic regression, trend.*

1. Introduction

Depuis le début de l'ère industrielle, la quantité de polluants rejetés dans l'air a considérablement augmenté. Cette pollution, en raison de son impact au niveau local et planétaire explique la mise en place d'actions de prévention individuelles et collectives. La région parisienne, avec plus de dix millions d'habitants représente depuis longtemps l'une des plus fortes concentrations d'activités d'Europe. C'est pourquoi, on s'y est préoccupé depuis longtemps des problèmes de pollution atmosphérique.

L'étude statistique de la pollution de l'air, quant à elle, date d'une trentaine d'années. Elle porte sur des échelles de temps et d'espace très différentes. Mais la plupart des travaux n'ont pas développé une méthodologie statistique originale. Il existe des exceptions notables, comme les publications de l'Université de Caroline du Nord (dans le cadre de l'Environmental Protection Agency), certains travaux australiens.

Le travail que nous présentons fait partie d'un projet plus vaste proposé par AIRPARIF (organisme chargé de la surveillance et de la qualité de l'air en Ile-de-France), publié par ailleurs dans Bellanger (1999).

Les conditions météorologiques telles la température journalière et la vitesse du vent jouent un grand rôle dans la prévision des pics de pollution. Les variations annuelles des conditions météorologiques peuvent donc masquer toute tendance à long(/moyen)-terme de l'ozone à relier à des changements dans les émissions de précurseurs d'ozone. De nombreux auteurs se sont attachés à modéliser la tendance à long terme du maximum d'ozone troposphérique journalier en tenant compte des conditions météorologiques, à l'aide de techniques variées. Ces modèles permettent entre autre d'estimer la part de la tendance de l'ozone qui n'est pas prise en compte par les tendances des variables météorologiques.

Par exemple, Cox et Shao-Hang (1993) décrivent une méthode probabiliste basée sur la distribution de Weibull où le paramètre échelle peut fluctuer jour après jour en fonction de l'année, du jour et des conditions météorologiques journalières favorables à la formation de l'ozone; elle peut être utilisée pour obtenir des tendances ajustées en fonction des conditions météorologiques. Bloomfield et *al.* (1993a, 1993b) utilisent des méthodes graphiques (étude empirique de l'association des niveaux d'ozone avec les différentes variables météorologiques) et des méthodes non-paramétriques pour déterminer les variables météorologiques significatives et choisir les formes fonctionnelles permettant de modéliser la dépendance de l'ozone. La méthode des moindres carrés non-linéaires permet alors d'estimer les coefficients du modèle contenant ces variables, leurs interactions et l'année. Gao et *al.* (1994) utilisent des techniques semi-paramétriques pour construire des modèles tenant compte de la météorologie et de l'année pour estimer les niveaux d'ozone.

Alors que l'idée première serait d'utiliser un maximum de données, du point de vue de la santé publique, il apparaît plus important d'essayer de dégager une tendance à moyen terme uniquement dans les épisodes de forte pollution. Les résultats obtenus pourront peut-être par la suite permettre une meilleure compréhension de la relation entre les épisodes de forte pollution et leurs effets à moyen terme sur la santé (augmentation/diminution du nombre de personnes atteintes d'allergies, d'insuffisances respiratoires, d'asthme...).

Dans cet article, nous nous attachons à mettre en évidence une tendance sur plusieurs années dans les épisodes de pollution aiguë, en tenant compte des conditions météorologiques à la fois dans la fréquence et la taille des dépassements d'un seuil fixé. On utilise un modèle basé sur le processus de Poisson non-homogène (PPNH) pour estimer les tendances dans l'intensité du processus qui génère les dépassements. L'approche statistique est basée sur le fait que l'on considère les dépassements d'un niveau élevé, se produisant dans le temps, comme des points d'un processus de Poisson. Des théorèmes limites pour de tels processus ont été développés par Pickands (1971) puis généralisés par Leadbetter et *al.* (1983). Ainsi, dans le cadre particulier de l'ozone troposphérique, Smith (1989), Shively (1991), et plus récemment Smith et Shively (1995) ont utilisé l'idée de considérer le nombre de dépassements de haut niveau comme généré suivant un PPNH, puisqu'une tendance peut exister.

2. Les données

Pour construire ce modèle et pour appliquer nos résultats théoriques, nous utilisons les données provenant de quatre stations de la région parisienne. Comme nous disposions aussi de données provenant de quelques stations de la région de Los Angeles, nous avons essayé de leur appliquer notre modèle.

Nous avons utilisé les valeurs des maxima journaliers d'ozone (fournis par AIRPARIF) enregistrées sur les sites de Neuilly/Seine, Champs/Marne, Aubervilliers et Créteil, ainsi que les valeurs journalières de variables météorologiques (fournies par le mât du commissariat à l'Énergie Atomique de Saclay) décrites plus bas, durant les mois de mai à septembre de la période 1988-1997. Les mois de mai à septembre forment la période de l'année dans laquelle la majorité des hautes valeurs d'ozone sont enregistrées.

La valeur du maximum journalier utilisée dans notre analyse correspond à la valeur maximale prise entre 6h00 et 18h00 TU.

Les covariables utilisées dans cette analyse sont :

• la température maximale mesurée au sol (TMAX) : maximum des valeurs horaires entre 6h00 TU et 18h00 TU. (correspondant à $w3$ dans le modèle),

• l'amplitude thermique (AMTEMP) : différence entre la valeur minimale et la valeur maximale mesurée entre 6h00 et 18h00. (correspondant à $w4$ dans le modèle),

• la vitesse moyenne du vent mesurée à 58 mètres (VENT) : moyenne entre 6h00 et 18h00. (correspondant à $w5$ dans le modèle),

• l'amplitude de vitesse de vent mesurée à 58 mètres (AMVENT) : différence entre la valeur minimale et la valeur maximale de la vitesse du vent sur la période 6h00-18h00. (correspondant à $w6$ dans le modèle),

• les variables dichotomiques t92 et t93 pour tenir compte des changements de capteurs intervenus en 1992 et 1993.

De plus, dès que pour un jour donné, au moins une observation horaire manquait sur cette période de 13 heures, la valeur de la variable journalière correspondante a été considérée manquante (*cf.* tableau 2.1). S'il n'y avait pas eu de données manquantes, nous aurions disposé de 1380 jours.

TABLEAU 2.1
Proportion de jours manquants par variables

Ozone	données manquantes (en %)
Neuilly/Seine	24.5
Champs/Marne	14.1
Aubervilliers	25.4
Créteil	25.0
Variables météorologiques	données manquantes (en %)
Température (TMAX et AMTEMP)	6.4

Le tableau 2.2 fournit pour chaque station, la proportion de jours dépassant le seuil u, pour u variant de 120 à 180 μgm^{-3}.

TABLEAU 2.2
Proportion de jours dépassant le seuil u par station
(Entre parenthèses nombre de jours)

	Seuil (en μgm^{-3})						
Station	% > 120	% > 130	% > 140	% > 150	% > 160	% > 170	% >180
Neuilly/Seine	12.4 *(129)*	8.5 *(89)*	6.3 *(66)*	4.2 *(44)*	3.4 *(35)*	2.0 *(21)*	1.6 *(17)*
Champs/Marne	6.1 *(72)*	4.2 *(50)*	2.5 *(30)*	1.8 *(21)*	1.3 *(15)*	0.5 *(6)*	0.3 *(3)*
Aubervilliers	11.0 *(113)*	7.4 *(76)*	5.4 *(56)*	4.2 *(43)*	2.9 *(30)*	1.7 *(17)*	1.2 *(12)*
Créteil	10.6 *(110)*	7.4 *(77)*	4.5 *(47)*	3.4 *(35)*	2.2 *(23)*	1.1 *(11)*	0.6 *(6)*

Les graphiques (2.1, 2.2) permettent d'observer les relations existant entre les variables. Le sens de variation de l'ozone en fonction des variables météorologiques, quand tous les jours sont pris en compte (graphiques de gauche), se dégage aisément. Il n'en est plus de même quand on ne conserve que les jours où le maximum d'ozone dépasse 130 μgm^{-3} (graphiques de droite); il est alors très difficile à mettre en évidence une quelconque variation à la seule lecture de ces graphiques.

On observe sur le graphique de gauche de la figure 2.1 que lorsque la température maximale est élevée, le maximum d'ozone journalier est grand. Cependant, la relation entre l'ozone et la température est beaucoup moins évidente pour les valeurs d'ozone élevées (graphique de droite de la figure 2.1). On peut obtenir deux graphiques analogues avec des conclusions identiques pour la valeur de l'amplitude thermique.

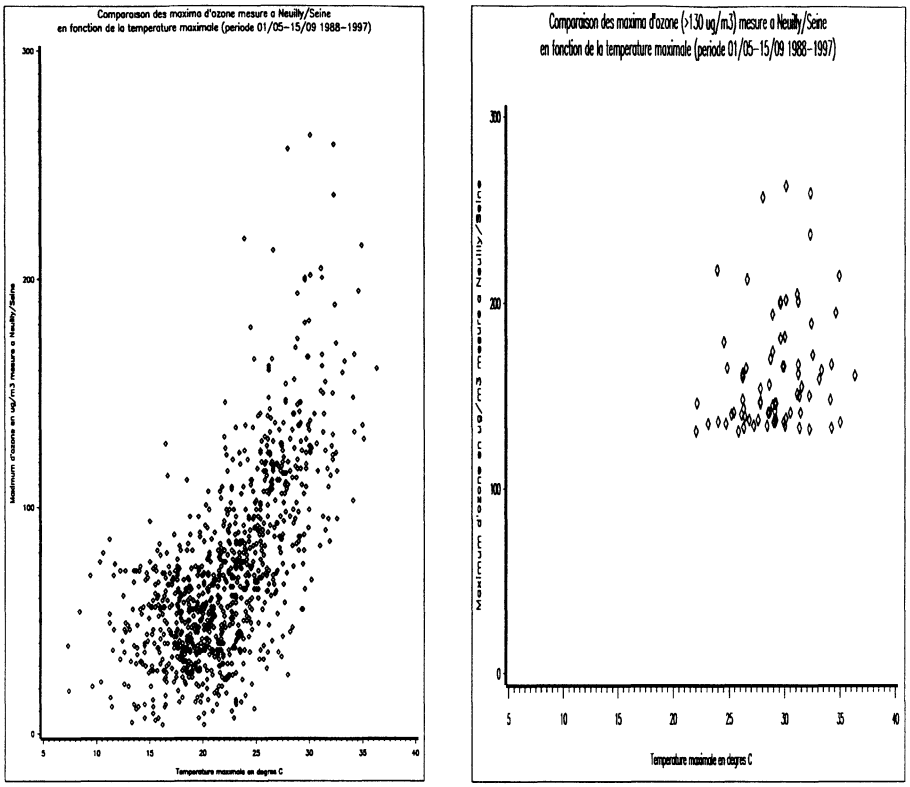

FIGURE 2.1
Relation entre les valeurs du maximum d'ozone mesurées sur le site de Neuilly/Seine et les valeurs de la température maximale.

Le graphique de gauche de la figure 2.2 met en évidence la diminution de la valeur du maximum d'ozone en fonction de l'augmentation de la vitesse moyenne du vent; l'existence de vent entraîne une diminution de l'ozone. Mais le graphique de droite ne traduit aucune relation évidente. Ces graphes donnent une idée des phénomènes de dépendance entre niveau d'ozone et covariables à prendre en compte, mais sont d'une aide très limitée quant au choix du modèle à adopter.

Le graphique de gauche de la figure 2.3 met en évidence la faible dispersion des maxima d'ozone journaliers en 1993, ainsi que de fortes valeurs adjacentes supérieures pour 1990 et 1991. Ceci reflète l'influence des régimes météorologiques dans le processus de formation de l'ozone (*cf.* figure 2.1). Le graphique de droite de la figure 2.3 (uniquement les jours où le niveau d'ozone est supérieur à 130 μgm^{-3}) montre des différences plus marquées entre les années. Les années 1990, 1991 et 1994 possèdent les percentiles 75% les plus grands, avec une valeur très élevée pour celui de 1991 (215 μgm^{-3}). De plus, l'année 1991 possède la valeur médiane la plus élevée, tandis que 1989 présente la médiane la plus faible de la période. Il est cependant très difficile d'en déduire une tendance.

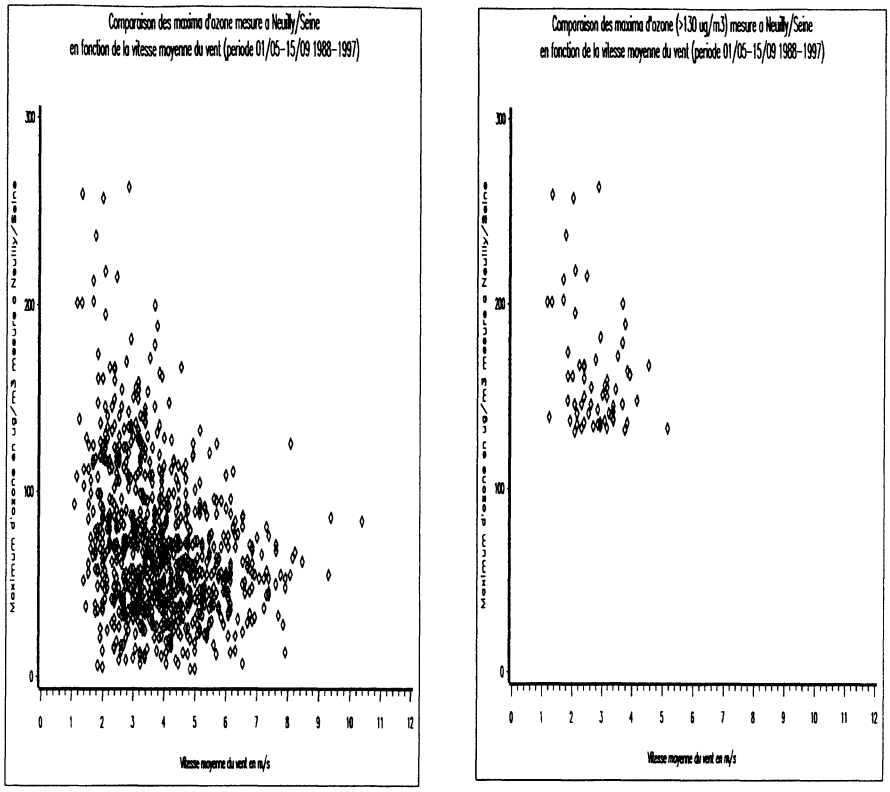

FIGURE 2.2
Relation entre les valeurs du maximum d'ozone mesurées sur le site de Neuilly/Seine et les valeurs de la vitesse moyenne du vent.

3. Modèles de valeurs extrêmes

Dans ce paragraphe, nous présentons un modèle permettant de dégager une tendance sur quelques années du phénomène de production de l'ozone ; pour plus de détails, se reporter à Leadbetter *et al.* (1983), Falk *et al.* (1994), Bellanger (1999).

3.1. Rappels sur la théorie classique des valeurs extrêmes

• Si nous nous restreignons au cas de variables aléatoires indépendantes et identiquement distribuées (iid), nous avons les résultats classiques suivants. Supposons $Y_1, Y_2, \ldots,$ une suite iid de distribution commune F, et $M_n = \max(Y_1, \ldots, Y_n)$. La théorie des valeurs extrêmes cherche les suites normalisantes $a_n > 0$, b_n telles que $(M_n - b_n)/a_n$ converge en distribution. On peut montrer que la distribution sous des suppositions assez larges converge vers ce que l'on appelle la *distribution généralisée des valeurs extrêmes, dont une distribution particulière est la distribution de Gumbel.*

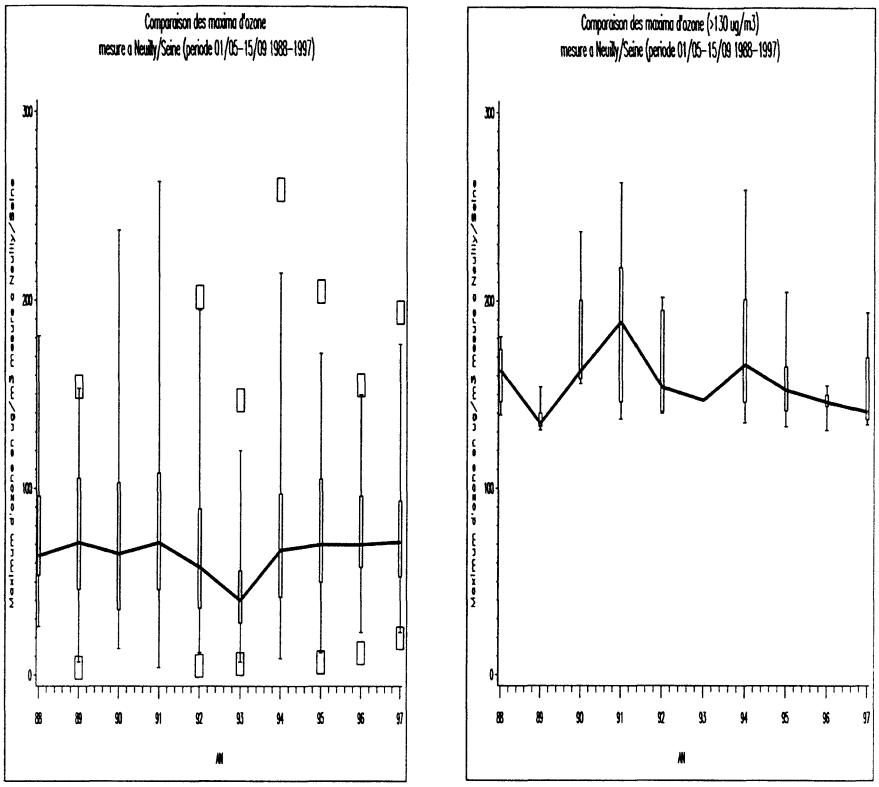

FIGURE 2.3

Boîtes à «moustaches» des valeurs du maximum d'ozone mesurées à Neuilly/Seine. Le rectangle central a pour longueur l'intervalle entre les quartiles 0.25 et 0.75 Les «moustaches» correspondent aux valeurs adjacentes supérieure et inférieure. Les points extérieurs aux «moustaches» sont des valeurs extrêmes. La ligne brisée relie les médianes de chaque année.

• On peut aussi étudier la densité conjointe de plusieurs statistiques d'ordre d'un échantillon, au lieu seulement du maximum. Si on note $M_n^{(k)}$ la k^{ème} plus grande variable aléatoire parmi $Y_1, Y_2, \ldots, Y_n$, $(M_n = M_n^{(1)})$, et si le maximum M_n possède une distribution limite H, alors $M_n^{(k)}$ possède aussi une distribution limite dont on sait calculer la forme. On peut aussi démontrer que le nombre de dépassements S_n d'un seuil u_n possède une distribution limite de Poisson de paramètre λ. Smith (1984, 1986) a utilisé cette approche pour analyser des données hydrologiques et il s'est limité aux 5 plus grandes valeurs annuelles.

• L'approche par seuil est basée sur l'approximation de la distribution des dépassements d'un seuil élevé u; Pickands (1975) a ainsi introduit la *distribution de Pareto Généralisée* (DPG), dont un cas particulier est la distribution exponentielle, très souvent utilisée dans les applications de la modélisation «Peaks Over Threshold»

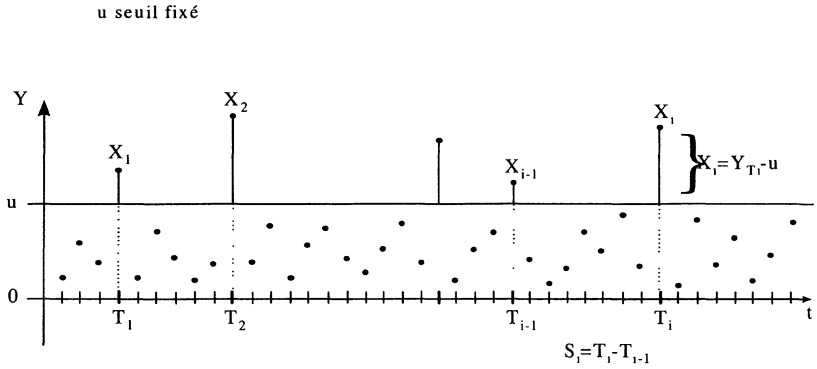

FIGURE 3.1
Représentation des dépassements.

(décrite par Smith, 1984 et Leadbetter, 1991) aux données hydrologiques ou de pollution de l'air (Smith, 1984, Davidson, 1984 et Hosking et Wallis, 1987).

On peut enfin généraliser les propriétés asymptotiques de Poisson du nombre de dépassements de u_n (notés S_n ci-dessus) en considérant le processus ponctuel N_n de dépassements de niveau u_n. Si $E \subset]0,1]$, alors $N_n(E) = \text{Card}\,(i : i/n \in E : Y_i > u_n,\ 1 \leqslant i \leqslant n) = \text{Card}\,(i \in nE : Y_i > u_n,\ 1 \leqslant i \leqslant n)$.

Si u_n satisfait $n\big(1 - F(u-n)\big) \xrightarrow[n\to\infty]{} \lambda$, alors $N_n(I)$ converge en loi vers $N(I)$ où N est un processus de Poisson défini sur $I \subset]0,1]$ d'intensité λ.

Si on considère non plus N_n comme un processus sur $]0,1]$, mais comme un processus dans le plan, pris aux points $(i/(n+1), X_{n,i}), i = 1, \ldots, n$, où $X_{n,i} = (Y_i - b_n)/a_n, i = 1, \ldots, n$, on a alors le résultat dû à Pickands (1971) de convergence en loi de N_n vers N, processus de Poisson dans le plan. Alors, tous les résultats de la théorie des valeurs extrêmes peuvent être obtenus à partir de cette représentation par processus.

3.2. Processus de Poisson non-homogène (PPNH) dans le plan

Pour étudier s'il existe une tendance à long terme, les différents auteurs utilisent :

• soit une approche distributionnelle, comme Shively (1990); très lourde à mettre en œuvre, elle a des défauts majeurs (en particulier la difficulté pour des données de respecter les suppositions de base et la non prise en compte des conditions météorologiques).

• soit une approche par processus ponctuel comme Smith (1989), Shively (1991), Smith et Shively (1995); c'est elle seule que nous aborderons.

Le principal problème de cette modélisation provient de ce qu'elle ne tient pas compte du regroupement des hautes valeurs, puisqu'elle est basée sur l'hypothèse

d'indépendance des intervalles de temps entre deux dépassements de seuil u fixé élevé. Donc, les modèles à une dimension paraissent peu réalistes. Ceci nous conduit à présenter un modèle, *a priori* plus réaliste, qui prend en compte simultanément fréquence et taille de dépassement, développé par Smith et Shively (1995).

Pour approcher le processus ponctuel du nombre et de la taille des dépassements, il est nécessaire d'étudier le PPNH dans le plan, en suivant la démarche suivante.

3.2.1. Vraisemblance associée aux dates et aux tailles de dépassements de seuil

Notons :

– Y variable aléatoire représentant le maximum d'ozone journalier

– $\Psi_t(y) = \begin{cases} P(Y > y) \text{ le jour } t \\ 0 \quad \text{si le jour } t \text{ est manquant} \end{cases}$

alors, la distribution de Y le jour t, si le jour t n'est pas manquant, est :

$$1 - \Psi_t(y)$$

Posons :

$$\psi_t(y) = -\frac{d}{dy}[\Psi_t(y)]$$

Si le processus est observé sur une période de temps $]0, T[$ et si les pics d'ozone dépassant le seuil fixé u sont représentés par :

$$(T_i; Y_i), 1 \leqslant i \leqslant N, \text{ où } T_i \text{ et } Y_i \text{ sont supposées indépendantes } \forall i$$

Le $i^{\text{ème}}$ pic se produit le jour T_i où il prend la valeur $Y_i \geqslant u$. Le nombre total des N pics étant lui aussi une variable aléatoire, la densité conjointe des pics observés peut être approchée par :

$$L = \left[(\prod_{i=1}^{N} \Psi_{t_i}(u)) \exp\left[-\int_0^T \Psi_t(u)dt \right] \right] * \left[\prod_{i=1}^{N} \frac{\psi_{t_i}(y_i)}{\Psi_{t_i}(u)} \right] = A * B \qquad (3.1)$$

où :

– Le premier terme A entre crochets dans (3.1) correspond à la densité d'un processus de Poisson non-homogène d'intensité $\Psi_t(u)$ qu'on notera aussi $\lambda(t)$ ultérieurement. Il correspond donc à la modélisation des jours de dépassement de niveau u (fréquence des dépassements).

– Le $i^{\text{ème}}$ terme du second terme B entre crochets correspond à la densité de Y_i sachant qu'un dépassement de seuil u a eu lieu.

Remarque : la densité conjointe (3.1) est une approximation de la vraie densité des $(T_i; Y_i)$, $1 \leqslant i \leqslant N$ puisque la fréquence des jours de dépassement est ici

modélisée par un processus stochastique *continu* dans le temps, alors que les données sont *discrètes* dans le temps (il se produit au plus un dépassement par jour car l'étude utilise le maximum d'ozone journalier). Cependant, il semble que la densité (3.1) est une approximation raisonnable au vu des résultats obtenus dans Smith et Shively (1995).

3.2.2. Modèles associés à la fréquence et aux tailles des dépassements

3.2.2.1. Fréquence des dépassements

Il existe plusieurs façons de modéliser l'intensité du processus de Poisson (Crow, 1974; Lee, 1980) que nous ne présenterons pas ici. Shively (1991), Vaquera-Huerta *et al.* (1997) utilisent l'idée que le nombre de dépassements d'un niveau élevé u est généré suivant un **PPNH**; ils estiment la tendance à long terme en tenant compte de la relation entre les très hautes valeurs d'ozone et les conditions météorologiques. Le processus de Poisson est non-homogène donc sa fonction d'intensité, représentant le taux d'apparition des événements par unité de temps, $\lambda(t) = P[Y > u \text{ le jour } t]$ n'est pas constante dans le temps. Shively (1991) a choisi un modèle exponentiel pour l'intensité (Cox et Lewis, 1966) ce qui lui a permis de tester l'existence d'une tendance :

$$\lambda(t) = \exp(\alpha_0 + \alpha_1 t) \tag{3.2}$$

ou, d'une tendance et de covariables :

$$\lambda(t) = \exp(\alpha(t)) = \exp\left(\alpha_0 + \alpha_1 s(t) + \sum_{j=2}^{p} \alpha_j w_j(t)\right) \tag{3.3}$$

où $w_j(t)$ représente la variable météorologique j le jour t, et $s(t)$ est un terme de tendance prenant la valeur 1 si le jour t appartient à l'année 1, 2 si le jour t appartient à l'année 2, etc. Ainsi, la fréquence dépend directement de $p-1$ variables météorologiques et de la tendance à long terme représentée par $s(t)$. Tester une tendance à long terme de la fréquence d'apparition des hautes valeurs d'ozone, en tenant compte des conditions météorologiques, revient à tester $H_0 : \alpha_1 = 0$ i.e. pas de tendance dans la fréquence des dépassements, contre l'alternative $H_A : \alpha_1 \neq 0$.

En pratique, on peut envisager d'autres formes paramétriques que (3.2); dans notre étude nous choisirons une forme logistique car elle présente des avantages intéressants.

Calcul et choix des estimateurs

Les estimateurs du maximum de vraisemblance des coefficients sont calculés en maximisant (3.1) par une procédure numérique classique comme l'algorithme de Newton-Raphson. Dans (3.3), les variables incluses dans $\alpha(t)$ peuvent être interprétées comme les variables reliées, dans le temps, à la fréquence des dépassements du seuil u. En particulier, le coefficient α_1 de $s(t)$ représente la tendance dans la fréquence des niveaux d'ozone à conditions météorologiques constantes. Les coefficients $\alpha_2, \ldots, \alpha_p$ fournissent une mesure de la relation entre conditions météorologiques et fréquence des valeurs élevées d'ozone.

Pour sélectionner les variables dans $\alpha(t)$, on commence par inclure toutes les variables ainsi qu'un terme constant. Ensuite, comme dans une régression classique, on élimine les variables statistiquement non significatives en effectuant un test sur les coefficients eux-mêmes en utilisant l'approximation normale des estimateurs $\widehat{\alpha}_j$ des α_j. En notant $\hat{\sigma}^2(\alpha_j)$ l'estimation de la variance de $\hat{\alpha}_j$, on effectue un test fondé sur la Normalité asymptotique des estimateurs du maximum de vraisemblance et l'on a :

$$\text{sous } H_0 : \alpha_j = 0, \frac{\hat{\alpha}_j - 0}{\hat{\sigma}(\alpha_j)} \approx N(0,1) \Rightarrow H_0 \text{ est rejetée au seuil 0.05 si :}$$

$$\left| \frac{\hat{\alpha}_j}{\hat{\sigma}(\alpha_j)} \right| \geqslant 1.96$$

L'intensité est alors ré-estimée avec les variables restantes. Cette procédure de sélection progressive descendante est répétée jusqu'à ce qu'aucun test statistique n'ait une valeur inférieure à 1.96; c'est la classique procédure «*backward*».

3.2.2.2. Taille des dépassements

La modélisation de la fréquence des dépassements du seuil u élevé utilise la même méthode que celle développée dans Shively (1991). Pour les tailles, d'après les résultats théoriques sur la taille des dépassements d'un seuil élevé du § 3.1.1, la distribution limite de $X = Y - u$ sachant que $Y \geqslant u$ quand $u \to \infty$ est une distribution de Pareto généralisée. Alors, on montre que le $i^{\text{ème}}$ terme du second terme B est la densité d'une loi exponentielle de paramètre $\beta(t)$; $1/\beta(t)$ correspond alors à l'espérance de la variable aléatoire taille du dépassement, sachant qu'un dépassement a eu lieu le jour t, de densité exponentielle. On peut donc l'interpréter comme l'estimation de la taille du dépassement (sachant qu'un dépassement s'est produit le jour t), conditionnellement aux valeurs atteintes par les variables météorologiques ce jour-là. De plus, les estimateurs du maximum de vraisemblance sont asymptotiquement gaussiens et efficaces et la matrice de covariance peut être estimée par l'inverse de la matrice d'information de Fisher. On prend pour $\beta(t)$ une forme linéaire analogue à celle de $\alpha(t)$:

$$\beta(t) = \beta_0 + \beta_1 s(t) + \sum_{j=2}^{p} \beta_j w_j(t) \tag{3.4}$$

Ainsi la taille dépend, comme la fréquence, directement des k variables météorologiques et de la tendance à long terme $s(t)$.

En utilisant une procédure similaire à celle utilisée pour le modèle approchant la fréquence des dépassements du seuil fixé, on élimine les variables statistiquement non significatives dans l'expression de l'estimateur $\hat{\beta}(t)$ de $\beta(t)$.

3.2.3. Validation du modèle

3.2.3.1. Fréquence des dépassements

Il y a deux hypothèses très importantes à vérifier pour s'assurer que les jours de dépassements T_i, $i = 1, \ldots, n$ peuvent être modélisés par un PPNH :

• i) *Distribution des intervalles* S_i : la distribution des intervalles $S_i = T_i - T_{i-1}$ peut être approchée par une distribution exponentielle. Suivant Shively (1991), si $h(s) = \int_t^{t+s} \lambda(r)dr$, la distribution est donnée par :

$$F_t(s) = P\{S \leqslant s\} = \int_0^s f_t(r)dr = 1 - \exp[-h(s)] \tag{3.5}$$

Comme les jours de dépassement sont discrets, $F_t(s)$ ne doit être calculée que pour les valeurs discrètes de t et s. On peut approcher (3.5) par :

$$F_t(s) = 1 - \exp\left\{-\sum_{k=1}^{s} \overline{\lambda}(k)\right\} \tag{3.6}$$

où $\overline{\lambda}(k) = \lambda(t+k)$

On transforme ensuite S_i en :

$$U_i = F_{t(i-1)}(Si) \tag{3.7}$$

où $t(i-1)$ correspond au jour du $(i-1)^{\text{ème}}$ événement.

Si les intervalles de temps S_i sont indépendants de distribution $F_t(s)$, cette transformation permet de réduire S_i à une variable aléatoire U_i de distribution uniforme sur $[0;1]$. A partir de la statistique d'ordre de U_i (notée $U_{(i)}$), on trace le graphique $(u(i);\, i/(n+1))$ pour $i = 1, \ldots, n$. Si les variables aléatoires S_i ont la distribution donnée par (3.5), alors les points doivent se situer au voisinage de la bissectrice.

• ii) *Indépendance mutuelle des intervalles* S_i : vérifier l'indépendance du nombre d'événements dans des intervalles de temps séparés est équivalent à vérifier que les intervalles de temps entre événements $S_i = T_i - T_{i-1}$ sont indépendants. En théorie, le fait que les variables S_i ne soient pas corrélées n'implique pas qu'elles soient indépendantes. Cependant, d'un point de vue pratique, s'il y a dépendance entre deux intervalles S_i successifs, cela signifie probablement qu'il existe une forte corrélation sérielle. Donc pour vérifier que les intervalles sont indépendants, on peut vérifier que les intervalles adjacents sont non corrélés, c'est-à-dire que S_i n'est pas corrélé à S_{i-1}. Shively (1991), propose de calculer le coefficient de corrélation entre intervalles adjacents.

3.2.3.2. Taille des dépassements

Pour valider le modèle relatif aux tailles de dépassements X_i, il faut vérifier les hypothèses suivantes :

• i) *Distribution de la taille des dépassements* : un test de Kolmogorov-Smirnov permet de vérifier que la distribution exponentielle est la distribution appropriée pour modéliser la taille des dépassements, et que l'expression de $\beta(t)$ dans (3.4) est correcte.

• ii) *Indépendance mutuelle des tailles de dépassement* X_i : pour vérifier l'indépendance, Smith et Shively (1995) calculent le coefficient de corrélation r entre tailles de dépassement survenu des jours consécutifs. Si n_{adj} représente le nombre de paires de tailles de dépassement adjacent, l'écart-type de r est $s(r) = \frac{1}{n_{adj}^{1/2}}$; on accepte l'hypothèse de non corrélation si $|r|/s(r) < 1.96$.

3.3. Stratégie, choix a priori d'un seuil u raisonnable

La procédure précédente est lourde, il paraît plus intéressant de choisir en début d'étude un seuil acceptable et de valider l'indépendance avant d'estimer les paramètres.

Pour un seuil u fixé, les intervalles entre deux dépassements s_i, $i = 1, \ldots, n$ étant connus, on calcule une estimation du coefficient de corrélation entre intervalles adjacents. Il suffit de calculer ce coefficient de corrélation par Bootstrap (Shao et Tu, 1996). Pour cela on considère l'ensemble des valeurs des t_i comme une population dans laquelle on effectue un grand nombre d'échantillonnages; on obtient ainsi une estimation du coefficient de corrélation avec sa distribution. On examine alors les percentiles (par exemple 2.5 % et 97.5 %) permettant d'estimer un intervalle de confiance avec un coefficient de confiance de 95 %. Si cet intervalle recouvre la valeur 0, nous décidons que ce coefficient n'est pas significativement différent de 0. Cette estimation ne nécessitant aucune connaissance sur la loi des S_i, elle peut donc être appliquée dès le début de la modélisation afin de déterminer une valeur raisonnable du seuil u en fonction de degré de corrélation entre intervalles adjacents, avant de passer à l'estimation et à la validation du modèle associé à la fréquence des dépassements.

Après avoir déterminé la valeur raisonnable du seuil u permettant d'utiliser un PPNH, il faut vérifier pour cette valeur que les tailles de dépassements correspondantes ne sont pas corrélées. On calcule donc une estimation du coefficient de corrélation uniquement pour les dépassements du seuil u se produisant des jours consécutifs, par la méthode Bootstrap. Ceci se justifie par le fait que, si les dépassements se produisant des jours successifs ne sont pas corrélés, alors les dépassements séparés par plus d'un jour seront probablement non corrélés. Les tailles de dépassements se produisant des jours successifs $x_i(t), i = 1, \ldots, n_{adj}$, elles sont connues ($n_{adj}$ nombre de paires de tailles de dépassement successifs : $\{x_{i-1}(t-1), x_i(t)\}$).

4. Construction d'un modèle pour les hautes valeurs d'ozone mesurées en région parisienne

Nous avons utilisé un PPNH pour modéliser la fréquence et la taille des dépassements d'un seuil élevé u, déterminé en suivant la stratégie des paragraphes 3.2 et 3.3. Nous n'évoquerons donc ici que les spécificités de ce modèle par rapport à celui développé par Smith et Shively (1995) et décrit dans la partie précédente. La densité (3.1) fournit donc une approximation de la vraie densité des $(T_i, Y_i), 1 \leqslant i \leqslant N$.

4.1. Spécificités du modèle associé à la fréquence des dépassements

4.1.1. L'intensité du PPNH

On remarque que $\Psi_t(u) = P[Y > u$, le jour $t]$ peut s'écrire en fonction d'une variable aléatoire dichotomique Z prenant la valeur 1 si un dépassement a eu lieu le jour t, 0 sinon :

$$P[Y > u, \text{ le jour } t] = P[Z = 1, \text{le jour } t/\text{covariables}] = E[Z/\text{covariables}]$$

Nous avons donc eu l'idée d'utiliser la distribution logistique pour modéliser $E[Z/\text{covariables}]$. Le modèle de régression logistique (Hosmer et Lemeshow, 1989) permet d'obtenir l'intensité λ du PPNH, différente de (3.2); elle a la forme suivante :

$$\lambda(t) = \Psi_t(u) = \frac{\exp(\alpha(t))}{1 + \exp(\alpha(t))} \tag{4.1}$$

où $\alpha(t)$ tient compte des interactions possibles entre covariables contrairement à l'écriture (3.3) :

$$\alpha(t) = \alpha_0 + \alpha_1 s(t) + \sum_{j=2}^{p} \alpha_j w_j(t) + \sum_{j=2}^{p} \alpha_{1j} s(t) w_j(t) + \sum_{i,j=2}^{p} \alpha_{ij} w_i(t) w_j(t) \tag{4.2}$$

Notre choix s'est porté vers ce modèle à cause de sa flexibilité, de l'interprétation relativement simple des estimations des paramètres et son implantation dans de nombreux logiciels de statistiques tel SAS (proc LOGISTIC) permettant d'éviter une programmation fastidieuse et source d'erreurs. Un autre avantage de la distribution logistique est l'utilisation d'une statistique facilement interprétable le *rapport des chances* (plus connue sous le vocable anglais de *odds ratio*), dont nous rappelons la définition dans le cas simple suivant. Si nous écrivons (4.1) dans le cas d'une variable dichotomique x (qui ne peut prendre que les valeurs 0 et 1) :

$$\pi(x) = \frac{\exp(\alpha_0 + \alpha_1 x)}{1 + \exp(\alpha_0 + \alpha_1 x)} \tag{4.3}$$

La réalisation d'une variable elle-même dichotomique y (comme le dépassement ou non d'un seuil), dans la régression logistique, en fonction de x est fournie dans le tableau suivant :

	$x = 1$	$x = 0$
$y = 1$	$\pi(1)$	$\pi(0)$
$y = 0$	$1 - \pi(1)$	$1 - \pi(0)$

On définit la chance (odds) d'une réalisation entre les observations pour lesquelles $x = 1$ par la valeur $\pi(1)/[1 - \pi(1)]$, de la même façon pour $x = 0$

on définit $\pi(0)/[1-\pi(0)]$. Le logarithme de chacun de ces rapports est appelé *logit* :

$$g(1) = \log\{\pi(1)/[1-\pi(1)]\} \text{ et } g(0) = \log \pi(0)/[1-\pi(0)]$$

Le rapport des chances φ est le rapport de ces deux quantités, et son logarithme (dans le jargon statistique le *log-odds*) :

$$\log(\varphi) = g(1) - g(0) = \log\{[\pi(1)/[1-\pi(1)]/[\pi(0)/[1-\pi(0)]\} = \alpha_1 \quad (4.4)$$

Cette statistique est très largement utilisée en épidémiologie, car elle a une interprétation simple; si la valeur $\varphi = \exp(\alpha_1) = 2$, ceci signifie que lorsque que l'on passe de la condition $x = 0$ à la condition $x = 1$, le dépassement du seuil se produit deux fois plus souvent. L'extension à des variables x qui ne sont plus dichotomiques est facile (Hosmer et Lemeshow, 1989).

4.1.2. Estimation des paramètres du modèle

La méthode du maximum de vraisemblance permet d'estimer les paramètres du modèle, puis une procédure de sélection « *backward* » permet de ne conserver que les variables significatives.

4.1.3. Validation du modèle

Après avoir calculé la distribution des intervalles de temps entre deux dépassements $S_i = T_i - T_{i-1}$, en fonction des estimations des paramètres des covariables dans l'intensité (4.1), on utilise comme dans Smith et Shively (1995), un test de Kolmogorov-Smirnov pour valider la distribution des S_i.

4.2. Spécificités du modèle associé à la taille des dépassements

On suppose, comme dans Smith et Shively (1995), que les tailles de dépassements X_i suivent une loi exponentielle. Le paramètre $\beta(t)$ devrait aussi prendre en compte les interactions possibles entre covariables, comme dans (4.2) mais, pour des raisons purement techniques, nous ne l'avons pas fait.

5. Résultats obtenus pour la région parisienne

5.1. Choix d'un seuil u raisonnable pour Neuilly/Seine

L'intervalle de confiance à 95 % de l'estimation du coefficient de corrélation entre tailles de dépassement ayant eu lieu des jours successifs pour le seuil 90 μgm^{-3} sur le site de Neuilly/Seine ne comprenant pas la valeur zéro (tableau 5.1), ce seuil ne peut pas être utilisé. Par conséquent, pour la station de Neuilly/Seine les seuils u raisonnables sont $\{100, 110, 120, 130, 140, 150\}$.

Des statistiques analogues à celles du tableau 5.1 ont été obtenues pour les trois stations, elles peuvent être consultées dans Bellanger (1999).

TABLEAU 5.1
Seuil u pour Neuilly/Seine, Nb : nombre d'observations, r : coefficient de corrélation et statistiques Bootstrap (med : médiane et quantiles 2.5 % et 97.5 %)

Seuil en μgm^{-3}	90	100	110	120	130	140	150
Estimation du coefficient de corrélation entre intervalles de temps successifs							
Nb	278	224	178	127	87	64	42
r	− 0.008	−0.03	0.10	0.17	0.06	0.00	−0.09
med(r_B)	−0.008	−0.04	0.10	0.17	0.05	−0.01	−0.10
$q_{0.025}$	−0.06	−0.08	−0.03	−0.06	−0.14	−0.16	−0.27
$q_{0.975}$	0.08	0.05	0.26	0.45	0.30	0.23	0.20
Estimation du coefficient de corrélation entre tailles de dépassement							
Nb	162	124	93	59	38	25	14
r	0.37	0.22	0.24	0.33	0.43	0.38	0.41
med(r_B)	0.37	0.21	0.23	0.32	0.45	0.37	0.42
$q_{0.025}$	0.18	−0.03	−0.05	−0.06	−0.03	−0.16	−0.29
$q_{0.975}$	0.54	0.46	0.51	0.65	0.75	0.75	0.91

Conclusion : les seuils u permettant à la fois d'utiliser un PPNH pour modéliser les dépassements et de comparer les résultats obtenus pour chacune des stations mesurant l'ozone en région parisienne sont {100, 110, 120, 130, 140 et 150 μgm^{-3}).

5.2. Modélisation sans interaction

5.2.1. Fréquence des dépassements

Neuilly/Seine, Champs/Marne et Créteil possèdent un effet année positif caractérisant le fait qu'à conditions météorologiques constantes, la fréquence des dépassements du seuil $u = 130$ μgm^{-3} a augmenté au cours de la période d'étude. On observe de plus des différences assez notables entre les estimations des coefficients des variables significatives prises sur chaque site de mesure :

• L'effet de la variable température maximale est plus faible à Créteil (0.380) que sur les autres sites de mesure.

• L'effet de la variable vitesse du vent est beaucoup plus important sur le site de Champs/Marne (-1.086) que sur les autres sites.

• La modélisation de la fréquence des dépassements du seuil $u = 120$ μgm^{-3} et $u = 130$ μgm^{-3} utilise les mêmes variables météorologiques quelle que soit

TABLEAU 5.2
Synthèse de la modélisation de la fréquence des dépassements
(Estimation de $\alpha(t)$, cf. (4.1) et (4.2) sans interaction)
Les chiffres entre parenthèses correspondent à des estimations non significatives

	Modèles					
Station-seuil	NEU120	NEU130	CS120	CS130	AUB130	CRE130
Cte *Ecart-type*	−11.755 1.239	−9.752 1.537	−18.966 2.283	−18.792 2.655	−23.098 2.828	−11.533 1.463
s *Ecart-type* Odd-ratio sur 10 ans int.de conf. 95% du Odd-ratio sur 10 ans	(0.152) (0.104)	0.376 0.128 43 [4-522]	0.715 0.124 1274 [112- 14478]	0.800 0.158 2981 [135- 65960]	(−0.211) (0.195)	0.150 0.075 4.5 [1-20]
t92 Ecart-type		−2.378 0.681				
t93 *Ecart-type*					2.418 0.522	
$t_{\max}$ *Ecart-type*	0.447 0.044	0.452 0.054	0.489 0.064	0.503 0.076	0.659 0.082	0.380 0.048
Vent *Ecart-type*	−0.508 0.131	−0.889 0.199	−0.518 0.186	−1.086 0.285	−0.461 0.202	−0.574 0.173
Test de Kolmogorov-Smirnov						
D_n	0.118	0.118	0.112	0.076	0.175	0.181
$d_{n,0.05}$	0.130	0.158	0.183	0.215	0.176	0.170
test K-S	H_0 acceptée pour $\alpha = 0.05$	H_0 acceptée pour $\alpha = 0.05$	H_0 acceptée pour $\alpha = 0.05$	H_0 acceptée pour $\alpha = 0.05$	H_0 acceptée pour $\alpha = 0.01$	H_0 acceptée pour $\alpha = 0.01$

la station de mesure. Cependant les coefficients affectés à ces variables diffèrent, traduisant l'influence de la position géographique du site de mesure.

• La valeur positive, entre parenthèses (car non significative dans le modèle pour $\alpha = 0.05$) de s, pour **NEU120** confirme la tendance à la hausse des dépassements, tandis que la valeur négative de s entre parenthèse pour **AUB130** met en évidence le caractère particulier de la station d'Aubervilliers.

• La valeur positive de l'estimation du coefficient de la variable année (s) est relativement importante (0.800) à Champs/Marne par rapport à celle de Créteil (0.150). Le signe positif de l'estimation de ce coefficient, nous permet d'affirmer qu'à conditions météorologiques constantes, la fréquence des dépassements du seuil $u = 130\ \mu\text{gm}^{-3}$ a augmenté au cours de la période d'étude pour les stations de Neuilly/Seine, Champs/Marne et Créteil. Cependant, l'estimation de l'intervalle de confiance à 95 % du **rapport des chances** ou «odd-ratio» sur la période 1988-1997 pour chacune de ces stations est très grand, traduisant la difficulté à quantifier cette augmentation de la fréquence des dépassements en utilisant ce modèle.

TABLEAU 5.3
Synthèse de la modélisation de la taille des dépassements
(Estimation de $\beta(t)$ du type de (4.2) sans interaction)

	Modèles					
Station-seuil	NEU120	NEU130	CS120	CS130	AUB130	CRE130
Nbre total de jours	856	856	985	985	857	855
Nbre dep.	95	62	43	32	48	50
Cte *Ecart-type*	0.0692 0.0297				0.141 0.0623	0.184 0.0780
s *Ecart-type*		−0.0006 0.0030				
t92 *Ecart-type*		−0.0358 0.0145				
$t_{\max}$ *Ecart-type*	−0.0024 0.0009	−0.0012 0.0004			−0.0044 0.0018	−0.0063 0.0024
Amtemp *Ecart-type*			0.0032 0.0004	0.0036 0.0005		
Vent *Ecart-type*	0.0137 0.0039	0.0188 0.0046			0.0143 0.0064	0.0237 0.0074
Test de Kolmogorov-Smirnov						
D_n	0.053	0.066	0.053	0.117	0.115	0.089
$d_{n,0.05}$	0.140	0.173	0.139	0.207	0.196	0.192
test K-S	H_0 acceptée pour $\alpha = 0.05$	H_0 acceptée pour $\alpha = 0.05$	H_0 acceptée pour $\alpha = 0.05$	H_0 acceptée pour $\alpha = 0.05$	H_0 acceptée pour $\alpha = 0.05$	H_0 acceptée pour $\alpha = 0.05$

5.2.2. Taille des dépassements

Cette modélisation permet d'observer des différences notables entre les stations. Pour le seuil $u = 130\ \mu\text{gm}^{-3}$:

• la variable année n'est significative que pour la station de Neuilly/Seine et le signe négatif de son coefficient nous permet de conclure que la taille des dépassements a augmenté au cours de la période d'étude.

• la modélisation de la taille des dépassements du seuil $u = 120\ \mu\text{gm}^{-3}$ et (ou) $u = 130\ \mu\text{gm}^{-3}$ sur les sites de Neuilly/Seine, Aubervilliers et Créteil, utilise les mêmes variables météorologiques.

• la seule variable significative pour la station de Champs/Marne est l'amplitude de thermique dont le coefficient prend la valeur 0.0036, c'est-à-dire que plus l'amplitude de température augmente, plus la taille du dépassement sera petite.

Par conséquent, il est difficile de conclure à une augmentation globale de la taille des dépassements dans la Région Parisienne, puisque la variable année n'est significative que sur un des sites modélisés. Cependant, cette modélisation permet d'observer des différences spatiales.

Dans le cadre de la modélisation de la fréquence des dépassements, les intervalles de confiance du rapport des chances sur la période 1988-1997 obtenus pour le seuil 130 μgm^{-3} à Neuilly/Seine, Champs/Marne et dans une moindre mesure Créteil, sont démesurés et instables (à cause d'une très grande variance). Ils ne permettent donc pas de quantifier précisément la tendance. Nous avons donc repris cette modélisation en introduisant les interactions entre les variables météorologiques et l'année.

5.3. Modélisation avec interaction : fréquence des dépassements

Le paramètre $\alpha(t)$ prend donc la forme (4.2) définie dans la section 4.1.1. De plus, dans toute la suite, nous avons utilisé un modèle commun aux quatre stations étudiées pour pouvoir comparer les résultats obtenus. Le modèle de régression logistique retenu dépend de l'année (**s**), de la vitesse moyenne du vent (**Vent**) et de l'interaction année*température maximale (**ttmax**). Après avoir estimé les paramètres des modèles pour chacune des stations, puis validé le modèle, nous avons calculé pour chaque station la température « critique » pour laquelle l'estimation du rapport des chances vaut 1; c'est-à-dire la température pour laquelle le risque d'observer un dépassement en 1997 est le même qu'en 1988. Elle nous permet au vu des signes des estimations des coefficients, de conclure pour chaque station de mesure, à une augmentation (respectivement diminution) du risque d'observer un dépassement entre 1988 et 1997 pour les jours où la température maximale mesurée est supérieure (respectivement inférieure) à cette valeur critique de température.

Pour Neuilly/Seine, le paramètre constant n'est plus significatif à partir du seuil 130 μgm^{-3}, nous l'avons cependant conservé par souci d'homogénéité. La phase de validation du modèle (test de Kolmogorov-Smirnov dans le tableau 5.4) permet de constater que les seuils 120 et 130 μgm^{-3} apparaissent les mieux adaptés.

TABLEAU 5.4
Synthèse de la modélisation de la fréquence des dépassements pour Neuilly/Seine (Estimation de $\alpha(t)$, cf. (4.1) et (4.2) avec interaction)

	Modèles				
Station-seuil	NEU110	NEU120	NEU130	NEU140	NEU150
Nbre total de jours de dep.	125	85	59	41	29
Cte *Ecart-type*	1.629 0.478	1.229 0.568	0.606 0.620	0.437 0.732	1.037 0.864
s *Ecart-type*	−2.672 0.288	−3.036 0.377	−2.604 0.371	−2.475 0.414	−2.840 0.534
Vent *Ecart-type*	−0.825 0.137	−0.879 0.173	−0.813 0.190	−1.007 0.244	−1.225 0.304
$\mathrm{tt_{max}} = \mathrm{s}^* \mathrm{t_{max}}$ *Ecart-type*	0.099 0.010	0.110 0.013	0.093 0.012	0.089 0.013	0.098 0.017
Test de Kolmogorov-Smirnov					
Nbre d'obs.	155	109	74	55	37
D_n	0.114	0.113	0.152	0.196	0.234
$d_{n,0.05}$	0.109	0.130	0.158	0.183	0.234
test K-S	H_0 acceptée pour $\alpha = 0.01$	H_0 acceptée pour $\alpha = 0.05$	H_0 acceptée pour $\alpha = 0.05$	H_0 acceptée pour $\alpha = 0.01$	H_0 acceptée pour $\alpha = 0.01$

La température critique (*cf.* tableau 5.5) est comprise entre 26.9 et 29.1 degrés. C'est-à-dire que le risque d'observer un dépassement sur la période d'étude a principalement augmenté pour les jours de pollution photochimique lors desquels la température enregistrée est élevée. L'intervalle de confiance du rapport des chances sur la période d'étude correspondant à une valeur estimée $\hat{\varphi}$ de φ égale à 1 est beaucoup plus raisonnable que dans le cadre de la modélisation sans interaction.

5.4. Contours de vraisemblance (Ross, 1990)

Afin de contrôler l'exactitude des résultats de notre programme pour estimer les paramètres β (tailles des dépassements), nous avons minimisé directement la log-vraisemblance (fonction $fmins$ de MATLAB) et nous avons obtenu pratiquement les mêmes valeurs pour les trois paramètres de NEU120 : $\{0.0760, -0.0026, 0.0141\}$ à comparer à notre premier résultat $\{0.0692, -0.0024, 0.0137\}$ (*cf.* tableau 5.3).

TABLEAU 5.5

Température maximale correspondant à un rapport de chances $\widehat{\varphi}(s = 1,\ s = 10,\ t_{\max})$ estimé à 1 pour Neuilly/Seine

Seuil	int. de conf. 95%	$t_{\max}$
110	[0.30-3.36]	26.9
120	[0.32-3.16]	27.5
130	[0.28-3.57]	28.0
140	[0.21-4.81]	27.7
150	[0.17-5.89]	29.1

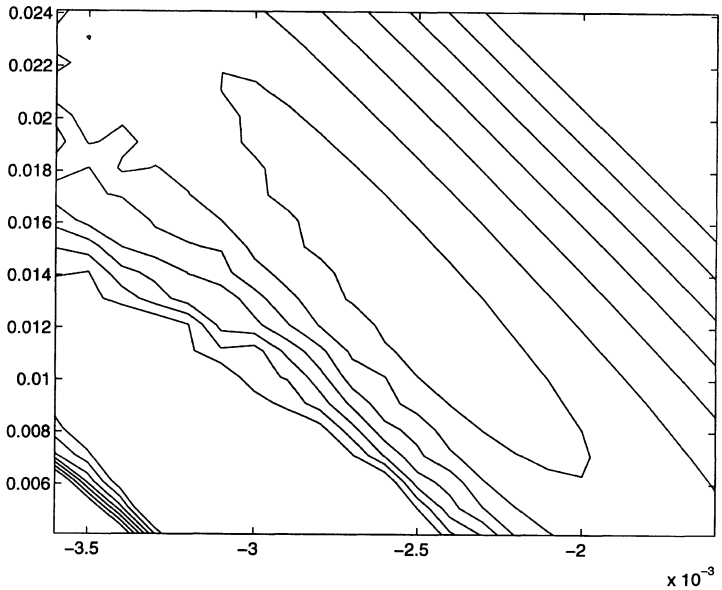

FIGURE 5.1

Contours de la log-vraisemblance pour les données de NEU120.
β_3 en abscisse et β_5 en ordonnée.
Le contour fermé au centre correspond au niveau 420.
Les niveaux des contours vont ensuite, des deux côtés, de 425 à 450.
Les valeurs du minimum sont $\beta_3 = -0.0026$ et $\beta_5 = 0.0141$.

Nous avons ensuite tracé les contours de la log-vraisemblance pour les paramètres β_3 ($T_{\max}$) et β_5 (Vent), en fixant la valeur β_1 (constante) à la valeur estimée 0.0760. Le résultat (*cf.* figure 5.1) paraît satisfaisant; et il traduit bien l'estimation négative (-0.38832) du coefficient de corrélation entre les paramètres.

6. Conclusion

Au terme de ce travail nous voulons faire un certain nombre de remarques. Tout d'abord, la méthode que nous avons présentée n'est qu'une partie de celles que nous avons utilisées et développées dans le cadre de l'étude menée par AirParif : des méthodes plus classiques d'analyse des données et de la statistique ont précédé ce que nous venons de présenter. Elles sont très voisines des études récentes faites sur la ville de Chicago (Davis *et al.*, 1998).

L'utilisation d'un modèle de régression logistique des dépassements d'ozone d'un seuil fixé permet d'utiliser des logiciels classiques pour estimer les paramètres α . Sans interaction le modèle conduit à des résultats difficilement utilisables, dans la mesure où la quantification de l'augmentation du risque (rapport des chances) est mal estimée. La prise en compte de l'interaction entre l'année et la température améliore l'estimation qui devient beaucoup plus plausible; mais alors l'augmentation fait intervenir la température maximale, à travers son interaction avec la tendance et crée de ce fait une nouvelle difficulté. Cette difficulté est inhérente à tout modèle avec interaction; elle traduit la complexité du phénomène physique étudié. Nous pouvons simplement en déduire que l'introduction d'une interaction est nécessaire pour assurer la stabilité du modèle et que le rapport des chances augmente avec la température maximale, ce qui n'était pas évident sur les graphiques (fig. 2.1 et 2.2) quand on ne conservait que les valeurs du maximum dépassant le seuil de 130 μgm^{-3}.

La modélisation de la taille des dépassements demande des moyens de calculs plus délicats à mettre en oeuvre pour estimer les paramètres β.

Les tendances sur une période de dix ans sur les valeurs élevées d'ozone sont complexes, fortement liées à la température, mais aussi à des phénomènes encore plus délicats à prendre en compte non seulement l'évolution du trafic routier et du parc automobile, mais aussi à l'évolution de la technologie des capteurs.

Quand nous avons essayé de transposer ces analyses sur les données de la région de Los Angeles, les suppositions nécessaires à l'utilisation de nos modèles n'étaient pas respectées. Il était difficile de conclure, même si les tendances «allaient dans le bon sens»! Faute d'une étude sur la robustesse des modèles, il était prudent de ne pas conclure.

Des modélisations plus fines dans lesquelles les covariables météorologiques pourront être considérées comme aléatoires, et non fixées comme dans ce que nous avons présenté sont sans doute à étudier. Des analyses théoriques, sans application immédiate, sont en cours et le lecteur pourra s'y reporter (Bellanger, 1999).

Remerciements

Nous remercions l'association AIRPARIF qui nous a donné l'autorisation d'utiliser ses données, ainsi que les auteurs des critiques qui nous ont permis d'améliorer le texte initial.

Bibliographie

BELLANGER L. (1999). *Statistique de la pollution de l'air. Méthodes mathématiques. Application au cas de la région parisienne*. Thèse présentée à l'Université de Paris XI (Paris-Sud Orsay).

BLOOMFIELD P., ROYLE A., YANG Q. (1993a). *Accounting for meteorological effects in measuring urban ozone levels and trends*. Technical Report 1, National Institute of Statistical Sciences, P.O. Box 14162, Research Triangle Park, N.C. 27709.

BLOOMFIELD P., ROYLE A., YANG Q. (1993b). *Rural ozone and meteorology Analysis and comparison with urban ozone*. Technical Report 5, National Institute of Statistical Sciences, P.O. Box 14162, Research Triangle Park, N.C. 27709.

COX D.R., LEWIS P.A. (1966). *The Statistical Analysis of Series of Events*. John Wiley, New York.

COX W.M., SHAO-HANG C. (1993). Meteorologically adjusted ozone trends in urban areas : A probabilistic approach. *Atmospheric Environment*, **27B** (4), 425-434.

CROW L.H. (1974). Reliability analysis for complex repairable systems. In *Reliability and Biometry* (ed. Proschan F. & Serfling R.J.), SIAM, Philadelphia, 379-410.

DAVIS, J.M., EDER, B.K., BLOOMFIELD, P. (1998). Modeling Ozone in the Chicago Urban Area. In *Case Studies in Environmental Statistics* (ed. D. Nychka, W.W. Piegorsch and L.H.Cox), Springer Verlag, New York.

DAVIDSON A.C. (1984). Modelling excesses over high thresholds, with an application In *Statistical extremes and applications* (ed. J. Tiago de Oliveira), Dordrecht, Reidel, 424-434.

DAVIDSON A.C., SMITH R.L. (1990). Models for exceedances over high thersholds (with discussion). *J.R. Statist. Soc.*, **52**, 393-442.

FALK M., HUSLER J., REISS R-D. (1994). *Law of Small Numbers : Extremes and Rare Events*. DMV Seminar 23, Birkhäuser-Verlag.

GAO F., SACKS J., WELCH W.J. (1994). *Predicting the ozone urban levels and trends with semiparametric modelling*. Technical Report 14, National Institute of Statistical Sciences, P.O. Box 14162, Research Triangle Park, N.C. 27709.

HOSKING J.M.R., WALLIS J.R. (1987). Parameter and quantile estimation for the generalized Pareto distribution. *Technometrics*, **29**, 339-349.

HOSMER D.W., LEMESHOW S. (1989). *Applied Logistic Regression*. John Wiley, New York.

LEADBETTER M.R. (1991). On a basis for "Peaks over Threshold". *Statistics and Probability Letters*, **12**, 357-362.

LEADBETTER M.R. (1993). *On exceedance based environmental criteria*. Technical Report 9, National Institute of Statistical Sciences, P.O. Box 14162, Research Triangle Park, N.C. 27709.

LEADBETTER M.R. (1995). On high level exceedance modelling and tail inference. *J. Stat. Plan. Inference*, **45** (1-2), 247-260.

LEADBETTER M.R., LINDGREN G., ROOTZEN H. (1983). *Extremes and Related Properties of Random Sequences and Series*. Springer Verlag, New York.

LEE L. (1980). Testing adequacy of the Weibull and log linear rate models for a Poisson process. *Technometrics*, **22** (2), 195-199.

PICKANDS J. (1971). The two-dimensional Poisson process and extremal processes. *J. Appl. Prob.* **8**, 745-756.

PICKANDS J. (1975). Statistical inference using extreme order statistics. *Ann. Statist.* **3** , 119-131.

PUGH, D., VASSIE, J.M. (1980). Applications of the joint probability method for extreme sea level computations. *Proc. Instn. Civ. Engrs*, Part 2, **69**, 959-975.

ROSS G.J.S. (1990). *Nonlinear estimation*. Springer-Verlag, London.

SHAO J., TU D. (1996). *The Jackknife and Bootstrap*. Springer Verlag, New York.

SHIVELY T.S. (1990). An analysis of the long-term trend in ozone data from two Houston, Texas monitoring sites. *Atmospheric Environment*, **24B** (4), 293-301.

SHIVELY T.S. (1991). An analysis of the trend in ground-level ozone using nonhomogeneous Poisson processes. *Atmospheric Environment*, **25B** (4), 387-396.

SMITH R.L. (1984). Threshold methods for sample extremes. In *Statistical extremes and applications* (ed. J. Tiago de Oliveira), Dordrecht, Reidel, 621-638.

SMITH R.L. (1985). Maximum likelihood estimation in a class of nonregular cases. *Biometrika*, **72** (1), 67-90.

SMITH R.L. (1986). Extreme value theory based on the r largest annual events. *J. Hydrology*, **86**, 27- 43.

SMITH R.L. (1987). Estimating tails of probability distributions. *The Annals of Statistics*, **15** (3), 1174-1207.

SMITH R.L. (1989). Extreme values analysis of environmental time series : An application to trend detection in ground-level ozone (with discussion). *Statistical Sciences*, **4**, 367-393.

SMITH R.L., SHIVELY T.S. (1995). Point process approach to modelling trends in tropospheric ozone based on exceedances of a high threshold. *Atmospheric Environment*, **29** (3), 3489-3499.

VAQUERA-HUERTA H., VILLASENOR J.A., HUGHES J. (1997). Statistical analysis of trends in urban ozone. *In Statistics for the Environment 3 : Pollution Assessment and Control* (ed. Barnett V. and Turkman K.F.), John Wiley, New York, 175-183.

WEISSMAN I. (1978). Estimation of parameters and large quantiles based on the k largest observations. *J. Amer. Statist. Assoc.*, **73**, 812-815.