

REVUE DE STATISTIQUE APPLIQUÉE

LISE BELLANGER

Une analyse globale de la tendance dans les hautes valeurs d'ozone mesurées en région parisienne

Revue de statistique appliquée, tome 49, n° 3 (2001), p. 73-92

http://www.numdam.org/item?id=RSA_2001__49_3_73_0

© Société française de statistique, 2001, 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/>

UNE ANALYSE GLOBALE DE LA TENDANCE DANS LES HAUTES VALEURS D'OZONE MESURÉES EN RÉGION PARISIENNE

Lise BELLANGER

CNRS UMR 8628, Probabilités, Statistique et Modélisation, Université Paris-Sud,
91405 Orsay Cédex. email : lise.bellanger@math.u-psud.fr

RÉSUMÉ

Dans cet article, nous complétons l'analyse des valeurs extrêmes sur les niveaux d'ozone enregistrés en région parisienne, décrite dans un précédent article ¹. Un processus de Poisson non-homogène, à paramètres dépendant de covariables, est utilisé pour modéliser les jours et les tailles de dépassement d'un niveau d'ozone très élevé. Nous appliquons cette approche à quatre stations de mesures. Nous construisons ensuite un modèle global regroupant les différentes stations étudiées, en introduisant des variables indicatrices les représentant. Enfin, nous comparons la modélisation par station à la modélisation globale.

Mots-clés : Ozone, Dépassements de seuil, Tendence, Application de la théorie des valeurs extrêmes, Processus de Poisson non-homogène, Régression logistique, Distribution de Pareto généralisée.

ABSTRACT

In this paper, we complete the extrem value analysis of ozone level exceedances registred in Paris area described in a previously one. A non-homogeneous Poisson process, which parameters depend on covariables, is used to model the occurences and sizes of exceedances over a very high level threshold. This approach is applied to four available sites. Then, we construct a global model gathering these sites, by introducing dummy variables representing sites. Finally, we compare station model and global model.

Keywords : Ozone, Level exceedance, Trend, Extreme value analysis, Non-homogeneous Poisson process, Logistic regression, Generalized Pareto distribution.

1. Introduction

La théorie des valeurs extrêmes s'est très largement développée ces dernières années; de nombreux ouvrages de synthèse ont été publiés comme Leadbetter *et al.* (1983), Falk *et al.* (1994) et plus récemment Embrechts *et al.* (1999). Ceci a conduit

¹ BELLANGER L., TOMASSONE R. (2000), La pollution de l'air dans la région parisienne : étude de la tendance dans les hautes valeurs d'ozone. *Revue de Statistiques Appliquées* XLVIII (1), 5-24.

à de nombreuses applications dans les domaines les plus variées : finance, assurance, hydrologie, météorologie, océanographie, pollution de l'air, etc.

Il existe un nombre important d'approches possibles permettant de modéliser les extrêmes de suites de données. Elles dépendent de la structure et de la complexité des données; nous pouvons les classer de la manière suivante :

- *Etude et modélisation du maximum annuel ou d'une statistique d'ordre* : si les suites sont assez longues, une méthode classique consiste à modéliser le maximum annuel de périodes consécutives de taille égale (par exemple les années, mois ou jours) des séries (supposées iid) par une des distributions de valeur extrême comme dans Gumbel (1958). Mais cette méthode présente un défaut majeur : elle demande un nombre important de données difficiles à obtenir, puisque caractéristiques de phénomènes rares. Une autre méthode d'analyse est basée sur un nombre fixé de statistiques d'ordre. L'estimation des paramètres est plus complexe, puisque les densités doivent prendre en compte la dépendance entre observations.

- *Etude des pics dépassant un seuil* : la méthode POT ("Peaks Over Threshold"), très utilisée, a été développée pour la première fois dans le Flood Studies Report (NERC, (1975)); elle a été décrite dans Leadbetter M.R. (1991) et Smith R.L. (1984), elle est basée sur l'estimation des paramètres d'un modèle stochastique représentant, soit des dépassements, soit des pics au-dessus d'un seuil.

- *Etude et modélisation des jours et des tailles de dépassement de très haut niveau par un processus de Poisson non homogène* : les suites d'observations étudiées peuvent être indépendantes ou présenter une tendance, un phénomène de saisonnalité, une dépendance à long ou à court terme, ceci ne fera que compliquer l'analyse.

Dans cet article, nous reprenons la méthodologie développée dans Bellanger et Tomassone (2000) pour étudier la tendance dans les hautes valeurs d'ozone enregistrées en région parisienne. Les valeurs d'ozone fournies par AIRPARIF, correspondent à celles mesurées sur les sites de Neuilly/Seine, Aubervilliers, Champs/Marne et Créteil, pour les mois de mai à septembre de la période 1988-1997. Le processus bi-dimensionnel des jours et des tailles de dépassement est modélisé par un processus de Poisson non-homogène, dont les paramètres peuvent dépendre de l'année et des variables météorologiques disponibles suivantes² : température maximale, amplitude des vitesses de vent, vitesse du vent et amplitude des vitesses de vent.

Dans la deuxième partie, nous rappelons brièvement le modèle mis en œuvre.

Dans la troisième partie, nous complétons les résultats obtenus pour les quatre stations étudiées et nous justifions l'utilisation d'une loi exponentielle pour approcher la loi des tailles de dépassement, sachant qu'un dépassement a eu lieu.

Enfin, dans la quatrième partie, nous construisons un modèle global regroupant les différentes stations étudiées. Leur faible nombre ne nous permettait pas d'utiliser des méthodes de statistique spatiale (comme celle développée par Coles et Tawn (1996) pour des données pluviométriques), nous ne pouvions pas espérer obtenir un véritable modèle régional, avec la possibilité d'estimer des valeurs sur des stations de mesures sans observations. Nous employons donc une technique, consistant à introduire des variables indicatrices, représentant les stations; puis à les prendre

² Fournies par le mât du Commissariat à l'Énergie Atomique de Saclay.

comme covariables dans les paramètres du processus de Poisson non-homogène. Nous comparons ensuite la modélisation par station à la modélisation globale.

2. Processus de Poisson non-homogènes (PPNH) dans le plan

Nous rappelons ici ce que nous avons déjà développé dans un précédent article (Bellanger et Tomassone, 2000), à savoir que les jours de dépassement d'un seuil u élevé et les tailles de ce dépassement peuvent être approchées par un processus de Poisson non-homogène (PPNH). Pour développer ce modèle, nous avons posé :

$$\Psi_i(y) = \begin{cases} P(Y > y \text{ le jour } i) \\ 0 \text{ si le jour } i \text{ est manquant.} \end{cases}$$

La distribution de la variable aléatoire Y (maximum d'ozone journalier) le jour i , s'écrit donc $1 - \Psi_i(y)$, et on notera $\psi_i(y) = -\frac{d}{dy}[\Psi_i(y)]$, sa densité.

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 $(T_i; Y_i)$, $1 \leq i \leq N$, où T_i et Y_i sont supposées indépendantes $\forall i$, le $i^{\text{ème}}$ pic se produit le jour T_i et le maximum d'ozone Y_i est supérieur ou égal à u . Le nombre total des N pics étant lui aussi une variable aléatoire, la densité conjointe des pics observés peut être approchée par :

$$L = \left[\left(\prod_{i=1}^N \Psi_{t_i}(u) \right) \exp \left[- \int_0^T \Psi_t(u) dt \right] \right] * \left[\prod_{i=1}^N \frac{\psi_{t_i}(y_i)}{\Psi_{t_i}(u)} \right] = A * B \quad (2.1)$$

où la signification de chacun des termes ci-dessus est décrite dans Bellanger et Tomassone (2000).

2.1. Modèle associé aux jours de dépassement

Il existe de nombreuses façons de modéliser l'intensité $\Psi_i(u)$ du processus de Poisson en tenant compte de la relation entre très hautes valeurs d'ozone et conditions météorologiques (Shively (1991), Smith et Shively (1995), Vaquera-Huerta *et al* (1997), Bellanger et Tomassone (2000)). Nous avons choisi de conserver le modèle de régression logistique (Hosmer et Lemeshow, 1989; Bellanger et Tomassone, 2000), à cause de sa flexibilité, de l'interprétation relativement simple des estimations des paramètres et de son implantation dans de nombreux logiciels de statistiques tel SAS (proc LOGISTIC dont la description se trouve dans SAS *Institute Inc.* SAS/STAT User's Guide (1994)).

L'intensité λ du PPNH s'écrit donc :

$$\lambda(i) = \Psi_i(u) = \frac{\exp(\alpha(i))}{1 + \exp(\alpha(i))} \quad (2.2)$$

où $\alpha(i)$ s'écrit :

$$\alpha(i) = \alpha_0 + \alpha_1 t(i) + \sum_{j=2}^p \alpha_j w_j(i) + \sum_{j=2}^p \alpha_{1j} t(i) w_j(i) + \sum_{k,j=2}^p \alpha_{kj} w_k(i) w_j(i) \quad (2.3)$$

$t(i)$ est un terme de tendance prenant la valeur k si le jour i appartient à l'année k (dans notre cas, $k \in \{1, 2, \dots, 10\}$), $w_j(i)$ représente la valeur de la variable météorologique j le jour i . L'écriture de $\alpha(i)$ tient donc compte des interactions possibles entre covariables.

2.2. Modèle associé aux tailles de dépassement

Les résultats théoriques dus à Pickands (1975) et Davison et Smith (1990) sur la distribution de la taille des dépassements d'un seuil élevé u (notée, $X = Y - u$), nous permettent d'approcher la distribution de $X = Y - u$ sachant que $Y \geq u$, par une distribution de Pareto généralisée notée G :

$$P[Y_i > u + x / Y_i > u] = 1 - G(x; \beta(i), \xi(i)) = \left(1 + \xi(i) \frac{x}{\beta(i)}\right)^{-\frac{1}{\xi(i)}} \quad (2.4)$$

avec $x = y - u$ où $\beta(i) > 0 \forall i$, Y_i étant la variable aléatoire maximum d'ozone le jour i .

La densité limite de X le jour i sachant qu'un dépassement du seuil u a eu lieu ce jour-là, s'écrit donc :

$$\lim_{u \rightarrow +\infty} \frac{\psi_i(y)}{\Psi_i(u)} = g(x; \beta(i), \xi(i)) = \frac{1}{\beta(i)} \left(1 + \xi(i) \frac{x}{\beta(i)}\right)^{-\left(\frac{1}{\xi(i)} + 1\right)} \quad (2.5)$$

Pour u suffisamment grand (2.5) permet d'approcher la densité de X . Par souci de simplification nous avons de plus supposé que le paramètre ξ ne dépendait pas de i et que $\beta(i)$, comme $\alpha(i)$ dans (2.3), prenait la forme analytique suivante :

$$\beta(i) = \beta_0 + \beta_1 t(i) + \sum_{j=2}^p \beta_j w_j(i) + \sum_{j=2}^p \beta_{1j} t(i) w_j(i) + \sum_{k,j=2}^p \beta_{kj} w_k(i) w_j(i) \quad (2.6)$$

afin d'étudier la dépendance des tailles avec des variables météorologiques, la tendance à long terme $t(i)$ et de possibles interactions. Si u est suffisamment grand, la densité de X peut donc être approchée par :

$$g(x; \beta(i), \xi) = \frac{1}{\beta(i)} \left(1 + \xi \cdot \frac{x}{\beta(i)}\right)^{-\left(\frac{1}{\xi} + 1\right)} \quad (2.7)$$

où $\beta(i)$ prend la forme (2.6).

Remarque : si $\xi \approx 0$ comme $\lim_{\xi \rightarrow 0} g(x; \beta(i), \xi) = \frac{1}{\beta(i)} \exp\left(-\frac{x}{\beta(i)}\right)$, on obtient la densité exponentielle de paramètre $\frac{1}{\beta(i)}$.

Pour chacun de ces deux modèles, la méthode du maximum de vraisemblance permet d'estimer les paramètres de (2.3) et (2.6), puis une procédure "backward" permet de ne conserver que les variables significatives. Les calculs ont été effectués avec le logiciel SAS pour les jours de dépassement et avec le logiciel SPLUS (Venables et Ripley (1997)) pour les tailles de dépassement. Nous avons aussi adapté certaines fonctions développées sous Splus (Coles (1999)) et disponibles sur Internet (<http://maths.lancs.ac.uk/> col).

3. Choix a priori des seuils raisonnables et tests de validation du modèle

3.1 Choix a priori des seuils raisonnables

Le choix du seuil u est complexe : trop faible nous ne pourrions pas utiliser les résultats asymptotiques, trop élevé nous aurons peu d'observations et une grande variabilité.

L'analyse de la vraisemblance est une première approche permettant de choisir les seuils raisonnables : les résultats asymptotiques suggèrent que le modèle est valide pour tous les seuils supérieurs à une certaine valeur inconnue! Donc, le calcul des estimations des paramètres du modèle pour plusieurs seuils doit nous conduire à observer une stabilité de ces estimations. Il suffit d'examiner les graphes respectifs de chaque estimation de paramètre avec son intervalle de confiance correspondant (± 1.96 écart-type) par rapport au seuil.

Une autre méthode graphique permet de sélectionner les seuils pour lesquels il est possible d'utiliser une distribution de Pareto généralisée à paramètres constants. En effet, si la distribution de $X = Y - u$ conditionnellement au seuil u , peut être approchée par distribution de Pareto généralisée définie en (2.4), son espérance prend la forme :

$$E[Y - u/Y > u] = \frac{\beta_0 + \xi \cdot u}{1 - \xi} \quad \text{où } \xi < 1 \quad (3.1)$$

De plus, si le modèle GPD est valide pour un certain seuil u , alors il devra l'être pour tout seuil supérieur. Par conséquent, le graphe de la taille moyenne observée des dépassements du seuil u par rapport à u devra être linéaire au-dessus d'une valeur du seuil pour laquelle les résultats asymptotiques sont valides. Ce graphe est généralement appelé *mean residual life plot* (Yang, 1978, Hall et Wellner, 1981). D'après (3.1), sa pente et son ordonnée à l'origine fournissent des estimations simples de $\xi/1 - \xi$ et $\beta_0/1 - \xi$.

Naturellement, ces techniques ne nous fournissent qu'une estimation graphique de β_0 et ξ , indications insuffisantes dans la mesure où β dépend de i (2.6). De plus, il faut naturellement ensuite vérifier les hypothèses d'indépendance des dates et des tailles respectivement, pour pouvoir utiliser un processus de Poisson non-homogène.

Nous avons déjà montré (Bellanger et Tomassone, 2000) qu'elle pouvait se faire en tout début d'étude.

3.2 Validation du modèle

3.2.1 Tests

Après avoir estimé les paramètres du modèle, il est nécessaire de s'assurer que les hypothèses distributionnelles effectuées sur les jours et les tailles de dépassement, sachant qu'un dépassement de seuil u a eu lieu, sont respectées. Un test de Kolmogorov-Smirnov, ainsi qu'un test d'une distribution exponentielle d'espérance inconnue (ou test d'exponentialité) décrit dans Pearson et Hartley (1976), permettent de vérifier que la distribution exponentielle est la distribution appropriée pour modéliser les intervalles de temps entre deux dépassements et les tailles de dépassement.

Nous avons aussi utilisé des *Percentage-percentage (P-P) probability Plots*, graphiques fréquemment employés pour valider une hypothèse distributionnelle. Ces graphes sont décrits dans de nombreux ouvrages de statistiques appliquées comme ceux de Snedecor et Cochran (1971), Lawless (1982), Draper et Smith (1981). Supposons que Y soit une variable aléatoire de distribution $F(y)$ et que $X = \sigma Y + \mu$, où μ est le paramètre de position et $\sigma (\sigma > 0)$ est le paramètre d'échelle. La variable aléatoire X possède donc la distribution $G(x) = F((x - \mu)/\sigma)$ et la variable $U = F((x - \mu)/\sigma)$ est uniformément distribuée sur $]0, 1]$. Par conséquent, si $x_{(1)}, \dots, x_{(n)}$ représentent les observations ordonnées d'un échantillon aléatoire de taille n de distribution G , un P-P probability plot standardisé est donc obtenu en traçant $F((x_{(i)} - \mu)/\sigma)$ par rapport à $i/(n + 1)$. Si l'hypothèse distributionnelle est raisonnable, le graphe correspondra approximativement à la première bissectrice. Les paramètres μ et σ sont généralement inconnus et peuvent être remplacés par leurs estimations obtenues en appliquant la méthode du maximum de vraisemblance. Gan et Koehler (1990) et Gan et Koehler (1992) ont ainsi développé un test basé sur le carré du coefficient de corrélation entre les points d'un P-P probability plot standardisé (k^2), dont ils calculent la probabilité observée pour tester la qualité des estimations en fonction de l'hypothèse distributionnelle choisie (Lois Normale, Gumbel ou exponentielle). Nous ne pouvons pas appliquer ce test. En effet, dans notre cas, les paramètres μ et σ ne sont pas constants.

3.2.2 Influence des observations individuelles

L'étude des résidus permet de diagnostiquer les observations influentes. Dans le cadre de la modélisation des jours de dépassement d'un seuil élevé, nous avons utilisé un modèle de régression logistique pour approcher l'intensité du processus de Poisson. Puis nous avons supprimé les observations trop influentes en utilisant les diagnostics de régression développés par Pregibon (1981). L'étude des observations influentes dans le cadre de la régression logistique ne pose aucun problème, puisqu'elle est implantée dans la procédure LOGISTIC du logiciel SAS (option INFLUENCE).

4. Application aux données d'ozone mesurées en région parisienne

Comme dans Bellanger (1999), nous n'avons conservé que les seuils u allant de 110 à 140 μgm^{-3} pour modéliser les jours et les tailles de dépassement. Nous allons d'abord donner les résultats par station, puis ceux fournis par un modèle global. Les stations seront notées NE, AU, CH et CR; ainsi par exemple le modèle correspondant à la station de Neuilly/Seine pour le seuil 130 μgm^{-3} s'écrira NE130.

4.1 Modèle par station

Dans toute la suite, le paramètre $\alpha(i)$ prend donc la forme (2.3). Dans la mesure où notre but est de comparer les résultats obtenus pour les différentes stations de mesures, nous avons utilisé un modèle commun aux quatre stations. Le modèle de régression logistique retenu dépend de l'année (t), de la vitesse moyenne du vent (Vent) et de l'interaction année*température maximale ($tt\text{max}$). Il s'écrit :

$$\alpha(i) = \alpha_0 + \alpha_1 t(i) + \alpha_2 \text{vent}(i) + \alpha_3 tt \max(i)$$

4.1.1 Modélisation des jours de dépassement

En utilisant les résultats de Prégibon (1981) pour diagnostiquer les observations influentes dans une régression logistique, nous avons supprimé 13 (respectivement 15, 15 et 24) observations pour Neuilly/Seine (respectivement Aubervilliers, Champs/Marne et Créteil).

Les résultats du test de Kolmogorov-Smirnov (Tab 4.1) et les P-P probability plot (Fig 4.1) permettent de valider l'hypothèse distributionnelle. Les résultats (Tab.4.1) montrent que l'effet du vent entraîne toujours une diminution significative de la probabilité d'observer un dépassement (coefficient négatif de cette variable). Pour la température, nous sommes confrontés à la difficulté de l'interprétation de l'interaction $tt\text{max}$: le calcul de la température critique $\hat{\varphi}$ dans Bellanger et Tomassone (2000) fournit toujours les valeurs les plus élevées pour Neuilly et, à l'opposé, les plus faibles pour Champs/Marne. Cette constance des résultats traduit des différences locales pouvant laisser supposer que le risque d'observer un dépassement a augmenté sur un ensemble de jours de la période d'étude plus grand à Champs/Marne qu'à Neuilly. : ces jours correspondant à ceux où, un dépassement est observé et la température est supérieure à la température critique $\hat{\varphi}$. Cette conjecture devrait, bien sûr, être validée par les spécialistes de l'étude de la pollution de l'air.

4.1.2 Modélisation des tailles de dépassement

Pour les tailles, nous devons d'abord nous assurer que leur distribution est bien exponentielle, puis faire des calculs analogues à ceux du paragraphe précédent.

TABLEAU 4.1
*Synthèse de la modélisation des jours de dépassement pour les quatre stations
 (Estimation des coefficients du paramètre $\alpha(i)$ par un modèle de régression logistique)*

Station-seuil # dépassements. (# jours)	NE110	NE120	NE130	NE140	AU110	AU120	AU130	AU140	CH110	CH120	CH130	CH140	CR110	CR120	CR130	CR140
-2 Log L	125 (843)	85 (843)	59 (843)	41 (843)	100 (845)	70 (845)	44 (845)	31 (845)	59 (970)	38 (970)	28 (970)	19 (970)	92 (836)	58 (836)	38 (836)	24 (836)
Constante	1.629	1.229	0.606	0.437	-1.646	-1.671	-1.306	-1.445	-3.194	-4.190	-2.708	-2.294	-0.508	-0.566	-0.208	-0.423
Ecart-type	0.478	0.568	0.620	0.732	0.517	0.648	0.800	0.936	0.692	0.908	1.170	1.197	0.505	0.659	0.834	1.093
t (année)	-2.672	-3.036	-2.604	-2.475	-1.815	-2.221	-2.532	-2.727	-1.835	-1.430	-1.789	-1.529	-1.767	-2.006	-1.846	-1.631
Ecart-type	0.288	0.377	0.371	0.414	0.224	0.313	0.415	0.499	0.309	0.466	0.466	0.471	0.226	0.312	0.365	0.443
Vent	-0.825	-0.879	-0.813	-1.007	-0.484	-0.767	-0.844	-0.847	-0.791	-0.881	-2.021	-1.685	-0.638	-0.940	-1.474	-2.212
Ecart-type	0.137	0.173	0.190	0.244	0.140	0.198	0.255	0.304	0.194	0.245	0.462	0.433	0.143	0.211	0.317	0.487
tmax=t*tmx	0.099	0.110	0.093	0.089	0.081	0.096	0.101	0.104	0.089	0.076	0.094	0.067	0.074	0.083	0.078	0.071
Ecart-type	0.010	0.013	0.012	0.013	0.008	0.011	0.014	0.016	0.011	0.011	0.017	0.016	0.008	0.010	0.012	0.015
ϕ	27.0	27.6	28.0	27.8	22.4	23.1	25.1	26.2	20.6	18.8	19.0	22.8	23.9	24.2	23.7	23.0
D _n	0.134(*)	0.140(NS)	0.126(NS)	0.153(NS)	0.141(*)	0.208(*)	0.122(NS)	0.110(NS)	0.137(NS)	0.097(NS)	0.109(NS)	0.158(NS)	0.140(NS)	0.145(NS)	0.081(NS)	0.136(NS)

Statistique D_n de Kolmogorov-Smirnov: (NS) non significatif à 5%, (*) significatif à 5%, (**) significatif à 1%

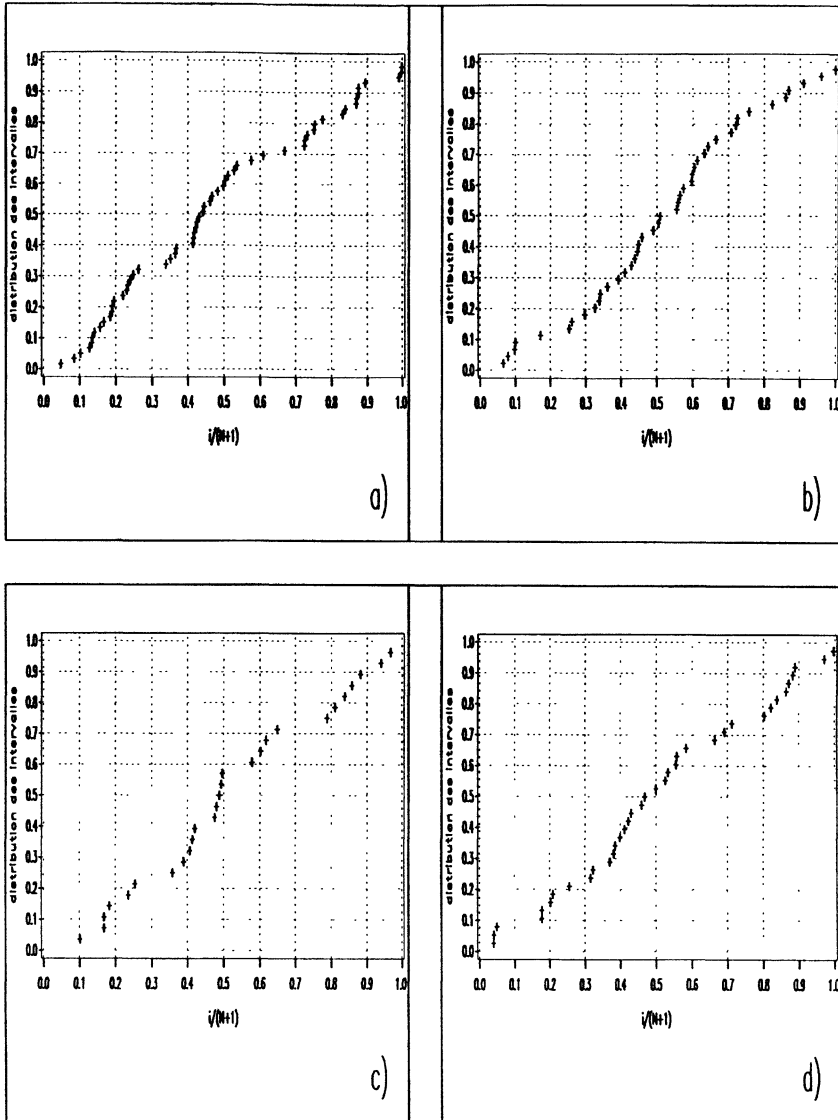


FIGURE 4.1
*P-P probability plot : modélisation des jours de dépassement
pour les quatre stations (seuil $130 \mu\text{mg}^{-3}$)*

TABLEAU 4.2
Estimation des paramètres de la loi de Pareto Généralisée pour les quatre stations.

Station-seuil	NE110	NE120	NE130	NE140	AU110	AU120	AU130	AU140	CH110	CH120	CH130	CH140	CR110	CR120	CR130	CR140
Nbre de dép.	138	95	62	43	108	78	48	32	65	43	32	20	110	73	50	26
log L	604.6	418.242	278.990	196.211	456.436	323.495	200.459	133.144	271.151	177.617	127.665	77.631	454.232	297.496	197.686	108.372
β	27.542	27.219	32.556	39.144	28.815	25.167	28.255	29.506	33.344	37.928	32.967	38.255	23.738	21.437	16.031	24.212
<i>Ecart-type</i>	3.497	4.458	6.777	9.784	3.686	4.115	5.467	6.565	5.967	7.492	7.984	11.615	2.913	3.400	3.811	6.301
ξ	-0.066	0.099	0.017	-0.104	-0.135	-0.083	-0.165	-0.224	-0.335	-0.505	-0.506	-0.763	-0.038	0.010	-0.179	-0.019
<i>Ecart-type</i>	0.094	0.128	0.165	0.199	0.085	0.118	0.131	0.141	0.136	0.147	0.188	0.267	0.077	0.107	0.193	0.171

TABLEAU 4.4

*Synthèse de la modélisation des tailles de dépassement pour les quatre stations
(Estimation des coefficients du paramètre $\beta(i)$ de la loi exponentielle)*

Station-seuil	NE110	NE120	NE130	NE140	AU110	AU120	AU130	AU140	CH110	CH120	CH130	CH140	CR110	CR120	CR130	CR140
# dépassements.	138	95	62	43	108	78	48	32	65	43	32	20	110	73	50	26
Log L	600.756	414.023	274.570	193.740	455.526	320.989	196.441	131.921	272.686	179.545	129.150	80.936	452.113	293.712	192.315	107.878
constante	41.429	43.580	66.506	67.957	34.626	33.557	45.259	47.489	29.810	34.417	29.215	17.408	30.460	31.395	33.075	37.578
<i>Ecart-type</i>	4.335	4.896	10.239	18.113	5.085	4.864	7.155	11.775	6.066	6.749	6.787	12.203	4.148	4.296	5.054	15.279
Vent	-4.221	-4.897	-12.253	-12.529	-3.009	-3.457	-7.009	-8.294	-1.700	-3.094	-2.819	-1.489	-2.516	-3.312	-5.017	-5.789
<i>Ecart-type</i>	0.867	0.779	2.377	5.484	1.187	0.971	1.226	2.929	1.487	1.239	1.436	4.805	0.919	0.717	0.865	5.374
Statistique D_n de Kolmogorov-Smirnov et E_n test d'exponentialité : (NS) non significatif à 5%, (*) significatif à 5%, (**) significatif à 1%.																
D_n	0.057(NS)	0.051(NS)	0.047(NS)	0.089(NS)	0.105(NS)	0.056(NS)	0.109(NS)	0.168(NS)	0.097(NS)	0.144(NS)	0.103(NS)	0.175(NS)	0.071(NS)	0.094(NS)	0.115(NS)	0.171(NS)
E_n	0.668(NS)	0.90(NS)	0.355(NS)	0.584(NS)	1.108(*)	0.491(NS)	0.764(NS)	0.973(NS)	0.790(NS)	0.961(NS)	0.579(NS)	0.801(NS)	0.746(NS)	0.811(NS)	0.821(NS)	0.890(NS)

4.1.2.1 Distribution de Pareto généralisée ou distribution exponentielle ?

Supposons que la taille de dépassement, quand un dépassement a eu lieu, suive une loi de Pareto généralisée de paramètres constants β et ξ .

On constate (Tab 4.2) que seule la station de Champs/Seine possède une estimation du paramètre ξ significative. Cependant, le test du rapport des vraisemblances maximales de l'hypothèse $\xi = 0$, pour les différents seuils étudiés, suggère qu'on ne peut pas rejeter cette hypothèse. Par conséquent, la distribution de Pareto généralisée peut être réduite à la distribution exponentielle ($\xi = 0$). Par exemple, pour le seuil 130, la statistique de déviance (à comparer à un χ_1^2) est :

TABLEAU 4.3
Calcul de la déviance $-2[\text{Log}(L(\hat{\beta}, 0)) - \text{Log}(L(\hat{\beta}, \hat{\xi}))]$ pour le seuil 130

Station-seuil	NE130	AU130	CH130	CR130
$-2[\text{Log}(L(\hat{\beta}, 0)) - \text{Log}(L(\hat{\beta}, \hat{\xi}))]$ ($P[\chi_1^2 > 3.841] = 5\%$).	0.0052	0.544	2.138	0.629

Où $\text{Log}(L(\hat{\beta}, \hat{\xi}))$ représente la Log-vraisemblance du modèle pour β et ξ estimés par la méthode du maximum de vraisemblance; $\text{Log}(L(\hat{\beta}, 0))$ celle sous l'hypothèse nulle.

Par souci d'homogénéité, nous supposons donc dans toute la suite que, pour toutes les stations, la taille de dépassement, quand un dépassement a eu lieu le jour i , suit une loi exponentielle de paramètre $\beta(i)$ prenant la forme (2.6).

4.1.2.2 Modèle exponentiel

Le tableau 4.4 synthétise les résultats de la modélisation des tailles de dépassement par une loi exponentielle de paramètre $1/\beta(i)$. En outre, à l'instar de la modélisation des jours de dépassement, nous avons décidé de prendre un modèle commun aux quatre stations étudiées de façon à pouvoir comparer et mieux appréhender les relations entre les stations. Le paramètre $\beta(i)$ retenu ne dépend que de la vitesse du vent, il s'écrit donc :

$$\beta(i) = \beta_0 + \beta_1 \text{vent}(i).$$

Les résultats du test de Kolmogorov-Smirnov (Tab 4.4) et les P-P probability plot (Fig 4.2) permettent de valider l'hypothèse distributionnelle. Le modèle ne permet pas de déceler de tendance à moyen terme dans les tailles de dépassement. Il semble en effet que le seul facteur influent commun aux quatre stations, soit de nature météorologique (Tab 4.4) : la vitesse du vent. Le modèle commun obtenu est plus simple que celui développé dans Bellanger & Tomassone (2000) : nous retrouvons un phénomène bien connu des experts : plus la vitesse du vent augmente, plus la taille moyenne d'un dépassement diminue, quand un dépassement a eu lieu. Nous observons

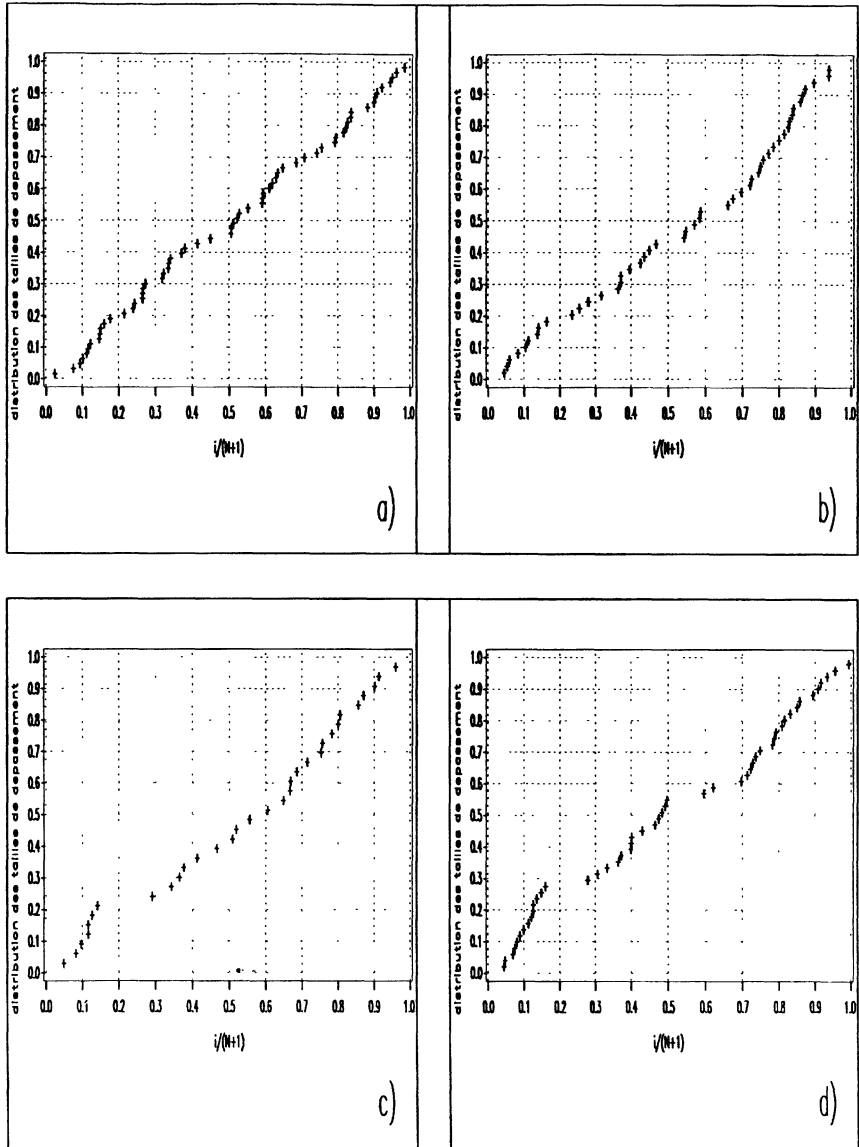


FIGURE 4.2
*P-P probability plot : modélisation des tailles de dépassement
pour les quatre stations (seuil $130 \mu\text{mg}^{-3}$)*

aussi une certaine instabilité entre les estimations du paramètre correspondant à la vitesse du vent, pour les seuils strictement inférieurs à $130 \mu\text{mg}^{-3}$ et les seuils 130 et $140 \mu\text{mg}^{-3}$:

- pour NE120, $\hat{\beta}_1 = -4.897$ avec un écart type de 0.779 et
- pour NE130, $\hat{\beta}_1 = -12.253$ avec un écart type bien supérieur 2.377

Le rôle particulier du seuil $130 \mu\text{mg}^{-3}$ dans le domaine de la surveillance de qualité de l'air en Ile-de-France, pourrait peut-être expliquer ce phénomène de rupture observé principalement pour les stations de Neuilly/Seine, Aubervilliers et Créteil. Il est cependant important de noter que les estimations obtenues pour le seuil 140 paraissent peu fiables au vu du faible nombre d'observations sur lequel elles se basent. En dernier lieu, au vu de l'ensemble des estimations de β_0 et β_1 obtenues, les tailles de dépassements d'un seuil fixé mesurées sur chacun des quatre stations semblent présenter des différences assez notables.

4.2 Modèle global

Les modèles développés au paragraphe précédent fournissent des résultats par station; il peut être intéressant de les regrouper en un modèle global qui pourra s'appuyer sur l'ensemble des observations, donc augmenter le nombre de dépassements pris en compte dans un modèle unique. La technique employée est tout à fait courante dans les applications du modèle linéaire : nous introduisons trois variables indicatrices D1, D2 et D3. Pour Neuilly nous avons le triplet $\{0, 0, 0\}$, et pour trois autres respectivement $\{1, 0, 0\}$ pour Aubervilliers, $\{0, 1, 0\}$ pour Champs sur Marne et $\{0, 0, 1\}$ pour Créteil.

4.2.1 Modélisation des jours de dépassement

4.2.1.1 Premier essai : modèle global pour les quatre stations

Parmi tous les modèles essayés, le suivant aboutit à des coefficients tous significatifs (Tab 4.5) :

$$\alpha(i) = \alpha_1 t(i) + \alpha_2 \text{vent}(i) + \alpha_3 tt \max(i) + \alpha_4 D_1 + \alpha_5 D_2 + \alpha_6 t(i) * D_1 + \alpha_7 t(i) * D_2$$

Les modèles pour Neuilly et Créteil sont sans coefficient constant, le coefficient de D3 étant nul. Ils ont le même coefficient pour le vent; la valeur négative du coefficient de plus en plus élevée en valeur absolue pourrait suggérer que le poids de la force du vent tend à croître quand le seuil augmente. L'interaction année*température maximale est la même pour les quatre stations (Tab 4.5).

Le modèle global traduit bien aussi les valeurs différentes de l'estimation de la température critique ($\hat{\varphi}$) dues à l'interaction $ttmax$, mais aussi aux deux termes – qui sont aussi des interactions – tD1 et tD2 (Tab 4.5).

Les modèles par station nous permettent de calculer les statistiques du test de Kolmogorov-Smirnov et du test d'une distribution exponentielle d'espérance inconnue Tableau 4.7.

TABLEAU 4.5
*Estimation des paramètres du modèle global
 (après élimination des jours influents des modèles par station)*

Seuil	Modèles			
	GL110	GL120	GL130	GL140
# dépassements/ # jours	376/3118	251/3243	169/3325	115/3379
-2 Log L	1191.517	839.562	655.808	512.512
t	-2.106	-2.343	-2.277	-2.199
Ecart-type	0.123	0.157	0.177	0.198
Vent	-0.581	-0.772	-0.974	-1.209
Ecart-type	0.054	0.074	0.098	0.128
ttmax=t*tmax	0.083	0.090	0.086	0.084
Ecart-type	0.004	0.005	0.006	0.007
D1	-1.447	-1.659	-1.157	-1.020
Ecart-type	0.442	0.554	0.657	0.774
D2	-3.491	-4.400	-3.684	-2.798
Ecart-type	0.607	0.896	0.969	1.045
tD1	0.237	0.284	0.174	0.155
Ecart-type	0.072	0.087	0.102	0.118
tD2	0.378	0.505	0.437	0.310
Ecart-type	0.090	0.125	0.137	0.151

TABLEAU 4.6
*Estimation des paramètres variant suivant les stations
 dans le modèle global après élimination des jours influents*

Seuil	110			120			130			
	Station	cte	t	$\hat{\phi}$	cte	t	$\xi\xi$	cte	t	$\hat{\phi}$
NE		0	-2.106	25.4	0	-2.343	28.2	0	-2.277	26.5
AU		-1.447	-1.869	22.5	-1.659	-2.059	24.8	-1.157	-2.103	24.5
CH		-3.491	-1.728	20.8	-4.400	-1.838	22.1	-3.684	-1.840	21.4
CR		0	-2.106	25.4	0	-2.343	28.2	0	-2.277	26.5

TABLEAU 4.7
Validation distributionnelle du modèle global pour le seuil 130

Test de Kolmogorov-Smirnov				
Station	NE130	AU130	CH130	CR130
D _n	0.208 (*)	0.122 (NS)	0.112 (NS)	0.102 (NS)
Test d'exponentialité				
E _n	1.623 (*)	0.810 (NS)	0.583 (NS)	0.623 (NS)

Après avoir déduit du tableau 4.5 le modèle correspondant à chaque station (Tab 4.6), les statistiques du test de Kolmogorov-Smirnov et du test d'une distribution exponentielle d'espérance inconnue sont calculées pour chaque station (Tab 4.7). Cette phase de validation distributionnelle permet de mettre en évidence le fait que le modèle n'est pas adapté à la station de Neuilly/Seine. Le modèle global n'est donc pas validé. Au vu des résultats précédents, il semblait intéressant de conserver le modèle par station pour Neuilly/Seine et de tenter de construire un modèle partiel pour les stations d'Aubervilliers, Champs/Marne et Créteil.

4.2.1.2 Modèle partiel pour les stations Aubervilliers, Champs/Marne et Créteil

Par une approche strictement identique sur ces trois stations (deux variables indicatrices P1 et P2. : pour Aubervilliers $\{0, 0\}$, pour Champs sur Marne $\{1, 0\}$ et pour Créteil $\{0, 1\}$) nous avons abouti au modèle :

$$\alpha(i) = \alpha_0 + \alpha_1 t(i) + \alpha_2 \text{vent}(i) + \alpha_3 t t \max(i) + \alpha_4 P_1 + \alpha_5 t P_1$$

Les tests permettent de valider les hypothèses distributionnelles; nous ne donnons que les résultats pour le seuil 130 (Tab 4.8).

TABLEAU 4.8
Seuil 130 : comparaison modèles par station et modèle partiel

Station-seuil	Station	Station	Partiel	Station	Partiel	Station	Partiel
	NE130	AU130		CH130		CR130	
# dépassements.	59	44		28		38	
Constante	0.606 ^(NS)	-1.306	0	-2.708	-3.227	-0.208 ^(NS)	0
t (année)	-2.604	-2.532	-2.235	-1.789	-1.857	-1.846	-2.235
Vent	-0.813	-0.844	-1.397	-2.021	-1.397	-1.474	-1.397
tmax=t*tmax	0.093	0.101	0.090	0.094	0.090	0.078	0.090
$\hat{\phi}$	28.0	25.1	24.8	19.0	20.6	23.7	24.8

L'hypothèse d'un modèle partiel représentant les trois stations ne peut pas être rejetée face à l'hypothèse d'une modélisation par station. Pour les jours de dépassement, nous avons donc un modèle pour Neuilly et un modèle global pour les trois stations.

Il n'est donc pas possible d'obtenir un modèle unique représentant toutes les stations. Cependant, ce problème peut être contourné, en conservant le modèle par station obtenu pour Neuilly/Seine (cf. §4.1.1) et en en construisant un autre regroupant les stations restantes.

Il est important de noter que la variable jour de dépassement contient moins d'informations que la variable taille de dépassement. Elle ne permet pas toujours de distinguer le comportement des stations d'observation (Créteil et Champs/Marne) de celui des stations appartenant au réseau de mesures de fond (réseau permettant de quantifier géographiquement la pollution atmosphérique; Neuilly/Seine et Aubervilliers) : lors d'un épisode de pollution aigu, il est fort probable que toutes ces stations dépassent le seuil 130; mais ce dépassement aura une durée et une amplitude

différente suivant le type de stations. Les experts ont ainsi pu observer que la station de Neuilly/Seine était la plus chargée en NOx dans l'agglomération parisienne depuis de nombreuses années et que la station de Créteil avait souvent tendance à se comporter comme une station de fond, même si ce n'en est pas une.

4.2.2 Modélisation des tailles de dépassement

Le modèle global choisi est identique à celui du paragraphe 4.2.1.2, seul le vent présente un effet significatif; seuls les seuils 120 et 130 révèlent un effet, et même pour 120 il n'y a aucun effet dû aux stations. Le codage des effets du vent sur la taille est identique à celui de la tendance

TABLEAU 4.9
Synthèse de la modélisation globale des tailles de dépassement

Seuil	Modèles		
	GL120a	GL120b	GL130
# dépassements.	289	289	192
Log L	1209.354	1210.995	793.683
cte	39.259	36.255	51.793
<i>Ecart-type</i>	3.416	2.533	5.241
Vent	-4.294	-3.750	-8.095
<i>Ecart-type</i>	0.588	0.448	0.900
D2	-4.842 (NS)		-22.578
<i>Ecart-type</i>	7.564		8.574
D3	-7.864 (NS)		-18.719
<i>Ecart-type</i>	5.488		7.280
vD2=Vent*D2	1.200 (NS)		5.276
<i>Ecart-type</i>	1.372		1.695
vD3=Vent*D3	0.982 (NS)		3.078
<i>Ecart-type</i>	0.928		1.248

Cette modélisation globale des tailles de dépassement ne permet pas de mettre en évidence des particularités spatiales notables au seuil 120; pour le seuil 130, Neuilly/Seine et Aubervilliers ont un comportement similaire (D1 et vD1 non significatifs), Champs/Marne et Créteil présentent des particularités locales (Tab 4.9). Nous retrouvons donc ici le même "phénomène de rupture" entre les seuils 120 et 130 mis en évidence dans le cadre de la modélisation par station (§4.1.2.2).

Pour le seuil 130, les modèles obtenus pour Neuilly/Seine et Aubervilliers présentent donc le même coefficient constant et le même coefficient pour le vent (Tab 4.10).

Les tests de Kolmogorov-Smirnov et d'exponentialité permettent de valider les hypothèses distributionnelles (Tab 4.11). Pour le seuil 130, la statistique du test du rapport des vraisemblances maximales s'écrit : $-2[\text{Log}L(\text{global}) - [\text{Log}L(\text{sta})]]$ (à comparer à $\chi_2^2(P[\chi_2^2 > 5.991] = 5\%)$) et vaut 2.414. L(global) représente la vraisemblance du modèle global à 6 ddl et L(sta) celle du modèle de la station sta à 2 ddl, d'où (2*4) ddl pour les quatre stations. Le nombre de degrés de liberté (2 ddl) est

TABLEAU 4.10
Déductions des estimations des paramètres locaux
du modèle global des tailles de dépassement

Seuil	120 (NS)		130	
Station	cte	vent	cte	vent
NE	39.259	-4.294	51.793	-8.095
AU	39.259	-4.294	51.793	-8.095
CH	34.417	-3.094	29.215	-2.819
CR	31.395	-3.312	33.074	-5.017

TABLEAU 4.11
Validation du modèle global des tailles de dépassement pour le seuil 130

Test de Kolmogorov-Smirnov				
Station	NE130	AU130s	CH130	CR130
# observations.	62	48	32	50
D _n	0.083 (NS)	0.081 (NS)	0.103 (NS)	0.115 (NS)
Test d'exponentialité				
E _n	0.656 (NS)	0.560 (NS)	0.579 (NS)	0.821 (NS)

égal à leur différence. Par conséquent, l'hypothèse d'une modélisation globale des tailles de dépassement du seuil 130 ne peut pas être rejetée contre l'hypothèse d'une modélisation par station.

Dans cette étude, le choix des stations repose sur la disponibilité des données, et non sur leur représentativité. Le modèle obtenu pour les tailles de dépassement du seuil 130 traduit principalement la différence de comportement entre les stations d'observation et stations de fond. Parmi les quatre stations étudiées, deux sont des sites d'observation, dont le comportement est difficile à prévoir du fait de l'influence ponctuelle, donc non systématique, d'axes routiers voisins et de parkings placés juste à côté du prélèvement. Il paraît donc raisonnable que les stations de Champs/Marne (av. J. Jaurès) et Créteil-Eglise se distinguent des deux stations urbaines de fond Neuilly/Seine et Aubervilliers. Les stations de Champs/Marne et de Créteil sont des stations d'observation et non des stations de fond, telles que définies par la classification nationale des stations, en tant que telles les valeurs enregistrées sur ces sites correspondent ponctuellement à la pollution de fond : par exemple lorsque le vent place la station hors d'influence de l'axe routier en question.

5. Conclusions

Le présent travail permet de compléter les résultats obtenus dans un article précédent (Bellanger et Tomassone, 2000). Pour étudier la tendance dans les hautes valeurs d'ozone enregistrées en région parisienne, nous utilisons un PPNH à paramètres dépendant des conditions météorologiques et de l'année, pour modéliser les jours et les tailles de dépassement d'un seuil élevé. Nous montrons ainsi que la distribution des tailles de dépassement peut être approchée par une loi exponentielle, puisque l'hypothèse $\xi = 0$ dans (2.7) ne peut pas être rejetée. Par conséquent, la dis-

tribution de Pareto peut être réduite à la distribution exponentielle. Le modèle obtenu ne permet pas de déceler de tendance à moyen terme dans les tailles de dépassement; la seule variable influent sur celles-ci étant la vitesse du vent.

Ensuite, nous avons construit un modèle unique pour les quatre stations. Dans notre contexte les méthodes de statistiques spatiales paraissent inappropriée, vu le faible nombre de stations dont nous disposons. Nous contourrons donc cette difficulté en utilisant une technique courante dans les applications du modèle linéaire : nous introduisons des variables indicatrices représentant les stations dans les paramètres du PPNH. Les résultats obtenus ne permettent pas d'obtenir un modèle unique pour les jours de dépassement du seuil 130. Cependant l'hypothèse d'un modèle partiel représentant les trois stations Aubervilliers, Champs/Marne et Créteil ne peut pas être rejetée. Neuilly/Seine paraît donc avoir un comportement très particulier, nous obligeant à conserver le modèle par station lui correspondant. Le modèle partiel permet d'observer le caractère particulier de la station de Champs/Marne. Dans le cadre de la modélisation des tailles de dépassement du seuil 130, l'hypothèse d'un modèle unique ne peut pas être rejetée. Ce modèle met en évidence les comportements similaires des stations de Neuilly/Seine et d'Aubervilliers. Les résultats de cette modélisation traduisent bien la complexité des relations entre ces quatre stations; relations qui diffèrent suivant que l'on étudie le processus des jours de dépassement ou celui des tailles de dépassement.

Remerciements : l'auteur remercie vivement Richard Tomassone pour ses précieux conseils, sa relecture attentive et ses encouragements; Véronique Bonneau et Christian Renaudot d' AIRPARIF, pour leur expertise des résultats ainsi qu' AIRPARIF pour la nouvelle utilisation des données

Références

- BELLANGER L. (1999). *Statistique de la Pollution de l'air. Méthodes Mathématiques. Applications au cas de la Région Parisienne*. Thèse, Université Paris-Sud, Orsay, 235p.
- BELLANGER L., TOMASSONE R. (2000). La pollution de l'air dans la région parisienne : étude de la tendance dans les hautes valeurs d'ozone. *Revue de Statistiques Appliquées* XLVIII(1), 5-24.
- COLES S. (1999). *Extreme value theory and applications*, notes de cours présentées lors de la 44ième Reunião Annual da RBRAS e 8th SEAGRO à Bucato, São Paulo, Brasil, 26-30 Juillet 1999.
- COLES S. and TAWN J. (1996). *Modelling Extremes of the Area Rainfall Process*. J. R. Statist. Soc. B, 2, 329-347.
- COX D.R., LEWIS P.A. (1966). *The Statistical Analysis of Series of Events*. John Wiley, 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 thresholds (with discussion). *J.R. Statist. Soc.*, 52, 393-442.
- DRAPER N.R. and SMITH H. (1981). *Applied Regression Analysis*. John Wiley & Sons, New York, NY, pp. 177-183.
- EMBRECHTS P., KLÜPPELBERG C., MIKOSCH T. (1999). *Modelling Extremal Events for Insurance and Finance*. Springer Verlag, New York, 2ième éd.
- FALK M., HUSLER J., REISS R-D. (1994). *Law of Small Numbers : Extremes and Rare Events*. DMV Seminar 23, Birkhäuser-Verlag.
- GAN F.F., KOEHLER K.J. (1990). Goodness-of-Fit Tests Based on P-P Probability Plots. *Technometrics*. 32(3), 289-303.
- GAN F.F., KOEHLER K.J. (1992). Goodness-of-Fit Tests Based on P-P Probability Plots (Computer Programs edited by Jack C. Ming Wang). *Journal of Quality Technology*. 24(2), 96-102.
- GUMBEL, E.J. (1958). *Statistics of Extremes*. Columbia Univ. Press, New York.
- HALL, W. J. and WELLNER, J. (1981). Mean residual life. In *Statistics and Related Topics*, (eds M. Csörgo, D. A. Dawson, J. N. K. Rao and A. K. Md E. Saleh), pp. 169-184. North-Holland, Amsterdam.
- HOSMER D.W., LEMESHOW S. (1989). *Applied Logistic Regression*. John Wiley, New York.
- LAWLESS J.F. (1982). *Statistical Models and Methods for Lifetime Data*. John Wiley & Sons, New York, NY, pp. 84-88.
- LEADBETTER M.R. (1991). On a basis for "Peaks over Threshold" modeling. *Statistics and Probability Letters*, 12, 357-362.
- LEADBETTER M.R., LINDGREN G., ROOTZ...N.H.(1983). *Extremes and Related Properties of Random Sequences and Series*. Springer Verlag, New York.
- PEARSON, E. S. and HARTLEY, H. O. (eds) (1976). *Biometrika Tables for Statisticians*, Vol. II. High Wycombe : Griffin.
- 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.
- PREGIBON, D. (1981). Logistic regression diagnostics. *Ann. Statist.*, 9, 705-724.
- MATH WORKS INC. (1995a). *MATLAB Référence Guide*, Cary, The Math Works Inc., 1995.
- MATH WORKS INC. (1995b). *MATLAB User's Guide*, Cary, The Math Works Inc., 1995.
- NERC (1975). *Flood Studies Reports*, Vol 1. National Environmental Research Council, London.
- SAS Institute Inc. (1994). *SAS/STAT User's Guide : Version 6*, Fourth Edition, Vol. 1 and 2. Cary, NC : SAS Institute Inc.

- SAPORTA G. (1990). *Probabilités Analyse des données et Statistique*, Editions Technip, Paris.
- 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. (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.
- SNEDECOR G.W., COCHRAN W.G. (1971). *Méthodes statistiques*. Association de Coordination Technique Agricole, Paris.
- 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.
- VENABLE W.N., RIPLEY B.D. (1997). Modern Applied Statistics with S-Plus. (Second Edition) In *Statistics and Computing* (Series Editors : J. Chambers, W. Eddy, W. Härdle, S. Sheather, L. Tierney), Springer-Verlag, New York.
- YANG, G. L. (1978). Estimation of a biometric function. *Ann. Statist.*, 6,112-116.