LISE BELLANGER

Statistique de la pollution de l'air. Méthode mathématiques. Applications au cas de la région parisienne, 1999

ORSAY
N° D'ORDRE

UNIVERSITÉ DE PARIS SUD
U.F.R SCIENTIFIQUE D'ORSAY

THÈSE

présentée

pour obtenir

Le TITRE de DOCTEUR EN SCIENCES
DE L'UNIVERSITÉ PARIS XI ORSAY

SPÉCIALITÉ : MATHÉMATIQUES

par

Lise BELLANGER

Sujet :

STATISTIQUE DE LA POLLUTION DE L'AIR. MÉTHODES MATHEMATIQUES. APPLICATIONS AU CAS DE LA RÉGION PARISIENNE.

Rapporteurs : M. Jean-Marc AZAIS
M. Byron MORGAN

Soutenue le 14 Janvier 1999, devant le jury composé de:

M. Jean-Marc AZAIS	Rapporteur
M. Didier DACUNHA-CASTELLE	Directeur
Mme. Monique GRAF-JACCOTTET	Examinateur
M. Philippe LAMELOISE	Invité
M. Byron MORGAN	Rapporteur
M. Gonzalo PERERA	Examinateur
M. Rémi STROEBEL	Invité
M. Richard TOMASSONE	Président

Abstract

As it occurs in all great cities, Paris has a serious photochemical ozone air pollution problem. Our work is inserted in the project on forecasting ozone episodes in the Paris area realised in coorperation with AIRPARIF, the Paris area air pollution agency.

In the first part, we biefly present the ozone formation phenomena.

In the second part, we analyse the influence of wind on pollution repartition; to achieve this aim, classical data analysis (Principal Component Analysis, Procustean Analysis, Multidimensional Scaling) and linear model were used.

In the field of air pollution control, the rare event is often more significance than the common one. This is evidenced by the content of air quality standards which define acceptable upper limits of air pollution concentrations. The purpose of the third part is to establish whether observed trends in the data of tropospheric ozone are real, meaning that they could be attributed to actual changes in the emissions of toxic gases into atmosphere, or whether they are the result of meteorological changes affecting the conditions under which ozone is generated. To investigate this question, we construct a regression model in which the level of ozone is represented as a function of both meteorological variables and time, in order to determine the significance of the time component when the meteorological variables are taken into account. Also, we propose to use a logistic regression to model the probability of an exceedance of a high threshold level every day, in accounting for the relationship between very high values of ozone and meteorological conditions.Then, we apply the results of the extreme value theory to model the point process consisting of the times and the sizes of high-level exceedances by a non-homogeneous Poisson process. We apply the method to data from the Paris and Los Angeles areas.

In the fourth one, we demonstrate the convergence to a Compound Poisson process of a high-level exceedances point process $N_n(B) = \sum_{\frac{i}{n} \in B} 1_{\{X_j > u_n\}}$, where $X_n = \varphi(\xi_n, Y_n)$, φ a (regular) regression function, u_n grows to infinity with n in a suitable way, ξ and Y are mutually independent, ξ is stationary and weakly dependent, and Y is non-stationary, satisfying some ergodic conditions. The basic technique is the study of high-level exceedances of stationary process over suitable collections of random sets.

Key words and phrases: Principal Component Analysis, Procustean analysis, Multidimensional Scaling, linear model, exceedances, point processes, convergence, logistic regression, diagnostic model testing, generalized Pareto distributions, meteorological conditions, non-homogeneous Poisson process, Bootstrap method, Compound Poisson process, level sets, mean occupation measures, asymptotically ponderable collections of sets.

Remerciements

J'adresse mes plus vifs remerciements à ceux qui m'ont permis de mener à bien cette recherche :

Didier Dacunha-Castelle, Richard Tomassone et Gonzalo Perera dont les qualités humaines et les conseils m'ont toujours encouragée à poursuivre ce travail. Ils m'ont soutenue de façon constante dans le choix de m'intéresser à des sujets parfois fondamentaux et parfois appliqués, orientant et enrichissant mes recherches. Ils ont su m'apporter la confiance qui m'a permis de les mener à bien.

Byron Morgan, Ross Leadbetter et Jean-Marc Azais pour avoir accepté d'être les rapporteurs de ma thèse.

Monique Graf-Jaccottet qui a accepté, malgré ses nombreuses charges, de participer à mon jury de thèse.

Rémi Stroebel de l'Agence de l'Environnement et de la Maîtrise de l'Energie (ADEME) et Philippe Lameloise directeur de l'organisme chargé de la surveillance de la qualité de l'air en Ile-de-France (AIRPARIF) qui ont cofinancé cette thèse et qui n'aurait pas vu le jour sans leur appui.

Toutes celles et tous ceux qui ont participé, ou participent encore, au projet "Prévision des pointes de pollution en Région Parisienne "réalisé à la demande d'AIRPARIF :

Lilianne Bel, Michel Bobbia, Véronique Bonneau, Gabriela Ciuperca, Jean Coursol, Claude Deniau, Elisabeth Gilibert, Badih Ghattas, Patrick Jacubowisz, Michel Misiti, Yves Misiti, Georges Oppenheim, Jean-Michel Poggi, pour les échanges fructueux et réguliers que nous avons eus, qui avaient alors, il y a un peu plus d'un an, débouché sur la mise en œuvre de quatre modèles de prévision à courte échéance des pointes d'ozone issus de méthodes variées.

Joe Cassmassi qui m'a gracieusement transmis les données de pollution et météorologiques de Los Angeles et qu'il possédait.

Denis Brion (METEO-FRANCE) pour ses analyses météorologiques des jours pollués.

L'équipe de Statistique de la faculté d'Orsay qui m'a accueillie durant ces années.

Catherine et Sabine pour leur soutien, leur aide et leurs qualités humaines.

Tous mes proches et mes amis, pour leur soutien et leur affection.

Et Philippe, pour ce qui est au delà des mots.

"(...) Merci mes maîtres, qui m'ont transmis les connaissances lentement accumulées
par l'humanité depuis qu'elle interroge l'univers.
Merci, vous qui m'avez aimé de votre irremplaçable amour.
Mais c'est à moi d'achever l'ouvrage, à moi de poser la poutre faîtière. Oubliez
[celle] que vous auriez voulu que je sois. Je n'ai pas pu réaliser le rêve que vous aviez fait
pour moi, ce serait trahir ma nature d'homme. Pour que je sois vraiment un
homme, vous me devez un dernier cadeau :
la liberté de devenir [celle] que je choisis d'être."
Albert Jacquard.

Table des matières

Liste des figures

Liste des tableaux

Introduction

Avant de présenter les travaux que nous avons réalisés, il nous semble indispensable de les replacer dans un contexte plus général. Pour cela, nous reprenons les termes de notre projet défini à la fois par un besoin de connaissances dans le domaine de la qualité de l'air et dans celui plus académique d'un travail universitaire.

Le projet qui nous a été proposé conjointement par AirParif (organisme chargé de la surveillance et de la qualité de l'air en Ile-de-France), le laboratoire de Modélisation Stochastique et Statistique en collaboration avec le laboratoire de Biométrie du CNRS (Université de Lyon I) commençait par un bref historique que nous reproduisons :

"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".

A partir de là, trois problèmes principaux étaient proposés pour notre travail :

- L'étude de l'évolution de la pollution sur le long et le moyen terme,
- La prévision,
- L'aspect spatio-temporel : typologie des stations et l'évolution au court de la journée.

Dans notre travail nous avons tenté de répondre à une partie des questions soulevées par ces problèmes. En fait, la logique de notre recherche nous conduit à adopter la présentation suivante :

Dans une première partie, nous avons rappelé quels étaient les phénomènes chimiques permettant la formation de l'ozone troposphérique (la partie de l'atmosphère comprise entre le sol et la stratosphère), quelles étaient les conditions favorables à l'apparition d'épisodes d'O_3 ; nous avons aussi donné quelques détails sur le corpus de données dont nous pouvions disposer.

Dans une seconde partie, nous avons étudié l'influence du transport en région parisienne. La formation d'O_3 à court terme est fonction de conditions atmosphériques particulières dans lesquelles le vent (sa direction, sa vitesse) joue un rôle important, mais aussi la température maximale, l'humidité de l'air. Nous avons d'abord étudié l'influence de la direction du vent à l'aide de techniques statistiques élémentaires mais, à notre connaissance, jamais encore utilisées dans ce contexte. Nous avons élargi ce premier modèle à d'autres facteurs à l'aide des techniques, elles aussi classiques, du modèle linéaire. Ensuite, pour étudier la variabilité et la représentativité des onze stations disponibles nous avons été conduite à mettre en oeuvre des méthodes d'analyse des données, elles aussi classiques, mais assemblées de façon à mettre en évidence ce qui les différencie et ce qui les rapproche selon la direction du vent, qu'il souffle de l'océan ou de l'intérieur.

Dans une troisième partie, nous nous sommes limités aux valeurs élevées d'O_3 ; en effet ce sont elles qui déterminent les règles d'alerte d'AirParif et qui, selon toute vraisemblance, sont les plus importantes dans le déclenchement d'allergies sur les personnes les plus sensibles. Nous avons

tenté de montrer qu'en nous limitant à ces valeurs élevées (dont nous préciserons les valeurs mais qui tournent autour des seuils d'alerte) les relations qui apparaissent grossièrement sur l'ensemble des données deviennent beaucoup moins évidentes. Pour les décrire, nous avons mis en oeuvre des modèles de valeurs extrêmes, leur nombre et leur intensité, en utilisant un processus de Poisson non-homogène. Cette approche nous a conduit à utiliser la régression logistique, bien connue dans le domaine médical mais largement sous-exploitée dans l'étude la pollution atmosphérique. En outre, il a fallu que nous mettions au point des algorithmes de minimisation adaptés à la recherche d'estimation des paramètres de l'intensité par la méthode du maximum de vraisemblance. Naturellement, les problèmes numériques ont été délicats à mettre en oeuvre. Ceci nous a permis de déceler des différences entre stations pour l'évolution à un horizon de dix ans. Bénéficiant d'un corpus de données beaucoup plus important et sur une période plus longue dans la région de Los Angeles, nous avons essayé d'appliquer le même type de modèle que dans la région parisienne, et nous avons soulevé plus de questions nouvelles que de réponses précises, ce qui est peut-être une façon d'envisager comment prolonger cette modélisation.

Enfin, dans une quatrième partie, nous avons abordé un point de vue entièrement théorique sans possibilité, dans sa phase actuelle, d'appliquer les résultats que nous avons obtenus à des données réelles. Nous pouvons envisager ultérieurement des applications mais il était difficile pendant le temps limité au cours duquel nous avons réalisé notre travail de les faire de manière correcte ; là aussi, sans doute, des prolongements dans un proche avenir.

Nous avons reporté en annexe un certain nombre de rappels théoriques (sans doute indispensables pour le lecteur non statisticien, et inutiles pour le statisticien !). Nous avons aussi fourni de nombreux listings de résultats afin que le lecteur curieux puisse aller les consulter. Nous avons aussi fourni les listes de quelques programmes (en SAS et en MATLAB) pour aider ceux qui voudraient appliquer nos résultats à d'autres corpus de données.

Enfin, si nous avons utilisé un grand nombre de modèles de natures différentes, nous avons tenu à ce qu'ils aient tous une capacité potentielle d'applications, sans nous faire trop d'illusion sur leur possibilité de décrire entièrement une réalité complexe.

Et maintenant, nous pouvons entrer dans le vif du sujet !

"Mon Dieu, délivrez-moi du modèle ! "
DIDEROT D., Essais sur la peinture, observations sur le Salon de peinture de 1765
Garnier, éd., Paris 1988.

Partie I

L'ozone troposphérique

Introduction

L'ozone (O_3) est un composant secondaire, c'est-à-dire qu'il n'existe pas d'émetteur direct d'O_3. C'est un constituant naturel de la troposphère, sa concentration de fond dans l'atmosphère est de l'ordre de 50 ppb (1 ppb correspond à environ 2 μg/m3). Comme le soulignent Carlier et Mouvier dans [11], "l'ozone, comme de nombreux composés minoritaires dans l'atmosphère,est la conséquence d'un ensemble complexe de phénomènes physiques et chimiques que l'on peut classer en 5 grandes catégories :

1. Emission de composés dits primaires, soit par des processus naturels, soit par des processus anthropogéniques (industriels, domestiques ou agricoles).

2. Transformation chimique conduisant à la formation de composés dits secondaires dans l'atmosphère à partir des composés primaires ou d'autres composés secondaires préalablement formés.

3. Transport et dispersion de l'ensemble des composés présents du fait du déplacement des masses d'air, de la turbulence et de la convection, sous l'influence des différents facteurs météorologiques.

4. Transfert entre phases, c'est-à-dire l'ensemble des phénomènes d'échanges de matière entre la phase gazeuse et des phases condensées solides ou liquides en suspension (phénomène de nucléation, condensation, absorption, etc...).

5. Elimination par dépôt sec ou humide, absorption par les végétaux, corrosion de matériaux divers."

Quels sont les mécanismes régissant la formation de ce polluant ? Qu'en est-il en Région Parisienne ? C'est ce que nous allons essayer de préciser dans les chapitres suivants.

Chapitre 1

Présentation brève des phénomènes chimiques permettant la formation de l'ozone troposphérique

La pollution par O_3 est le résultat de phénomènes nombreux et complexes, dont le moteur est le cycle induit par la photolyse du dioxyde d'azote (NO_2), appelé *cycle de Chapman*, mettant en jeu les composants NO, NO_2 et O_3. En effet, ce cycle met en lumière les principaux mécanismes de production-consommation d'O_3 :

$$O_3 + NO \longrightarrow NO_2 + O_2 \tag{1.1}$$

$$O(^3P) + O_2 \longrightarrow O_3 \tag{1.2}$$

$$NO_2 + h\nu \longrightarrow NO + O(^3P) \text{ Photolyse du } NO_2 \tag{1.3}$$

où $O(^3P)$ représente l'atome d'oxygène dans l'état fondamental. Par conséquent, en situation normale, la concentration d'ozone est modérée, puisque la formation d'ozone dans (1.2) implique réaction chimique (1.3) qui produit du NO, piège de l'ozone dans (1.1).

1.1 Production par transport vertical

L'ozone est un oxydant réagissant rapidement avec les réducteurs tels que le NO dans l'atmosphère (réaction de titration (1.1)). Par conséquent, si l'ozone est en excès, NO est totalement oxydé en NO_2 et il reste un résidu d'O_3; si au contraire NO est en excès, O_3 est totalement détruit et reste un résidu de NO. Cette réaction explique entre autre, pourquoi, au voisinage des autoroutes ou des grands centres urbains, producteurs de NO (gaz d'échappement), on enregistre en général, des taux d'ozone relativement faibles.
(1.1) permet d'interpréter un grand nombre de situations, notamment celles où la production photochimique d'O_3 au cours de la journée reste négligeable (zones rurales qui ne sont pas sous le vent de l'agglomération par exemple). Les phénomènes observés résultent alors simplement de la compétition entre les mécanismes de consommation d'O_3, par émission de NO notamment et ceux d'apport d'O_3 par transport à partir de la stratosphère. Les équations (1.2) et (1.3) correspondent à des périodes de pollution photochimique, c'est par conséquent en période d'intense irradiation solaire que des taux d'ozone élevés seront observés.

Qu'en est-il de la production directe d'ozone?

1.2 Production directe dans la basse troposphère

Seule l'équation (1.2) permets la production d'ozone. Par conséquent, la formation d'ozone en grande quantité fait appel à la réaction d'oxydation du NO en NO_2 suivante :

$$H_2O + NO \longrightarrow OH + NO_2 \tag{1.4}$$

(1.4) court-circuite (1.1) et permet ainsi à l'ozone de s'accumuler. schématiquement et très succinctement, on a :

$$(1.4) \leftrightarrow (1.3) \rightarrow (1.2)$$

Chapitre 2

Conditions favorables à l'apparition d'épisodes d'O_3

2.1 Conditions météorologiques et émissions anthropiques

Il est évident que la réaction de photolyse du NO_2 (1.3) n'aura lieu que lors de période d'intense irradiation solaire. Mais l'apparition d'un épisode d'O_3 nécessite aussi des concentrations en précurseurs (NO) suffisamment grandes, une vitesse de vent faible, une température élevée au sol, une humidité relative faible et une pression atmosphérique élevée. C'est en période anticyclonique que ces épisodes sont les plus fréquents. En effet, les masses d'air sont peu mobiles et les précurseurs de l'O_3 ont donc tendance à s'accumuler, notamment lors des phénomènes d'inversion thermique.

Rappel : Inversion thermique
Théoriquement, la température baisse avec l'altitude dans les premières couches de l'atmosphère. On dit qu'il y a inversion thermique lorsque la température augmente avec l'altitude. L'air froid situé en-dessous est alors bloqué par l'air chaud situé au-dessus. Les polluants ne peuvent alors pas se disperser dans l'atmosphère, ils sont piégés par ce que l'on appelle alors la couche d'inversion.

2.2 Le rôle du transport

"Aux échelles locales et régionales, les concentrations [de pollution] varient rapidement dans le temps et l'espace, si bien que les échelles pertinentes pour décrire les phénomènes sont typiquement l'heure et la dizaine de kilomètres. "(cf. [1], chapître V). Les processus physico-chimiques y sont caractérisés par :

- La nature des constituants : la chimie organique doit prendre en compte à cette échelle un très grand nombre de composés réunis sous le terme de composés organiques volatils (COV) qui ont une importance capitale dans la production de polluants secondaires tels que l'ozone.

- La proximité des sources de pollution, conduisant à des niveaux de concentration bien supérieurs à ceux observés dans l'atmosphère libre et variant rapidement en fonction de la distance des sources.

Cinq échelles d'espace sont nécessaires à l'analyse du comportement de l'ozone troposphérique (National Research Council (1992)):

- Le noyau urbain : Zone d'habitat très dense dans laquelle les sources de précurseurs sont très intenses, où le piégeage de l'ozone domine (consommation d'ozone). La concentration d'ozone y est donc presque toujours faible.

- Le périmètre urbain : Zone encore très densément habitée dans laquelle le trafic automobile est intense. C'est la première zone à être touchée par l'ozone formé dans le panache du

noyau urbain, même si ce n'est pas dans cette zone que les concentrations les plus élevées s'observent, un piégeage notable subsistant du fait de l'importance des émissions.

- La région de transport mésoéchelle : région dans laquelle les concentrations d'ozone sont les plus élevées.
- La région de transport d'échelle synoptique : échelle des grands systèmes météorologiques susceptibles d'intervenir de façon déterminante sur la concentration d'ozone dans les basses couches (dépression, anticyclone). Les processus frontaux ou les subsidences anticycloniques peuvent disperser ou au contraire accumuler l'ozone dans un rayon de 300 à 500 km.
- Le réservoir de fond constitué par le reste de l'atmosphère, au-delà des limites de l'échelle précédente.

Par conséquent, différents types d'épisodes de pollution peuvent avoir lieu (cf. [56]) :

- **Episode local :**
  Episode de forte pollution en O_3 lors duquel la vitesse du vent est très faible, la direction du vent est variable et où l'on observe une grande disparité entre les taux enregistrés sur les différents sites de mesure, reflétant le manque d'homogénéité de la masse d'air (due à une forte concentration en précurseurs permettant une production locale d'ozone).
- **Episode régional :**
  Episode lors duquel la direction du vent est assez stable, la vitesse du vent existe et où les taux d'ozone enregistrés sur les différents sites de mesure sont relativement homogènes. Ceci reflète le peu de production locale et le rôle du transport. Dans ce cas, l'ozone n'est pas (ou peu) produit localement. Il est dû au transport d'une masse d'air de forte concentration en ozone sur l'agglomération.
- **Episode de recirculation d'air pollué ou smog âgé :**
  Les particularités de ces épisodes ne seront pas prises en compte dans l'étude statistique à cause de leur complexité. Pour plus de détails, nous renvoyons les personnes intéressées à l'article [53] dans lequel on trouvera la description des phénomènes régissant de tels épisodes.

Chapitre 3

Corpus de données disponible en Région Parisienne

3.1 Caractéristiques géographiques et climatiques

Paris se situe dans une plaine à peu près entièrement sous influence océanique avec une vitesse de vent suffisante pour disperser la pollution.

3.2 Données d'ozone et données météorologiques

3.2.1 Données pollution : AIRPARIF

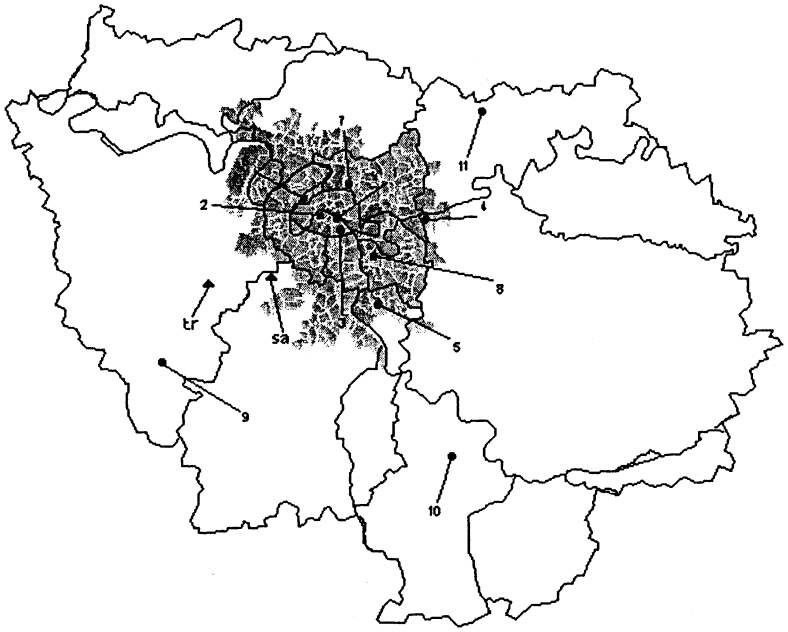
En Ile-de-France, l'exploitation, la gestion du dispositifs de surveillance de la qualité de l'air et la diffusion des informations auprès du public sont sous la responsabilité d'AIRPARIF, association regroupant au sein de son conseil d'administration quatre collèges regroupant des représentants de l'Etat, des collectivités territoriales d'Ile-de-France, des industriels et des associations. Il existe deux types de sites mesurant les concentrations d'ozone (cf. tableau 3.1) :

- les stations de fond urbaines, éloignées de toute source directe de pollution d'origine industrielle ou automobile. Elles mesurent l'exposition de la population francilienne quelle que soit son activité.
- les stations de fond rurales, installées en périphérie de l'agglomération parisienne. Elles permettent de suivre les phénomènes de transport de pollution. Elles forment un bon indicateur du niveau d'ozone créé par réaction chimique à partir de polluants émis directement par les véhicules de l'agglomération.

	Station	Code
Urbaine	Neuilly/Seine	1F92
	Aubervilliers	1F93
	Créteil	1F94
	Paris 04	4F75
	Champs/Marne	1F77
	Paris 13	13F75
	Paris 07	71F75
	Montgeron	4F91
Rurale	Fontainebleau	3F77
	Montgé-en-Goële	4F77
	Rambouillet	2F78

Tableau 3.1: Stations de mesure en région parisienne

[2] fournit un bilan des mesures d'ozone effectuées entre 1992 et 1996 et rappelle les principales caractéristiques de ce polluant secondaire :



Carte d'implantation des sites météorologiques et de mesure de l'ozone

(1) Paris 4ème	(2) Paris 7ème	(3) Paris 13ème
(4) Champs sur Marne	(5) Montgeron	(6) Neuilly sur Seine
(7) Aubervilliers	(8) Créteil	(9) Rambouillet
(10) Fontainebleau	(11) Montgé-en-Goële	
(tr) Trappes	(sa) Saclay	

- Cycle annuel :
  On observe une saisonnalité marquée due au fait que les conditions météorologiques favorables à la formation d'ozone s'observent uniquement durant une période allant environ du 1er **Mai au 15 Septembre 1994-1996**.

- Cycle journalier :
  Pendant la période estivale, le profil d'ozone classique possède la forme caractéristique suivante (cf. [53], où l'on trouvera les explications "chimiques "nécessaires à la compréhension de la forme de cette courbe) :

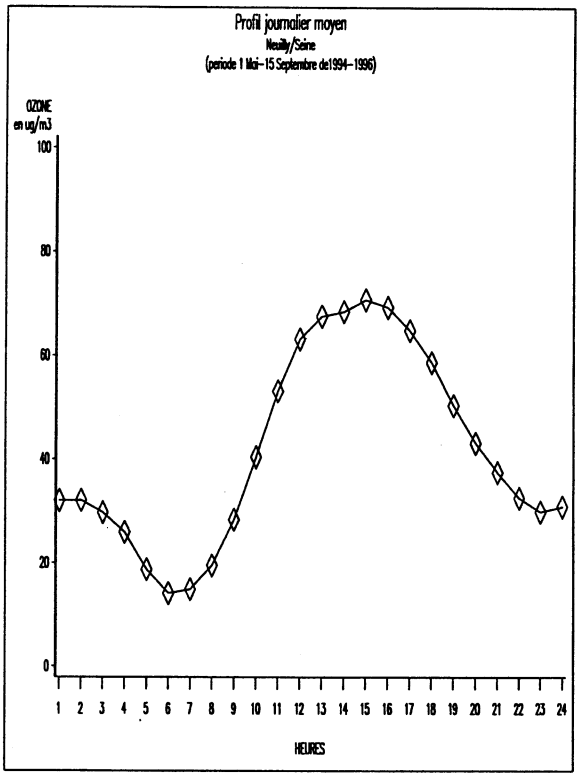


Figure 3.1: Profil d'ozone journalier moyen pour la station de Neuilly/Seine (1994-1996)

- les taux d'ozone mesurés en zone rurale sont en général plus élevés que ceux mesurés en zone urbaine, du fait du déplacement des masses d'air chargées en précurseurs d'ozone.
- En dehors des épisodes photochimiques d'ozone, le niveau de fond moyen d'ozone relevé sur les stations de fond rurales ou au sommet de la Tour Eiffel, est de l'ordre de 60 $\mu g/m^3$

Les seuils de la procédure d'alerte en Ile-de-France :

	Niveau 1 Mise en éveil des services techniques	Niveau 2 Information des autorités et du public	Niveau 3 Alerte
O_3 en $\mu g/m^3$	130	180	360

Tableau 3.2: Seuils de la procédure d'alerte en Ile-de-France pour l'ozone

3.2.2 Données météorologiques

Deux types de données météorologiques mesurées en région parisienne seront utilisées par la suite :

- les données du sondage de Trappes fournies par Météo France, comprenant :
  - la direction du vent mesurée à Trappes à 00h00 TU et à 12h00 TU, à 900hPa (environ 1000 mètres) et à 850hPa (environ 1500 mètres). La direction du vent (en degrés) mesurée à 950 hPa (environ 500 mètres) a été exclue de cette étude, jugée peu significative par D. Brion (Météo-France, communication personnelle),
  - la vitesse du vent (en m/s) mesurée à 00h00 TU et à 12h00 TU, à 850 hPa et à 900 hPa,
  - l'humidité relative (en pourcentage) mesurée à 00h00 TU à 900 hPa.

- les données fournies par le mât du commissariat à l'Energie Atomique de Saclay , mesurées au sol :
  - les mesures horaires de la température au sol,
  - les mesures horaires de la vitesse du vent à 58 mètres d'altitude.

Partie II

Influence du transport en région parisienne

Chapitre 4

Direction de vent et concentration d'ozone

Contrairement à de nombreux articles et rapports estimant la concentration d'ozone journalière tels [34], [6], [4], [5], [10], [8], [9], [18], [20] ou [56] (méthode non statistique), notre but est l'analyse de l'influence du transport régional. Nous nous sommes donc concentrés sur des épisodes relativement pollués lors desquels la vitesse du vent était suffisamment élevée (c'est-à-dire supérieure à 4 mètres par seconde) pour permettre d'obtenir une mesure fiable de la direction du vent. En effet, il nous a semblé intéressant, dans un premier temps d'étudier la relation entre épisodes de pollution et direction du vent (cf. [23], [51] et [7]).

4.1 Les données utilisées

Nous avons considéré les variables suivantes mesurées pendant la période 1^{er} **Mai-15 Septembre 1994-1996** :

4.2 Données pollution

Sur les onze stations définies en tableau 3.1de la partie 3.2.1, nous disposons des relevés des maxima d'ozone et de l'heure à laquelle ces maxima sont atteints, qui sont notés respectivement max*sta* et TM*sta*.

4.3 Données météorologiques

Nous avons utilisé les données météorologiques mesurées à Trappes, décrites dans la partie 3.2.2, qui seront notées :

- DVx direction du vent à 00h00 à x hPa où $x \in \{850, 900\}$
- FVx force du vent à 00h00 à x hPa où $x \in \{850, 900\}$.

4.4 Présentation de la méthode d'estimation du maximum d'ozone en fonction de la direction du vent

Dans une première étape, nous avons essayé de déterminer une mesure de dépendance entre θ, la variable aléatoire direction du vent prenant des valeurs angulaires et X la variable aléatoire maximum d'ozone journalier prenant ses valeurs sur $\mathbb{R}$, en utilisant le modèle développé par Johnson et Wehrly dans [23]. On définit la corrélation angulaire-linéaire ρ_{AL} par :

$$\rho_{AL} = \max_{\alpha} \rho\{\cos(\theta - \alpha), X\}.$$

ρ_{AL} correspond, d'après les résultats de l'analyse des corrélations canoniques, au coefficient de corrélation canonique dominant entre $(\cos\theta, \sin\theta)$ et X. Ce qui par construction, puisque $X \in \mathbb{R}$, est équivalent au modèle de régression suivant :

$$X_i = \beta_0 + \beta_1 \sin\theta_i + \beta_2 \cos\theta_i + \varepsilon_i$$

où

- X_i correspond au maximum d'ozone le jour i ($\mu g/m3$)
- θ_i correspond à la direction du vent le jour i (Rad)
- ε_i correspond à l'erreur aléatoire le jour i, sous les hypothèses classiques de centrage, normalité et homoscédasticité.
- β_0, β_1 et β_2 sont les paramètres à estimer.

Il est alors possible d'introduire la valeur $\hat{\alpha}_{max}$ de la direction du vent conduisant à la valeur de la direction du vent permettant d'obtenir la meilleure estimation de X_i et fournissant donc la valeur de ρ_{AL}, en transformant le modèle précédent [51] :

$$X_i = \beta_0 + \gamma \cos(\theta_i - \hat{\alpha}_{max}) + \varepsilon_i$$

et alors

$$\rho^2_{AL} = R^2.$$

4.5 Résultats obtenus

Les calculs ont été réalisés à l'aide du logiciel SAS [42].
Nos conclusions sont les suivantes :

- L'angle $\hat{\alpha}_{max}$ fournissant la corrélation angulaire-linéaire entre direction du vent et maximum d'ozone se situe quelles que soient l'heure et l'altitude à l'Est, Sud-est.
- $\hat{\alpha}_{max}$ de Fontainebleau (3F77) est celui situé le plus à l'Est, tandis que celui de Montgé-en-Goële (4F77) est celui situé le plus au Sud-est.
- Le R^2 maximum est faible ($\simeq 0.3$) traduisant le faible pouvoir explicatif de la direction du vent seule.
- On observe un groupe de stations pour lequel le pouvoir prédictif de la direction du vent varie très peu en fonction de l'heure de mesure de la direction du vent : Champs/Marne (1F77), Paris 04 (4F75), Paris 13 (13F75), Paris 07 (71F75), Montgé-en-Goële (4F77) et Rambouillet (2F78).
- On observe un groupe de stations pour lequel le pouvoir prédictif de la direction du vent est très variable en fonction de l'heure de mesure de la direction du vent : Neuilly/Seine (1F92), Aubervilliers (1F93), Créteil (1F94), Montgeron (4F91) et Fontainebleau (3F77).
- Toupance et al. (1986) dans [54] ont montré que les concentrations les plus élevées s'observent régulièrement dans le Sud-ouest de l'Ile-de-France, avec des valeurs très élevées à Rambouillet. Selon ces auteurs, l'impact de l'agglomération parisienne sur son environnement étant dissymétrique, les épisodes de pollution les plus marqués se rencontrent dans un secteur qui va du Sud à l'Ouest. Cette dissymétrie de l'impact de l'agglomération met en évidence l'importance des effets météorologiques : par période anticyclonique marquée par des vents de Nord-est, le Sud-ouest est sous le vent de l'agglomération. Les résultats précédents suggèrent quant à eux une direction privilégiée des fortes valeurs d'ozone située dans le **secteur Est** (pour des directions de vent mesurées à haute altitude)....

	00h00 TU						12h00 TU					
	900 hPa			850 hPa			900 hPa			850 hPa		
Station	Nbre d'obs.	$\hat{\alpha_{max}}$	R^2	Nbre d'obs.	$\hat{\alpha_{max}}$	R^2	Nbre d'obs.	$\hat{\alpha_{max}}$	R^2	Nbre d'obs.	$\hat{\alpha_{max}}$	R^2
13F75	197	338	0.27	204	339	0.21	65	326	0.13	72	313	0.14
1F77	198	315	0.22	205	317	0.17	65	283	0.16	72	273	0.12
1F92	197	342	0.18	203	346	0.14	65	330	0.04	72	316	0.08
1F93	193	340	0.20	199	344	0.17	65	325	0.02	72	350	0.04
1F94	188	346	0.23	196	349	0.16	65	333	0.05	72	320	0.06
2F78	191	342	0.28	196	354	0.18	65	347	0.20	72	335	0.23
3F77	185	356	0.21	193	3	0.14	65	14	0.02	72	3	0.03
4F75	189	346	0.29	193	348	0.23	65	348	0.15	72	333	0.15
4F77	189	298	0.22	196	300	0.14	65	278	0.21	72	272	0.19
4F91	191	342	0.29	199	347	0.20	65	324	0.04	72	318	0.04
71F75	192	321	0.30	199	323	0.27	65	306	0.25	72	305	0.29
	uniquement les jours où au moins un des maxima d'ozone $> 90\mu g/m3$ et $VV > 4m/s$						uniquement les jours où au moins un des maxima d'ozone $> 90\mu g/m3$, $VV > 4m/s$ et tous les maxima d'ozone sont atteints après 12h00TU					

Tableau 4.1: Corrélation angulaire-linéaire entre la direction du vent mesurée à Trappes et le maximum d'ozone journalier

- Les résultats obtenus en utilisant les mesures de la direction du vent à 00h00 TU sont meilleurs que ceux utilisant les mesures à 12h00TU. Ceci peut paraître surprenant au premier abord, puisque le maximum d'ozone journalier a en général lieu entre 12h00 et 17h00 TU, mais pourrait peut-être s'expliquer par la plus grande stabilité de l'atmosphère à 00h00 qu'à 12h00. Par conséquent, *l'estimation de l'ozone par la direction du vent peut être effectuée à partir des données directionnelles mesurées à 00h00 Tu.*

Cependant, le pouvoir explicatif de la direction du vent reste faible. Qu'en est-il si on ajoute des variables exogènes supplémentaires? L'angle $\hat{\alpha}_{max}$ est-il modifié?

4.6 Estimation du maximum d'ozone en fonction de la direction et de la vitesse du vent, de la température et de l'humidité relative

Comme nous avons vu dans la partie précédente que la direction du vent à 00h00 avait un pouvoir explicatif plus grand que celle mesurée à 12h00, nous n'avons considéré ici uniquement les mesures à 00h00 à 900hPa.

4.6.1 Modèle linéaire utilisé

Les quinze premiers jours de Septembre de chaque année (en général très peu pollués) ; après une analyse des observations influentes, les jours suivants ont été supprimés :

- 2 Juillet 1994
- 3 Juillet 1994
- 4 Juillet 1994
- 18 Juillet 1994
- 27 Juillet 1994
- 21 Juin 1995
- 11 Août 1995
- 4 Juin 1996
- 17 Juin 1996 pour la station Rambouillet (2F78)
- 24 Juillet 1994 pour la station Fontainebleau (3F77)
- 4 Août 1996 Neuilly/Seine (1F92).

Les notations suivantes seront utilisées pour caractériser les variables significatives dans le modèle linéaire conservé :

- tb2 : valeur journalière **au carré** de la température maximale mesurée à 00h00, 900 hPa (l'analyse préalable du modèle utilisant la température non transformée donnait de moins bons résultats),
- hub00 : valeur journalière de l'humidité mesurée à 00h00, 900 hPa,
- C1000 (respectivement S1000) : valeur journalière du cosinus (respectivement sinus) de la direction du vent mesurée à 00h00, 900 hPa,
- sta : station de mesure,
- fv : vitesse du vent journalière mesurée à 00h00, 900 hPa.

De plus, pour accentuer l'importance des valeurs de pollution élevées, nous avons utilisé la pondération suivante des jours :

$$poids = ((2/3)(max/90))^{3/2}$$

où max correspond à la valeur journalière du maximum d'ozone pour chaque site de mesure. On conserve le modèle linéaire suivant :

$$\begin{aligned} X_{i,sta} = \beta'_0 + \beta'_1 tb2_i + \beta'_2 sta + \beta'_3 tb2_i * sta + \beta'_4 hub00_i \\ + \beta'_5 tb2_i * hub00_i + \beta'_6 fv_i + \beta'_7 tb2_i * fv_i \\ + \beta'_8 C1000_i + \beta'_9 tb2_i * C1000_i + \beta'_{10} S1000_i + \beta'_{11} tb2_i * S1000_i \\ + \beta'_{12} C1000_i * sta + \beta'_{13} S1000_i * sta + \varepsilon_i \end{aligned}$$

```
                    General Linear Models Procedure
                       Class Level Information

       Class     Levels    Values
       STA           11    13F75 1F77 1F92 1F93 1F94 2F78 3F77 4F75 4F77 4F91 71F75
Number of observations in data set = 2142
NOTE: Due to missing values, only 1997 observations can be used in this analysis.

Dependent Variable: MAX
Weight:             POIDS

Source               DF    Sum of Squares     Mean Square  F Value   Pr > F
Model                49  1016560.25912940  20746.12773733    77.43   0.0001
Error              1947   521658.03189310    267.92913811
Corrected Total    1996  1538218.29102250

          R-Square       C.V.     Root MSE       MAX Mean
          0.660869   14.29186  16.36854111   114.53054149
```

Nous ne fournissons dans le tableau 4.2 que les sommes de type III qui sont les mieux adaptées à notre étude, confirmant ainsi les conclusions de [3] :

Source	DF	F value	$Pr > F$	Interprétation
TB2	1	170.74	0.0001	Effet de la température au carré
STA	10	15.47	0.0001	Différence entre stations
TB2*STA	10	4.43	0.0001	La température n'a pas le même effet selon les stations
HUB00	1	5.17	0.0231	Effet de l'humidité
TB2*HUB00	1	23.84	0.0001	Interaction température-humidité
FV	1	10.89	0.0010	Effet de la vitesse du vent
TB2*FV	1	37.69	0.0001	Interaction température- vitesse du vent
C1000	1	31.41	0.0001	Effet de la direction du vent
TB2*C1000	1	19.71	0.0001	Interaction température- direction du vent
S1000	1	0.01	0.9156	Effet de la direction du vent
TB2*S1000	1	12.84	0.0003	Interaction température- direction du vent
C1000*STA	10	8.75	0.0001	Interaction station- direction du vent
S1000*STA	10	3.04	0.0008	Interaction station- direction du vent

Tableau 4.2: Sommes de types III

Nous ne donnons ci-dessous que les estimations utiles pour mettre en évidence les différences entre stations (tableau 4.2) :

Paramètre	Estimation	Ecart-type	Paramètres	Estimation	Ecart-type
C1000	12.6128	2.6094	**S1000**	-7.2571	3.1803
C1000*13F75	12.7653	2.4917	**S1000*13F75**	3.7551	3.0725
C1000*1F77	-0.3990	2.5057	**S1000*1F77**	-4.3675	3.0644
C1000*1F92	6.0270	2.5464	**S1000*1F92**	3.3659	3.0471
C1000*1F93	5.9906	2.3683	**S1000*1F93**	2.2201	2.9064
C1000*1F94	5.8472	2.5057	**S1000*1F94**	5.8943	3.0144
C1000*2F78	17.1486	2.2166	**S1000*2F78**	1.7095	2.6084
C1000*3F77	7.2703	2.4562	**S1000*3F77**	0.4720	3.0628
C1000*4F75	11.9394	2.9648	**S1000*4F75**	2.2962	3.5709
C1000*4F77	-4.4402	2.4000	**S1000*4F77**	-9.1936	2.9121
C1000*4F91	9.2570	2.3411	**S1000*4F91**	2.7586	2.8521
C1000*71F75	12.6128	2.6130	**S1000*71F75**	-7.2571	3.1940

Tableau 4.3: Estimation des valeurs des paramètres C1000 et S1000 par stations

Ainsi, après avoir obtenu une estimation des écart-types des paramètres de chaque station, comme précédemment on standardise et on normalise les coefficients C1000 et S1000 de chaque station de façon à obtenir une estimation de la direction privilégiée pour chaque station (cf. tableau 4.4). Ce modèle étant meilleur que le premier (valeur du R^2 beaucoup plus grande et écart-types des coefficients C1000 et S1000 plus petits) les directions obtenues sont plus précises et donc les différences entre stations sont plus marquées.

4.6.2 $\hat{\alpha_{max}}$ obtenus

Nous déduisons ci-dessous (tableau 4.4) l'estimation de l'angle α_{max} pour chacune des stations. Il faut noter que d'autres valeurs peuvent être fournies pour d'autres termes du modèle, mais qu'elles ne présentent moins d'intérêt pour notre propos.

Station	$\hat{\alpha_{max}}$ en deg
13F75	13
1F77	264
1F92	25
1F93	17
1F94	40
2F78	5
3F77	3
4F75	9
4F77	240
4F91	14
71F75	335

Tableau 4.4: Estimation de $\hat{\alpha_{max}}$

On estime de plus une direction privilégiée moyenne en prenant les moyennes des estimations des coefficients C1000 et S1000 correspondants aux différentes stations de mesure. On obtient une estimation de la direction privilégiée moyenne du vent de 359 degrés, plein Est .
Remarque :
Le modèle linéaire par station :

$$X_i = \beta'_0 + \beta'_1 tb2_i + \beta'_4 hub00_i + \beta'_5 tb2_i * hub00_i + \beta'_6 fv_i + \beta'_7 tb2_i * fv_i$$
$$+\beta'_8 C1000_i + \beta'_9 tb2_i * C1000_i + \beta'_{10} S1000_i + \beta'_{11} tb2_i * S1000_i + \varepsilon_i$$

conduit à des estimations de α_{max} légèrement différentes. Il faut noter le caractère particulier de la station 4F77 (Fontainebleau) où l'on obtient une valeur estimée de 356 degrés, soit 116 degrés de différence avec le modèle complet ! Notre choix s'est porté sur le modèle complet à cause de sa meilleure précision sur les estimations des paramètres C1000 et S1000 de chaque station. Ce modèle traduit en effet mieux les phénomènes de formations globales de l'ozone :

- l'estimation de la direction privilégiée moyenne (Est) correspond à la direction du vent correspondant à des épisodes régionaux,
- les estimations des directions privilégiées pour chaque station fournissent les directions de vent respectives des épisodes de forte pollution locale.

En conclusion, il existe des différences notables entre certaines stations ; 1F77 (Champs/Marne) et 4F77 (Fontainebleau) sont particulières ; leur examen (si ces résultats se confirmaient avec d'autres observations) mériterait une attention particulière.

Chapitre 5

Existence d'une relation caractéristique des 11 stations pour le facteur contrôlé direction

L'influence du transport régional s'observe-t-elle aussi au niveau temporel? En d'autres termes, existe-t-il des différences notables dans les heures des maxima d'ozone des 11 stations étudiées entre classes de directions de vent?
L'Analyse en Composantes Principales par classe de directions de vent (OCEAN, CONTI et TC) nous permettra d'obtenir les relations caractéristiques de l'ensemble des 11 stations, représentées par les facteurs principaux. Les résultats de cette ACP nous conduiront à conserver 3 facteurs principaux par classe de directions de vent. Puis, nous comparerons ces "structures factorielles" (3 tableaux 11*3 des 3 premiers facteurs principaux), c'est-à-dire nous rechercherons la ressemblance entre ces 3 tableaux, en utilisant l'analyse procustéenne fournissant une matrice (3*3) des mesures de proximité entre tableaux. Dans un dernier temps, nous représenterons graphiquement ces trois structures factorielles en appliquant la méthode du positionnement multidimensionnel et nous essaierons d'expliquer les différences entre ces structures en revenant aux stations.

Remarque:
Nous n'avons conservé dans toute la suite que les jours où au moins un des onze maxima d'ozone était strictement supérieur à 90 $\mu g/m3$, de façon à supprimer le bruit engendré par les jours très peu pollués.

5.1 Regroupement des directions de vent par classe

Dans un premier temps, nous associons un secteur de vent à chaque jour: la direction du vent mesurée à 00h00 est classée en trois groupes (TC, OCEAN et CONTI) construits de la manière suivante:
Les jours où la vitesse du vent est strictement inférieure à 4m/s dans une classe appelée TC (Temps calme). L'analyse des graphes en annexe A.1, nous a permis de diviser les jours où la vitesse du vent à 00h00 est supérieure ou égale à 4 m/s en deux catégories en fonction de la direction du vent, (suivant le cercle trigonométrique):

- Influence océanique (**OCEAN**): DV $\in [120, 300[$ degrés.
- Influence continentale (**CONTI**): DV $\in [0, 120[\bigcup[300, 360[$ degrés.

Nous avons alors appliqué cette classification aux directions de vent mesurées à 850 hPa et à 900 hPa. Nous avons alors obtenu la méthode de classification heuristique suivante:

- TC: les jours pour lesquels FV850 **et** FV900 sont strictement inférieures à 4 m/s
- OCEAN: les jours pour lesquels

- $DV850 \in OCEAN$ et $DV900 \in OCEAN$
- $DV850 \in OCEAN$ et $FV900 \leq 4$
- $FV850 \leq 4$ et $DV900 \in OCEAN$

- CONTI : les jours pour lesquels
  - $DV850 \in CONTI$ et $DV900 \in CONTI$
  - $DV850 \in CONTI$ et $FV900 \leq 4$
  - $FV850 \leq 4$ et $DV900 \in CONTI$

- FRONT : les jours pour lesquels
  - $DV850 \in CONTI$ et $DV900 \in OCEAN$
  - $DV850 \in OCEAN$ et $DV900 \in CONTI$
  - $FV850 \leq 4$ et $DV900 \in CONTI$

Remarque :
Cette classification n'engendre aucun problème de classement entre CONTI et OCEAN suivant l'altitude : la classe FRONT ne sera donc pas utilisée.

Nous obtenons finalement le corpus de données suivant d'un total de 215 jours, où au moins une des stations de mesure atteint une valeur maximale supérieure à 90 $\mu g/m3$ (tableau 5.1) :

Secteur	OCEAN	CONTI	TC
Nbe jours	85	85	45
Fréquence	39.5	39.5	21.0

Tableau 5.1: Synthèse des classes de directions de vent

Mais des données d'ozone n'ont pas pu être mesurées certains de ces jours, le corpus définitif ne comprend que 118 jours, respectivement 39, 56 et 23.

5.2 Recherche de la structure factorielle

On commence par effectuer une Analyse en Composantes Principales (ACP) (cf. [40]) sur les heures auxquelles le maximum d'ozone est atteint sur chaque site de mesure, par classe de directions de vent. Cette ACP est faite sur les **covariances** (et non comme il est classique de le faire sur les corrélations) car les variables, les heures par station, sont mesurées dans la même unité.

5.2.1 Statistiques simples

Le tableau 5.2 indique que :

- pour la station de Mongé-en-Goële, l'estimation de l'heure du maximum d'ozone est toujours plus tardive que sur les autres sites, quelle que soit la direction du vent.
- pour la station de Paris 07, l'estimation de l'heure du maximum d'ozone est toujours plus tôt que sur les autres sites, pour les directions de vent OCEAN et CONTI.
- les écart-types des heures auxquelles le maximum d'ozone est atteint sont très élevés sur les sites de Paris 04, Paris 07 et Montgé-en-Goële.

5.2.2 Valeurs propres

Avec une grande variance (104.350), la contribution du premier facteur principal du groupe OCEAN est plus élevée (62 %), les stations se ressemblent donc plus par ce type de directions de vent que par vent venant du continent ou encore par vent très faible (TC) (cf. tableau 5.3). Ceci peut apparaître comme une évidence pour un météorologue ; elle peut toutefois appuyer notre démarche utilisant des variables peu fiables, mais contenant pourtant une certaine information.

		OCEAN	CONTI	TC
Nombre d'observations		39	56	23
Station	Code	Valeur Moyenne Ecart-type		
Neuilly/Seine	1F92	14.2 3.7	15.0 2.6	14.6 1.9
Aubervilliers	1F93	14.6 3.7	15.4 2.8	15.2 2.1
Créteil	1F94	14.6 4.0	15.2 3.3	15.3 2.2
Paris 04	4F75	14.2 **5.1**	14.9 **4.4**	**15.7** **2.8**
Champs/Marne	1F77	14.9 4.0	15.7 2.7	16.2 2.6
Paris 13	13F75	14.4 3.5	15.6 2.7	15.6 1.7
Paris 07	71F75	**13.2** **4.6**	14.4 **4.4**	**15.6** **2.1**
Montgeron	4F91	14.6 3.8	15.2 2.6	15.4 1.8
Fontainebleau	3F77	15.3 3.2	15.7 1.7	15.0 1.9
Montgé-en-Goële	4F77	**15.4** **3.9**	**17.3** **4.5**	**16.7** **5.0**
Rambouillet	2F78	15.1 3.1	15.7 1.7	14.8 2.1

Tableau 5.2: Heure moyennes et écart-types des heures auxquelles le maximum d'ozone journalier est atteint sur les différents sites de mesure, par direction de vent

	OCEAN		CONTI		TC	
	Valeur propre	%	Valeur propre	%	Valeur propre	%
PRIN1	104.350	62.3	57.9304	51.8	30.7715	43.6
PRIN2	17.369	10.4	17.6338	15.8	16.1424	22.9
PRIN3	12.693	7.6	12.5632	11.2	5.9866	8.5
Cumul		80.3		73.8		75.0

Tableau 5.3: Premières valeurs propres par direction de vent, obtenues à partir de la matrice de covariances

5.2.3 Facteurs principaux

Les trois ACP fournissent intrinsèquement les moyens de juger de l'importance de chacune des stations dans l'analyse des heures des maxima (tableau 5.4). Mais, avant de le faire, nous pouvons nous demander s'il n'existe pas une structure commune à ces trois ensembles de facteurs, ce qui faciliterait alors l'interprétation.

	Premiers facteurs principaux								
	OCEAN			CONTI			TC		
Code station	PRIN1	PRIN2	PRIN3	PRIN1	PRIN2	PRIN3	PRIN1	PRIN2	PRIN3
1F92	0.330678	0.021434	-0.076547	0.237745	0.187183	0.128183	0.182815	0.172660	-0. 313240
1F93	0.334382	-0 .091174	-0.240173	0.317598	0.093560	0.125651	0.250741	0.240078	-0.048740
1F94	0.331129	-0.165048	-0.129724	0.362924	-0.260974	-0.110650	0.208357	0.356019	-0.052063
4F75	0.421096	-0.112743	0.474344	0.418161	-0.190293	-0.814588	0.337701	0.391026	0.125577
1F77	0.326104	-0.051812	-0.375571	0.248782	0.176263	0.103926	0.250019	0.185665	-0.650013
13F75	0.302234	-0.066921	-0.254870	0.325917	0.041062	0.100259	0.148769	0.293327	0.152549
71F75	0.335385	-0.316904	0.570732	0.380572	-0.599817	0.503360	0.201698	0.214124	0.176370
4F91	0.260574	0.517715	-0.123947	0.271501	0.037193	0.104109	0.073610	0.297851	0.397569
3F77	0.143459	0.526887	0.094018	0.039238	0.064638	0.015269	0.043277	0.125509	0.433510
4F77	0.281239	-0.022699	-0.214828	0.380120	0.671111	0.080903	0.783340	-0.586138	0.121268
2F78	0.118849	0.545464	0.304181	0.057821	0.058344	0.011952	0.004058	0.136061	-0.208405

Tableau 5.4: Premiers facteurs principaux calculés à partir de la matrice de covariances des heures auxquelles le maximum d'ozone journalier est atteint sur les différents sites de mesure, par direction de vent

5.2.4 Test de Flury

Après avoir obtenu les axes principaux de chacune des trois classes, nous allons utiliser le test peu connu de Flury (cf. [30] et annexe A.3) pour déterminer si les axes principaux des trois populations peuvent être considérés comme identiques. Nous allons donc tester l'hypothèse H_0 d'existence d'une matrice Φ commune d'axes principaux (qui sont parallèles ou confondus) ; Δ_g étant la matrice diagonale des valeurs propres du groupe g, Σ_g matrice de variance-covariance du groupe g :

$$H_0 \ : \ \Phi' \Sigma_g \Phi = \Delta_g$$

La matrice de covariances-variances moyenne S est :

```
8.5177159 6.8100973  6.3606691 6.7811847  5.9831742 6.0983418  5.2156614 5.040287  1.9559141 6.3855819 2.3170989
6.8100973 8.9657075  7.7395656 8.1076871  7.1524968 7.1794316  6.6363117 5.5738014 1.5850167 7.3039118 1.2221298
6.3606691 7.7395656 11.251129  10.364733   5.7579603 7.1856572  8.104004  5.4688579 1.4389144 5.2202507 1.1930919
6.7811847 8.1076871 10.364733  19.27712    6.4525032 7.3857236  8.7226085 6.1724703 2.2331353 7.6380866 2.2575053
5.9831742 7.1524968  5.7579603 6.4525032 10.320223  6.6578597  5.2562114 5.7378482 1.7980601 6.2823399 1.7189526
6.0983418 7.1794316  7.1856572 7.3857236  6.6578597 8.0600634  6.9578417 5.6606167 2.0252723 6.33426   1.3833117
5.2156614 6.6363117  8.104004  8.7226085  5.2562114 6.9578417 17.197187  4.7277213 1.3824152 5.2798644 1.5696232
5.040287  5.5738014  5.4688579 6.1724703  5.7378482 5.6606167  4.7277213 8.5873767 2.9645828 4.5298983 3.1049269
1.9559141 1.5850167  1.4389144 2.2331353  1.7980601 2.0252723  1.3824152 2.9645828 5.4768476 1.8825502 2.4091502
6.3855819 7.3039118  5.2202507 7.6380866  6.2823399 6.33426    5.2798644 4.5298983 1.8825502 19.306944 1.2841624
2.3170989 1.2221298  1.1930919 2.2575053  1.7189526 1.3833117  1.5696232 3.1049269 2.4091502 1.2841624 5.3196118
```

Ses valeurs propres sont :

```
66.563283
14.228939
10.265408
 9.3416644
 7.1837689
 3.8936506
 3.3046553
 2.9613313
 2.0426432
 1.3162008
 1.1783824
```

et ses vecteurs propres, correspondant à l'estimation de la matrice commune Φ :

```
0.2878073  0.0803456  0.200679  0.0779774 -0.126349   0.370954   0.5193647  0.3174703 -0.467632  -0.291629  -0.191138
0.3292456  0.0446632  0.075911  0.0692654 -0.294013   0.1194145 -0.007946   0.0680176 -0.059325   0.3023334  0.8238474
0.3437559 -0.236003  -0.010745 -0.124507  -0.24961    0.5173324 -0.16069   -0.12085    0.5883808 -0.28058   -0.123431
0.4248727 -0.251886  -0.101369 -0.778621   0.1737345 -0.279914   0.0461081  0.0568157 -0.155071   0.03877   -0.008405
0.2966318  0.1074787  0.2910995 0.2058549 -0.344595  -0.680042   0.0378081  0.200583   0.2847826 -0.244065  -0.099034
0.30956   -0.011636   0.0955937 0.1457324 -0.174565   0.0734099 -0.190127   0.0257236 -0.090886   0.7378197 -0.498348
0.3488142 -0.456506  -0.526196  0.5288678  0.2863211 -0.123492   0.0423612  0.0280771 -0.070981  -0.085723   0.0338859
```

```
0.2590868 -0.008228   0.422865  0.1260462  0.1569386 -0.035168  -0.24144   -0.705722  -0.331553  -0.213581   0.0188885
0.0950337  0.0393615  0.3323354 0.0607486  0.4876945  0.1271517 -0.539371   0.5612365  0.0066457 -0.104111   0.0669202
0.3473442  0.8066409 -0.40357  -0.018399   0.1971171  0.035747  -0.0398    -0.098859   0.0936552 -0.058174  -0.042776
0.0882313 -0.00586    0.3473981 0.0643214  0.5220864  0.0192252  0.5586401 -0.110757   0.4409298  0.2686724  0.05886
```

Comme d'après la théorie générale des tests de rapport de vraisemblance, la statistique du rapport des log-vraisemblances X^2 (cf. annexe A.3), sous l'hypothèse d'existence d'axes principaux communs, suit asymptotiquement une loi du χ^2 à 110 degrés de liberté et que $X^2 = 199.5$, on rejette l'hypothèse H_0 ($p < 0.01$). Les trois populations possèdent donc des directions différentes : les relations entre stations (structures factorielles) sont différentes selon les trois classes. Nous devons donc les analyser séparément (tableau 5.4).

- OCEAN/CONTI :
  - les coefficients du deuxième facteur principal de chacune de ces deux directions correspondant aux stations de Montgeron et Montgé-en-Goële sont très différents.
  - les coefficients du troisième facteur principal correspondant aux stations de Paris 04 et Paris 07 sont opposés pour la direction CONTI et de même signe pour OCEAN.

  OCEAN et CONTI se différencient principalement par le comportement des stations de Montgeron, Montgé-en-Goële, Paris 07 et Paris 04.
- CONTI/TC :
  - Seuls les coefficients du deuxième facteur principal correspondant aux stations de Neuilly/Seine et Champs/marne sont proches. Mais les coefficients du troisième facteur principal correspondant à ces stations sont de signe opposé.

  Les différences entre ces deux directions touchent toutes les stations.
- OCEAN/TC :
  - les coefficients du premier facteur principal aux stations de Montgeron, Montgé-en-Goële et Rambouillet sont très différents.
  - les coefficients du deuxième facteur principal correspondant aux stations de Montgé-en-Goële et Rambouillet sont très différents.
  - les coefficients du troisième facteur principal correspondant aux stations de Paris 07 et Montgeron sont très différents.

  OCEAN et TC se différencient principalement par le comportement des stations de Montgeron, Montgé-en-Goële, Paris 07 et Rambouillet.

Nous allons maintenant utiliser la méthode de l'analyse procustéenne (cf. [52]) pour obtenir une mesure un peu plus précise de ressemblance entre directions de vent.

5.3 Calcul d'une ressemblance entre triplets de structures factorielles

5.3.1 Analyse procustéenne

La méthode de l'**analyse procustéenne** (cf. [52]), dont on rappelle en annexe A.4 le principe, permet d'obtenir une matrice D^2 des mesures de proximité entre triplets de facteurs.

La matrice des triplets de facteurs standardisés est la suivante :

OBS	STA	FOOCEAN1	FOOCEAN2	FOOCEAN3	FOCONTI1	FOCONTI2	FOCONTI3	FOTC1	FOTC2	FOTC3
1	TM1692	0.14750	-0.05131	-0.07906	-0.09674	0.16246	0.10590	-0.06496	0.00796	-0.32572
2	TM1693	0.16078	-0.16721	-0.24269	0.10311	0.06850	0.10336	0.03757	0.08872	-0.06101
3	TM1694	0.14912	-0.24324	-0.13224	0.21655	-0.28728	-0.13360	-0.02641	0.22761	-0.06433
4	TM4675	0.47182	-0.18940	0.47185	0.35480	-0.21635	-0.83953	0.16883	0.26955	0.11345
5	TM1677	0.13109	-0.12670	-0.37809	-0.06912	0.15150	0.08158	0.03648	0.02354	-0.66277

6	TM1367	0.04547	-0.14225	-0.25739	0.12394	0.01582	0.07790	-0.11636	0.15251	0.14045
7	TM7167	0.16438	-0.39952	0.56824	0.26072	-0.62733	0.48213	-0.03646	0.05763	0.16429
8	TM4691	-0.10396	0.45945	-0.12646	-0.01226	0.01194	0.08176	-0.22980	0.15793	0.38567
9	TM3677	-0.52405	0.46889	0.09151	-0.59356	0.03948	-0.00733	-0.27559	-0.04853	0.42164
10	TM4677	-0.02984	-0.09673	-0.21735	0.25959	0.64810	0.05849	0.84150	-0.90104	0.10914
11	TM2678	-0.61232	0.48801	0.30168	-0.54705	0.03316	-0.01066	-0.33479	-0.03589	-0.22080

L'analyse procustéenne fournit alors la matrice D des mesures de proximité entre triplets de facteurs suivante :

OBS	OCEAN	CONTI	TC
OCEAN	0.00000	1.30402	1.89124
CONTI	1.30402	0.00000	1.69773
TC	1.89124	1.69773	0.00000

Par conséquent, les directions OCEAN et TC sont les plus différentes, puis viennent CONTI/TC, puis finalement CONTI/OCEAN.
Le calcul dans la matrice 11*3 des différences entre les X et Y, pour les trois groupes de comparaison, après transformation procustéenne, des contributions (en pourcentage) de chaque ligne des stations aux distances au carré entre groupe fournit les résultats suivants (tableau 5.5) :

Station	Code	OCEAN/CONTI	OCEAN/TC	CONTI/TC
Distance(D^2)		1.7005	3.5768	2.8823
Neuilly/Seine	1F92	5.78	1.78	4.12
Aubervilliers	1F93	2.68	2.36	2.18
Créteil	1F94	9.56	4.25	2.71
Paris 04	4F75	**17.77**	3.60	**17.13**
Champs/Marne	1F77	8.07	3.36	**14.45**
Paris 13	13F75	2.22	7.88	2.76
Paris 07	71F75	**15.48**	**10.73**	**14.93**
Montgeron	4F91	**15.17**	9.88	7.05
Fontainebleau	3F77	1.06	4.79	**21.55**
Montgé-en-Goële	4F77	**15.58**	**33.34**	**12.33**
Rambouillet	2F78	6.63	**18.03**	0.80
Total		100	100	100

Tableau 5.5: Différences entre structures factorielles, rôle des stations (% D^2)

La prédominance de certaines stations dans les comparaisons (en gras dans le tableau 5.5) est justifiée par le fait que les axes principaux des trois populations sont statistiquement distincts (comme l'a indiqué le test de Flury).

5.3.2 Comportement des coefficients de corrélation entre les stations prédominantes dans les % de D^2 et les autres

Nous pouvons regarder de façon encore plus fine comment ces stations " différentes "sont reliées aux autres ; pour cela nous revenons aux données de base qui ont servi à estimer les structures factorielles, à savoir les covariances des matrices diagonalisées plus haut ; mais pour faciliter l'interprétation nous donnons les coefficients de corrélation de chacune de ces stations en fonction des dix autres, pour les trois directions.

Pour Paris 04

Le tableau 5.6 permet d'observer :

- OCEAN/CONTI : les valeurs du coefficient de corrélation moyen avec les autres stations (r_{moy}) montre que la station de Paris 04 est plus liée aux autres stations par vent OCEAN que par vent CONTI (0.575 contre 0.374)

Station	Code	CONTI	OCEAN	TC
Neuilly/Seine	1F92	0.3223	0.7136	0.5433
Aubervilliers	1F93	0.4824	0.7378	0.6783
Créteil	1F94	0.6932	0.7189	0.6839
Paris 04	4F75	1.0000	1.0000	1.0000
Champs/Marne	1F77	0.3374	0.5784	0.4387
Paris 13	13F75	0.5406	0.6361	0.6543
Paris 07	71F75	0.3428	0.6504	0.4454
Montgeron	4F91	0.4414	0.5128	0.5245
Fontainebleau	3F77	**0.0918**	**0.3068**	**0.2461**
Montgé-en-Goële	4F77	0.3348	0.5539	0.3186
Rambouillet	2F78	**0.1493**	**0.3435**	**-0.0310**
r_{moy}		*0.374*	*0.575*	*0.450*

Tableau 5.6: Corrélations entre Paris 04 et les autres stations par classe de directions de vent

- CONTI/TC : même résultat moins marqué
- Fontainebleau et Rambouillet sont les stations les moins liées à Paris 04 quelle que soit la direction du vent.

Pour Champs/Marne

Station	Code	CONTI	OCEAN	TC
Neuilly/Seine	1F92	0.5749	0.7307	0.4152
Aubervilliers	1F93	0.6989	0.8577	0.3863
Créteil	1F94	0.3905	0.6885	0.4258
Paris 04	4F75	0.3374	0.5784	0.4387
Champs/Marne	1F77	1.0000	1.0000	1.0000
Paris 13	13F75	0.6570	0.8575	0.4076
Paris 07	71F75	0.2442	0.5702	0.3423
Montgeron	4F91	**0.7438**	0.6158	**0.0685**
Fontainebleau	3F77	**0.1779**	0.3477	**-0.0721**
Montgé-en-Goële	4F77	0.3924	0.6160	0.2994
Rambouillet	2F78	0.3988	**0.1573**	0.1728
r_{moy}		*0.462*	*0.602*	*0.288*

Tableau 5.7: Corrélations entre Champs/Marne et les autres stations par classe de directions de vent

Le tableau 5.7 permet d'observer :

- CONTI/TC : les valeurs du coefficient de corrélation moyen (r_{moy}) montre que la station de Champs/Marne est moins liée aux autres stations par vent TC que par vent CONTI (0.288 contre 0.462)

Pour Paris 07

Le tableau 5.8 permet d'observer que selon la direction dominante des vents :

- La station de Paris 07 est plus liée au autres stations par vent OCEAN,
- Paris 07 est plus ou moins liée aux autres stations (les coefficients assez différents de autres selon la direction sont notés en gras).

Station	Code	CONTI	OCEAN	TC
Neuilly/Seine	1F92	0.2836	**0.6109**	0.2386
Aubervilliers	1F93	0.4969	0.5921	0.4858
Créteil	1F94	0.6422	0.5626	**0.2949**
Paris 04	4F75	0.3428	0.6504	0.4454
Champs/Marne	1F77	0.2442	**0.5702**	0.3423
Paris 13	13F75	0.5579	0.6032	**0.8240**
Paris 07	71F75	1.0000	1.0000	1.0000
Montgeron	4F91	0.4419	0.3323	0.5441
Fontainebleau	3F77	**0.0348**	0.2405	0.0959
Montgé-en-Goële	4F77	0.1923	**0.4839**	0.2813
Rambouillet	2F78	**0.0810**	0.2344	0.2135
r_{moy}		*0.332*	*0.488*	*0.377*

Tableau 5.8: Corrélations entre Paris 07 et les autres stations par classe de directions de vent

Station	Code	CONTI	OCEAN	TC
Neuilly/Seine	1F92	0.6626	0.6160	0.0584
Aubervilliers	1F93	0.7375	0.6065	0.3496
Créteil	1F94	0.6634	0.4900	0.4597
Paris 04	4F75	0.4414	0.5128	0.5245
Champs/Marne	1F77	0.7438	0.6158	0.0685
Paris 13	13F75	**0.8394**	0.5658	0.6519
Paris 07	71F75	0.4419	0.3323	0.5441
Montgeron	4F91	1.0000	1.0000	1.0000
Fontainebleau	3F77	0.2150	**0.5475**	0.4459
Montgé-en-Goële	4F77	0.4113	0.4851	-0.0715
Rambouillet	2F78	0.3240	**0.6012**	0.1572
r_{moy}		*0.548*	*0.537*	*0.319*

Tableau 5.9: Corrélations entre Montgeron et les autres stations par classe de directions de vent

Pour Montgeron

Le tableau 5.9 permet d'observer que :

- la station de Montgeron est liée en moyenne de la même façon par vent OCEAN ou par vent CONTI
- les coefficients de corrélations entre Montgeron et les autres stations, obtenus par vent CONTI ou par vent OCEAN, sont cependant assez différents.

Pour Fontainebleau

Le tableau 5.10 permet d'observer que :

- les coefficients de corrélation sont relativement faibles par vent CONTI et par vent TC,
- les coefficients de corrélation quasi nuls sont obtenus pour des stations différentes en fonction de la direction CONTI ou TC.

Pour Montgé-en-Goële

Le tableau 5.11 indique que les stations sont corrélées par vent OCEAN ($r_{moy} = 0.508$), moins par vent CONTI ($r_{moy} = 0.366$) et pratiquement pas par vent TC ($r_{moy} = 0.167$).

Station	Code	CONTI	OCEAN	TC
Neuilly/Seine	1F92	0.1875	0.3856	**0.0444**
Aubervilliers	1F93	0.1701	0.3173	**-0.0434**
Créteil	1F94	**0.0340**	0.2762	0.2290
Paris 04	4F75	**0.0918**	0.3068	0.2461
Champs/Marne	1F77	0.1779	0.3477	**-0.0721**
Paris 13	13F75	0.2206	0.3630	0.3293
Paris 07	71F75	**0.0348**	0.2405	**0.0959**
Montgeron	4F91	0.2150	0.5475	0.4459
Fontainebleau	3F77	1.0000	1.0000	1.0000
Montgé-en-Goële	4F77	0.1632	0.2943	**0.0368**
Rambouillet	2F78	0.1843	0.6043	0.2300
r_{moy}		*0.148*	*0.368*	*0.154*

Tableau 5.10: Corrélations entre Fontainebleau et les autres stations par classe de directions de vent

Station	Code	CONTI	OCEAN	TC
Neuilly/Seine	1F92	0.5213	0.6208	0.2728
Aubervilliers	1F93	0.5728	0.6739	0.3528
Créteil	1F94	0.3243	0.5132	0.1493
Paris 04	4F75	0.3348	0.5539	0.3186
Champs/Marne	1F77	0.3924	**0.6160**	0.2994
Paris 13	13F75	0.5679	0.6359	0.1093
Paris 07	71F75	0.1923	**0.4839**	0.2813
Montgeron	4F91	0.4113	0.4851	**-0.0715**
Fontainebleau	3F77	0.1632	0.2943	0.0368
Montgé-en-Goële	4F77	1.0000	1.0000	1.0000
Rambouillet	2F78	0.1821	0.2022	**-0.0821**
r_{moy}		*0.366*	*0.508*	*0.167*

Tableau 5.11: Corrélations entre Mongé-en-Goële et les autres stations par classe de directions de vent

Pour Rambouillet

A Rambouillet, les coefficients de corrélations sont nettement plus élevés par vent OCEAN que par vent TC (cf. 5.12).

En conclusion les relations entre stations sont différentes selon la direction du vent ; globalement plus homogènes par vent OCEAN que par vent CONTI et par TC. Ceci concorde avec les résultats des trois ACP obtenus auparavant. Cette analyse, assez fine, a permis de déceler quelques stations plus importantes que les autres dans la mesure où elles reflétaient une certaine spécificité. A contrario, les autres peuvent être considérées comme plus homogènes (Neuilly/Seine, Aubervilliers, Créteil et Paris 13).Naturellement, ceci ne concerne que la relation entre heure du maximum d'ozone et direction des vents ; rien ne nous indique que ces caractéristiques se retrouveraient pour d'autres critères. Néanmoins, *ces ressemblances seront à nouveau mises en évidence au cours de l'étude sur les dépassements de seuil.*

5.3.3 Méthode du positionnement multidimensionnel appliquée à la matrice des distances entre couples de facteurs principaux

Les résultats de l'application de la méthode du positionnement multiple (cf. annexe A.5) aux données de pollution: Les valeurs propres de B sont:

```
M
1.9076352
0.8122202
1.884E-16
```

On a donc une représentation plane.

Station	Code	CONTI	OCEAN	TC
Neuilly/Seine	1F92	0.2610	0.3927	0.3680
Aubervilliers	1F93	0.1884	0.2332	**-0.0731**
Créteil	1F94	0.1518	0.1900	0.0389
Paris 04	4F75	0.1493	0.3435	**-0.0310**
Champs/Marne	1F77	0.3988	0.1573	0.1728
Paris 13	13F75	0.1998	0.2349	0.1800
Paris 07	71F75	0.0810	0.2344	0.2135
Montgeron	4F91	0.3240	**0.6012**	0.1572
Fontainebleau	3F77	0.1843	**0.6043**	0.2300
Montgé-en-Goële	4F77	0.1821	0.2022	**-0.0821**
Rambouillet	2F78	1.0000	1.0000	1.0000
r_{moy}		*0.212*	*0.319*	*0.117*

Tableau 5.12: Corrélations entre Rambouillet et les autres stations par classe de directions de vent

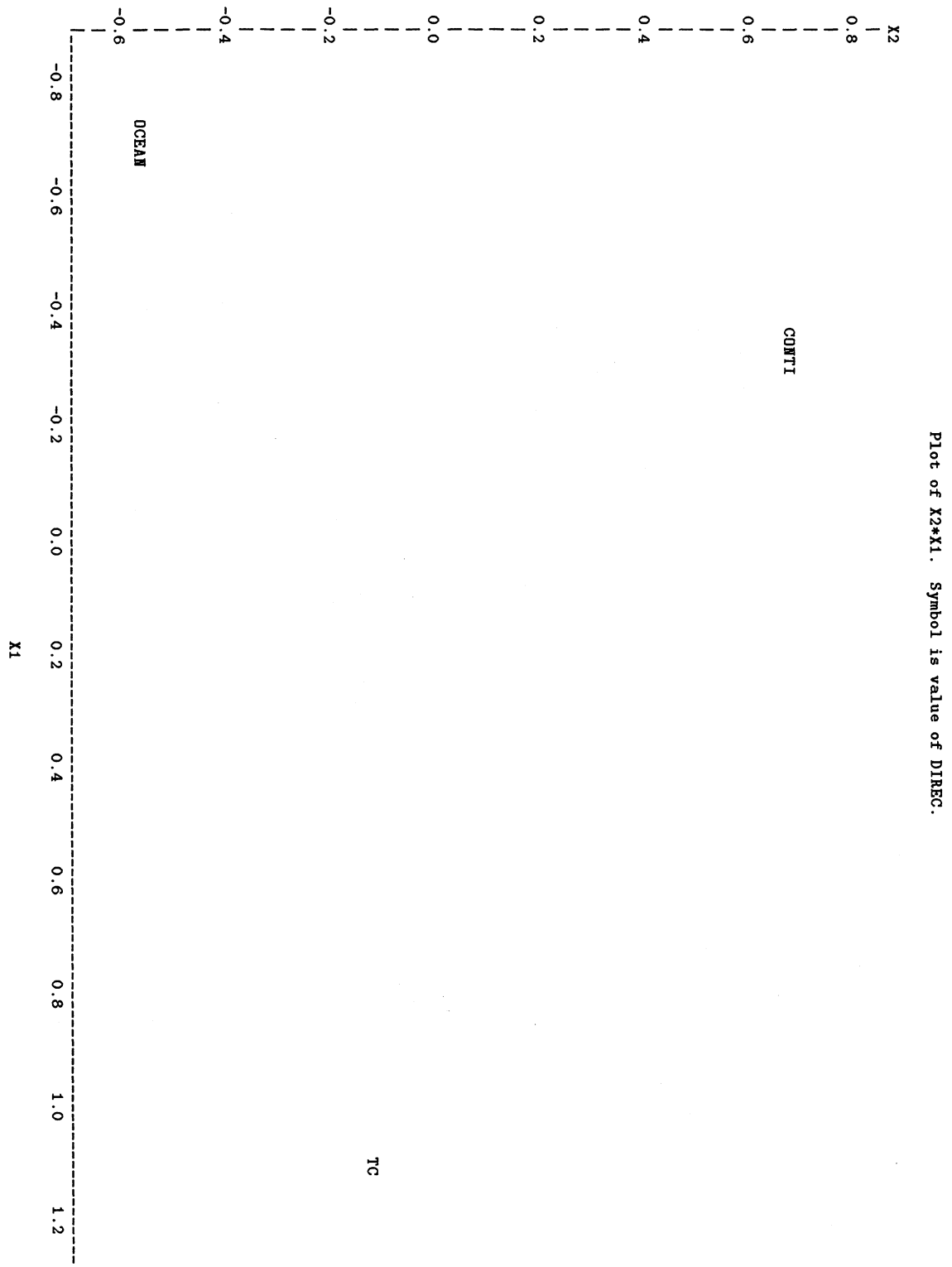
Les vecteurs propres normés de B sont :

```
                    E
-0.532827  -0.618678  0.5773503
-0.269378   0.7707804 0.5773503
 0.8022042 -0.152102  0.5773503
```

coordonnées des points correspondant a une direction donnée (OCEAN, CONTI et TC) a 00h00 TU a 900 hPa

OBS	DIREC	X1	X2
1	OCEAN	-0.73593	-0.55757
2	CONTI	-0.37206	0.69465
3	TC	1.10798	-0.13708

La représentation par carte de trois groupes de directions de vent, permet donc de conforter l'existence de différences importantes entre les triplets de structures factorielles en fonction du facteur contrôlé direction (CONTI, OCEAN et TC).


Plot of X2*X1. Symbol is value of DIREC.
X2
-0.6
-0.4
-0.2
0.0
0.2
0.4
0.6
0.8
OCEAN
CONTI
TC
X1
-0.8
-0.6
-0.4
-0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2

En conclusion, dans ce chapitre nous avons pu mettre en évidence certaines différences sur les heures des maxima d'ozone selon la direction des vents. Les techniques que nous avons employées, pour élémentaires qu'elles soient, nous ont permis de déceler des spécificités propres à la majorité des stations ; a posteriori, ceci est une justification partielle de leur choix.

Partie III

Etude de la tendance dans les hautes valeurs d'ozone troposphérique

Introduction

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 dans [13] 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 tendance ajustées en fonction des conditions météorologiques. Bloomfield et al. dans [8] 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. dans [19] 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...).

Cette partie s'attache donc à mettre en évidence une tendance sur plusieurs années, dans les épisodes de pollution aigüe 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 Pickands ([35]) et améliorés par Leadbetter et al. ([28]). Ainsi, dans le cadre particulier de l'ozone troposphérique, Smith ([49]), Shively ([44]) et plus récemment les deux auteurs précédemment cités ([50]) ont utilisé l'idée de considérer le nombre de dépassements de haut niveau comme généré suivant un PPNH, puisqu'une tendance peut exister.

Chapitre 6

Les données

Pour construire ce modèle et pouvoir confronter les résultats obtenus, les données utilisées proviennent de la région parisienne et de la région de Los Angeles.

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'Energie 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 (TRANGE) : 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 (WSAVG) : moyenne entre 6h00 et 18h00. (correspondant à w5 dans le modèle),
- l'amplitude de vitesse de vent mesurée à 58 mètres (WSRANGE) : 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 6.1). S'il n'y avait pas eu de données manquantes, nous aurions disposé de 1380 jours.

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 TRANGE)	6.4
Vitesse du vent (WSAVG et WSRANGE)	15.0

Tableau 6.1: Proportion de jours manquants par variables

	Seuil						
Station	% > 120$\mu g/m3$	% > 130$\mu g/m3$	% > 140$\mu g/m3$	% > 150$\mu g/m3$	% > 160$\mu g/m3$	% > 170$\mu g/m3$	% > 180$\mu g/m3$
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 *(31)*	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)*

Tableau 6.2: Proportion de jours dépassant le seuil u par station

Le tableau 6.2 fournit pour chaque station, la proportion de jours dépassant le seuil u, pour u variant de 120 à 180 $\mu g/m3$. Les graphiques (6.1),(6.2) et (6.3)) permettent d'observer les relations existantes entre les variables. Ainsi, autant le sens de variation de l'ozone en fonction des variables météorologiques quand tous les jours sont pris en compte se dégage aisément, autant quand on ne conserve que les jours où le maximum d'ozone dépasse 130 μgm^{-3} il est très difficile à mettre en évidence à la seule lecture de ces graphiques.

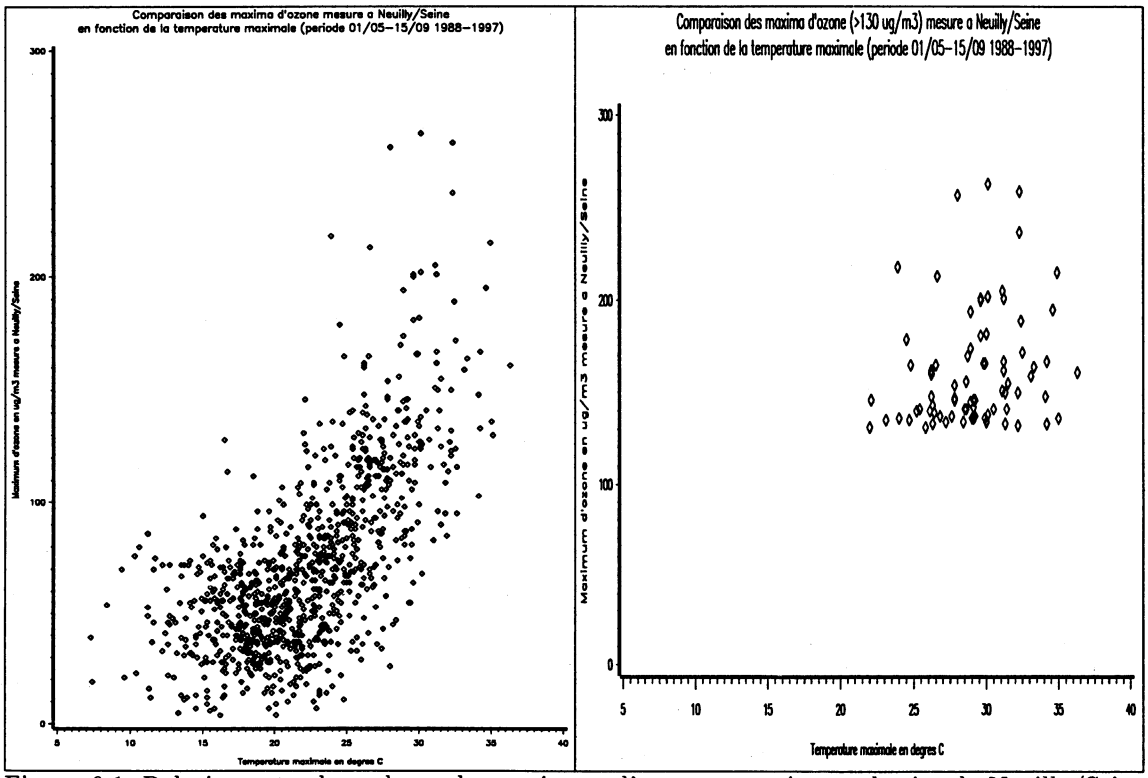


Figure 6.1: Relation entre les valeurs du maximum d'ozone mesurées sur le site de Neuilly/Seine et les valeurs de la température maximale

On observe sur le premier graphique de la figure 6.1 que plus les mesures des variables températures augmentent, plus le maximum d'ozone journalier correspondant est grand. Cependant, la relation entre l'ozone et la température est beaucoup moins évidente pour les valeurs d'ozone élevées (cf. deuxième graphique de la figure 6.1)

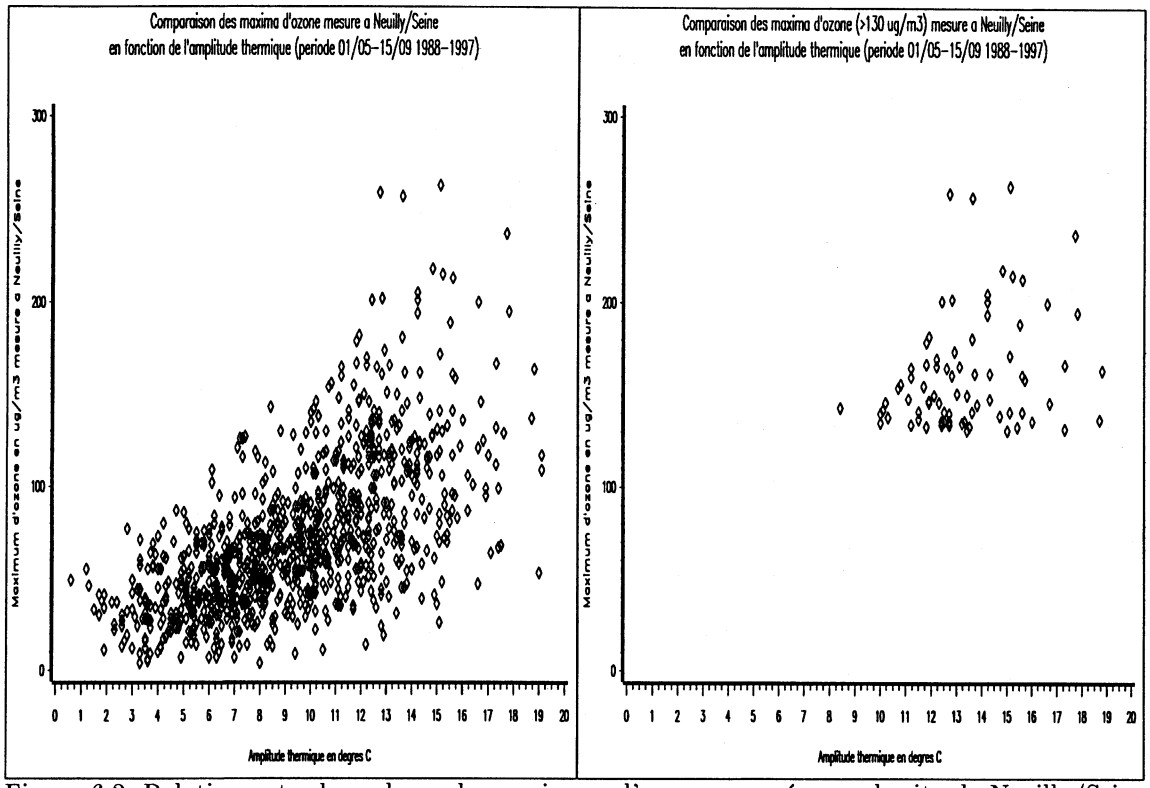


Figure 6.2: Relation entre les valeurs du maximum d'ozone mesurées sur le site de Neuilly/Seine et les valeurs de l'amplitude thermique

Quand tous les jours sont pris en compte, plus la valeur de l'amplitude thermique est élevée, plus les valeurs d'ozone sont hautes (cf. premier graphique de la figure 6.2). Cependant, la relation entre l'ozone et l'amplitude thermique est beaucoup moins évidente pour les valeurs d'ozone élevées (cf. deuxième graphique de la figure 6.2).

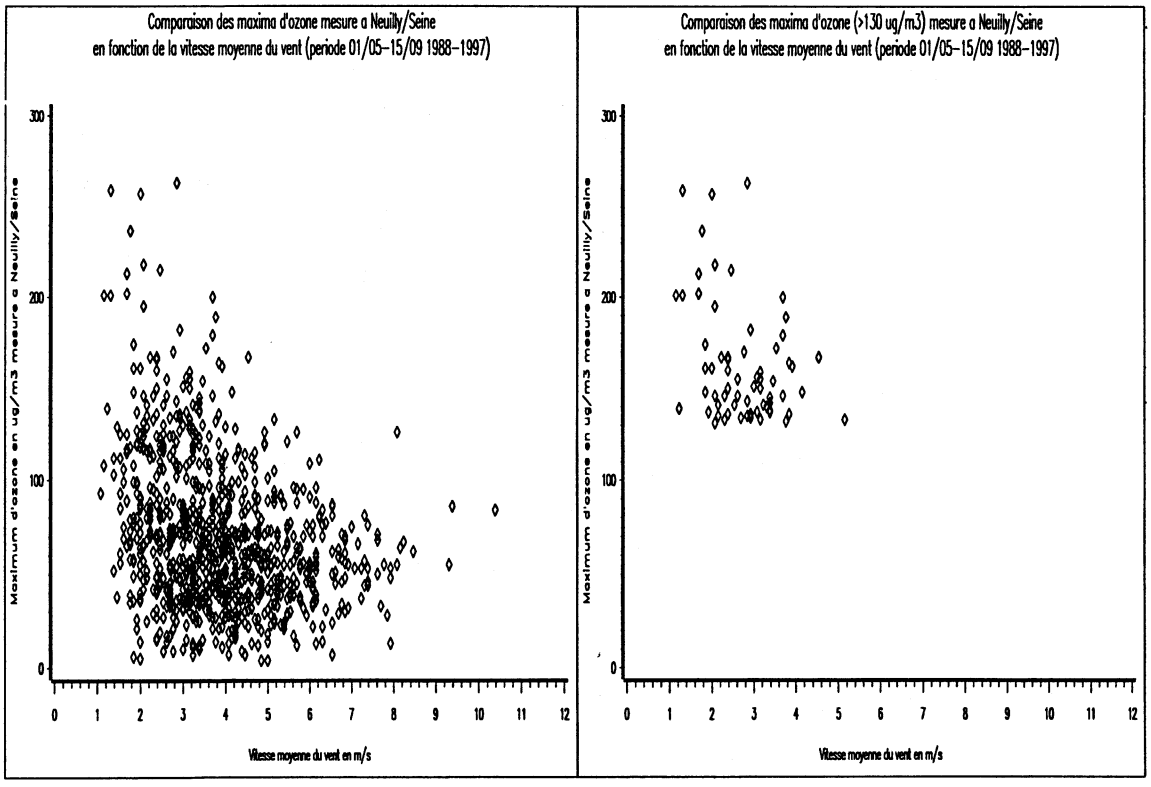


Figure 6.3: Relation entre les valeurs du maximum d'ozone mesurées sur le site de Neuilly/Seine et les valeurs de la vitesse moyenne du vent

Le premier graphique de la figure 6.3 met en évidence la diminution de la valeur du maximum d'ozone en fonction de l'augmentation de la vitesse moyenne du vent.
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.

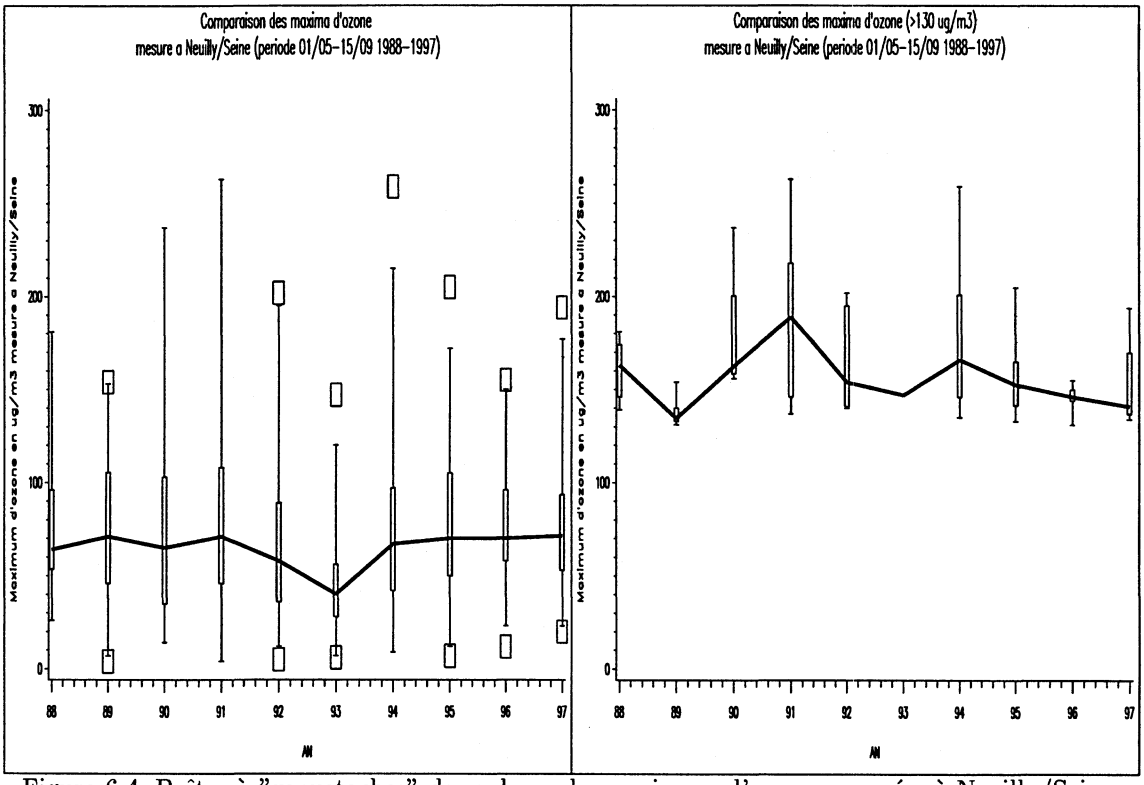


Figure 6.4: Boîtes à "moustaches" des valeurs du maximum d'ozone mesurées à Neuilly/Seine

Le premier graphe de la figure 6.4 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 (6.1)). Le deuxième graphe de la figure 6.4 (uniquement les jours où le niveau d'ozone est supérieur à 130 $\mu g/m3$) 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(200 $\mu g/m3$). 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.

Chapitre 7

Modèles de valeurs extrêmes

7.1 Cas iid

7.1.1 Théorie classique des valeurs extrêmes

Cette partie expose les résultats de la théorie des valeurs extrêmes utilisés pour modéliser des séries environnementales. Tout d'abord, restreignons nous au cas de variables aléatoires iid. Supposons $Y_1, Y_2, \cdots$, une suite iid de distribution commune F, et $M_n = \max(Y_1, \cdots, Y_n)$, la théorie classique des valeurs extrêmes cherche les suites normalisantes $a_n > 0, b_n$ telles que $(M_n - b_n)/a_n$ converge en distribution, ie.

$$P\left[(M_n - b_n)/a_n \leq y\right] = F^n(a_n y + b_n) \to H(y), \tag{7.1}$$

où H est une distribution non dégénérée.
La convergence dans (7.1) a lieu si et seulement si

$$n(1 - F(a_n y + b_n)) \to -\log H(y). \tag{7.2}$$

Dans ce cas, H doit appartenir à un des trois types de distributions limites, qui peuvent se combiner sous la forme de la *distribution généralisée des valeurs extrêmes*:

$$H(y; \mu, \sigma, k) = \exp\left[-(1 - k(y-\mu)/\sigma)^{1/k}\right] \tag{7.3}$$

pour

y tel que $1 - k(y-\mu)/\sigma > 0$,

$\sigma > 0$ et μ, k quelconques .

Le cas $k = 0$ est interprété comme la limite $k \to 0$,

$$H(y; \mu, \sigma, 0) = \exp\left[-\exp\left(-(y-\mu)/\sigma\right)\right]$$ souvent appelée distribution de Gumbel.

La méthode du maximum de vraisemblance permet d'estimer les paramètres de la famille (7.3) en pratique. Cependant les conditions de régularité du maximum de vraisemblance ne sont satisfaites que si $k < 0.5$ (cf. [16], [46]) ; la condition $k < 0.5$ est vraie dans la plupart des applications environnementales.

Une autre approche est basée sur la densité conjointe de plusieurs statistiques d'ordre d'un échantillon, au lieu seulement du maximum. En effet, notant $M_n^{(k)}$ la k-ième plus grande variable aléatoire parmi $Y_1, Y_2, \cdots, Y_n$, si $M_n = M_n^{(1)}$ satisfait (7.1), alors en identifiant $u_n = a_n y + b_n, \lambda = -\log(H(x))$, (7.2) est vrai. Si on note S_n le nombre de dépassements de u_n par $Y_1, Y_2, \cdots, Y_n$, i.e. le nombre de i, $; 1 \leq i \leq n$, tels que $Y_i > u_n$, S_n suit une loi Binomiale de paramètres $(n, p_n = 1 - F(u_n))$ avec $np_n \to \lambda$. Par conséquent, S_n possède une distribution limite de Poisson de paramètre λ. De plus, l'équivalence des événements $\{M_n^{(k)} \leq u_n\}$ et $\{S_n < k\}$ conduit à la

relation

$$P\left[a_n(M_n^{(k)} - b_n) \leq y\right] \to H(y) \sum_{s=0}^{k-1} (-\log H(y))^s / s!.$$

Par conséquent, si le maximum M_n possède une distribution limite H, alors $M_n^{(k)}$ possède la distribution limite donnée par l'équation ci-dessus.

Weissman ([57]) proposa une procédure utilisant cette distribution conjointe dans le cas d'un échantillon iid. Smith ([47]) utilisa cette approche pour analyser des données hydrologiques dans lesquelles il tenait compte des r plus grandes valeurs de chaque année (r fixé), la distribution des r plus grandes valeurs était estimée par la méthode du maximum de vraisemblance. Cette approche améliore le cas classique $r = 1$, mais pour obtenir une bonne estimation des paramètres de la distribution, r ne doit pas être pris trop grand (Smith suggérait $r = 5$ pour les données d'hydrologie étudiées dans [45]).

L'approche par seuil est basée sur l'approximation de la distribution des dépassements au dessus d'un seuil élevé u en utilisant le résultat de Pickands ([36]) suivant: Posant $u_n = a_n y + b_n$, si (7.2) est vraie alors $P[Y_1 > u_n + x / Y_1 > u_n] = \frac{1-F(u_n+x)}{1-F(u_n)} \to G(x; \sigma, k)$. où

$$G(x; \sigma, k) = 1 - (1 - kx/\sigma)^{1/k} \tag{7.4}$$

$\sigma > 0$, et $0 < x < \infty (k \leq 0)$ ou $0 < x < \sigma/k (k > 0)$.
C'est la *distribution de Pareto Généralisée (DPG)* introduite par Pickands ([36]). Le cas $k = 0$ correspond à la distribution exponentielle, très souvent utilisée dans les applications de la modélisation "Peaks Over Threshold "(décrite dans [45] et [25]) aux données hydrologiques ou de pollution de l'air (cf. [45], [15] et [21]).

7.1.2 Les processus ponctuels associés aux extrêmes

Les résultats mentionnés précédemment sur les propriétés asymptotiques de Poisson du nombre de dépassements de u_n satisfaisant (7.2, noté S_n ci-dessus) peuvent être généralisés en considérant le processus ponctuel N_n de dépassements de niveau u_n. Si $E \subset]0, 1]$, alors $N_n(E) = Card(i/n \in E : Y_i > u_n, 1 \leq i \leq n) = Card(i \in nE : Y_i > u_n, 1 \leq i \leq n)$.

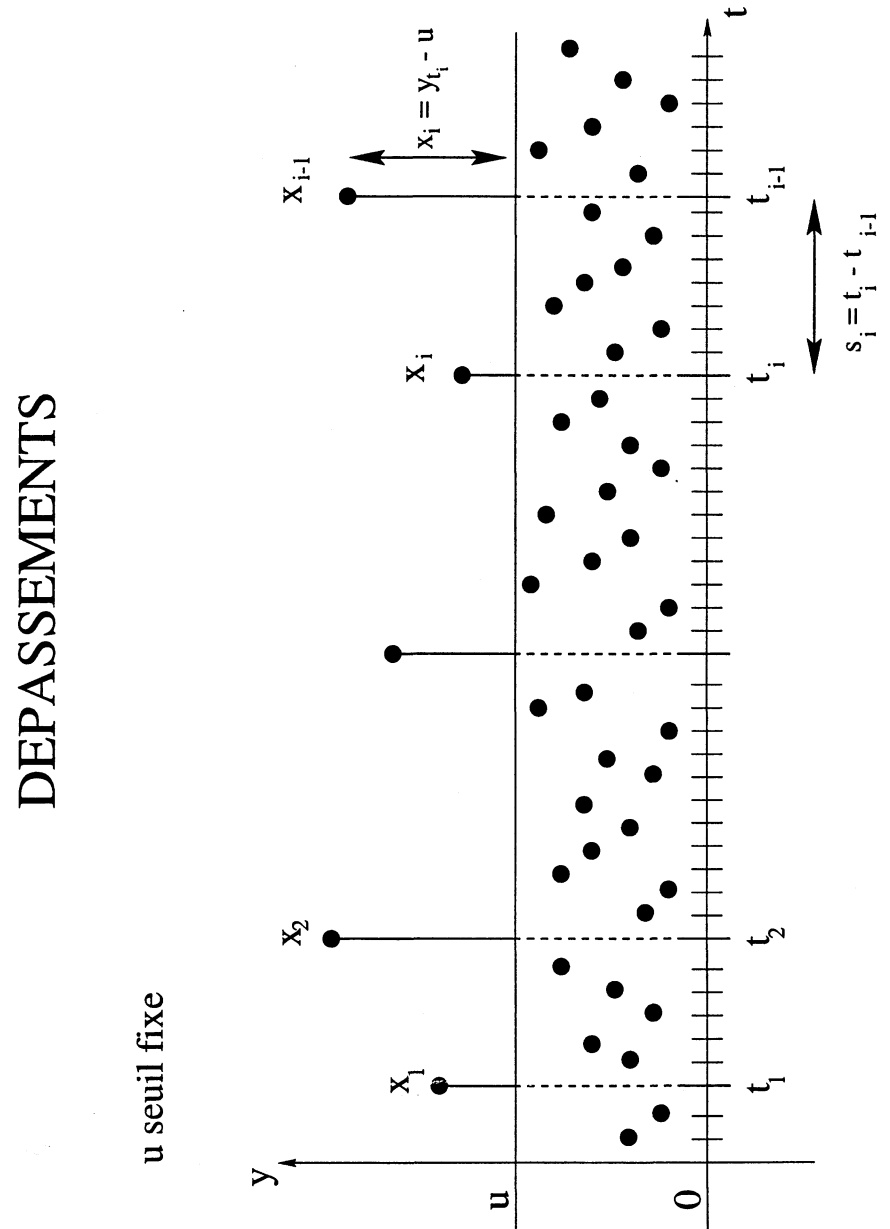


Figure 7.1: Représentation des dépassements

7.1.3 Les processus limites

Pour u_n satisfaisant $n(1-F(u_n)) \to \lambda$, on obtient immédiatement comme précédemment que $N_n(I) \to^d N(I)$, où N est un processus de Poisson sur $]0,1]$ d'intensité λ quel que soit l'intervalle $I \subset]0,1]$. Par indépendance, pour $I_1, \cdots, I_k$ intervalles indépendants, on obtient la convergence de la distribution conjointe de $N_n(I_1), \cdots, N_n(I_k)$. Ceci est en fait suffisant pour démontrer la convergence en distribution $N_n \to^d N$ des processus ponctuels N_n vers N. Les résultats sur les distributions asymptotiques du maximum et des autres statistiques d'ordre peuvent être obtenus à partir de ce résultat. Ils peuvent être étendus en considérant le processus ponctuel vectoriel des dépassements de niveaux multiples, pour obtenir les distributions asymptotiques conjointes d'un nombre fini de statistiques d'ordre. Par exemple, si le maximum possède une distribution asymptotique H donnée par (7.1), la considération de deux niveaux conduit à la distribution asymptotique des deux premières statistiques d'ordre $M_n = M_n^1$ et M_n^2 :

$$P\left[(M_n - b_n)/a_n \leq y_1, (M_n^{(2)} - b_n)/a_n \leq y_2\right] \to H(y_2)\left(\log H(y_1) - \log H(y_2) + 1\right)$$

quand $n \to \infty$, ; $y_1 > y_2$. Si on considère non plus N_n comme un processus sur $]0,1]$, mais comme un processus dans le plan, pris aux points $(i/(n+1), X_{i,n})$, $i = 1, \cdots, n$, où $X_{i,n} = (Y_i - b_n)/a_n$, $i = 1, \cdots, n$, on a alors le résultat suivant ([35]) :

Théorème 1 *Soit $Y_1, \cdots, Y_n$ un échantillon aléatoire iid de distribution F, $M_n = \max\{Y_1, \cdots, Y_n\}$, N_n le processus ponctuel dans le plan pris aux points $(i/(n+1), X_{i,n})$, $i = 1, \cdots, n$, où $X_{i,n} = (Y_i - b_n)/a_n, i = 1, \cdots, n$. Supposons (7.1) vraie, i.e. $P[(M_n - b_n)/a_n \leq y] \to H(y)$, H distribution non dégénérée, $y_0 = \inf\{y; H(y) > 0\}$.*
Alors

$$N_n \to^d N \text{ sur }]0,1[\times]y_0, \infty[.$$

où N est un processus de Poisson de mesure d'intensité, le produit de la mesure de Lebesgue et de celle définie par la fonction croissante $\log H(x)$.

Remarque :
La mesure d'intensité du processus limite se déduit de (7.2) et (7.3) :

$$\Lambda\left(]t_1, t_2[\times]y, \infty[\right) = (t_2 - t_1)\left(1 - k(y-\mu)/\sigma\right)^{1/k} \tag{7.5}$$

pour $0 \leq t_1 \leq t_2 \leq 1$, $y > y_0$ et $1 - k(y-\mu)/\sigma > 0$.
Tous les résultats de la théorie des valeurs extrêmes mentionnés précédemment peuvent être obtenus à partir de cette représentation. Par exemple, la probabilité que $(M_n - b_n)/a_n$ soit plus petit que y est simplement la probabilité que N_n n'ait pas de point dans $]0,1[\times]y, \infty[$. Si le processus limite est un processus de Poisson d'intensité (7.5), cette probabilité correspond précisément à (7.3). La distribution de la r-ième plus grande statistique d'ordre, ou la distribution conjointe des r plus grandes statistiques d'ordre (pour r fixé et n tendant vers l'infini), peut aussi être obtenue.
La distribution de Pareto Généralisée peut aussi en être déduite. En effet, la probabilité conditionnelle limite que $Y_{n,i} > u_n + x$ sachant que $Y_{n,i} > u_n$ est

$$\frac{\Lambda\left(]0,1[\times]u+x, \infty[\right)}{\Lambda\left(]0,1[\times]u, \infty[\right)} = \left[1 - \frac{kx}{\sigma - ku + k\mu}\right]^{1/k} \tag{7.6}$$

qui correspond à la distribution de Pareto Généralisée en remplaçant σ dans (7.4) par $\sigma - ku + k\mu$. La fonction d'intensité dépend des paramètres μ, σ et k qui peuvent être estimés en utilisant la méthode du maximum de vraisemblance dans le cadre d'échantillons grands. Les estimateurs du maximum de vraisemblance sont asymptotiquement normaux et efficaces si $k < 0.5$ [46]. La matrice de covariance limite est alors obtenue à partir de l'inverse de la matrice d'information de Fisher.

7.2 Cas plus général

Les données environnementales peuvent comporter une forte saisonnalité ou exhiber une dépendance qui conduit au regroupement des très grandes valeurs. La théorie des valeurs extrêmes dans le cas de

variables aléatoires dépendantes stationnaires ou non stationnaires sera abordée dans la partie IV. Nous nous contenterons ici d'évoquer quelques résultats utiles pour la compréhension des solutions mises en oeuvre en pratique pour résoudre ces problèmes de non-stationnarité et de dépendance. Pour les processus stationnaires, les lois classiques de valeurs extrêmes restent vraies sous une condition de mélange (condition D de Leadbetter). Il existe alors une relation générale de la forme

$$P[M_n \leq y] \approx [F(y)]^{n\theta}$$

où $0 \leq \theta \leq 1$ est un paramètre appelé *index extrême* du processus. C'est une mesure de la quantité de regroupements dans le processus, $1/\theta$ étant la taille moyenne du regroupement. Si $\theta > 0$, la connaissance de θ et de F permet de déterminer la distribution limite du maximum de l'échantillon. Dans des problèmes comme celui lié à la pollution de l'air où on ne connaît ni F, ni la structure de dépendance, une solution consiste à procéder directement par identification des regroupements de dépassements et étude de la distribution du maximum d'un regroupement. En effet, sous des hypothèses générales, le processus ponctuel limite du nombre de maxima de regroupement dépassant un seuil u_n tend vers un processus de Poisson.
Dans le cas de suites non-stationnaires, il n'existe pas de théorie générale. Par conséquent, en statistique appliquée deux approches sont considérées :

- décomposer les séries entières en une somme d'une composante saisonnière effectivement déterministe et d'un bruit aléatoire, supposé stationnaire. Cette idée a été développée sous le nom de " méthode de probabilité conjointe " pour séparer les effets de la marée et de la houle dans les études du niveau de la mer (cf. [37]) ; et sous une forme différente dans [45]. Cette méthode semble raisonnable quand il existe un fort mécanisme physique sous-jacent...
- supposer les paramètres du processus dépendant d'un facteur saisonnier et de covariables.

Avant d'aborder plus en détails les modélisations permettant de tenir compte de la non-stationnarité. Une question importante, reste le choix du seuil u à partir duquel en pratique l'approximation Poissonnienne peut avoir lieu.

7.3 Les différents théorèmes limites en fonction du seuil u_n

7.3.1 Les niveaux élevés et modérés

Si on note par $Y_1, Y_2, \cdots, Y_n$ les n valeurs mesurées et $X_i = (Y_i - u_n)_+, 1 \leq i \leq n$, les tailles de dépassement du seuil u_n

$$N_n = \sum_{i=1}^{n} 1_{(Y_i - u_n)_+} \tag{7.7}$$

alors (cf. [27] et [38]), on obtient sous des hypothèses générales sur la nature statistique des variables environnementales Y_i deux types de distributions asymptotiques pour N_n :

- Poisson Composé (CP) pour les niveaux très élevés
- Gaussienne pour les niveaux modérés.

Ainsi, la distinction dans les théorèmes limites entre ces deux asymptotiques provient de la croissance différente du seuil u_n avec n, le nombre de valeurs observées.
Brièvement (pour plus de détails, cf. [38]) si F représente la distribution de la variable aléatoire environnementale Y_i, $F(x) = P[Y_i \leq x]$, les niveaux u_n sont considérés comme *élevés* si la probabilité de dépassement individuelle $(1 - F(u_n))$ est petite et l'espérance du nombre de dépassements $c_n = n(1 - F(u_n))$ prend une valeur " modérée ". D'un autre côté, si la probabilité de dépassement $(1 - F(u_n))$ est petite, mais l'espérance du nombre de dépassements c_n est grande, le niveau u_n est considéré comme *modéré*. Ces idées intuitives correspondent aux conditions respectives $n(1 - F(u_n)) \to \lambda$ (pour λ fini fixé) et $n(1 - F(u_n)) \to \infty$ dans les théorèmes limites. On peut donc résumer ceci dans le tableau suivant (cf. [27]) :

Seuil	Condition sur le seuil u_n dans le théorème limite	Implications pratiques
très élevé Modèle Poisson Composé	$1-F(u_n)\to 0$ $n(1-F(u_n))\to\lambda<\infty$	Nombre de dépassements petit ou modéré.
modéré Modèle Gaussien	$1-F(u_n)\to 0$ $n(1-F(u_n))\to\infty$	Grand nombre de dépassements.

Tableau 7.1: Seuil u_n et théorème limite

7.3.2 Regroupements de dépassements: dépendance

Tout modèle statistique réaliste doit tenir compte de la dépendance entre les Y_i. Souvent, cela entraîne une très forte corrélation positive entre les valeurs voisines, traduisant le regroupement des dépassements.
Pour les seuils très élevés, les regroupements sont souvent définis du premier au dernier dépassement dans un groupe. Pour les niveaux plus bas, les regroupements sont moins faciles à définir de cette façon. On définit donc dans ce cas (cf. [27] et [26]) des " block clusters", obtenus en choisissant une taille de bloc r_n et en divisant les valeurs observées $Y_1,\cdots,Y_n$ en blocs successifs de taille r_n, ainsi le bloc B_1 contient les r_n premières valeurs $Y_1,\cdots,Y_{r_n}$, B_2 contient $Y_{r_n+1},\cdots,Y_{2r_n}$, etc.
L'importance des blocs est qu'ils tendent à exhiber certaines propriétés d'indépendance statistique même pour des seuils modérés, ce qui n'est pas nécessairement le cas pour les regroupements de dépassements (dans le cas de seuils très élevés ces deux concepts sont confondus).

Seuils élevés

Pour les seuils u_n élevés, les dépassements tendent à se produire dans des groupes bien séparés. L'espérance de la taille de groupes (i.e. le nombre de dépassements dans un groupe) est notée $\theta^{-1}, 0<\theta\leq 1$. Si $n(1-F(u_n))\approx\lambda$, le nombre C de groupes est d'après [38] approximativement une variable de Poisson d'espérance $\theta\lambda$ i.e.

$$P[C=r]\approx e^{-\theta\lambda}(\theta\lambda)^r/r! \tag{7.8}$$

Leadbetter , Rootzén et de Haan dans [38] montrent que les contributions de chaque groupe à N_n sont approximativement indépendantes de distribution G représentant donc la distribution de la taille d'un groupe (θ^{-1} est son espérance). Donc N_n en entier est un processus de Poisson Composé basé sur l'espérance $\theta\lambda$ de la loi de Poisson de C et sur la distribution G (noté $N_n=CP(\theta\lambda,G)$). La distribution de N_n s'écrit donc:

$$P[N_n\leq x]=e^{-\theta\lambda}\sum_{s=0}^{x}(\theta\lambda)^s G_s(x)/s! \tag{7.9}$$

où $x\in\mathbb{N}$ et G_s est la convolution de G s fois par elle-même.

Seuils modérés

Comme nous l'avons vu précédemment, si le seuil u_n est modérément élevé, l'espérance du nombre de dépassements $c_n=n(1-F(u_n))$ est grande et les regroupements de dépassements sont trop fréquents pour exhiber une structure Poissonnienne,ceci même dans le cas d'indépendance. Dans le cas dépendant, comme les " blocks clusters " sont asymptotiquement indépendants, sous les conditions classiques standard (incluant une " condition de Lindeberg "), on obtient pour N_n une distribution limite gaussienne et donc N_n est approximativement gausssien:

$$N_n\approx\mathcal{N}(\mu_n,\sigma_n) \tag{7.10}$$

où μ_n et σ_n sont sa moyenne et son écart-type. Dans [38], les auteurs obtiennent les valeurs des estimateurs s_n et m_n respectifs de σ_n et μ_n :

$$s_n^2 = Var(N_n^2) = \sum_{i=1}^{k_n} \left(\sum_{j \in B_i} 1_{(Y_j - u_n)_+} - r_n m_n \right)^2 \tag{7.11}$$

$$= \sum_{i=1}^{k_n} N_n^2(B_i) - N_n^2/k_n \tag{7.12}$$

où $N_n(B_i)$ correspond au nombre de dépassements dans le ième des k_n blocs de taille r_n.

$$m_n = n^{-1} \sum 1_{(Y_j - u_n)_+} = n\,(1 - F(u_n)) \tag{7.13}$$

Il est important de noter qu'en pratique le nombre n de valeurs observées est le seul élément fixe dont on dispose, il faut ensuite déterminer le seuil en fonction de l'approximation que l'on désire effectuer. Si on désire appliquer ces résultats théoriques, il faut donc choisir une taille de bloc r_n, et ensuite estimer m_n et s_n^2 dans le cas Gaussien, le paramètre θ et la distribution G dans le cas Poissonnien. Cette méthode est donc très lourde en pratique.

7.3.3 Pour l'approximation Pareto Généralisée

Smith obtient dans [48] des résultats théoriques sur le choix d'un seuil u suffisamment élevé à partir duquel, la distribution de la taille des dépassements de seuil u sachant qu'un dépassement a eu lieu peut être approchée par une distribution de Pareto Généralisée. Cependant, ils sont difficilement utilisables en pratique. Il est donc préférable d'adopter une approche plus pragmatique dans laquelle une gamme de techniques de diagnostic est utilisée pour évaluer l'estimation du modèle.

En conclusion, en pratique la détermination du seuil u permettant d'effectuer l'approximation Poissonnienne est obtenue par étude des quantiles (Nbre de jours dépassant le seuil u/ Nbre de jours total) et l'estimation du degré de dépendance (calcul du coefficient de corrélation entre intervalles de temps successifs par la méthode Bootstrap).

7.4 Modèles de valeurs extrêmes pour les données de pollution : étude de tendance

7.4.1 Approche distributionnelle

Shively, dans [43], utilise une approche distributionnelle pour déterminer s'il existe une tendance à long terme dans les valeurs du maximum d'ozone annuel mesuré à Houston (Texas). En effet, la théorie des valeurs extrêmes permet d'identifier sous des conditions faibles, la distribution de la valeur maximale d'une suite de variables aléatoires et d'estimer les paramètres de cette distribution. De plus, on peut montrer de façon empirique que la distribution de type I des valeurs extrêmes ($f(y) = \frac{1}{\sigma} \exp\left[-\exp\left(\frac{y-\mu}{\sigma}\right) - \frac{y-\mu}{\sigma}\right]$, où σ est une mesure de dispersion et μ est une mesure de l'emplacement de la distribution) fournit une estimation excellente de la valeur du maximum d'ozone annuel. Par conséquent, la probabilité de dépasser un niveau fixé élevé peut être calculée à partir de cette distribution. μ peut donc être estimé et on peut alors effectuer un test d'hypothèses pour tester si la distribution a changé au cours de la période d'étude et donc si la probabilité d'avoir un dépassement a changé dans le temps. Ainsi, dans [43], pour obtenir plus d'information sur le paramètre μ de la distribution de la valeur maximale annuelle d'O_3, les k plus grandes observations annuelles d'O_3 sont prises en compte. En particulier, ceci lui permet d'obtenir un test d'hypothèses plus puissant pour tester s'il y a un changement dans le paramètre μ sur la période d'étude. Le fait qu'il y ait plus d'information dans les k plus grandes observations en ce qui concerne la distribution du maximum (noté Y_1) est assez clair d'un point de vue intuitif, puisque étudier la forme probabiliste de $Y_1 = \max(X_1, \cdots, X_n)$ est très lié à l'étude de la forme de la queue de la distribution F générant la suite de variables aléatoires $X_1, \cdots, X_n$. Weissman , dans [57],

montre que la densité conjointe limite ($n \to \infty$) des k plus grandes observations, notées $Y_1, \cdots, Y_k$ de la suite $X_1, \cdots, X_n$ est

$$f(y_1, \cdots, y_k) = \frac{1}{\sigma^k} \exp\left[-\exp\left(\frac{y_k - \mu}{\sigma}\right) - \sum_{j=1}^{k} \frac{y_j - \mu}{\sigma}\right] \tag{7.14}$$

Donc pour n grand (7.14) fournit une bonne approximation de la densité conjointe de $Y_1, \cdots, Y_k$, si k est relativement petit par rapport à n.

Supposant que la période d'étude comporte T années et que les k plus grandes valeurs de différentes années soient indépendantes, Shively ([43]) obtient l'approximation de la densité conjointe suivante, pour n grand :

$$f(y_{11}, \cdots, y_{1k}, \cdots, y_{T1}, \cdots, y_{Tk}) = \tag{7.15}$$
$$\frac{1}{\sigma^{Tk}} \exp\left[-\sum_{t=1}^{T}\left[\exp\left(\frac{y_{tk} - \mu_t}{\sigma}\right) - \sum_{j=1}^{k} \frac{y_{tj} - \mu_t}{\sigma}\right]\right]$$

où

$$\mu_t = \alpha + \beta t/T$$

il fait donc l'hypothèse que ce paramètre change linéairement avec le temps. Après avoir estimé α, β et σ par la méthode du maximum de vraisemblance, l'hypothèse $H_0 : \beta = 0$ i.e il n'y a pas de tendance peut être testée contre l'alternative $H_A : \beta \neq 0$. Cette méthode comporte des défauts majeurs :

- Le choix de k pose problème. En effet, d'un point de vue théorique la valeur que doit prendre k par rapport à celle de n n'est pas claire. Il faut donc en pratique essayer différentes valeurs pour k.
- (7.15) est vraie s'il n'y a pas de regroupement des très hautes valeurs d'O_3. Or la validité de cette hypothèse est très contestable pour les données d'ozone.
- Les phénomènes météorologiques qui influent sur les taux d'ozone mesurés ne sont pas pris en compte.
- Ce modèle ne permet pas de quantifier la tendance.

7.4.2 Approche processus ponctuel

Modèle introduit par Smith , [49]

Smith , dans [49], tient compte du regroupement des mesures élevées d'O_3 (application aux données mesurées à Houston, Texas) en étudiant les valeurs du maximum mesuré sur des intervalles de regroupement. Ainsi, les années sont divisées en périodes de M jours, pour par exemple $M = 31$ et un modèle est estimé sur chaque période. Si N_{ij} représente le nombre de dépassements dans la période j de l'année i, Y_{ijm} $(1 \leq m \leq N_{ij})$ les tailles de dépassements individuelles, p_{ij} le nombre d'observations dans la période j de l'année i (les p_{ij} ne sont pas égaux à cause des données manquantes, en % du nombre de jours que comporte l'année i). L'auteur définit alors les paramètres μ_{ij}, σ_{ij} et k_{ij} de la distribution de valeurs extrêmes de la période j de l'année i, tels que :

$$\mu_{ij} = \alpha_j + i\beta_j, \ \sigma_{ij} = \sigma_j, \ k_{ij} = k_j, \tag{7.16}$$

les paramètres dépendent donc linéairement de l'année.
Utilisant la théorie des valeurs extrêmes pour les données à l'intérieur de la période j de l'année i et un choix du seuil suffisamment élevé, les dates de dépassement et les tailles de dépassements forment un processus de Poisson non-homogène de mesure d'intensité donnée par l'équation (7.5) dans laquelle $\mu = \mu_{ij}$, $\sigma = \sigma_j$, $k = k_j$. La vraisemblance sur l'ensemble des périodes s'écrit donc

$$L = \prod_{i,j}\left[\exp\left(-p_{ij}(1 + k_j\mu_{ij}/\sigma_j)^{1/k_j}\right) \prod_{m=1}^{N_{ij}}\left((1 - k_j(y_{ijm} - \mu_{ij})/\sigma_j)^{1/k_j - 1}/\sigma_j\right)\right] \tag{7.17}$$

La méthode du maximum de vraisemblance est ensuite utilisée pour obtenir les estimations des paramètres et de leur écart-type.
Cette méthode est très lourde. En effet, avant d'arriver à l'étape de détermination de tendance, il faut choisir le seuil, l'intervalle de regroupement et la longueur des périodes divisant l'année (par exemple seuil de 240 μgm^{-3}, intervalle de regroupement de 72 heures, période de $M = 31$ jours). De plus, comme dans [43], elle ne tient pas compte des facteurs extérieurs pouvant modifier la tendance tels que les conditions météorologiques, les erreurs de mesures...

Utilisation d'un PPNH pour approcher le processus ponctuel du nombre de dépassements

- LES DIFFÉRENTES MODÉLISATIONS DE L'INTENSITÉ

Shively, dans [44], Vaquera-Huerta , Villasenor et Hughes dans [55] utilisent l'idée considérant le nombre de dépassements d'un niveau élevé u comme généré suivant un processus de Poisson non-homogène (PPNH) pour estimer la tendance à long terme en tenant compte de la relation entre les très hautes valeurs d'ozone et les conditions météorologiques. Un processus de Poisson est non-homogène si la fonction d'intensité $\lambda(t) = P[Y > u$ le jour t] n'est pas constante dans le temps, $\lambda(t)$ représentant le taux d'apparition des événements par unité de temps. De nombreux modèles paramétriques ont été proposés pour modéliser $\lambda(t)$.
Le modèle exponentiel pour l'intensité (cf. [12]) a été introduit par Shively dans [44] (données Houston, Texas) pour tester l'existence d'une tendance et tenir compte de covariables. Le modèle est donné par :

$$\lambda(t) = \exp(\alpha + \beta t) \tag{7.18}$$

soit sous forme plus explicite :

$$\lambda(t) = \exp(\alpha(t)) = \exp\left(\alpha_0 + \alpha_1 s(t) + \sum_{j=2}^{p} \alpha_j w_j(t)\right) \tag{7.19}$$

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 des k variables météorologiques et de la tendance à long terme représentée par $s(t)$. Tester une tendance à long terme dans la fréquence à laquelle les hautes valeurs d'ozone apparaissent, 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 $H_A : \alpha_1 \neq 0$.
Une autre famille de modèles paramétrique pour l'intensité d'un PPNH a été proposée par Crow dans [14], l'intensité est une fonction de Weibull de la forme :

$$\lambda(t) = (\alpha/\mu)(t/\mu)^{\alpha-1},\ \alpha > 0, \mu > 0. \tag{7.20}$$

Un modèle plus général permettant des formes de tendance variées a été proposé par Lee dans [29] et la fonction d'intensité est donnée par

$$\lambda(t) = \mu\alpha t^{\alpha-1} \exp(\beta t),\ \alpha > 0,\ -\infty < \beta < \infty,\ \mu > 0. \tag{7.21}$$

Le modèle (7.21) se réduit au modèle exponentiel si $\alpha = 1$; si $\beta = 0$, on retrouve le modèle de type Weibull. Si $\alpha = 1$ et $\beta = 0$, on obtient une intensité constante (i.e. un processus de Poisson homogène), qui correspond au cas où il n'y a pas de tendance.
[55] propose un modèle moins général que (7.21), mais qui permet de tester de nombreuses combinaisons de tendance :

$$\lambda(t) = \mu\alpha t^{\alpha-1} \exp(\beta t^{\alpha}),\ \alpha > 0,\ -\infty < \beta < \infty,\ \mu > 0. \tag{7.22}$$

- LA VRAISEMBLANCE ASSOCIÉE AUX VARIABLES ALÉATOIRES DATES DE DÉPASSEMENTS DE SEUIL $T_i, i = 1, \cdots, N$

Supposons que le nombre de dépassements N se produit suivant un processus de Poisson de fonction d'intensité $\lambda(t)$ sur l'intervalle de temps $]0, T[$, alors pour une date r donnée appartenant à $]0, T[$, la densité de l'intervalle de temps jusqu'au prochain événement est donnée par :

$$f_r(z; \alpha, \beta, \gamma) = \exp\left[-\left(h(r+z) - h(r)\right)\right] \lambda(r+z) \tag{7.23}$$

où

$$h(r) = \int_0^r \lambda(t)dt.$$

La densité $f_r(z; \alpha, \beta, \gamma)$ caractérise la probabilité qu'un dépassement se produise dans un intervalle de temps donné. Cette probabilité dépend de la fonction d'intensité, et bien sûr des conditions météorologiques dans cet intervalle de temps. Supposant que les événements se produisent aux dates $t_1, \cdots, t_n; 0 < t_1 < t_2 < \cdots < t_n < T$ sur la période $]0, T[$, la vraisemblance est alors :

$$\begin{aligned} L(t_1, \cdots, t_n, n) = \lambda(t_1) \exp\left[-\int_0^{t_1} \lambda(u)du\right] \lambda(t_2) \exp\left[-\int_{t_1}^{t_2} \lambda(u)du\right] \\ \cdots \lambda(t_n) \exp\left[-\int_{t_{n-1}}^{t_n} \lambda(u)du\right] \exp\left[-\int_{t_n}^{T} \lambda(u)du\right] \\ = \left[\prod_{i=1}^{n} \lambda(t_i)\right] \exp\left[-\int_0^T \lambda(u)du\right]. \end{aligned} \tag{7.24}$$

Alors, la densité conditionnelle par rapport à n est

$$f(t_1, \cdots, t_n | n) = n! \left(\int_0^T \lambda(u)du\right)^{-n} \prod_{i=1}^{n} \lambda(t_i) \tag{7.25}$$

où $0 < t_1 < t_2 < \cdots < t_n < T$. Par conséquent, les variables aléatoires dates de dépassement $T_1, \cdots, T_n$ sachant que n dépassements ont eu lieu en tout, sont distribuées comme les statistiques d'ordre dans un échantillon de taille n obtenues à partir d'une distribution de densité

$$f(t) = \left(\int_0^T \lambda(u)du\right)^{-1} \lambda(t) \tag{7.26}$$

où $0 < t < T$. Par conséquent, si les données disponibles correspondent aux n premières dates de dépassements (ordonnées) $t_1, \cdots, t_n$, en utilisant (7.26) leur vraisemblance est donc

$$L(\alpha, \beta | t_1, \cdots, t_n) = \prod_{i=1}^{n} f(t_i) \tag{7.27}$$

- ESTIMATION DES PARAMÈTRES DU MODÈLE

Les estimateurs du maximum de vraisemblance des coefficients sont alors calculés en maximisant (7.27) par une procédure numérique (l'algorithme de Newton-Raphson peut être appliqué dans ce cas pour trouver les estimateurs du maximum de vraisemblance).
Dans [44], les variables dans $\alpha(t)$ (cf. (7.19) peuvent être interprétées comme les variables reliées dans le temps à la fréquence des dépassements du seuil u. En particulier, le coefficient α_1 associé à la variable de tendance représente la tendance dans la fréquence des niveaux d'ozone à conditions météorologiques constantes. Les coefficients $\alpha_3, ..., \alpha_p$ fournissent une mesure de la relation entre conditions météorologiques et fréquence des valeurs élevées

d'ozone. Pour sélectionner les variables dans $\alpha(t)$, les jours de dépassement sont modélisés en utilisant un processus de Poisson non-homogène. La vraisemblance de ce processus est donnée par (7.27). On commence par inclure un terme constant, et toutes les variables dans l'expression de $\alpha(t)$, ensuite on élimine les variables statistiquement non significatives en effectuant, sous l'hypothèse que l'échantillon de données soit suffisamment grand, un test sur les coefficients eux-mêmes en utilisant l'approximation normale des $\hat{\alpha_j}$. On a :
sous $H_0 : \alpha_j = 0$

$$\frac{\hat{\alpha_j} - 0}{\hat{\sigma}(\alpha_j)} \sim N(0,1)$$

$\Rightarrow H_0$ est rejetée au seuil 0.05 si :

$$|\frac{\hat{\alpha_j}}{\hat{\sigma}(\alpha_j)}| \geq 1.96$$

Si les valeurs absolues des rapports des coefficients estimés sur leur écart-type estimés sont toutes supérieures à 1.96, aucune variable n'est éliminée. Sinon, la variable pour laquelle le test a la plus petite valeur est éliminée. L'intensité $\Psi(t)$ 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. (Si la distribution des statistiques de test est approximativement normale, alors la valeur 1.96 est approximativement la valeur critique au seuil 0.05 du test bilatéral de l'hypothèse nulle qu'un coefficient donné soit nul contre l'hypothèse que le coefficient soit non nul).
Une autre alternative pour tester la présence d'une tendance à celle présentée ci-dessus et développée dans [44] est d'utiliser un test du rapport de vraisemblance et tester l'hypothèse $H_0 : \alpha = \alpha_0,\ \beta = \beta_0$ contre l'alternative $H_a : \alpha = \hat{\alpha},\ \beta = \hat{\beta}$, ($\hat{\alpha}$ et $\hat{\beta}$ estimateurs du maximum de vraisemblance).En effet, sous H_0 la distribution de $-2\log\frac{L(\alpha_0,\beta_0|t_1,\cdots,t_n)}{L(\hat{\alpha},\hat{\beta}|t_1,\cdots,t_n)}$ est asymptotiquement celle d'un χ^2 à deux degrés de liberté.

- VALIDATION DU MODÈLE

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

- l'indépendance du nombre d'événements dans des intervalles de temps séparés, ce qui est équivalent à vérifier que les intervalles de temps entre événements $S_i = T_i - T_{i-1}$ sont indépendants.
- la distribution des intervalles $S_i = T_i - T_{i-1}$ peut être approchée par une distribution exponentielle de la forme (7.23).
- *Test sur la distribution des intervalles S_i*

  Dans [44], Shively propose de calculer la distribution des intervalles S_i, après avoir estimé les paramètres de l'intensité $\lambda(t)$ prenant la forme exponentielle (7.19)
  La densité de l'intervalle S (s=(t+s)-t) jusqu'au prochain événement se produisant le jour $t + s$ s'écrit :

$$f_t(s) = \lambda(t+s)\exp[-h(s)] \qquad (7.28)$$
$$\text{où } h(s) = \int_t^{t+s} \lambda(r)dr$$

  et donc la distribution est donnée par :

$$F_t(s) = P\{S \leq s\} = \int_0^s f_t(r)dr = 1 - \exp[-h(s)] \qquad (7.29)$$

Or, les jours de dépassement pour les données d'ozone étant discrets (le premier dépassement se produit le jour t_1, le deuxième le jour t_2,..), $F_t(s)$ ne doit être calculée que pour les valeurs discrètes de t et s. On peut donc approcher (7.29) par :

$$F_t(s) = 1 - \exp\left\{-\sum_{k=1}^{s} \overline{\lambda}(k)\right\} \tag{7.30}$$

où

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

On transforme ensuite S_i en :

$$U_i = F_{t(i-1)}(S_i)$$

où (7.31)

$t(i-1)$correspond au jour du (i-1)ième événement.

Ainsi, si les intervalles de temps S_i sont indépendants de distribution $F_t(s)$, alors cette transformation permet de réduire S_i à une variable aléatoire U_i de distribution uniforme sur $[0;1]$. On ordonne ensuite les valeurs de U_i (statistique d'ordre de U_i notée $U_{(i)}$), puis on effectue le graphique $(u_{(i)}; i/(n+1))$ pour $i = 1, \cdots, n$. Si les variables aléatoires S_i ont la distribution donnée par (7.29), alors les points doivent se situer au voisinage de la bissectrice.

- *Indépendance mutuelle des intervalles S_i*
En théorie, le fait que les variables S_i ne soient pas corrélées n'implique pas qu'elles sont indépendantes. Cependant, d'un point de vue pratique, s'il y a dépendance entre deux intervalles S_i successifs, cela signifiera 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, $S_i = T_i - T_{i-1}$ n'est pas corrélé avec $S_{i-1} = T_{i-1} - T_{i-2}$.
Shively , dans [44], propose de calculer le coefficient de corrélation entre intervalles adjacents, les intervalles s_i $i = 1, \cdots, n$ étant connus :

$$r = \frac{1}{n-2} \sum_{i=2}^{n} \left[\frac{s_i - E_{t(i-1)}(S_i)}{\sigma_{t(i-1)}(S_i)}\right] \left[\frac{s_{i-1} - E_{t(i-2)}(S_{i-1})}{\sigma_{t(i-2)}(S_{i-1})}\right] \tag{7.32}$$

où, $E_{t(i-1)}(S_i)$ [$\sigma_{t(i-1)}(S_i)$] est l'intervalle moyen [l'écart-type] entre le $(i-1)$ième événement et le ième.

Notations :
Dans un but de simplification, les indices $(i-1)$ et i seront omis.
Pour la fonction d'intensité $\lambda(r)$ et un jour donné t, la densité de l'intervalle S est donnée par (7.28).
D'où :

$$E_t(S) = \int_0^{+\infty} s f_t(s) ds = \int_0^{+\infty} s\lambda(t+s) \exp\left\{-\int_t^{t+s} \lambda(r) dr\right\} ds \tag{7.33}$$

L'intégrale dans (7.33) est difficilement calculable en utilisant des techniques d'intégration standard, à cause de la fonction exponentielle. On utilise donc la technique suivante :
Etant donné que :

$$E_t(S) = \int_0^{+\infty} s f_t(s) ds = \sum_{k=1}^{+\infty} \left(\int_{k-1}^{k} s f_t(s) ds\right) \tag{7.34}$$

Notant $\overline{\lambda}(k) = \lambda(t+k)$, si le jour $t+k$ n'est pas manquant, on obtient alors :

$$\int_{k-1}^{k} s f_t(s)ds = \int_{k-1}^{k} s\overline{\lambda}(k) \exp\left\{ - \left[\sum_{j=1}^{k-1} (\overline{\lambda}(j)) + (s-(k-1))\overline{\lambda}(k) \right] \right\} ds \qquad (7.35)$$

Puisque :

$$\forall s \in (k-1;k],\ \lambda(t+s) = \overline{\lambda}(k)$$
$$\Rightarrow \int_{k-1}^{k} s f_t(s)ds = \int_{k-1}^{k} s\lambda(t+s) \exp\left[-h(s)\right] ds$$

où,

$$\begin{aligned} h(s) &= \int_{t}^{t+s} \lambda(r)dr = \int_{0}^{s} \lambda(t+v)dv \\ &= \int_{0}^{k-1} \lambda(t+v)dv + \int_{k-1}^{s} \lambda(t+v)dv \\ &= \sum_{j=1}^{k-1} \overline{\lambda}(j) + (s-(k-1))\overline{\lambda}(k). \end{aligned}$$

D'où, en intégrant par parties (7.35) :

$$\begin{aligned} \int_{k-1}^{k} s f_t(s)ds &= \left[(k-1) + \frac{1}{\overline{\lambda}(k)} \right] \exp\left\{ - \sum_{j=1}^{k-1} \lambda(j) \right\} \\ &\quad + \left[k + \frac{1}{\overline{\lambda}(k)} \right] \exp\left\{ - \sum_{j=1}^{k} \lambda(j) - \overline{\lambda}(k) \right\} \end{aligned} \qquad (7.36)$$

Remarque :
Si le jour $t+k$ est manquant, alors $\int_{k-1}^{k} s f_t(s)ds = 0$.

Les équations (7.34) et (7.36) permettent d'obtenir une expression de $E_t(S)$ facilement calculable. En pratique, la somme infinie dans (7.34) doit être tronquée après un nombre suffisant de termes. Le nombre $k = 60$ paraît raisonnable, la densité étant très petite après 60 termes, puisque la probabilité que les intervalles soient supérieurs ou égaux à 60 jours est négligeable.
De même, pour calculer $\sigma_t(S)$, on utilise l'expression :

$$\sigma_t^2(S) = Var_t(S) = E_t(S^2) - [E_t(S)]^2$$

où

$$E_t(S^2) = \int_{0}^{+\infty} s^2 f_t(s)ds = \int_{0}^{+\infty} s^2\lambda(t+s) \exp\left\{ \int_{t}^{t+s} \lambda(r)dr \right\} ds$$

$E_t(S^2)$ peut être calculée en utilisant une technique similaire à celle développée pour $E_t(S)$.

Le principal problème de ce genre de modélisation est qu'il ne tient pas compte du regroupement des hautes valeurs, puisqu'il est basé sur l'hypothèse d'indépendance des intervalles de temps entre deux dépassements de seuil u fixé élevé. Ces modèles à une dimension paraissent peut réalistes. Le modèle prenant en compte à la fois la fréquence et la taille de dépassement développé par Smith et Shively , dans [50], paraît plus réaliste.

Utilisation d'un PPNH dans le plan pour approcher le processus ponctuel du nombre et de la taille des dépassements

- LA VRAISEMBLANCE ASSOCIÉE AUX DATES ET AUX TAILLES DE DÉPASSEMENTS DE SEUIL
Notons :

  - Y variable aléatoire représentant le maximum d'ozone journalier
  -
$$\Psi_t(y) = \begin{cases} P(Y > y & \text{le jour t)} \\ 0 & \text{si le jour t est manquant} \end{cases}$$

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

$$1 - \Psi_t(y)$$

  Par conséquent :

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

  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 :

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

  Le ième pic se produit donc le jour T_i et prend la valeur $Y_i \geq u$. Le nombre total des N pics étant lui aussi une variable aléatoire, la densité conjointe des pics observés peut être approchée par :

$$L = \left[(\prod_{i=1}^{N} \Psi_{t_i}(u)) \exp\left[-\int_0^T \Psi_t(u)dt \right] \right] \left[\prod_{i=1}^{N} \frac{\psi_{t_i}(y_i)}{\Psi_{t_i}(u)} \right] = A * B \qquad (7.37)$$

  Interprétation de chaque terme entre crochets :

  - Le premier terme entre crochets dans (7.37) correspond à la densité d'un processus de Poisson non-homogène de fonction de fréquence (intensité) $\Psi_t(u)$ correspondant $\lambda(t)$ dans (7.24). Il correspond donc à la modélisation des jours de dépassement de niveau u (fréquence des dépassements).
  - Le ième terme du second terme entre crochets correspond à la densité de Y_i sachant qu'un dépassement de seuil u a eu lieu.

  Remarque :
  La densité conjointe (7.37) est une approximation de la vraie densité des $(T_i; Y_i), 1 \leq i \leq N$ 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 (par définition, il se produit au plus un dépassement par jour car l'étude utilise le maximum d'ozone journalier). Cependant, il semble que la densité (7.37) soit une approximation raisonnable au vu des résultats obtenus dans [50].

- MODÈLE ASSOCIÉ À LA FRÉQUENCE DES DÉPASSEMENTS
La modélisation de la fréquence des dépassements du seuil u élevé utilise la même méthode que celle développée dans [44].

- MODÈLE ASSOCIÉ AUX TAILLES DES DÉPASSEMENTS

Les résultats théoriques sur la taille des dépassements d'un seuil élevé, dans le paragraphe 7.1.1, permettent d'obtenir que la distribution limite de $X = Y - u$ sachant que $Y \geq u$ quand

$u \to \infty$ est une distribution de Pareto généralisée notée G, dans laquelle les paramètres dépendent des covariables météorologiques et de tendance :

$$G(x;\beta(t),\xi(t)) = 1 - (1+\xi(t)\beta(t)x)^{-\frac{1}{\xi(t)}} \tag{7.38}$$

avec $x = y - u$ où $\beta(t) > 0$, $\forall t$. La densité limite de X est donc :

$$\lim_{u\to+\infty} \frac{\psi_t(y)}{\Psi_t(u)} = g(x;\beta(t),\xi(t)) = \beta(t)\,(1+\xi(t)\beta(t)x)^{-\left(\frac{1}{\xi(t)}+1\right)} \tag{7.39}$$

pour u suffisamment grand (7.39) permet d'approcher la densité de X. De plus, comme les moments de la distribution (7.38) n'existent pas pour toutes les valeurs de ξ, généralement on suppose $\xi(t)$ indépendant de t et on modélise les variations dans β. [50] suppose que ξ ne dépend pas du jour, c'est-à-dire, $\xi(t) = \xi$ et que comme $\xi \approx 0$, dans ce cas puisque :

$$\lim_{\xi\to 0} g(x;\beta(t),\xi) = \beta(t)\exp(-\beta(t)x), \tag{7.40}$$

densité d'une loi exponentielle de paramètre $\beta(t)$.

$\frac{1}{\beta(t)}$ 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 (7.40). $\frac{1}{\beta(t)}$ peut donc être interprété 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 pour $\xi = 0$ d'après [46], 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. $\beta(t)$ prend la forme analytique suivante :

$$\beta(t) = \beta_0 + \beta_1 s(t) + \sum_{j=2}^{p} \beta_j w_j(t) \tag{7.41}$$

Remarque :
Comme par hypothèse, $\beta(t)$ doit être strictement supérieur à zéro, quel que soit t, une formulation plus naturelle (cf. [16]) pourrait être la suivante :

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

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 des k variables météorologiques et de la tendance à long terme représentée par $s(t)$.

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 $\hat{\beta}$.

En utilisant les propriétés de convergence asymptotique de l'estimateur du maximum de vraisemblance, la matrice de covariance peut alors être estimée par l'inverse de la matrice d'information de Fisher :

$$\hat{\sum} = \left[I(\hat{\beta}_0,..,\hat{\beta}_p)\right]^{-1} = \left[-E\left(\frac{\partial^2 l_2}{\partial\beta_j\partial\beta_k}\right)\right]^{-1}$$

- VALIDATION DU MODÈLE ASSOCIÉ AUX TAILLES DE DÉPASSEMENTS

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

- (i) La densité :

$$g(x;\beta(t)) = \lim_{u\to+\infty} \frac{\psi_t(x+u)}{\Psi_t(u)} = \beta(t)\exp(-\beta(t)x) \tag{7.42}$$

est le modèle approprié aux tailles de dépassement $X_i = Y_i - u$ sachant que $Y_i \geq u$. Mais, vérifier cette hypothèse nécessite en fait de vérifier :

  * (a) L'adéquation de l'équation (7.42) pour $\beta(t)$ décrit dans (7.41).
  * (b) Si u est un seuil de dépassement suffisamment élevé pour que la densité exponentielle $g(x;\beta(t))$ fournisse une bonne approximation de la densité de X.
  * (c) Si (7.42) est la densité appropriée à la variable aléatoire X représentant la taille du dépassement et si elle n'est pas acceptable, alors la densité plus générale (7.38) doit être utilisée.

- (ii) Les dépassements qui se produisent les jours T_i et T_j sont indépendants : $X_i = Y_i - u$ sachant que $Y_i \geq u$ est indépendant de $X_j = Y_j - u$ sachant que $Y_j \geq u$, pour $i \neq j$.

- *Distribution de la taille des dépassements*
  Le test de Kolmogorov-Smirnov permet de vérifier que la distribution exponentielle est la distribution appropriée pour modéliser la taille des dépassement, et que l'expression de $\beta(t)$ dans (7.41) est correcte.

- *Indépendance mutuelle des tailles de dépassement X_i*
  Pour vérifier l'indépendance , Smith et Shively ([50]) calculent le coefficient de corrélation entre tailles de dépassement survenu des jours consécutifs. Si n_{adj} représente le nombre de paires de tailles de dépassement adjacent, $\{x_{i-1}(t-1), x_i(t)\}$ où $x_i(t)$ correspond à la valeur du ième dépassement survenu le jour t, le coefficient de corrélation entre tailles de dépassement adjacent s'écrit :

$$r = \frac{1}{n_{adj}-1} \sum_{i=1}^{i=n_{adj}} \left[\frac{x_i(t) - E\left(X_i(t)\right)}{\sigma\left(X_i(t)\right)}\right]\left[\frac{x_{i-1}(t-1) - E\left(X_{i-1}(t-1)\right)}{\sigma\left(X_{i-1}(t-1)\right)}\right]$$

où

  * L'écart-type de r est approximativement $1/n_{adj}^{\frac{1}{2}}$
  * $\begin{cases} E(X_i(t)) = \frac{1}{\beta(t)} \\ Var(X_i(t)) = \frac{1}{\beta(t)^2} \end{cases}$

Chapitre 8

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. Nous n'évoquerons donc ici que les spécificités de ce modèle par rapport à celui développé par Smith et Shively dans [50] et décrit dans la partie précédente. La densité (7.37) fournit donc une approximation de la vraie densité des $(T_i, Y_i), 1 \leq i \leq N$.

8.1 Le choix d'un seuil u raisonnable

Pour un seuil u fixé, les intervalles s_i $i = 1, \cdots, N$ étant connus, on calcule une estimation du coefficient de corrélation entre intervalles adjacents par la méthode Bootstrap. Cette estimation ne nécessitant aucune connaissance sur la loi des $S_i = T_i - T_{i-1}$ (cf. annexe B.4), 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 pour modéliser les dates de dépassements, 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 $x_i(t)$, $i = 1, \cdots, n_{adj}$ étant connues (n_{adj} nombre de paires de tailles de dépassement successifs : $\{x_{i-1}(t-1), x_i(t)\}$).

8.2 Spécificités du modèle associé à la fréquence des dépassements

8.2.1 L'intensité du PPNH

On remarque que $\Psi_t(u) = P[Y > u \text{ le jour t }]$ 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 $E[Z/$ covariables $]$. Le modèle de régression logistique (cf. [22]) permettant d'obtenir l'intensité du PPNH, est le suivant :

$$\lambda(t) = \Psi_t(u) = \frac{\exp(\alpha(t))}{1 + \exp(\alpha(t))} \tag{8.1}$$

où $\alpha(t)$ tient compte des interactions possibles entre covariables contrairement à l'écriture (7.19) :

$$\alpha(t) = \alpha_0 + \alpha_1 s(t) + \sum_{j=2}^{p} \alpha_j w_j(t) + \sum_{j=2}^{p} \alpha_{1j} s(t) w_j(t) + \sum_{i,j=2}^{p} \alpha_{ij} w_i(t) w_j(t) \tag{8.2}$$

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.

8.2.2 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 (cf. annexe B.6).

8.2.3 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 (8.1), on utilise comme dans [50], un test de Kolmogorov-Smirnov pour valider la distribution des S_i.

8.3 Spécificités du modèle associé à la taille des dépassements

On suppose comme dans [50] que les tailles de dépassements X_i suivent une loi exponentielle de densité (7.40). Mais le paramètre $\beta(t)$ prend en compte les interactions possibles entre covariables :

$$\beta(t) = \beta_0 + \beta_1 s(t) + \sum_{j=2}^{p} \beta_j w_j(t) + \sum_{j=2}^{p} \beta_{1j} s(t) w_j(t) + \sum_{i,j=2}^{p} \beta_{ij} w_i(t) w_j(t) \tag{8.3}$$

Pour estimer les paramètres de ce modèle, nous avons dû écrire (en SAS) un programme spécifique fourni en annexe B.10.2.

Chapitre 9

Résultats obtenus pour la Région Parisienne

9.1 Choix d'un seuil u raisonnable

9.1.1 Neuilly/Seine

Seuil en μg/m3	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
Cor	-0.008	-0.03	0.10	0.17	0.06	0.0001	-0.09
Cormoy	-0.004	-0.03	0.10	0.17	0.06	-0.00009	-0.08
Corstd	0.03	0.03	0.07	0.14	0.12	0.12	0.12
Corinf	-0.08	-0.12	-0.08	-0.11	-0.24	-0.22	-0.34
Corsup	0.15	0.16	0.44	0.63	0.52	0.58	0.51
Cormed	-0.008	-0.04	0.10	0.17	0.05	-0.01	-0.10
q0025	-0.06	-0.08	-0.03	-0.06	-0.14	-0.16	-0.27
q0975	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
Cor	0.37	0.22	0.24	0.33	0.43	0.38	0.41
Cormoy	0.37	0.21	0.23	0.31	0.42	0.33	0.38
Corstd	0.09	0.13	0.14	0.18	0.21	0.26	0.12
Corinf	0.10	-0.17	-0.18	-0.20	-0.32	-0.39	-0.34
Corsup	0.62	0.60	0.66	0.81	0.92	0.89	0.98
Cormed	0.37	0.21	0.23	0.32	0.45	0.37	0.42
q0025	*0.18*	-0.03	-0.05	-0.06	-0.03	-0.16	-0.29
q0975	*0.54*	0.46	0.51	0.65	0.75	0.75	0.91

Tableau 9.1: Seuil u pour Neuilly/Seine (les notations des symboles sont indiquées en annexe B.4)

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 (cf. tableau 9.1), ce seuil ne peut pas être utilisé. Par conséquent, pour la station de Neuilly/Seine les seuil u raisonnables sont $\{100, 110, 120, 130, 140, 150\}$.

9.1.2 Champs/Marne

Les intervalles de confiance à 95 % des estimations des coefficients de corrélation correspondant au seuil 90 μgm^{-3} ne comprenant pas la valeur zéro (cf. tableau 9.2), ce seuil ne peut pas être utilisé pour modéliser le processus des dépassements. Par conséquent, pour la station de Champs/Marne les seuils u raisonnables sont $\{100, 110, 120, 130, 140\}$.

Seuil en μg/m3	90	100	110	120	130	140	150
Estimation du coefficient de corrélation entre intervalles de temps successifs							
Nb	205	141	101	70	48	28	19
Cor	0.25	0.10	0.04	-0.07	-0.08	-0.06	-0.07
Cormoy	0.25	0.10	0.05	-0.07	-0.06	-0.04	-0.05
Corstd	0.12	0.07	0.07	0.06	0.09	0.16	0.20
Corinf	-0.06	-0.11	-0.20	-0.27	-0.24	-0.31	-0.44
Corsup	0.60	0.30	0.32	0.42	0.55	0.91	0.78
Cormed	0.26	0.03	0.05	-0.08	-0.08	-0.07	-0.09
q0025	*0.01*	-0.05	-0.11	-0.17	-0.17	-0.26	-0.33
q00925	*0.49*	0.19	0.19	0.12	0.22	0.36	0.49
Estimation du coefficient de corrélation entre tailles de dépassement							
Nb	115	80	57	38	19	6	1
Cor	0.29	0.11	0.11	-0.07	-0.37	-0.61	////
Cormoy	0.29	0.11	0.12	-0.07	-0.35		////
Corstd	0.07	0.09	0.11	0.14	0.17		////
Corinf	0.06	-0.18	-0.16	-0.48	-0.84		////
Corsup	0.52	0.39	0.47	0.49	0.42		////
Cormed	0.29	0.11	0.12	-0.07	-0.37		////
q0025	*0.14*	-0.06	-0.08	-0.31	-0.66		////
q00925	*0.43*	0.30	0.35	0.24	0.08		////

Tableau 9.2: Seuil u pour Champs/Marne (les notations des symboles sont indiquées en annexe B.4)

9.1.3 Aubervilliers

Seuil en μg/m3	90	100	110	120	130	140	150
Estimation du coefficient de corrélation entre intervalles de temps successifs							
Nb	258	212	155	111	74	54	41
Cor	0.06	0.12	0.008	0.08	-0.06	0.04	-0.01
Cormoy	0.09	0.13	0.007	0.08	-0.06	0.08	0.01
Corstd	0.08	0.08	0.04	0.08	0.04	0.12	0.13
Corinf	-0.05	-0.06	-0.10	-0.17	-0.19	-0.15	-0.33
Corsup	0.42	0.48	0.23	0.48	0.07	0.68	0.70
Cormed	0.07	0.12	0.005	0.07	-0.06	0.06	-0.02
q0025	-0.02	*0.004*	-0.06	-0.06	-0.14	-0.09	-0.018
q00925	0.30	*0.31*	0.26	0.26	0.014	0.09	0.34
Estimation du coefficient de corrélation entre tailles de dépassement							
Nb	166	129	92	59	38	23	15
Cor	0.33	0.23	0.18	0.03	-0.06	-0.004	-0.09
Cormoy	0.33	0.24	0.18	0.03	-0.04	-0.002	-0.08
Corstd	0.07	0.08	0.10	0.12	0.15	0.18	0.23
Corinf	0.11	-0.02	-0.12	-0.39	-0.53	-0.61	-0.69
Corsup	0.53	0.47	0.45	0.48	0.49	0.66	0.84
Cormed	0.33	0.23	0.18	0.03	-0.04	0.001	-0.12
q0025	0.19	0.09	-0.02	-0.20	-0.33	-0.39	-0.45
q0975	0.46	0.39	0.37	0.26	0.26	0.35	0.39

Tableau 9.3: Seuil u pour Aubervilliers (les notations des symboles sont indiquées en annexe B.4)

L'intervalle de confiance à 95 % de l'estimation du coefficient de corrélation entre tailles de dépassement ayant eu lieu des jours consécutifs pour le seuil 90 μgm^{-3} ne comprend pas la valeur zéro (cf. tableau 9.3). De même, Les intervalles de confiance à 95 % des estimations des coefficients de corrélation correspondant au seuil 100 μgm^{-3} ne comprennent pas la valeur zéro. Ces deux seuils ne peuvent donc pas être utilisés pour modéliser le processus des dépassements. Pour la station d'Aubervilliers les seuil u raisonnables sont $\{110, 120, 130, 140, 150\}$.

Seuil en $\mu g/m3$	90	100	110	120	130	140	150
Estimation du coefficient de corrélation entre intervalles de temps successifs							
Nb	252	205	152	108	75	45	33
Cor	0.05	0.04	0.01	0.22	0.03	0.01	0.06
Cormoy	0.06	0.05	0.01	0.21	0.04	0.005	0.09
Corstd	0.04	0.04	0.04	0.19	0.08	0.12	0.15
Corinf	-0.03	-0.04	-0.12	-0.09	-0.14	-0.29	-0.27
Corsup	0.23	0.24	0.32	0.71	0.39	0.71	0.80
Cormed	0.05	0.05	0.01	0.22	0.03	-0.01	0.07
q0025	-0.01	-0.02	-0.05	-0.06	-0.09	-0.17	-0.15
q0975	0.16	0.15	0.09	0.58	0.24	0.34	0.44
Estimation du coefficient de corrélation entre tailles de dépassement							
Nb	156	123	84	46	27	16	12
Cor	0.27	0.18	0.13	0.07	-0.09	-0.10	-0.30
Cormoy	0.27	0.18	0.14	0.08	-0.05	-0.10	-0.34
Corstd	0.06	0.07	0.09	0.13	0.20	0.22	0.14
Corinf	0.07	-0.10	-0.14	-0.27	-0.59	-0.74	-0.87
Corsup	0.46	0.43	0.47	0.52	0.64	0.76	0.42
Cormed	0.27	0.18	0.14	0.08	-0.07	-0.11	-0.33
q0025	*0.14*	*0.04*	-0.02	-0.16	-0.43	-0.56	*-0.69*
q0975	*0.39*	*0.32*	0.32	0.38	0.41	0.42	*-0.11*

Tableau 9.4: Seuil u pour Créteil (les notations des symboles sont indiquées en annexe B.4)

9.1.4 Créteil

Les intervalles de confiance à 95 % des estimations du coefficient de corrélation entre tailles de dépassement ayant eu lieu des jours consécutifs pour les seuils 90, 100 et 150 μgm^{-3} ne comprennent pas la valeur zéro (cf. tableau 9.4). Pour la station de Créteil, les seuil u raisonnables sont donc $\{110, 120, 130, 140\}$.

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 **110, 120, 130 et 140** μgm^{-3}.

9.2 Modélisation sans interaction

9.2.1 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 $u = 130\ \mu g/m3$ 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 $u = 120\ \mu g/m3$ et $u = 130\ \mu g/m3$ utilise les mêmes variables météorologiques quelle que soit 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** confirment 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.

	Modèles					
Station-seuil	**NEU120**	**NEU130**	**CS120**	**CS130**	**AUB130**	**CRE130**
Cte	-11.755	-9.752	-18.966	-18.792	-23.098	-11.533
Ecart-type	1.239	1.537	2.283	2.655	2.828	1.463
s	*(0.152)*	0.376	0.715	0.800	*(-0.211)*	0.150
Ecart-type	*(0.104)*	*0.128*	*0.124*	*0.158*	*(0.195)*	*0.075*
Odd-ratio sur 10 ans	///	**43**	**1274**	**2981**	///	**4.5**
int. de conf. 95 %						
du Odd-ratio sur 10 ans	///	[3.5;521.6]	[112.1;14478.2]	[134.7;65959.8]	///	[1.0;19.5]
t92		-2.378				
Ecart-type		*0.681*				
t93					2.418	
Ecart-type					*0.522*	
tmax	0.447	0.452	0.489	0.503	0.659	0.380
Ecart-type	*0.044*	*0.054*	*0.064*	*0.076*	*0.082*	*0.048*
trange						
Ecart-type						
wsavg	-0.508	-0.889	-0.518	-1.086	-0.461	-0.574
Ecart-type	*0.131*	*0.199*	*0.186*	*0.285*	*0.202*	*0.173*
wsrange						
Ecart-type						
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 $\alpha = 0.01$

Tableau 9.5: Synthèse de la modélisation de la fréquence des dépassements

- 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 $u = 130\ \mu g/m3$ 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 (cf. annexe B.5) 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.

9.2.2 Taille des dépassements

Cette modélisation permet d'observer des différences notables entre les stations. Pour le seuil $u = 130\ \mu g/m3$:

- 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 $u = 120\ \mu g/m3$ et(ou) $u = 130\ \mu g/m3$ 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 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.

	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	0.0692				0.141	0.184
Ecart-type	*0.0297*				*0.0623*	*0.0780*
s		-0.006				
Ecart-type		*0.0030*				
t92		0.0358				
Ecart-type		*0.0145*				
t93						
Ecart-type						
tmax	-0.0024	-0.0012			-0.0044	-0.0063
Ecart-type	*0.0009*	*0.0004*			*0.0018*	*0.0024*
trange			0.0032	0.0036		
Ecart-type			*0.0004*	*0.0005*		
wsavg	0.0137	0.0188			0.0143	0.0237
Ecart-type	0.0039	*0.0046*			*0.0064*	*0.0074*
wsrange						
Ecart-type						
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 $\alpha = 0.05$

Tableau 9.6: Synthèse de la modélisation de la taille des dépassements

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.

9.3 Modélisation avec interaction : fréquence des dépassements

Le paramètre $\alpha(t)$ prend donc la forme (8.2) définie dans la section 8.2. 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 (**wsavg**) 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 vue 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.

9.3.1 Neuilly/Seine

	Modèles				
Station-seuil	**NEU110**	**NEU120**	**NEU130**	**NEU140**	**NEU150**
Nbre de dep.	125	85	59	41	29
Cte	1.629	1.229	0.606	0.437	1.037
Ecart-type	*0.478*	*0.568*	*0.620*	*0.732*	*0.864*
s	-2.672	-3.036	-2.604	-2.475	-2.840
Ecart-type	*0.288*	*0.377*	*0.371*	*0.414*	*0.534*
wsavg	-0.825	-0.879	-0.813	-1.007	-1.225
Ecart-type	*0.137*	*0.173*	*0.190*	*0.244*	*0.304*
ttmax	0.099	0.110	0.093	0.089	0.098
Ecart-type	*0.010*	*0.013*	*0.012*	*0.013*	*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 $\alpha = 0.01$

Tableau 9.7: Synthèse de la modélisation de la fréquence des dépassements 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 9.7) permet de constater que les seuils 120 et 130 μgm^{-3} apparaissent les mieux adaptés.

Seuil	$\hat{\psi}(s = 1, s = 10, tmax)$	**int. de conf. 95 %**	**tmax**
110	1	[0.30; 3.36]	26.9
120	1	[0.32; 3.16]	27.5
130	1	[0.28; 3.57]	28.0
140	1	[0.21; 4.81]	27.7
150	1	[0.17; 5.89]	29.1

Tableau 9.8: Température maximale correspondant à un rapport de chances estimé à 1 pour Neuilly/Seine

La température critique (cf. tableau 9.8) est donc 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 $\hat{\psi} = 1$ est beaucoup plus raisonnable que dans le cadre de la modélisation sans interaction.

9.3.2 Champs/Marne

	Modèles				
Station-seuil	CSM110	CSM120	CSM130	CSM140	CSM150
Nbre de dep.	35	38	28	19	13
Cte	-3.194	-4.190	-2.708	-2.294	-3.956
Ecart-type	*0.692*	*0.908*	*1.170*	*1.197*	*1.517*
s	-1.835	-1.430	-1.789	-1.529	-1.195
Ecart-type	*0.309*	*0.329*	*0.466*	*0.471*	*0.510*
wsavg	-0.791	-0.881	-2.021	-1.685	-1.530
Ecart-type	*0.194*	*0.245*	*0.462*	*0.433*	*0.466*
ttmax	0.089	0.076	0.094	0.075	0.067
Ecart-type	*0.011*	*0.011*	*0.017*	*0.016*	*0.016*
Test de Kolmogorov-Smirnov					
Nbre d'obs.	80	55	40	26	17
D_n	0.177	0.136	0.116	0.206	0.113
$d_{n,0.05}$	0.152	0.183	0.215		
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 $\alpha =$	H_0 acceptée pour $\alpha =$

Tableau 9.9: Synthèse de la modélisation de la fréquence des dépassements pour Champs/Marne

La phase de validation du modèle (test de Kolmogorov-Smirnov dans le tableau 9.9) permet de constater que les seuils 120 et 130 μgm^{-3} apparaissent les mieux adaptés.

Seuil	$\hat{\psi}(s = 1, s = 10, tmax)$	int. de conf. 95 %	tmax
110	1	[0.12; 8.44]	20.7
120	1	[0.06; 17.89]	18.7
130	1	[0.02; 44.37]	19.1
140	1	[0.02; 43.55]	20.3
150	1	[0.007; 143.44]	17.8

Tableau 9.10: Température maximale correspondant à un rapport de chances estimé à 1 pour Champs/Marne

La température critique (cf. tableau 9.10) est donc comprise entre 17.8 et 20.7 degrés. Cette température basse indique que le risque d'observer un dépassement sur la période d'étude a augmenté pour la majorité des jours de pollution (photochimique ou importée).

9.3.3 Aubervilliers

	Modèles				
Station-seuil	**AUB110**	**AUB120**	**AUB130**	**AUB140**	**AUB150**
Nbre de dep.	100	70	44	31	22
Cte	-1.646	-1.671	-1.306	-1.445	-1.571
Ecart-type	*0.517*	*0.648*	*0.800*	*0.936*	*1.136*
s	-1.815	-2.221	-2.562	-2.727	-2.610
Ecart-type	*0.224*	*0.313*	*0.415*	*0.499*	*0.566*
wsavg	-0.484	-0.767	-0.844	-0.847	-1.293
Ecart-type	*0.140*	*0.198*	*0.255*	*0.304*	*0.424*
ttmax	0.081	0.096	0.101	0.104	0.104
Ecart-type	*0.008*	*0.011*	*0.014*	*0.016*	*0.018*
Test de Kolmogorov-Smirnov					
Nbre d'obs.	121	87	60	46	35
D_n	0.149	0.144	0.206	0.205	0.229
$d_{n,0.05}$	0.121	0.146	0.176	0.201	0.228
test K-S	H_0 refusée	H_0 acceptée pour $\alpha = 0.05$	H_0 acceptée pour $\alpha = 0.01$	H_0 acceptée pour $\alpha = 0.01$	H_0 acceptée pour $\alpha = 0.01$

Tableau 9.11: Synthèse de la modélisation de la fréquence des dépassements pour Aubervilliers

La phase de validation du modèle (test de Kolmogorov-Smirnov dans le tableau 9.11) permet de rejeter le seuil 110 μgm^{-3} et de constater que le seuil 120 apparaît le mieux adapté.

Seuil	$\hat{\psi}(s = 1, s = 10, tmax)$	**int. de conf. 95 %**	**tmax**
120	1	[0.17; 5.79]	23.2
130	1	[0.13; 7.98]	25.3
140	1	[0.09; 11.10]	26.3
150	1	[0.04; 24.13]	25.19

Tableau 9.12: Température maximale correspondant à un rapport de chances estimé à 1 pour Aubervilliers

La température critique (cf. tableau 9.12) est donc comprise entre 23.2 et 26.3 degrés.

9.3.4 Créteil

	Modèles				
Station-seuil	**CRE110**	**CRE120**	**CRE130**	**CRE140**	**CRE150**
Nbre de dep.	92	58	38	24	18
Cte	-0.508	-0.566	-0.208	-0.423	-1.098
Ecart-type	*0.505*	*0.659*	*0.834*	*1.093*	*1.105*
s	-1.767	-2.006	-1.846	-1.631	-1.661
Ecart-type	*0.226*	*0.312*	*0.365*	*0.443*	*0.478*
wsavg	-0.638	-0.940	-1.474	-2.212	-1.467
Ecart-type	*0.143*	*0.211*	*0.317*	*0.487*	*0.406*
ttmax	0.074	0.083	0.078	0.078	0.071
Ecart-type	*0.008*	*0.010*	*0.012*	*0.015*	*0.015*
Test de Kolmogorov-Smirnov					
Nbre d'obs.	126	90	64	37	28
D_n	0.164	0.177	0.213	0.194	0.234
$d_{n,0.05}$	0.121	0.143	0.169	0.222	
test K-S	H_0 rejetée	H_0 rejetée	H_0 rejetée	H_0 acceptée pour $\alpha = 0.01$	H_0 acceptée pour $\alpha =$

Tableau 9.13: Synthèse de la modélisation de la fréquence des dépassements pour Créteil

La phase de validation du modèle (test de Kolmogorov-Smirnov dans le tableau 9.13) permet de rejeter les seuils 110, 120 et 130 μgm^{-3} et de ne retenir que le seuil $140\mu gm^{-3}$.

Seuil	$\hat{\psi}(s=1, s=10, tmax)$	int. de conf. 95 %	tmax
140	1	[0.03; 30.48]	21.0

Tableau 9.14: Température maximale correspondant à un rapport de chances estimé à 1 pour Créteil

La température critique (cf. tableau 9.14) pour le seuil 140 μgm^{-3} vaut 21 degrés.

9.4 Contours de vraisemblance ([39])

Afin de contrôler l'exactitude des résultats de notre programme pour estimer les paramètres β (tailles des dépassements), nous avons minimisé directement la 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 $(0.0692, -0.0024, 0.0137)$. Nous avons ensuite tracé les contours de la vraisemblance pour les paramètres β_3 (Tmax) et β_5 (WSAVG), en fixant la valeur β_1 (constante) à la valeur estimée 0.0760. Le résultat (cf. figure 9.1) paraît satisfaisant ; et il traduit bien l'estimation négative (-0.38832 annexe B.7) du coefficient de corrélation entre les paramètres.

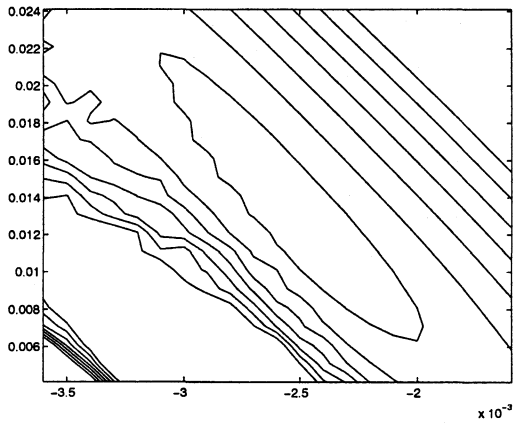


Figure 9.1: Contours de la vraisemblance pour les données de NEU120. Les niveaux des contours sont 420(5)450.

Chapitre 10

Résultats obtenus pour la Région de Los Angeles

Dans ce chapitre, nous allons essayer d'appliquer la modélisation utilisée pour l'Ile-de-France à des données fort aimablement fournies par Joe Cassmassi, Senior Meteorologist in the South Coast Air Quality Management District El Monte, Californie (cf. [58]),; elles proviennent de deux stations, Azusa et Long Beach, près de Los Angeles (Californie).

10.1 Choix d'un seuil u raisonnable

10.1.1 Azusa

Seuil en μg/m3	300	400	500
Estimation du coefficient de corrélation entre intervalles de temps successifs			
Nb	817	336	101
Cor	-0.003	0.12	0.20
Cormoy	-0.002	0.12	0.20
Corstd	0.01	0.10	0.17
Corinf	-0.03	-0.04	-0.10
Corsup	0.05	0.48	0.63
Cormed	-0.003	0.11	0.19
q0025	-0.03	-0.04	-0.10
q0975	0.02	0.33	0.52
Estimation du coefficient de corrélation entre tailles de dépassement			
Nb	559	192	41
Cor	0.43	0.36	0.03
Cormoy	0.43	0.36	0.02
Corstd	0.04	0.07	0.16
Corinf	0.27	0.15	-0.41
Corsup	0.56	0.54	0.56
Cormed	0.43	0.37	0.02
q0025	*0.35*	*0.21*	-0.29
q0975	*0.51*	*0.48*	0.35

Tableau 10.1: Seuil u pour Azusa (Los Angeles)

Seul 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 500 μgm^{-3} sur le site d'Azusa comprend la valeur zéro (cf. tableau 10.1). Par conséquent, pour cette station de mesure le seul seuil u raisonnable parmi les seuils étudiés est 500 μgm^{-3}.
Cependant, ce seuil est problématique puisque aucun jour depuis 1992 n'a dépassé ce niveau de pollution ! Azusa présente en effet une très forte tendance à la baisse dans les valeurs d'ozone enregistrées (cf. annexe B.9.2 graphe B.13) entre 1981 et 1996. L'existence de cette très forte

tendance induit une auto-corrélation entre dépassements du seuil 500 μgm^{-3} sur la première partie de la période 1981-1992 (cf. annexe B.9.2: jours de dépassement du seuil 500 μgm^{-3}) et remet donc fortement en cause la validité des hypothèses d'indépendance des dépassements. Il semble alors peu raisonnable d'utiliser un PPNH pour modéliser les dépassements de seuil 500 μgm^{-3} enregistrés à Azusa.

10.1.2 Long Beach

Seuil en μg/m3	160	180	200
Estimation du coefficient de corrélation entre intervalles de temps successifs			
Nb	263	171	118
Cor	0.03	0.08	0.09
Cormoy	0.03	0.08	0.09
Corstd	0.04	0.10	0.10
Corinf	-0.06	-0.10	-0.13
Corsup	0.16	0.41	0.42
Cormed	0.03	0.06	0.08
q0025	-0.03	-0.07	-0.06
q0975	0.10	0.28	0.30
Estimation du coefficient de corrélation entre tailles de dépassement			
Nb	108	57	35
Cor	0.26	0.11	0.04
Cormoy	0.26	0.11	0.05
Corstd	0.10	0.15	0.20
Corinf	-0.008	-0.26	-0.41
Corsup	0.55	0.57	0.68
Cormed	0.26	0.11	0.04
q0025	*0.08*	- 0.14	-0.31
q0975	*0.47*	0.42	0.45

Tableau 10.2: Seuil u pour Long Beach (Los Angeles)

Pour la station de Long Beach, parmi les seuils étudiés, les seuil u raisonnables correspondent à 180 et 200 μgm^{-3}.

10.2 Modélisation de la fréquence des dépassements ayant eu lieu sur le site de Long Beach

Les covariables météorologiques sont décrites en annexe.

	Modèles	
Seuil	**180**	**200**
Cte	-5.853	-7.871
Ecart-type	1.656	1.935
s	-0.367	-0.416
Ecart-type	*0.049*	*0.058*
Odd-ratio sur 15 ans	**1/245.18**	**1/512.85**
int. de conf. 95 % du Odd-ratio sur 15 ans	[1/1025.57; 1/58.63]	[1/2661.38; 1/98.83]
PRE5	-0.552	-0.646
Ecart-type	0.110	0.128
PRE6	-0.321	-0.327
Ecart-type	*0.085*	*0.087*
t2	-0.310	-0.398
Ecart-type	*0.105*	*0.123*
t4	0.631	0.745
Ecart-type	0.077	0.105
Test de Kolmogorov-Smirnov		
D_n	0.529	0.525
$d_{n,0.05}$	0.104	0.126
test K-S	H_0 rejetée	H_0 rejetée

Tableau 10.3: Synthèse de la modélisation de la fréquence des dépassements

Il n'est donc pas possible d'utiliser ce modèle pour modéliser les dépassements de seuils 180 et 200. De plus si on prend un seuil plus grand, 240 par exemple, on rencontre alors les mêmes problèmes que sur le site d'Azusa : beaucoup de dépassements avant 1992, puis ensuite un unique en 1994 !

Nous constatons donc que l'application du modèle, valable en Ile-de-France, à deux stations californiennes est impossible dans la mesure où les hypothèses à la base du modèle n'y sont pas respectées. De plus la baisse importante, notée après 1992, des dépassements est troublante. Il nous est difficile d'en donner une justification rigoureuse ; les mesures anti-pollution prises en Californie en sont vraisemblablement la cause. Mais, nous voudrions être sûre qu'aucun élément extérieur ne puisse aussi la justifier. Ainsi, nous ne savons rien sur la qualité métrologique des mesures qui ont été faites. A-t-on modifié les capteurs? Peut-on imaginer qu'une telle modification, si elle a été faite, introduise un artefact?

Conclusion

Au terme de ce travail, nous voulons faire un certain nombre de remarques.
Tout d'abord, la méthode que nous avons présentée est voisine de celle utilisée pour la ville de Chicago, publiée très récemment par Davis et *al.* dans [17].
La modélisation de la taille du dépassement du seuil u le jour t sachant qu'un dépassement de seuil u a eu lieu par une loi exponentielle de paramètre $\beta(t)$ est très lourde. En effet, les variables non significatives sont supprimées pas à pas ; le risque de conserver un maximum local au lieu de l'estimateur du maximum de vraisemblance est réel. Chaque étape est donc une source d'erreur potentielle. Par conséquent, il est indispensable de vérifier en traçant les contours de la log-vraisemblance que les estimations des paramètres retenus sont acceptables.
L'utilisation d'un modèle de régression logistique pour modéliser les dépassements d'ozone de seuil u restreint le temps de programmation à l'utilisation de la procédure LOGISTIC de SAS et simplifie l'analyse des coefficients correspondant aux variables significatives. Elle permet ainsi d'essayer de quantifier l'augmentation du risque d'observer un dépassement entre 1988 et 1997. Dans un premier temps les résultats de la modélisation sans interaction nous ont permis d'estimer une tendance. Mais les intervalles de confiance à 95 % de l'estimation du rapport des chances des différents sites étudiés étaient démesurés et traduisaient la très grande imprécision des estimations obtenues. Dans un second temps, la prise en compte de l'interaction entre l'année et la température maximale nous a permis d'obtenir des intervalles de confiance à 95 % des rapports des chances estimés à 1 pour chaque station plus précis et de mettre en évidence grâce à la valeur de la température critique des différences assez remarquables entre la station de Champs/Marne et les trois autres stations étudiées. Cependant, l'introduction de cette interaction engendre une nouvelle difficulté : l'impossibilité de pouvoir quantifier l'augmentation du risque d'observer un dépassement entre 1988 et 1997 sans faire intervenir une valeur donnée de la température maximale.
Les tendances à moyen terme (10 ans) dans les valeurs élevées d'ozone en région parisienne sont complexes, fortement liées à la température, mais aussi à des phénomènes plus difficiles à prendre en compte comme le changement des appareils de mesure, le changement dans le suivi et l'entretien des analyseurs, l'évolution de la fréquentation des axes routiers, l'augmentation du parc automobile et le nombre très important de données manquantes avant 1992.
Cependant, le traitement des données d'ozone troposphériques provenant de la région de Los Angeles nous a conduit à constater la difficulté d'appliquer cette méthodologie pour détecter des tendances à long terme (16 ans). En effet, les sites de mesures d'ozone d'Asuza et de Long Beach enregistrent une tendance forte à la baisse dans les dépassements de niveau élevé sur la période 1981-1996. Par conséquent, celle-ci induit une auto-corrélation entre ces dépassements et ne permet pas de valider les hypothèses sur lesquelles est construit le modèle.
Dans ce type de modélisation (modélisation à effets fixes), les hypothèses permettant d'aboutir à une telle approximation sont toutes basées sur les propriétés statistiques du processus de l'ozone (non-stationnarité, regroupement des très hautes valeurs..), les covariables météorologiques sont supposées fixées. Or les phénomènes de formation de l'ozone sont fortement liés aux conditions météorologiques, dont les propriétés statistiques sont mieux connues que celles régissant l'ozone.

Nous présentons donc dans la dernière partie de ce travail un résultat théorique sur la convergence en distribution du processus des dépassements de seuil u_n vers un processus de Poisson Composé, dans lequel les hypothèses ne sont pas basées sur le processus dont on étudie les dépassements de seuil.

Partie IV

Théorèmes Limite vers les Processus de Poisson Composé (CPLT)

Chapitre 11

Quelques rappels concernant les processus ponctuels

Toutes les références de la partie IV ont été regroupées à la fin de celle-ci.

11.1 Généralités

Une fonction aléatoire réelle est une application de Ω dans l'espace des fonctions réelles : à tout $\omega \in \Omega$ correspond une fonction d'une variable réelle t :

$$\omega \rightarrow X(t;\omega) = X_t(\omega)$$

En général l'argument t représente le temps et X s'appelle alors *processus aléatoire* ou *processus stochastique*.Soit T l'ensemble des temps t, un processus est donc une application X de $\Omega \times T$ dans $\mathbb{R}$, de valeurs $x(t;\omega)$.
Si ω est fixé $X(.,\omega)$ définit une fonction du temps appelée *trajectoire* de ω. Un processus peut donc être assimilé à l'ensemble de ses trajectoires lorsque ω décrit Ω. Si t est fixe $X(t;.)$ définit une variable aléatoire réelle X_t. Un processus peut donc être considéré comme une famille (infinie) de variables aléatoires indicées par le temps (X_t). C'est cette notation que nous utiliserons.
Si la variable aléatoire X_t admet pour tout t une espérance m_t et une variance σ_t^2 finies, celles-ci définissent des fonctions certaines du temps. La structure des dépendances temporelles définit l'essentielle de la régularité des trajectoires. On dit de plus, qu'un processus est *stationnaire au sens strict* si sa loi de probabilité est invariante par translation sur t : (X_t) et $(X_{t+\tau})$ ont les mêmes caractéristiques. La stationnarité au sens strict implique $m_t = m$ et $\sigma_t^2 = \sigma^2$. Lorsque l'on a uniquement les conditions $m_t = m$ et $\sigma_t^2 = \sigma^2$, on dit que le processus (X_t) est *stationnaire au sens large*. Dans toute la suite pour simplifier, nous utiliserons le terme stationnarité pour désigner la stationnaire au sens strict.
Un processus tel que $X_{t+h} - X_t$ soit stationnaire pour tout h est dit *à accroissements stationnaires.*

11.2 Définitions et propriétés des processus ponctuels

Si I est un intervalle fini sur $\mathbb{R}$, le nombre d'événements, $N(I)$, d'un processus ponctuel se produisant dans I doit être une variable aléatoire à valeurs entières. Plus généralement, pour tout ensemble Borélien borné B, $N(B)$ doit être une variable aléatoire à valeurs entières. De plus, le nombre d'événements dans la réunion finie ou dénombrable d'ensembles disjoints est la somme des nombres dans chaque ensemble, i.e. $N(B) = \sum_1^\infty N(B_i)$ si les B_i sont des ensembles Boréliens disjoints dont la réunion est B. Par conséquent, $N(.)$ est une mesure sur les ensembles Boréliens. La valeur de $N(B)$ doit être entière.

Pour un espace avec une structure suffisante (tels que $\mathbb{R}$ ou $\mathbb{R}^2$), on peut montrer que tout processus ponctuel N peut être représenté de la façon suivante :

$$N = \sum_j \beta_j 1_{\tau_j}$$

où

- τ_j sont des éléments aléatoires distincts
- 1_{τ_j} processus ponctuel où $\forall j,\ ; 1_{\tau_j}(B) = 1$ si $\tau \in B$ et $1_{\tau_j}(B) = 0$ sinon.
- β_j , variables aléatoires à valeurs entières, non-négatives. Dans le cas où les β_j sont tous égaux à 1, on dit que le processus ponctuel n'a pas d'évènements multiples ou qu'il est simple.

Si $B_1, \cdots, B_k$ sont des ensembles Boréliens bornés, $N(B_1), \cdots, N(B_k)$ sont des variables aléatoires et ont une distributions conjointe appelée une distribution fini-dimensionnelle du processus ponctuel. En fait, les propriétés probabilistes intéressantes du processus ponctuel sont spécifiées uniquement par la collection de telles distributions fini-dimensionnelles, i.e. pour tous les choix de k et des ensembles $B_1, \cdots, B_k$. Bien sûr pour définir un processus ponctuel en commençant par les distributions fini-dimensionnelles, il faut choisir ces distributions telles que la variable aléatoire $N(B)$ soit définie, non-négative, countably additive, etc. (cf. Kallenberg (1983) pour plus de détails).
Si N est un processus ponctuel, la mesure λ définie sur les ensembles Boréliens de l'espace considéré par

$$\lambda(B) = E\left(N\left(B\right)\right)$$

est appelée *mesure d'intensité* du processus ponctuel.
Il est intéressant de noter que les propriétés probabilistes d'un processus ponctuel N peuvent aussi être résumées par des fonctions génératrices variées telles que la transformée de Laplace. Pour des fonctions mesurables f non-négatives, la transformée de Laplace $L_N(f)$ est définie par :

$$L_N(f) = E\left[\exp\left(-\int f dN\right)\right] = E\left[\exp\left(-\sum \beta_j f(\tau_j)\right)\right]$$

où N est représenté par $\sum \beta_j \delta_{\tau_j}$. De telles fonctions génératrices possèdent des propriétés analogues à celles des fonctions caractéristiques et des transformées de Laplace de variables aléatoires. En particulier, si $f(x) = t\chi_B(x)$ (où $\chi_B(x) = 1$ si $x \in B$, 0 sinon), on a $L_N(f) = E\left[e^{-tN(B)}\right]$. C'est simplement la transformée de Laplace (ou la fonction génératrice évaluée en $-t$) de la variable aléatoire $N(B)$, et elle spécifie de façon unique la distribution de $N(B)$. De même, les transformées de Laplace conjointes des variables aléatoire $N(B_1), \cdots, N(B_k)$ peuvent être écrites en prenant $f = \sum_{i=1}^{k} t_i \chi_{B_i}$.

11.3 Les processus de Poisson

11.3.1 Le processus de Poisson simple

Supposons que la variable aléatoire X suive une loi exponentielle de paramètre λ :

$$P\left[X > x\right] = e^{-\lambda x},\ ; x \geq 0 \tag{11.1}$$

et donc

$$P\left[X > x + y | X > x\right] = P\left[X > y\right],\ ; x, y \geq 0 \tag{11.2}$$

, par définition des probabilités conditionnelles. X peut par exemple représenter le temps d'attente entre l'apparition d'événements. On démontre que la condition (11.2) implique que X suit la distribution définie par (11.1) pour un certain λ positif. Par conséquent, s'il n'y a pas d'effet après dans le mécanisme des temps d'attente, au sens de (11.2), alors la variable aléatoire X temps

d'attente suit nécessairement une loi exponentielle.
Soient X_1 le temps d'attente jusqu'au premier événement, X_2 le temps d'attente entre le premier et le second événement, ainsi de suite. Le modèle formel est donc formé d'une suite infinie de variables aléatoires $X_1, X_2, \cdots$ définies sur un certain espace probabilisé, et $S_n = X_1 + \cdots + X_n$ représente la date d'apparition du $n^{\text{ième}}$ événement ; par convention $S_0 = 0$. S'il n'y a pas d'événement multiple, S_n doit être strictement croissante et si seulement si un nombre fini d'événements se produisent dans chaque intervalle de temps fini, S_n doit tendre vers l'infini :

$$0 = S_0 < S_1 < S_2 < \cdots, \ ; \sup_n S_n = \infty. \tag{11.3}$$

Cette condition est équivalente à

$$X_1 > 0, X_2 > 0, \cdots, \ ; \sum_n X_n = \infty. \tag{11.4}$$

On suppose donc la condition :
Condition 0 :
Les événements définis par (11.3) et (11.4) admettent la probabilité 1.
Le nombre N_t d'événements se produisant dans l'intervalle de temps $[0, t]$ est le plus grand entier n tel que $S_n \leq t$:

$$N_t = \max_n \{S_n \leq t\} \tag{11.5}$$

Sur l'ensemble des ω définit par (11.3), on définit N_t par (11.5) ; et partout ailleurs $N_t = 0$. Notons que $N_t = 0$ si $t < S_1 = X_1$; en particulier, $N_0 = 0$. Le nombre d'événements dans $]s, t]$ est égal à $N_t - N_s$.
(11.5) conduit à la relation liant N_t et S_n suivante :

$$[N_t \geq n] = [S_n \leq t]\,. \tag{11.6}$$

Donc,

$$[N_t = n] = [S_n \leq t < S_{n+1}] \tag{11.7}$$

Chaque N_t est donc une variable aléatoire.
On dit qu'un processus temporel est un processus de Poisson si ce processus représente l'apparition d'événements aléatoires $E_1, E_2, \cdots$ satisfaisant aux trois conditions suivantes :

- deux événements ne peuvent arriver simultanément,
- la loi du nombre d'événements arrivant dans l'intervalle $[t; t + T]$ ne dépend que de T,
- les temps d'attente $X_1, X_2, \cdots$ sont des variables aléatoires indépendantes (processus sans mémoire)

On démontre alors (cf. Billingsley 1979) que si la condition 0 est satisfaite, un processus de Poisson peut être caractérisé par l'une des trois conditions équivalentes suivantes :
Condition 1 :
Les variables aléatoires X_n sont indépendantes et chacune suit une distribution exponentielle de paramètre λ.

Dans ce cas $P[X_n > 0] = 1$ pour chaque n, la loi forte des grands nombres donne $n^{-1}S_n \to \lambda^{-1}$, et donc la condition 0 est vraie. Sous la condition 1, en appliquant (11.6), $P[N_t \geq n] = \sum_{i=1}^{\infty} e^{-\lambda t}(\lambda t)^i / i!$ et

$$P[N_t = n] = e^{-\lambda t}\frac{(\lambda t)^n}{n!} \ ; n = 0, 1, \cdots \tag{11.8}$$

N_t possède la distribution de Poisson de paramètre λt.
Condition 2 :

- *(i) Pour* $0 < t_1 < \cdots < t_k$ *, les variables aléatoires* $N_{t_1}, N_{t_2} - N_{t_1}, \cdots, N_{t_k} - N_{t_{k-1}}$ *sont indépendantes.*
- *(ii)* $P[N_t - N_s = n] = e^{-(t-s)}\frac{(\lambda(t-s))^n}{n!}$, $; n = 0, 1, \cdots, 0 \leq s < t$; *avec* $N_0 = 0$ *avec la probabilité 1.*

Condition 3 :

- *(i) Pour* $0 < t_1 < \cdots < t_k$, *les variables aléatoires* $N_{t_1}, N_{t_2} - N_{t_1}, \cdots, N_{t_k} - N_{t_{k-1}}$ *sont indépendantes.*
- *(ii) La distribution de* $N_t - N_s$ *dépend seulement de la différence* $t - s$, *et* $N_0 = 0$ *avec la probabilité 1.*

Condition 4 :
Si $0 < t_1 < \cdots < t_k$ *et si* $n_1, \cdots, n_k$ *sont des entiers non-négatifs, alors*

$$P\left[N_{t_k+h} - N_{t_k} = 1 | N_{t_j} = n_j, j \leq k\right] = \lambda h + o(h)$$

et

$$P\left[N_{t_k+h} - N_{t_k} \geq 2 | N_{t_j} = n_j, j \leq k\right] = o(h)$$

quand h tend vers 0.
De plus,

$$\lim_{s \to t} P[N_s \neq N_t] = 0, \text{ et } N_0 = 0 \text{ avec la probabilité } 1.$$

11.3.2 Processus de Poisson Composé

Soit $\{Y_i\}_{i=1}^{\infty}$ une suite de variables iid ayant pour distribution commune F et de fonction caractéristique φ. Soit $\{N(t), t \geq 0\}$ un processus de Poisson de paramètre $\lambda > 0$, indépendant de la suite $\{Y_i\}$, le processus stochastique

$$X(t) = \sum_{k=1}^{N(t)} Y_k, \; ; t \geq 0$$

est appelé *processus de Poisson composé.* Les processus de Poisson Composé sont utilisés pour modéliser une large variété de situations physiques.
Exemples :

- (a) Soit $N(t)$ le nombre de déclarations de sinistres contre une compagnie d'assurance jusqu'au temps t et supposons que Y_i soit le montant du ième sinistre, alors $X(t)$ est le montant cumulé des demandes de remboursement jusqu'au temps t.
- (b) Soit $N(t)$ le nombre de chocs apparus dans un système jusqu'au temps t, supposons que Y_i soit le dommage causé par le ième choc. En postulant que le dommage est additif, $X(t)$ est le dommage cumulé par le système jusqu'au temps t.

Un processus de Poisson Composé possède des accroissements stationnaires indépendants.
Puisque la distribution conjointe d'ensembles finis arbitraires $(X(t_1), \cdots, X(t_n))$ caractérise complètement le processus stochastique, on a établi le théorème suivant :

Théorème 2 *Soit* $\{X(t), t \geq 0\}$ *un processus stochastique ayant des accroissements stationnaires, indépendants et pour leque! X(0) = 0. ALORS*
$\{X(t), t \geq 0\}$ *est un processus de Poisson composé*
si et seulement si
la fonction caractéristique $\varphi_{X(t)}$ *de* $X(t)$ *est de la forme*

$$\varphi_{X(t)}(u) = \exp\{-\lambda t\,[1 - \varphi(u)]\}, \; ; -\infty < u < +\infty$$

où $\lambda > 0$ *et* φ *est une fonction caractéristique.*

Théorème 3 *Soit* $X(t) = \sum_{k=1}^{N(t)} Y_k, t \geq 0$ *un processus de Poisson Composé, où* $\{N(t), t \geq 0\}$ *est un processus de Poisson de paramètre* λ *et* $\{Y_k\}$ *est une suite de variables aléatoires iid. Soit* $A_1, \cdots, A_m$ *une partition de l'espace des valeurs possibles pour* Y_k, *c'est-à-dire* $A_i \cap A_j = \emptyset$ *si* $i \neq j$ *et* $P[Y_k \in A_1 \cup \cdots \cup A_m] = 1$. *Posons*

$$Y_k^i = \begin{cases} Y_k & si\ Y_k \in A_i, \\ 0 & si\ Y_k \notin A_i, \end{cases}$$

pour $i = 1, \cdots, m$, *et définissons*

$$X_i(t) = \sum_{k=1}^{N(t)} Y_k^i,\ i = 1, \cdots, m.$$

Alors chaque $\{X_i(t), t \geq 0\}$ *est un processus de Poisson Composé et les processus*

$$\{X_1(t), t \geq 0\}, \{X_2(t), t \geq 0\}, \cdots, \{X_m(t), t \geq 0\}$$

sont mutuellement indépendants.

11.4 Convergence de processus ponctuels

11.4.1 Approche simple

Supposons que $\{N_n\}$ soit une suite de processus ponctuels sur un rectangle $S \subset R^n$ et N un processus ponctuel. On dit que N_n **converge en distribution** vers N si la suite du vecteur aléatoire $(N_n(B_1), \cdots, N_n(B_k))$ converge en distribution vers $(N(B_1), \cdots, N(B_k))$ quel que soit le choix de k, et des ensembles Boréliens bornés $B_j \subset S$ tels que $N(\partial B_i) = 0$ a. s., $i = 1, \cdots, k$. Un processus ponctuel peut être vu comme un élément aléatoire d'un certain espace métrique (dont les points sont des mesures) et la convergence en distribution de N_n vers N devient la convergence faible des distributions de N_n vers celles de N. La définition précédente est équivalente à cette vision plus générale. Le résultat souvent utilisé (Leadbetter et al. 1983) pour démontrer la convergence en distribution du processus ponctuel des dépassements vers un processus de Poisson est une application d'un résultat de Kallenberg (1983) :

Théorème 4 *Soient* $N, N_n, n \geq 1$*des processus ponctuels sur l'intervalle semi-fermé* $S \cup \mathbb{R}$. *Supposons que* N *soit simple. Supposons que :*

- *(a)* $E[N_n(]c,d])] \to E[N(]c,d])], \forall -\infty < c < d < \infty$ *tels que* $]c,d] \subset S$, *et*
- *(b)* $P[N_n(B) = 0] \to P[N(B) = 0], \forall B$, *de la forme* $\cup_{i=1}^k]c_j, d_j]$ *avec* $]c_j, d_j] \subset S$, *pour* $i = 1, \cdots, k; k = 1, 2, \cdots$.

Alors

$$N_n \xrightarrow{d} N.$$

La forme remarquable de ce résultat est que la convergence de la probabilité d'apparition d'aucun événement sur certains ensembles donnés est suffisante pour garantir la convergence des quantités $P[N_n(B) = r]$ et des probabilités conjointes correspondantes. La proposition suivante est une étape importante de la preuve de Kallenberg du théorème ci-dessus : (a) assure l'existence pour chaque suite d'entiers d'une sous-suite telle que N_n converge vers un certain processus ponctuel simple et la proposition ci-dessous montre alors que la variable aléatoire limite admet la même distribution que N, ce qui complète la preuve.

Proposition 1 *Supposons que* N *et* N' *soient des processus ponctuels simples sur un intervalle semi-fermé* $S \subset \mathbb{R}$ *et que* $P[N(B) = 0] = P[N'(B) = 0], \forall B$ *de la forme* $\bigcup_{i=1}^k]c_i, d_i]$ *avec* $[c_i, d_i] \subset S$ *pour* $i = 1, \cdots, k; k = 1, 2, \cdots$.
Alors N *et* N' *possèdent la même distribution.*

11.4.2 Approche plus générale

Celle-ci nécessite quelques rappels sur la convergence faible des mesures de probabilité sur les espaces métriques. On note $D([0,1])$, l'espace des fonctions $x = (x_t)$, $; t \in [0,1]$, *cadlag*, c'est-à-dire, continues à droite ($x_t = x_{t+}$, $; \forall t < 1$) et définies à gauche (pour tout $t > 0$). D([0,1]) est un espace metrique pour la métrique définie par Skorohod suivante :

$$d(x,y) = \inf\{\varepsilon > 0 : \exists \lambda \in \Delta : \sup_t |x_t - y_{\lambda(t)}| + \sup_t |t - \lambda(t)| \leq \varepsilon\}$$

où Δ est l'ensemble des fonctions strictement croissantes $\lambda = \lambda(t)$ qui sont continues sur $[0,1]$ et sont telles que $\lambda(0) = 0, \lambda(1) = 1$.
Soit T un sous-ensemble de $\mathbb{R}$. Un ensemble de variables aléatoires $X = (\xi_t)_{t \in T}$ est appelé *processus aléatoire de domaine de temps* T. Ici T=[0,1], X est donc appelé processus aléatoire à temps continu.

Définition 1 *Soit $X = (\xi_t)_{t \in [0,1]}$ un processus aléatoire.*

- *Pour tout $\omega \in \Omega$, la fonction $(\xi_t(\omega))_{t \in [0,1]}$ est appelée trajectoire du processus correspondant à la réalisation ω. On montre que tout élément aléatoire $X = \{\xi_t\}_{t \in [0,1]}$ sur $(D[0,1], \mathcal{B}_0(D([0,1])))$ peut être considéré comme un processus aléatoire dont les trajectoires appartiennent à l'espace des fonctions sans discontinuité de deuxième espèce.*

- *La mesure de probabilité P_X sur $(R^{[0,1]}, \mathcal{B}(R^{[0,1]}))$ définie par*

$$P_X(B) = P[\omega : X(\omega) \in B], B \in (R^{[0,1]})$$

  est appelée distribution de X.

- *Les probabilités*

$$P_{t_1,\cdots,t_n}(B) \equiv P[\omega : (\xi_{t_1}, \cdots, \xi_{t_n}) \in B]$$

  avec $t_1 < t_2 < \cdots < t_n, t_i \in [0,1]$ sont appelées les distributions fini-dimensionnelles.

On définit la convergence faible de la manière suivante :

Définition 2 *Sur l'espace métrique $(D([0,1]), \mathcal{B}_0(D([0,1])), d)$, soient $P, P_1, P_2, \cdots$ des mesures de probabilité sur $(D([0,1]), \mathcal{B}_0(D([0,1])), d)$.*

- *On dit que la suite de mesures de probabilité $\{P_n\}$ converge faiblement vers la mesure de probabilité P si*

$$\int_{D([0,1])} f(x) P_n(dx) \to \int_{D([0,1])} f(x) P(dx)$$

  pour toute fonction $f = f(x)$ continue, bornée de $D([0,1])$. On écrit : $P_n \overset{w}{\Rightarrow} P$.

- *Une suite de mesures de probabilité $\{P_n\}$ converge en général vers la mesure de probabilité P (notation : $P_n \Rightarrow P$) si*

$$P_n(A) \to P(A)$$

  pour tout ensemble $A \in \mathcal{B}_0(D([0,1]))$ pour lequel $P[\partial A] = 0$ ($\partial A = A \cap \bar{A}$)

Convergence faible ou convergence en loi

Compacité relative et tension des familles de distributions
Supposons que toutes les mesures soient définies sur l'espace métrique $(D([0,1]), \mathcal{B}_0(D([0,1])), d)$

Définition 3 • *Une famille de mesures de probabilité $\mathcal{P} = \{P_\alpha; \alpha \in \mathcal{U}\}$ est relativement compacte si toute suite de mesures de $\mathcal{P}$ contient une sous-suite qui converge faiblement vers une mesure de probabilité.*

- *La suite de variables aléatoires $\{X_n\}$ est relativement compacte (ou pré-compacte) si toute suite $\{X_{n_k}\}$ admet une sous-suite $\{X_{n_{k_m}}\}$ telle que $X_{n_{k_m}} \xrightarrow{d} X$.*

Dans cette définition, la mesure limite doit être une mesure de probabilité, mais elle peut ne pas appartenir à $\mathcal{P}$. Il est loin d'être facile de vérifier qu'une famille donnée de mesures de probabilité est relativement compacte en utilisant cette définition. Par conséquent, il est nécessaire d'avoir des critères simples pour tester cette propriété.

Définition 4 • *Une famille de mesures de probabilité $\mathcal{P} = \{P_\alpha; \alpha \in \mathcal{U}\}$ est tendue si, pour tout $\varepsilon > 0$, il existe un ensemble compact $K \subseteq D([0,1])$ tel que :*

$$\sup_{\alpha \in \mathcal{U}} P_\alpha [E \backslash K] \leq \varepsilon.$$

- *La suite de variables aléatoires définie sur un espace métrique séparable et complet $(E, \mathcal{E}, \rho)$ est tendue si : $\forall \varepsilon > 0$, $\exists$ un compact $K_\varepsilon \subset E$ tel que $P[X_n \in K_\varepsilon] \geq 1 - \varepsilon \ \forall n$.*

Le résultat suivant est fondamental dans l'étude de la convergence faible des mesures de probabilité :

Théorème 5 *(Théorème de Prohorov)*
Soit $\mathcal{P} = \{P_\alpha; \alpha \in \mathcal{U}\}$ une famille de mesures de probabilité définie sur un espace métrique séparable et complet $(E, \mathcal{E}, \rho)$. Alors $\mathcal{P}$ est relativement compacte si et seulement si elle est tendue.

Remarque :
Si on montre qu'une suite de variables aléatoires $\{X_n\}$ définie sur un espace métrique séparable et complet $(E, \mathcal{E}, \rho)$ est tendue, grâce au le théorème de Prohorov, on obtient alors la compacité relative de $\{X_n\}$.Il suffit donc ensuite de montrer que les limites possibles sont identiques et on aura alors unicité de la limite.

Par conséquent montrer la convergence en loi d'un processus défini sur l'espace métrique séparable et complet $(E, \mathcal{E}, \rho)$ équivaut à montrer la convergence fini-dimensionnelle et la tension de processus.
La convergence d'un processus se montre donc en deux étapes :

- convergence fidi (cv au sens des répartitions finies dim.),
- tension.

Conditions simples à vérifier pour la tension

- $E = \mathbb{R}^k$
  Si $\{X_n\}$ est équi intégrable ($\lim_{\lambda \to \infty} \sup_n E\left[||X_n|| 1_{\{||X_n|| \geq \lambda\}}\right] = 0$) alors $\{X_n\}$ est tendue.
  Or s'il existe $p >$, tel que $\sup_n E[||X_n||^p] < \infty$, alors $\{X_n\}$ est équi intégrable.
- $E = \mathcal{C}([0,1])$ (espace des fonctions continues sur $[0,1]$, c'est-à-dire suite de processus $\{X_n(t)\}$

  Théorème 6 *(Aizela-Ascoli)*
  Si il existe $\alpha > 1$, $\beta > 0$ tels que :

  $$E\left[|X_n(t) - X_n(s)|^\beta\right] \leq c^{te} |t - s|^\alpha, \ \forall n, t, s$$

  alors $\{X_n\}$ est tendue.

Le cas particulier du processus $N(]0,t])$

$X = (N(]0,t]))_{t \in [0,1]}$, où N est un processus ponctuel défini sur l'espace métrique $(D([0,1]), \mathcal{B}(D([0,1])), d)$.
Définissons :

- σ espace métrique polonais (i.e. complet et séparable),
- $\mathcal{S}$: σ-algèbre des Boréliens,

- $\mathcal{B} = \{A \in \mathcal{S} \ : \bar{A}$ compact $\}$,
- Soit $\mathcal{U} \subset \mathcal{B}$, on dira que $\mathcal{U}$ est un DC anneau (Dissecting Covering), si :
  - c'est un anneau
  - $\forall \varepsilon > 0$ et $B \in \mathcal{B}, \exists U_1, U_2, \cdots, U_n \in \mathcal{U}$tels que : $(U_i) < \varepsilon$ et $B \subset \cup_{i=1}^{n} U_i$.

Le théorème de Kallenberg 4.2 (1983) fournit le résultat suivant :

Théorème 7 *Soient $\xi_1, \xi_2, \cdots$ des mesures aléatoires définies sur σ et soit $\mathcal{I} \subset \mathcal{B}_\xi$ un DC anneau. Alors on a les équivallences suivantes :*

- *(i)* $\xi_n \xrightarrow{d} \xi$,
- *(ii)* $(\xi_n I_1, \cdots, \xi_n I_k) \xrightarrow{d} (\xi I_1, \cdots, \xi I_k)$, $k \in \mathbb{N}$, $I_1, \cdots, I_k \in \mathcal{I}$.

où $\mathcal{B}_\xi = \{B \in \mathcal{B} : \xi \partial B = 0 \ a.s.\}$.

Comme $(D([0,1]), \mathcal{B}(D([0,1])), d)$ est un espace polonais, le résultat précédent s'applique au processus $N\,(]0,t])$. Ainsi, pour montrer la convergence du processus en distribution $N_n\,(]0,t])_{t\in[0,1]}$, il suffit de montrer la convergence en distribution et la tension de $N_n\,(]0,1])$.

Chapitre 12

Quelques rappels sur les CPLT

12.1 Convergence vers les distributions de Poisson Composé

12.1.1 Caractérisation des distributions de Poisson Composé: Renyi (1951)

Théorème 8 *La classe des distributions de Poisson Composé peut être caractérisée comme la classe des distributions infiniment divisibles de variables aléatoires à valeurs entières, non-négatives, qui prennent la valeur 0 avec une probabilité* > 0.

La preuve de ce théorème se trouve dans Renyi (1951).

Corollaire 1 *Si une distribution de Poisson Composé* $F(x)$ *est la convolée de deux distributions infiniment divisibles,* $F_1(x)$ *et* $F_2(x)$ *qui possèdent un saut positif à* $x = 0$, *elles doivent aussi être des distributions de Poisson Composé de degré ne dépassant pas celui de* $F(x)$.

12.1.2 Sommes de variables aléatoires indépendantes (Renyi (1951))

Dans le théorème suivant Renyi (1951), démontre que, sous certaines hypothèses, la distribution limite de sommes de variables aléatoires indépendantes à valeurs entières est la distribution générale de Poisson Composée:

Théorème 9 *Soient* $\xi_{n1}, \xi_{n2}, \cdots, \xi_{nk_n}$, *des variables aléatoires indépendantes à valeurs entières non-négatives* $(n = 1, 2, \cdots)$ *"infiniment petites ", c'est-à-dire, supposons que*

$$\lim_{n\to\infty} \max_{1\leq k\leq k_n} P(\xi_{nk} \neq 0) = 0 \tag{12.1}$$

Posons,

$$\eta_n = \xi_{n1} + \xi_{n2} + \cdots + \xi_{nk_n}, \tag{12.2}$$

De plus, $p_{nks} = P(\xi_{nk} = s)$ *et* $c_{ns} = \sum_{k=0}^{k_n} p_{nks}$.
La condition nécessaire et suffisante pour la convergence des distributions de la somme η_n *est l'existence d'une suite de nombres non-négatifs* $c_1, c_2, \cdots, c_s, \cdots$ *ayant la propriété suivante :* $\sum_{s=1}^{\infty} c_s$ *converge, et* $\sum_{s=1}^{\infty} c_s > 0$, *et de plus*

$$\lim_{n\to\infty} \sum_{s=1}^{\infty} |c_{ns} - c_s| = 0 \tag{12.3}$$

Si (12.3) est satisfaite, la distribution de η_n *tend pour* $n \to \infty$ *vers la distribution de Poisson Composé de fonction génératrice*

$$\varphi(z) = \exp\left(\sum_{s=1}^{\infty} c_s (z^s - 1)\right) \tag{12.4}$$

12.1.3 Sommes de variables aléatoires stationnaires mélangeantes (Dziubdziela (1988))

Soit $\{X_t, -\infty < t < +\infty\}$ une suite strictement stationnaire de variables aléatoires. Soit $\mathcal{F}_i^j$, la σ-algèbre des événements générés par $X_i, X_{i+1}, \cdots, X_j, -\infty \leq i \leq j \leq \infty$. Pour $k = 1, 2, \cdots$, on définit les coefficients de mélange suivants par :

$$\alpha(k) = \sup_m \sup_{A \in \mathcal{F}_{-\infty}^m, B \in \mathcal{F}_{m+k}^\infty} |P[A \cap B] - P[A]P[B]| \tag{12.5}$$

$$\psi(k) = \sup_m \sup_{A \in \mathcal{F}_{-\infty}^m, B \in \mathcal{F}_{m+k}^\infty} \frac{1}{P[A]P[B]} |P[A \cap B] - P[A]P[B]| \text{ où } P[A]P[B] \neq 0 \tag{12.6}$$

Remarque :
Si le processus est stationnaire, les coefficients de mélange définis ci-dessus prennent donc les formes simplifiées suivantes :

$$\alpha(k) = \sup_{A \in \mathcal{F}_{-\infty}^0, B \in \mathcal{F}_k^\infty} |P[A \cap B] - P[A]P[B]|$$

$$\psi(k) = \sup_{A \in \mathcal{F}_{-\infty}^0, B \in \mathcal{F}_k^\infty} \frac{1}{P[A]P[B]} |P[A \cap B] - P[A]P[B]| \text{ où } P[A]P[B] \neq 0$$

Définition 5 *On dit que :*

- $\{X_t\}$ *est fortement mélangeante (ou α-mélangeante) si* $\alpha(k) \to 0$ *quand* $k \to \infty$
- $\{X_t\}$ *est ψ-mélangeante si* $\psi(k) \to 0$ *quand* $k \to \infty$

Remarque :
Si $\{X_t\}$ est ψ-mélangeante alors elle est fortement mélangeante.
Dziubdziela (1988) obtient pour des suites fortement mélangeantes le théorème suivant :

Théorème 10 *Soit $\{X_t, -\infty < t < \infty\}$ une suite strictement stationnaire fortement mélangeante de variables aléatoires de distribution commune F telle qu'il existe une constante $\lambda, 0 < \lambda < \infty$ et une suite de nombres réels $\{u_n, n \geq 1\}$ telles que $\lim_{n\to\infty} n(1 - F(u_n)) = \lambda$. Alors*
la distribution des sommes $J_n = \sum_{j=1}^n 1_{\{X_j > u_n\}}$ converge faiblement, quand $n \to \infty$, vers une distribution de Poisson Composée de fonction caractéristique $\phi(t) = \exp\left(\lambda \sum_{n=1}^\infty d_n(e^{itn} - 1)\right)$

$$\Leftrightarrow$$

$$\lim_{n\to\infty} \sum_{s=1}^{\infty} |k_n P\left[\sum_{j=1}^{p_n} 1_{\{X_j > u_n\}}\right] - \lambda d_s| = 0$$

pour $\{k_n, p_n, n \geq\}$ suites d'entiers positifs satisfaisants :

$$\lim_{n\to\infty} k_n = \infty, \quad \lim_{n\to\infty} p_n = \infty,$$

$$\lim_{n\to\infty} \frac{k_n p_n}{n} = 1$$

$$\lim_{n\to\infty} k_n \alpha(q_n) = 0$$

où α est donné par (12.5) et $q_n = \left[\frac{n - k_n p_n}{k_n}\right]$, $n = 1, 2, \cdots$ ([.] représente la partie entière).

Et pour des suites ψ-mélangeantes, Dziubdziela (1988) obtient le théorème suivant :

Théorème 11 *Soit* $\{X_t, -\infty < t < \infty\}$ *une suite strictement stationnaire,* ψ*-mélangeante de variables aléatoires de distribution commune* F *telle qu'il existe une constante* $\lambda, 0 < \lambda < \infty$ *et une suite de nombres réels* $\{u_n, n \geq 1\}$ *telles que* $\lim_{n\to\infty} n(1 - F(u_n)) = \lambda$.
Alors
Si $\psi(1) < \infty$ *où* ψ *est donné par (12.6), la distribution des sommes* $J_n = \sum_{j=1}^{n} 1_{\{X_j > u_n\}}$ *converge faiblement, quand* $n \to \infty$, *vers une distribution de Poisson de paramètre* λ.

12.1.4 Cas particulier des sommes de variables aléatoires stationnaires de Bernoulli

m-dépendantes

Hudson, Tucker et Veeh(1989) obtiennent tout d'abord des conditions nécessaires et suffisantes pour qu'une suite de sommes de variables aléatoires de Bernoulli strictement stationnaires et m-dépendantes dans chaque ligne converge en distribution vers la distribution de Poisson :

Théorème 12 *Soit* $\{X_{n,i}, 1 \leq i \leq n, n \geq 1\}$ *un tableau triangulaire de variables aléatoires de Bernoulli qui est strictement stationnaire et m-dépendant dans chaque ligne. Alors* $S_n = \sum_{i=1}^{n} X_{n,i}$ *converge en distribution vers la distribution de Poisson de paramètre* λ

$$\Leftrightarrow$$

$nP[X_{n,1} = 1] \to \lambda$ *et* $nCov(X_{n,1}, X_{n,j}) \to 0, 2 \leq j \leq m+1$.

Puis,ils présentent un résultat plus général :

Théorème 13 *Soit* $\{X_{n,i}, 1 \leq i \leq n, n \geq 1\}$ *un tableau triangulaire de variables aléatoires de Bernoulli strictement stationnaire et m-dépendant dans chaque ligne. Alors* $S_n = \sum_{i=1}^{n} X_{n,i}$ *converge en distribution*

$$\Leftrightarrow$$

$nP[X_{n,i} = 1]$ *converge et*

$$nP\left[\sum_{i=1}^{m} X_{n,i} = 0, X_{n,m+1} = 1, \sum_{i=m+2}^{2m+1} X_{n,i} = j-1\right] \to \lambda_j,\ 1 \leq j \leq m+1,$$

De plus la distribution limite est celle de $\sum_{j=1}^{m+1} jY_j$ *où* $Y_1, \cdots, Y_{m+1}$ *sont indépendantes et* Y_j *est une variable de Poisson de paramètre* λ_j.

Δ-mélangeantes

Soit $\{X_{n,i}, 1 \leq i \leq n, n \geq 1\}$ un tableau triangulaire de variables aléatoires de Bernoulli strictement stationnaire dans chaque ligne et satisfaisant la condition de mélange Δ définie de la façon suivante (cf. Hsing, Hüsler et Leadbetter (1988)) :
Pour chaque n, i, j avec $1 \leq i \leq j \leq n$, définissons :

- $\mathcal{B}_i^j(n)$ comme la σ-algèbre générée par les variables aléatoires $X_{n,s}, i \leq s \leq j$.
- pour chaque n et $1 \leq l \leq n-1$,

$$\alpha(n,l) = \max\left(|P[A \cap B] - P[A]P[B]| : A \in \mathcal{B}_1^k(n), B \in \mathcal{B}_{k+l}^n(n), 1 \leq k \leq n-l\right).$$

Le tableau $\{X_{n,i}\}$ est dit satisfaire la condition Δ si $\alpha(n, l_n) \to 0$ quand $n \to \infty$ pour une certaine suite $\{l_n\}$ telle que $l_n = o(n)$.
Définissons $\{k_n, n \geq 1\}$ comme la suite d'entiers positifs satisfaisant :

$$k_n \to \infty,\ k_n l_n / n \to 0 \text{ quand } n \to \infty. \tag{12.7}$$

et

$$k_n \alpha(n, l_n) \to 0 \text{ quand } n \to \infty. \tag{12.8}$$

Grâce à la condition Δ, $\{k_n\}$ peut-être choisie telle que (12.7) et (12.8) soient satisfaites (cf. Leadbetter et Nandagopalan (1989)). Notant $S_n = \sum_{i=1}^{n} X_{n,i}, n \geq 1$, Dziubdziela (1995) en utilisant les résultats de Rényi (1951) donne des conditions nécessaires et suffisantes pour que S_n converge en distribution vers la distribution de Poisson Composé:

Théorème 14 *Supposons que Δ soit vraie pour $\{X_{n,i}, 1 \leq i \leq n, n \geq 1\}$ tableau triangulaire de variables aléatoires de Bernoulli strictement stationnaire dans chaque ligne. Alors $S_n = \sum_{i=1}^{n} X_{n,i}$ converge en distribution vers une distribution de Poisson Composé de transformée de Laplace $\exp\left(\sum_{j=1}^{\infty} \lambda_j (e^{-sj} - 1)\right)$ où $\lambda_j \geq 0, j \geq 1$*

$$\Leftrightarrow$$

il existe une suite de nombres réels non-négatifs $\{\lambda_n\}$ tels que $\sum_{j=1}^{\infty} \lambda_j$ converge, $\sum_{j=1}^{\infty} \lambda_j > 0$ et

$$\sum_{s=1}^{\infty} |k_n P\left[\sum_{i=1}^{r_n} X_{n,i} = s\right] - \lambda_s| \to 0 \text{ quand } n \to \infty$$

pour $\{k_n\}$ suite d'entiers positifs satisfaisant (12.7) et (12.8) et $r_n = [n/k_n]$.

12.2 Modélisation des dépassements de très haut niveau

12.2.1 Le processus ponctuel des dépassements de très haut niveau

Pour modéliser les dépassements, il est pratique de normaliser les dates d'apparition $\{t : X_t > u_n, 1 \leq t \leq n\}$ par un facteur n pour obtenir un processus ponctuel sur $]0,1]$ formé des points t/n pour lesquels $X_t > u_n$. C'est-à-dire pour $B \subset]0,1]$, ; $N_n(B)$ est le nombre de points $t/n \in B$ de dépassements normalisés, i.e.

$$\begin{aligned} N_n(B) &= Card\,(t/n \in B : X_t > u_n, 1 \leq t \leq n) \\ &= Card\,(t \in nB : X_t > u_n, 1 \leq t \leq n,) \\ &= \sum_{t \in [0,n] \,:\, t/n \in B} 1_{\{X_t > u_n\}}, \; ; B \in]0,1]. \end{aligned}$$

N_n sera appelé *processus ponctuel des dépassements sur* $]0,1]$, formé des dépassements normalisés parmi les variables aléatoires $X_1, X_2, \cdots, X_n$.
Ainsi, si les variables aléatoires X_t sont identiquement distribuées de distribution commune F, l'intensité de N_n est donc:

$$E\,(N_n\,(]0,1])) = n\,(1 - F\,(u_n))$$

12.2.2 Théorèmes limite pour le processus ponctuel des dépassements de très haut niveau: Théorèmes Limites Poisson Composé (CPLT)

Les modèles pour les dépassements de très hautes valeurs résultent de théorèmes limites (CPLT) quand $u_n \to \infty$. L'obtention de distributions limites Poissoniennes nécessite une convergence rapide (i.e. dépassements de très haut niveau), telle que si les X_t sont identiquement distribuées, on ait $n\,(1 - F(u_n))$ converge vers une limite finie et dans le cas non identiquement distribuées $\sum_{t \leq n} (1 - F_{X_t}(u_n))$ converge quand $n \to \infty$ vers une limite finie.

Suites aléatoires indépendantes et identiquement distribuées

Notant N_n^* le processus ponctuel de dépassement non-normalisé, c'est-à-dire

$$N_n^* = \sum_{t \in [1,n]} 1_{\{X_t > u_n\}}$$

Si $\{X_t\}$ est une suite iid de distribution F, alors u_n tel que $\lim_{n \to \infty} n\,(1 - F\,(u_n)) = \lambda$ est équivalent à $P\,[N_n^* \leq k] \to e^{-\lambda |B|} \sum_{s=0}^{k} \frac{\lambda^s}{s!}$. En effet si $\{X_t\}$ est une suite iid de distribution F, N_n^* suit une

loi binomiale $bin\,(n, 1 - F(u_n))$ et donc après un calcul élémentaire, on vérifie l'équivalence entre la convergence de $n\,(1 - F\,(u_n))$ quand n tend vers l'infini, vers λ et la convergence faible de N_n^* vers une loi de Poisson de paramètre λ.

Notons alors que si on effectue un changement d'échelle et que l'on considère donc le processus ponctuel des dépassements normalisé N_n défini sur $]0, 1]$, on démontre que N_n a une distribution limite de Poisson. De même, $N_n\,(B)$ quel que soit $B \subset]0, 1]$ possède une distribution limite de Poisson et si les ensembles bornés B_i sont disjoints les $N_n(B_i)$ sont clairement indépendantes. Ceci suggère donc que les dépassements de u_n, s'ils sont décrits par le processus ponctuel normalisé (c'est-à-dire, pris aux points j/n), se comportent comme un processus de Poisson quand n est grand. On a alors le théorème suivant :

Théorème 15 *$\{X_t\}$ est une suite iid de distribution F, $0 \leq \lambda \leq \infty$, et u_n satisfaisant :*

$$\lim_{n\to\infty} n\,(1 - F\,(u_n)) = \lambda$$

$$\Rightarrow$$

N_n converge en loi vers un processus de Poisson d'intensité λ sur $]0, 1]$.

Le théorème obtenu dans le cadre de suites de v.a. iid peut être généralisé en autorisant la dépendance ou en permettant aux X_t à d'avoir des distributions différentes ou les deux :

- On peut supposer les variables de la suite aléatoire indépendantes et leurs distributions respectives non-identiques. De telles suites sont dites **indépendantes**. Elles ont déjà été utilisées par Shively (1990) pour des valeurs extrêmes d'ozone. On peut supposer approximativement l'indépendance des ensembles des données suffisamment séparés dans le temps.

- Le cas particulier le plus connu de suites dépendantes est le cas **stationnaire**. On rappelle qu'une suite aléatoire est dite (strictement) **stationnaire**, si les distributions finidimensionnelles de la suite aléatoire sont telles que :

$$F_{X_{t_1}, X_{t_2}, \cdots, X_{t_k}}(.) = F_{X_{t_1+m}, X_{t_2+m}, \cdots, X_{t_k+m}}(.) \tag{12.9}$$

pour tout $\{t_i \in \mathbb{N}, i = 1, \cdots, k\}$ et $k, m \in \mathbb{N}$. Evidemment, ceci implique que $F_{X_i}(.) = F_{X_1}(.)$ pour tout $i \geq 1$ (en prenant $k = 1$ dans (12.9)).
Si (12.9) n'est pas vraie, alors la suite aléatoire est habituellement dite **non-stationnaire**.

Suites aléatoires indépendantes de variables aléatoires non identiquement distribuées

On considère des variables aléatoires X_t indépendantes de distribution $F_{X_t}(.)$, en général non identiques.

Définition 6 *On dira qu'une suite aléatoire est* **uniformément asymptotiquement négligeable (uan)**, *si*

$$\sup_{t\leq n} P\,[X_t < u_n] = \sup_{t\leq n} [1 - F_{X_t}(u_n)] \to 0 \text{ quand } n \to \infty. \tag{12.10}$$

Sous la condition uan (12.10), le nombre de dépassements $N_n = \sum_{t<n} 1_{X_t > u_n}$ converge en loi vers une variable de Poisson de paramètre $\lambda \in [0, \infty[$, si et seulement si

$$\sum_{t\leq n} [1 - F_{X_t}(u_n)] \to \lambda \text{ quand } n \to \infty. \tag{12.11}$$

Remarque :
La condition (12.11) généralise la condition $n(1 - F_X(u_n)) \to \lambda$.

et on a le CPLT suivant :

Théorème 16 *$\{X_t\}$ est une suite de variables aléatoires indépendantes, de distributions respectives F_{X_t}, $0 \leq \lambda \leq \infty$, et u_n satisfaisant:*
$\lim_{n\to\infty} \sum_{t\leq n} (1 - F_{X_t}(u_n)) = \lambda$

$\Rightarrow$

N_n converge en loi vers un processus de Poisson d'intensité λ sur $]0,1]$.

Suites aléatoires stationnaires et dépendantes

Dans cette partie, on rappellera que sous certaines restrictions limitant la structure de dépendance de la suite, les loi limites sont précisément les mêmes que dans le cas iid. Les suites stationnaires concernées sont celles exhibant une structure de dépendance qui n'est pas "trop forte ". Pour démontrer ces résultats, Leadbetter a utilisé des hypothèses de mélange de type distributionnelles plus faibles que les formes usuelles de restriction de dépendance telle que le mélange fort (Loynes(1965)).

- RESTRICTIONS DE DÉPENDANCE POUR LES SUITES STATIONNAIRES

  - **La condition de type mélange** $D(u_n)$ (Leadbetter(1974))
    Pour affaiblir la condition de mélange, on note que les évènements nous intéressant dans la théorie des valeurs extrêmes sont ceux du type $\{X_t \geq u_n\}$ ou leurs intersections. Pour simplifier les notations, dans la suite on écrira $F_{t_1\ldots t_n}(u)$ pour $F_{t_1\ldots t_n}(u,\ldots,u)$, si $F_{t_1\ldots t_n}(x_1,\ldots,x_n)$ représente la distribution conjointe de $X_1,\ldots,X_n$.
    La condition $D(u_n)$:
    Posons:

$$\alpha_{n,l} = \max\left\{\left|F_{t_1\ldots t_p,j_1\ldots j_{p'}}(u_n) - F_{t_1\ldots t_p}(u_n)\,F_{j_1\ldots j_{p'}}(u_n)\right| : 1 \leq t_1 \leq \cdots \leq t_p < j_1 \leq \cdots \leq j_{p'}, j_1 - t_p \geq l\right\}$$

    Définition 7 *$D(u_n)$ est vérifiée pour une suite aléatoire $\{X_t\}$ et u_n suite tendant vers l'infini convenablement, si:*
    il existe une suite $\{l_n\}$ telle que $\alpha_{n,l_n} \to 0$ et $l_n\overline{F}_{n,\max} \to 0$ quand $n \to \infty$, où $\overline{F}_{n,\max} = \sup_{t\leq n}(1 - F_{X_t}(u_n))$.

    Remarque:
    * Si les variables aléatoires sont indépendantes $\alpha_{n,l_n} = l_n = 0$.
    * Le fort mélange implique $D(u_n)$, pour toute suite $\{u_n\}$.

    Comme, on suppose $\lambda(]0,1]) > 0$, on a: $\liminf n\overline{F}_{n,\max} > 0$. Par conséquent, l'hypothèse $l_n\overline{F}_{n,\max} \to 0$ implique que $l_n = o(n)$.
    On peut toujours choisir une suite croissante $\{k_n\}$ d'entiers telle que:

$$\lim_{n\to\infty} k_n l_n \overline{F}_{n,\max} = 0 \text{ et } \lim_{n\to\infty} k_n \alpha_{n,l_n} = 0 \tag{12.12}$$

    Notons que $\{k_n\}$ peut être bornée ou tendre vers l'infini, mais on a toujours $k_n l_n = o(n)$.
    Exemple:
    $k_n = \left[\min\left(l_n\overline{F}_{n,\max}, \alpha_{n,l_n}\right)\right]^{-1/2}$

    L'importance de pouvoir définir une suite $\{k_n\}$ d'entiers vérifiant (12.12) réside dans le lemme suivant qui démontre comment la condition $D(u_n)$ donne le degré d'indépendance approprié pour une discussion sur les extrêmes:

    Lemme 1 *Supposons que $D(u_n)$ soit satisfaite pour la suite aléatoire $\{X_t, t \geq 1\}$ et la suite $\{u_n\}$. Soient $B_j(= B_{j,n}), j \leq k_n$, des intervalles disjoints de $]0,1]$, où (12.12) est satisfaite pour k_n. Alors*

$$P\left[N_n\left(\bigcup_{j\leq k_n} B_j\right) = 0\right] - \prod_{j\leq k_n} P\left[N_n(B_j) = 0\right] \to 0$$

Ce lemme a été démontré de façon habituelle, en utilisant la propriété de mélange $D(u_n)$, $k_n - 1$ fois pour approcher pour chaque

$$l, 2 \leq l \leq k_n, P\left[N_n\left(\bigcup_{j\leq l-1} B_j \bigcup B_l\right) = 0\right]$$

par

$$P\left[N_n\left(\bigcup_{j\leq l-1} B_j\right) = 0\right] P\left[N_n\left(B_l\right)\right]$$

Si les nB_j sont séparés par l_n, le résultat est vrai, puisque $\{k_n\}$ est choisie telle que $k_n \alpha_{n,l_n} \to 0$ quand $n \to \infty$. Si ils ne sont pas séparés par l_n, alors les B_j sont approchés par B_j^* qui sont séparés par l_n, en supprimant un petit intervalle de longueur l_n/n à l'extrémité droite de chaque B_j. Ainsi nB_j^* sont séparés par l_n et les erreurs d'approximations tendent vers 0 puisque $n \to \infty \lim k_n l_n \overline{F}_{n,\max} = 0$ (cf. Leadbetter(1983), Leadbetter et Nandagopalan (1989)).

En utilisant ce lemme, la vérification de la condition (b) du théorème (4) se réduit à montrer la convergence :

$$P\left[N_n\left(B\right) = 0\right] \to P\left[N\left(b\right) = 0\right]$$

pour tout $B =]c, d] \subset]0, 1]$, B peut être découpé en k_n intervalles disjoints $B_{j,n}, j \leq k_n$, tels que $m\left(B_{j,n}\right) \to 0$. Appliquant le lemme précédent encore une fois :

$$P\left[N_n\left(B\right) = 0\right] - \prod_{j\leq k_n} P\left[N_n\left(B_{j,n}\right) = 0\right] \to 0$$

- **La condition de dépendance locale**

Le problème se réduit donc à considérer $P\left[N_n\left(B_{j,n}\right) = 0\right] = P\left[X_t \leq u_n, t \in nB_{j,n}\right]$. Notons que seule la forme locale de la suite aléatoire $\{X_t\}$ est nécessaire dans cette probabilité. Cependant, la forme locale de cette suite n'est pas restreinte par la condition de mélange $D\left(u_n\right)$ et on doit donc ajouter une condition $D'\left(u_n\right)$ qui restreigne la dépendance locale. cf. Leadbetter et al. (1983)

$D'(u_n)$ sera vérifiée si :

$$\limsup_{n\to\infty} n \sum_{j=2}^{[n/k]} P\left[X_1 > u_n, X_j > u_n\right] \to 0 \text{ quand } k \to \infty$$

([.] représente la partie entière)

Cette condition limite la possiblité de regroupement des dépassements, les évènements multiples sont par conséquent exclus à la limite.

- LES LIMITES POISSON

Théorème 17 *Supposons que la suite de variables aléatoires stationnaires $\{X_n\}$ satisfasse les conditions $D(u_n)$ et $D'(u_n)$, u_n satisfaisant $nP\left[X_1 > u_n\right] \to \lambda$, quand $n \to \infty$.*
Alors
N_n converge en loi vers un processus de Poisson N sur $]0, 1]$ de paramètre λ.

preuve :
cf. Leadbetter et al. (1983), Leadbetter et Nandagopalan (1989).

- LES LIMITES POISSON COMPOSÉ

Dans Leadbetter et Hsing (1990), Leadbetter (1995) la condition de dépendance suivante, notée $\Delta(u_n)$ est utilisée pour démontrer un CPLT :

Si $\{u_n\}$ est une suite de constantes, pour chaque n, t, j avec $1 \leq t \leq j \leq n$, définissons $\sum_n(t,j) = \sigma(\{\{X_s > u_n\} : t \leq s \leq j\})$, σ-algèbre engendrée par les évènements $\{X_s > u_n\}, t \leq s \leq j$.. Donc pour n et $1 \leq l \leq n-1$, on écrit :

$$\alpha_{n,l} = \sup\left(|P(A \cap B) - P(A)P(B)| : A \in \sum_n(1,k), \ ; B \in \sum_n(k+l,n), 1 \leq k \leq n-l\right)$$

$\{X_t\}$ est dite satisfaire la condition $\Delta(u_n)$ si $\alpha_{n,l_n} \to 0$, quand $n \to \infty$, pour une certaine suite $\{l_n\}$ avec $l_n = o(n)$.
Notons que la condition $\Delta(u_n)$ est plus forte que la condition de mélange distributionnelle $D(u_n)$, mais plus faible que la condition de fort mélange.

La condition $\Delta(u_n)$ permet d'obtenir une suite $\{k_n\}$, $k_n \to \infty, k_n = o(n)$ d'entiers , $r_n = [n/k_n]$, les entiers $1, 2, \cdots, k_n r_n (\sim n)$ sont alors divisés en k_n groupes consécutifs ou (blocs) $((i-1)r_n + 1, (i-1)r_n + 2, \cdots, ir_n)$, $1 \leq i \leq k_n$. Les dépassements (s'il y en a) dans un tel bloc seront appelés classe. Si $\{X_n\}$ est stationnaire et satisfait la condition $\Delta(u_n)$, les groupes de dépassements convergeant chacun vers un point unique, après la normalisation du temps ($r_n/n = [n/k_n]/n \sim 1/k_n \to 0$). On définit donc :

- les classes normalisées, correspondant aux groupes de dépassements dans les intervalles de temps normalisés $J_i =](i-1)r_n/n, ir_n/n], 1 \leq i \leq k_n$, qui avec l'intervalle $]k_n r_n/n, 1]$ forment une k_n-partition de $]0, 1]$.
- La distribution de la taille de classe π_n comme

$$\pi_n\{r\} = P\left[\{N_n(J_1) = r\}/\{N_n(J_1) > 0\}\right], r = 1, 2, \cdots \tag{12.13}$$

Leadbetter obtient le théorème :

Théorème 18 *Soit $(X_n, n = 1, 2, \cdots)$ stationnaire, satisfaisant $\Delta(u_n)$ et u_n une suite telle que $n(1 - F(u_n)) \to \lambda$ quand $n \to \infty$. Supposons que $\pi_n \to^w \pi$, une distribution et la taille moyenne de classe $\mu_n = \sum_{j=1}^{\infty} j\pi_n(j) \to \theta^{-1}, 0 > \theta \leq 1$. Alors N_n converge en distribution vers un processus de Poisson Composé $N = CP(\theta\lambda, \pi)$ basé sur un processus de Poisson d'intensité $\theta\lambda$ et de distribution des événements multiples π.*

Le paramètre $\theta(\leq 1)$ qui a un rôle important en théorie des valeurs extrêmes est appelé *index extrémal* de la suite $\{X_t\}$. Il est relié au regroupement des dépassements de cette suite. Si $\theta = 1$, alors les dépassements ne sont pas regroupés, c'est-à -dire, les tailles de classe sont asymptotiquement égales à 1 avec la probabilité 1.

La classe des suites aléatoires non-stationnaires est plus grande, une théorie des valeurs extrêmes pour la classe générale des suites aléatoires non-stationnaires n'existe pas à ce jour (cf. Falk, Hüsler et Reiss , 1994).

Suites aléatoires non-stationnaires et dépendantes

Dépendance locale (cf. Falk, Hüsler et Reiss (1994))

Soit α_n^* tel que :

$$\sum_{t<j<j+1 \in I} P\left[X_t > u_n, X_j \leq u_n, X_{j+1} > u_n\right] \leq \alpha_n^*$$

pour tout intervalle $I = \{t_1 \leq t \leq t_2 \leq n\} \subset \mathbb{N}$, avec

$$\sum_{t \in I} P\left[X_t > u_n\right] \leq \sum_{t \leq n} P\left[X_t > u_n\right]/k_n,$$

où $\{k_n\}$ satisfait (12.12).

Définition 8 *$D^*(u_n)$ est vérifiée par la suite aléatoire $\{X_t\}$ et $\{u_n\}$ si*

$$k_n \alpha_n^* \to 0 \text{ quand } n \to \infty.$$

Notons que $D^*(u_n)$, comme $D'(u_n)$ dans le cas stationnaire, exclut la possibilité de regroupement de dépassements dans un petit intervalle I, car elle exclut les cas où la suite aléatoire $\{X_t\}$ oscille rapidement autour de $\{u_n\}$.
On a le théorème suivant :

Théorème 19 *Supposons que les conditions $D(u_n)$ et $D^*(u_n)$ soient vérifiées par la suite aléatoire $\{X_t, t \geq 1\}$ et $\{u_n, n \geq 1\}$.Si*

(a) $\sum_{t \leq nT} P[X_t > u_n] = \sum_{t \leq nT} (1 - F_t(u_n)) \to \lambda(T) = \lambda([0,T])$,
(b) $\sum_{t \leq nT-1} P[X_t \leq u_n, X_{t+1} > u_n] \to \mu(T)$, *pour* $T \leq 1$ *et* $\mu(.)$ *fonction bornée*

sont vraies, avec $\mu(.) \equiv \lambda(.)$ continues, Alors
N_n converge en loi vers un processus de Poisson non-homogène N sur $]0,1]$ d'intensité $\lambda(.)$.

Remarque :
On peut montrer que les conditions (a), (b) ensembles avec $\mu(.) \equiv \lambda(.)$ impliquent la condition $D'(u_n)$ (cf. Leadbetter et al. (1983) pour le cas stationnaire et Hüsler (1983) pour le cas non-stationnaire) :

$$\lim_{n \to \infty} k_n \sum_{t < j \in I} P[X_t > u_n, X_j > u_n] = 0$$

pour les mêmes ensembles I que dans la condition $D^*(u_n)$. Inversement, si $D'(u_n)$ et (a) sont vraies, alors $D^*(u_n)$ et (b) sont vraies avec $\mu(1) = \lambda(1)$.

Chapitre 13

CPLT for high-level exceedances of non-stationnary processes

Compound Poisson Limit theorems for high-level exceedances of some non-stationary processes

Lise Bellanger *and Gonzalo Perera ◇†
Laboratoire Modélisation Stochastique et Statistique, Université de Paris-Sud, ORSAY
◇ and Universidad de la República, Montevideo, Uruguay.

Abstract
We show the convergence to a Compound Poisson process of the high-level exceedances point process $N_n(B) = \sum_{\frac{i}{n} \in B} 1_{\{X_i > u_n\}}$, where $X_n = \varphi(\xi_n, Y_n)$, φ a (regular) regression function, u_n grows to infinity with n in a suitable way, ξ and Y are mutually independent, ξ is stationary and weakly dependent, and Y is non-stationary, satisfying some ergodic conditions. The basic technique is the study of high-level exceedances of stationary process over suitable collections of random sets.

AMS 1991 subject classifications: 60F05, 60G44, 60G55, 60J55.
Key words and phrases: exceedances, point processes, convergence, Compound Poisson process, level sets, mean occupation measures, asymptotically ponderable collections of sets

1 Introduction

In many meteorological or hydrological problems, relevant features are related to exceedances of high levels by some time series. In particular, current standards for ozone regulation involve the exceedances of high levels. In this case it is clear that the time series of actual ozone level depends on some non-stationary, meteorological variables, like temperature or wind speed. Therefore, a reasonable model for the ozone level at time t (say X_t) should be of the form

$$X_t = \varphi(\xi_t, Y_t) \tag{1}$$

where ξ_t is "pure noise", corresponding to local fluctuations of measurements systems, Y_t is a vector that contains the values of all the "explicative" variables at time t and φ is some suitable regression function. Indeed, as we will see later on, Y may contain not only actual values, but also previous values (for instance, temperatures of the last q days). We may also think t as a d-dimensional parameter, corresponding to space and time; all the models and results of this paper are valid in this context, but, for the sake of simplicity, we will only present here the case $d = 1$. One important remark is that we may assume that ξ is a "nice" process in the sense that it is stationary and very weakly dependent (say mixing), but Y may not be so "nice ". Y may not be stationary: for instance in the case of temperature, besides seasonal effects that affect the time scale , spatial variations due to differences between urban and rural areas make the assumption of stationarity not reasonable. Furthermore, even if Y may satisfy some ergodic properties (some Law of Large

*partially supported by ADEME and AIRPARIF
†to whom correspondence should be adressed

Numbers) it is not reasonable to expect mixing, association or any particular weak dependence structure.

We will deal with discrete-time observations, so we will observe the exceedances of $X_1, \cdots, X_n$ of a level u_n that grows to infinity with n in a suitable way. When $X = (X_t : t \in \mathbb{N})$ is *iid* a very simple computation shows that the point process

$$N_n(B) = \sum_{\frac{t}{n} \in B} 1_{\{X_t > u_n\}}, \; B \in \mathcal{B} \tag{2}$$

(where $\mathcal{B}$ stands for the Borel σ-algebra of $[0,1]$) converges to a Poisson process of intensity

$$\lambda = \lim_n nP(\{X_0 > u_n\})$$

If X is stationary and weakly dependent (if it satisfies some mixing conditions, for instance), clustering of exceedances may occur and one obtains a Compound Poisson Process. In the sequel, if X is a random process such that (for some sequence u_n) the point process of (2) converges to a Compound Poisson process, we shall say that X satisfies a Compound Poisson Limit Theorem (CPLT, for short). CPLT for stationary processes satisfying some mixing conditions are known (see [Cohen (1989)], [Dziubdziela (1988)], [Ferreira (1993)], [Hsing, Hüsler & Leadbetter (1988)], [Leadbetter & Nandagopalan (1989)], [Leadbetter & Hsing (1990)], [Leadbetter (1991)], [Leadbetter (1995)]) as well as for Markov Chains (see [Hsiau (1997)]). Some results are also available for X weakly dependent but non-stationary: see [Alpuim, Catkan & Hüsler (1995)], [Dziubdziela (1995)], [Hudson, Tucker & Veeh (1989)], [Hüsler (1993)]. For a nice summary of many related results see [Falk, Hüsler & Reiss (1994)]. The authoritative text [Leadbetter, Lindgren & Rootzén (1983)] is a basic reference for exceedances, extremes and related topics, as well as [Leadbetter & Rootzén (1988)]. For continuous-time results see [Volkonskii & Rozanov (1959)], [Volkonskii & Rozanov (1961)], [Wschebor (1985)] and the very nice monograph [Berman (1992)]. In some cases, rates of convergence can also be obtained, by means of Stein-Chein method: see an extensive account in [Barbour, Holst & Janson (1992)] and see also [Brown & Xia (1995)]. For the application of point process exceedances to practical modelling of ozone data, see for instance [Smith & Shively (1994)].

However, models like (1), where Y is not "nice", can fail to satisfy the weak-dependence hypotheses required on those results. The aim of this paper is to prove that for the model (1), the point process defined in (2) still has a Compound Poisson Distribution. Our result generalizes the preceeding ones; at first, our assumptions do not imply that X has a particular weak-dependence structure (like mixing, association, Markov, etc.), hence previous results do not apply to our models. For instance, we may have X defined by (1) where ξ is mixing and Y is merely ergodic but non-associated, nor mixing nor Markov and X is neither mixing, associated nor Markov while our assumptions still hold. At second, without additional effort we also obtain the limit distribution of N_n when Y (hence X) presents long-range dependence: in that case the limit distribution is no longer Compound Poisson but a mixture of several Compound Poisson distributions. Finally, we consider our approach interesting by itself, because the technique is based on the study of the high-level exceedances that belong to an "irregular "set, and it is found that the geometry of this set plays a key role.

More precisely, in this paper we first prove a CPLT for $N_n^*(B) = \sum_{j=1}^h \sum_{m \in B \cap [1,n]} 1_{\{\xi_m^j > u_n\}} 1_{\{m \in A^j\}}$, $B \subset \mathbb{N}$, where $(A^1, ..., A^h)$ is a collection of subsets of $\mathbb{N}$ satisfying a condition called asymptotic ponderability (that is satisfied, for instance, by level sets of ergodic processes), and $\vec{\xi} = (\xi^1, ..., \xi^h)$ a stationary and weakly dependent $\mathbb{R}^h$-valued random process; assuming that for $B_1, ..., B_k$ Borel sets of $\mathbb{R}$ we have that $Y^{-1}(B_1), ..., Y^{-1}(B_h)$ is an asymptotically ponderable collection, by conditioning with respect to Y, the limit distribution of N_n can be deduced from that of N_n^* for a suitable $\vec{\xi}$.

Roughly speaking, what we show here is that the addition of a component Y whose mean occupation measure has a limit (i.e., for large samples we can control in the mean how much time does the process Y spend on any set) on a weakly dependent model just averages the limits that are obtained for the weakly dependent case over irregular sets; if Y is ergodic, averaging will be non-random and a CPLT will hold; if Y is non-ergodic, a mixture of Compound Poisson will be obtained. If we look just to the ergodic case, it is clear that we require that the regression model really depends on the weakly dependent component that makes a CPLT possible. If we consider that this weakly dependent component (what we called "noise ") is negligible, then our results will fail to hold because the asymptotic will be driven just by Y. This is of course a limitation of our approach, but we must also empashize that we are only requiring that "noise "is not negligible, what seems to be reasonable in many situations.

The results presented here concerning the asymptotic distribution of the high-level exceedances over a collection of sets of irregular shape are, up to our knowledge, new; we do not know previous results determining the role played by the geometry of the collection. We extend here to the context of CPLT the results of [Perera (1994)], [Perera (1997)a], [Perera (1997)b] for Central Limit Theorems. We think that the study of the asymptotic distribution of additive functionals defined over "irregular" sets, showing the relevance of the geometric or arithmetic properties of these sets on the final result, could be a new ingredient of the asymptotic theory of aditive functionals.

This paper is organized as follows: Section 2 presents some basic notations and definitions and the statement of main result whose proof is presented in Section 5. Section 3 and 4 contain the basic ingredients: CPLT's over "irregular"sets; in Section 3 we present a detailed proof of such a CPLT, in Section 4 we give examples where this result applies. After the CPLT for the model (1), presentd in Section 5, we also inlude two appendices. In the first we present a set that is "too irregular "for a CPLT but regular enough for a Central Limit Theorem like those of [Perera (1997)a]; the second one presents the analysis of some real ozone data where this type of results may be applied.

2 Definitions and main results.

We will start by setting some definitions and notations.

All along this paper, we will consider $\mathbb{R}^d$ equipped with the sup-norm and C will denote a generic constant that may change from line to line. We shall also denote by C_s^d the combinatorial coefficients $C_s^d = \frac{d!}{(d-s)!s!}$, $0 \leq s \leq d$. An important role will be played by the coefficients

$$\Theta(s;d) = (-1)^{d-1} \sum_{j=0}^{j=s-1} C_j^d (-1)^j,\ 1 \leq s \leq d.$$

Recall that a point process N is a Compound Poisson Process with intensity measure ν (what we will denote by $CP(\nu)$), where ν is a positive finite measure on $\mathbb{N}$, if:

- For any $h \in \mathbb{N}$, if $B_1, \cdots, B_h$ are disjoint Borel sets, then $N(B_1), \cdots, N(B_h)$ are independent.
- For any Borel set B, the Laplace transform of $N(B)$ is :

$$L(B;s) = \exp\left(m(B) \sum_{j=1}^{+\infty} \nu_j \left(\exp(-sj) - 1\right) \right),$$

where m denotes Lebesgue measure, and $\nu_j = \nu(\{j\})\ \forall j \in \mathbb{N}$.

Given any $A \subset \mathbb{N}$ and $n \in \mathbb{N}$ we will set $A_n = A \cap [1, n]$. If B is a subset of $\mathbb{N}$ and $\vec{r} \in \mathbb{N}^d$, $d \geq 1$, then set

$$T_n(\vec{r}; B) = \bigcap_{i=1}^{i=d} (B_n - r_i)\ \forall n \in \mathbb{N}$$

More in general, if $\mathcal{A} = (A^1, \cdots, A^d)$, is any (ordered) finite collection of subsets of $\mathbb{N}$ and $\vec{r} \in \mathbb{N}^d$, $d \geq 1$, we define

$$T_n(\vec{r}; \mathcal{A}) = \bigcap_{i=1}^{i=d} (A_n^i - r_i) \forall n \in \mathbb{N}$$

Definition 1 *Let A be a subset of $\mathbb{N}$ we will say that A is* **asymptotically ponderable set** *(APS, for short) if:*

- *For any $d \geq 1$, $\vec{r} \in \mathbb{N}^d$, the following limit exists:*

$$\lim_n \frac{card\,(T_n(\vec{r}; A))}{n} := F(\vec{r}; A)$$

Definition 2 *Let $(A^1, \cdots, A^h)$ be a collection of subsets of $\mathbb{N}$; we will say that $(A^1, \cdots, A^h)$ is an* **asymptotically ponderable collection** *(APC, for short), if:*

- *For any $d \geq 1$, $\vec{r} \in \mathbb{N}^d$, $\{i_1, \cdots, i_d\} \in \{1, \cdots, h\}$ and any sub-collection $\mathcal{A} = (A^{i_1}, \cdots, A^{i_d})$, the following limit exists:*

$$\lim_n \frac{card\,(T_n(\vec{r}; \mathcal{A}))}{n} := F(\vec{r}; \mathcal{A}).$$

Remark 1

(a) Indeed, A is APS if and only if $\mathcal{A} = (A)$ is an APC.
(b) It is clear from the definition that asymptotic ponderability is an hereditary property: if a collection is an APC, so does any sub-collection.
(c) If $(A^1, \cdots, A^k)$ is a partition of $\mathbb{Z}^d$ and an APC with $F(0, A^j) > 0\ \forall j$, then $(A^1, \cdots, A^h)$ is an asymptotically measurable partition, in the sense of [Perera (1994)], [Perera (1997)a]. Indeed, a set A is called asymptotically measurable if the convergence of Definition 1 holds for $d = 1, 2$. Therefore, the sets of [Perera (1997)a], Lemma 3.2, that are not asymptotically measurable, provide an example of sets that are not APS. Let us present here one of this sets; define, for $n \in \mathbb{N}$, $I(n) = [100^{2^{n-1}}, 100^{2^n})$ and $A(n, 0) = I(n) \cap (5\mathbb{N})$, $A(n, 1) = I(n) \cap [(10\mathbb{N}) \cup (10\mathbb{N} + 1)]$, and set $A = \bigcup_{i=1}^{\infty} (A(2i, 0) \cup A(2i + 1, 1))$. After a straightforward computation we can check that this set A, based on the alternance of two differents patterns, satisfies that $F(1; A)$ does not exist, and hence, A it is not an APS. We can also give an example of an asymptotically measurable set that is not an APS (see Appendix 1).
(d) Consider $\{Y_t : t \in \mathbb{N}\}$ a stationary and ergodic random process, such that Y_0 takes values on $\{1, \cdots, k\}$ and let $A^j(\omega) = \{t \in \mathbb{N} : Y_t(\omega) = j\}$. Then, by the Ergodic Theorem, $\mathcal{A} = (A^1, \cdots, A^k)$ is, almost surely, an APC and $F(\vec{r}; \mathcal{A}) = E\left(\prod_{i=1}^{i=m} 1_{\{Y_{r_i} = j\}}\right)\ \forall \vec{r} \in \mathbb{N}^m$, $\forall m$.

Remark 2 Let us try to explain what is the intuitive meaning of the definition of an *APS*. Assume that we are trying to study the asymptotic behaviour of a functional of the form

$$Z_n = \sum_{t \in A_n} f_n(X_t)$$

where X is stationary, f_N is a real function ($f_n(x) = 1_{\{x>u_n\}}$ in our case of high-level exceedances, $f_n(x) = \frac{x}{\sqrt{n}}$ in the case of averages of centered processes, considered in [Perera (1997)a], [Perera (1997)b]).

If we are trying to show that Z_n is asymptotically gaussian, in order to identify its limit we only need to compute asymptotic moments up to order two. If Z_n is centered, we only need to deal with the second moment; using the stationarity of X, this can be computed as

$$E(Z_n^2) = \sum_{t=0}^{\infty} E\left(f_n(X_0)f_n(X_t)\right) card\{A_n \cap (A_n - t)\} = \sum_{t=0}^{\infty} nE\left(f_n(X_0)f_n(X_t)\right) \frac{card(T_n\left((0,t),A\right))}{n}$$

If we assume that

$$\lim_n nE\left(f_n(X_0)f_n(X_t)\right) = \rho(t) \forall t$$

and that there exist $(g(t))_{t\in\mathbb{N}}$ such that $\sum_{t=0}^{\infty} g(t) < \infty$ and

$$\left|nE\left(f_n(X_0)f_n(X_t)\right)\right| \leq g(t)\ \forall t$$

then, by the Dominated Convergence Theorem, the asymptotic variance can be computed if the convergence of Definition 1 holds for $d = 1, 2$. Therefore, we deduce what is presented in [Perera (1994)] and [Perera (1997)a]: to get gaussian limits for averages of stationary and weakly dependent process over irregular sets A, we only need to control (in the mean) the arithmetic distribution of couples of points of A.

But assume now that we are trying to obtain a non-gaussian limit (like in this paper). Computation of moments up to order two may not be enough to identify the limit distribution of Z_n and we can try to compute the limit of all the moments of Z_n. Let us take a glace of what happens with moments of order three: we will obtain now

$$E(Z_n^3) = \sum_{t=0}^{n}\sum_{s=t}^{n} E\left(f_n(X_0)f_n(X_t)f_n(X_s)\right) card\{A_n \cap (A_n - t) \cap (A_n - s)\}$$

$$= \sum_{t=0}^{n}\sum_{s=t}^{n} nE\left(f_n(X_0)f_n(X_t)f_n(X_s)\right) \frac{card(T_n\left((0,t,s),A\right))}{n}$$

under similar hypotheses to those listed above we will conclude that for the computation of asymptotic moments up to order three, we will need to handle the convergence of Definition 1 for $d = 1, 2, 3$.

After this rough description, we hope the reader will be convinced that for asymptotics involving all the moments of Z_n, Definition 1 appears naturally as a condition allowing to identify possible limits.

The application of Definition 2 to level sets of random process leads to the following definition.

Definition 3 *We will say that a real-valued process Y is* **ponderable** *if for every $d \in \mathbb{N}$, $\vec{r} \in \mathbb{N}^d$, there exists a (random) probability measure $\mu^{\vec{r}}(.)(\omega)$ defined on the Borel sets of $\mathbb{R}^d$ such that if $B_1, \cdots, B_d$ are Borel real sets then the (random) collection $\mathcal{A}((B_1, \cdots, B_d))(\omega) = \{A^1(\omega), \cdots, A^d(\omega))$ defined by*

$$A^j(\omega) = \{t \in \mathbb{N} : Y_t(\omega) \in B_j\}$$

is an APC almost surely with $F\left(\vec{r}, \mathcal{A}((B_1, \cdots, B_d))(\omega)\right) = \mu^{\vec{r}}\left(B_1 \times B_2 \times \cdots \times B_d\right)(\omega)$.

If, in addition, the measures $\mu^{\vec{r}}$ are not-random, we will say that Y is **regular**.

Remark 3 Let Y and fix $k \in \mathbb{N}$ and $\vec{r} \in \mathbb{N}^k$, $h \in \mathbb{N}$; it is clear from its definition that $F\left(\vec{r}, \mathcal{A}((B_1, \cdots, B_k))(\omega)\right)$ does not depend on $(Y_t : \|t\| \leq h)$ (a finite set of coordinates does not affect averages) and hence, we deduce that $\mu^{\vec{r}}$ is measurable with respect to the σ-algebra

$$\sigma_\infty^Y = \bigcap_{h=1}^{\infty} \sigma\left(Y_t : \|t\| \geq h\right)$$

Therefore, if σ_∞^Y is trivial, Y is regular.

Remark 4 Observe that Y ponderable means that for any $B_1, \cdots, B_d$ Borel real sets its (mean) asymptotic occupation measure is defined $a.s.$, i.e.,

$$\mu^{\vec{r}}(B_1 \times B_2 \times \cdots \times B_k)(\omega) = \lim_n \frac{1}{n} \sum_{t=0}^{n} 1_{\{Y_{t+r_j}(\omega) \in B_j \; j=1,\ldots,h\}} \; a.s.$$

In this way, a process is regular when a deterministic mean occupation measure exists.

Example 1

By the Ergodic Theorem (see [Guyon (1995)], p. 108), if Y is stationary, then Y is ponderable. We have already seen that if, in addition, σ_∞^Y is trivial, then Y is regular. In particular, that is the case if Y satisfies a Marcinkiewicz-Zygmund inequality. More precisely, we say that a centered random process $Y = \{Y_t : t \in \mathbb{N}\}$ satisfies a Marcinkiewicz-Zygmund inequality of order $q > 2$ if for any $d \geq 1, \vec{r} \in \mathbb{N}^d$ there exist a constant $\mathbb{C}(\vec{r}, q)$ such that for any function $f : \mathbb{R}^d \to \mathbb{R}$ bounded by 1 one has:

$$E\{\left(\sum_{t=1}^{N}[f(Y_t(\vec{r})) - E\{f(Y_t(\vec{r}))\}]\right)^q\} \leq \mathbb{C}(\vec{r}, q) N^{q/2}$$

where $Y_t(\vec{r}) = (Y_{t+r_1}, Y_{t+r_2}, ..., Y_{t+r_d})$.

We refer to [Doukhan & Louichi (1996)] for a syntethical overview of different contexts where inequalities apply; see also [Bryc & Smolenski(1993)].

Further, if Y is non-stationary, but it satisfies a Marcinkiewicz-Zygmund inequality of order $q > 2$ and there exists a probability measure $\mu^{\vec{r}}$ such that for any Borel sets $B_1, \cdots, B_k$ we have :

$$\lim_n \frac{1}{n} \sum_{m=1}^{m=n} E\left(\prod_{i=1}^{i=m} 1_{\{Y_{m+r_i} \in B_j\}}\right) = \mu^{\vec{r}}(B_1 \times B_2 \times \cdots \times B_k),$$

then a simple Borel-Cantelli argument proves that

$$F(\vec{r}; \mathcal{A}((B_1, \cdots, B_k)) = \mu^{\vec{r}}(B_1 \times B_2 \times \cdots \times B_k) \; \forall \vec{r} \in \mathbb{N}^m, \; \forall m,$$

and that Y is regular.

Some additional notation: if $J \subset \mathbb{R}$ and $d \geq 1$ then

$$J_L^d = \{(j_1, \cdots, j_d) : j_i \in J, j_i < j_{i+1} \; \forall i\}$$

If $\vec{r} \in \mathbb{N}^d$, V any random process and $u > 0$, set:

$$\{V(\vec{r}) > u\} := \bigcap_{i=1}^{d} \{V_{r_i} > u\}$$

In a similar way, if $J \subset \mathbb{N}$, set

$$\{V(J) > u\} := \bigcap_{t \in J} \{V_t > u\}$$

Now we turn the attention to the "ξ"component of the model (1).

Definition 4 *Let ξ be a real-valued random process and $\varphi : \mathbb{R}^2 \to \mathbb{R}$ a measurable function. We will say that ξ is* φ**-noise** *if for every finite $h \in \mathbb{N}$, any vector $(y_1, \cdots, y_h) \in \mathbb{R}^h$ and any APC $\mathcal{A} = (A^1, \cdots, A^h)$ the random process*

$$X_t = \sum_{j=1}^{j=h} \varphi(\xi_t, y^j) 1_{A^j}(t) \tag{3}$$

satisfies the CPLT.

We will prove a CPLT for random processes of the form (3) to provide examples of the previous definition. That result is presented in Section 3, and several examples of weak dependence structures where this definition applies will be derived in Section 4. For the moment, we only need to keep in mind that ξ is a "weakly dependent "random process, in a sense to be precised later on.

Finally, introduce the notation :

$$\varphi(\xi(\vec{r}), \vec{y}) = (\varphi(\xi_{r_1}, y_1), \cdots, \varphi(\xi_{r_d}, y_d)) \ \forall \vec{r} \in \mathbb{N}_L^d, \ \vec{y} \in \mathbb{R}^d$$

Definition 5 Let X be a real-valued random process. We will say that X admits an **I-decomposable regression** if there exist a ponderable process Y, a measurable function φ and a $\varphi - noise$ ξ independent of Y such that $X_t = \varphi(\xi_t, Y_t)$ $\forall t \in \mathbb{N}$ and

(a) $\forall K > 0 \ \limsup_{\delta \to 0} \limsup_{n} \sup_{|x-z| \leq \delta, |x| \leq K} nP(\{\varphi(\xi_0, x) > u_n\} \nabla \{\varphi(\xi_0, z) > u_n\}) = 0$ and
$\limsup_{K} \limsup_{n} \sup_{|x| > K} nP(\{\varphi(\xi_0, x) > u_n\}) < \infty$

(b) $\forall x \in \mathbb{R}, \lim_n nP(\{\varphi(\xi_0, x) > u_n\}) := \lambda(x)$.

(c) $\forall d \in \mathbb{N}, \ \vec{y} \in \mathbb{R}^d, \vec{r} \in \mathbb{N}_L^d, \lim_n P(\{\varphi(\xi(\vec{r}), \vec{y}) > u_n\} / \{\varphi(\xi_0, y_0) > u_n\}) = a(\vec{r}, \vec{y})$ and $a(\vec{r}, .)$ continuous $\forall \vec{r}$.

(d) $\sum_{d=s}^{\infty} |\Theta(s;d)| \sum_{\vec{r} \in \mathbb{N}_L^d} \sup_{y \in \mathbb{R}^d} a(\vec{r}, \vec{y}) < \infty \ \forall s \in \mathbb{N}, \ \lim_{s \to \infty} \sum_{d=s}^{\infty} |\Theta(s;d)| \sum_{\vec{r} \in \mathbb{N}_L^d} \sup_{y \in \mathbb{R}^d} a(\vec{r}, \vec{y}) = 0.$

Remark 5 A strightforward computation shows that condition (a) implies that the function λ defined in (b) is uniformly continuous and bounded

Remark 6 Let us briefly explain these assumptions. **(b)** assumes that for fixed x, the limit intensity of $\varphi(\xi_m, x)$ is well-defined. Its continuity, as well as **(c)**, guarantees that if we approximate Y by another process Y^*, their limit Laplace transforms are also close. **(a)** allows to approximate Y by, first, its restriction to a compact set $[-K, K]$ and, second, it discretization. **(d)** is the technical hypothesis required to apply a suitable CPLT (Corollary 2 of the next section) for the approximation of Y.

Remark 7 The reader may ask himself where does the "I "of "I-decomposable "comes from. It comes from "Independent "and we are just trying to keep in mind that X is a process that we can decompose on two **independent** random components, one that is mainly "local ", "weakly dependent "(ξ) and the other that we can control "in the mean "(the ponderable process Y). The reader can also wonder if there is an intrinsic description of what kind of processes X admits an I-decomposable regression (i.e., what are the conditions required to X to know that there is a function φ and two processes ξ, Y as in Definition 5 such that $X = \varphi(\xi, Y)$. Unfortunately, authors do not know the answer for that question.

Example 2 Let us present here a very simple example of a process that satisfies Definition 5. Consider $U = (U_t)_{t\in\mathbb{N}}$ *iid* and such that its common probability distribution μ is absolutely continuous. Let V be a random variable assuming a finite number of values (say $V(\omega) \in \{1, ..., S\}\ \forall\omega$) and independent of U and let $(a_t)_{t\in\mathbb{N}}$ be a sequence of real numbers satisfying $\lim_t a_t = a$. Define $Y_t = U_t + a_t V$.

Let us check that Y is ponderable:

consider $B_1, ..., B_d$ Borel real sets, $d \geq 1$, $\vec{r} \in \mathbb{N}^d$ and set $\Omega_i = \{\omega : V(\omega) = i\}$, $i = 1, ..., S$. Then, if $\omega \in \Omega_i$, we have

$$\frac{1}{n}\sum_{t=0}^{n} 1_{\{Y_{t+r_j}(\omega)\in B_j\ \forall j\}} = \frac{1}{n}\sum_{t=0}^{n} 1_{\{U_{t+r_j}(\omega)\in(B_j - a_t i)\ \forall j\}}$$

Without loss of generality we can assume that $r_j \neq r_h\ \forall j \neq h$; observe then that

$$E\left(1_{\{U_{t+r_j}(\omega)\in(B_j - a_t i)\ \forall j\}}\right) = \prod_{j=1}^{d} \mu(B_j - a_t i)$$

and hence, by an elementary computation

$$\lim_t E\left(1_{\{U_{t+r_j}(\omega)\in(B_j - a_t i)\ \forall j\}}\right) = \prod_{j=1}^{d} \mu(B_j - ai)$$

since U is *iid* a Borel-Cantelli argument shows that, over Ω_i,

$$\lim_n \frac{1}{n}\sum_{t=0}^{n} 1_{\{Y_{t+r_j}(\omega)\in B_j\ \forall j\}} = \prod_{j=1}^{d} \mu(B_j - ai)\ a.s.$$

and thus

$$\lim_n \frac{1}{n}\sum_{t=0}^{n} 1_{\{Y_{t+r_j}(\omega)\in B_j\ \forall j\}} = \prod_{j=1}^{d} \mu(B_j - aV(\omega))\ a.s.$$

what means that Y is ponderable and that $\mu^{\vec{r}}(B_1 \times B_2 \times \cdots \times B_d)(\omega) = \prod_{j=1}^{d} \mu(B_j - aV(\omega))$; in particular, if $a = 0$, then Y is regular.

Consider now $\varphi(\xi, y) = \xi g(y)$ with g a real, bounded and positive function and ξ a moving average of *iid* Cauchy variables.

It is easy to check that $X_t = \varphi(\xi_t, Y_t)$ satisfies Definition 5. Indeed, the reader can find in Section 4, and in particular, in the Example 4 of Section 4, a precise guide to show that ξ is a φ-noise; conditions (a) to (d) of Definition are obtained by elementary computations.

This concrete example shows clearly what are the essential properties required to ξ, Y and φ to make $X_t = \varphi(\xi_t, Y_t)$ an example of Definition 5.

Concerning Y we need a sort of "stationarity in the mean", in the sense that the occupation measures

$$\mu_{\vec{r}n}(C) = \frac{1}{n}\sum_{t=1}^{n} 1_{\{Y_t(\vec{r})\in C\}}$$

must converge. Those asymptotic occupation measures will be not-random if Y has "short-range memory".

Concerning ξ we need weak dependence, like the conditions presented on Section 4, and a good knowledge of the conditional distribution of ξ given ξ_0.

Finally, φ must really depend on the first variable and it must be a smooth function (more precisely, we require a smooth control on the second variable of the probability tails of $\varphi(\xi_0, y)$).

The main result of this paper is the following

Theorem 1 *If X admits an I-decomposable regression, then:*
(a) If Y is regular, X satisfies the CPLT. More precisely, if $X = \varphi(\xi, Y)$, and

$$N_n(B) = \sum_{\frac{t}{n} \in B} 1_{\{X_t > u_n\}},\ B \in \mathcal{B},$$

then N_n converges in law to N, Compound Poisson process with Laplace transform

$$L(B;x) = exp\left(m(B) \sum_{j=1}^{\infty} \nu_j (e^{-xj} - 1) \right),\ with\ \nu_j = \sum_{d=j}^{\infty} (-1)^{j+d} C_j^d \int_{\mathbb{R}^d} \sum_{\vec{r} \in \mathbb{N}_L^{d-1}} a(\vec{r}, \vec{y}) \lambda(y_0) \mu^{\vec{r}}(dy)\ \forall j \in \mathbb{N}$$

(b) If Y is not regular, then N_n/Y satisfies the CPLT and N_n converges weakly to a mixture of Compound Poisson processes

3 CPLT over asymptotically ponderable collections.

In this sections we will obtain the CPLT that we need to check that Definition 4 holds.
Consider now a stationary process $\xi = \{\xi_n : n \in \mathbb{N}\}$, an APS A and set

$$X_t = \xi_t 1_A(t) \tag{4}$$

Let us remind that the sequence u_n satisfies that

$$\lim_n nP(\{\xi_0 > u_n\}) = \lambda > 0 \tag{5}$$

A little bit more of notation: if B is any set and $k \in \mathbb{N}$, then $\mathcal{C}_k(B)$ will denote the collection of all the subsets of B with k elements, i.e.,

$$\mathcal{C}_k(B) = \{D \subset B : card(D) = k\}$$

We will present now some auxiliary results; in particular, we will introduce in the following two lemmata the coefficients $\Theta(s,d)$.

Lemma 1 *Let B be a subset of $\mathbb{N}$, X as in (4), $u > 0$ and define*

$$N_n^*(B) = \sum_{t \in B_n} 1_{\{X_t > u\}} \tag{6}$$

Then

$$P(\{N_n^*(B) \geq s\}) = \sum_{d=s}^{d=card(B_n \cap A)} \Theta(s;d) \sum_{\vec{r} \in [1, card(B_n \cap A)]_L^{d-1}} P(\{\xi(\vec{r}) > u\})\, card\,(T_n(\vec{r}; B \cap A)) \tag{7}$$

where

$$\Theta(s;d) = \sum_{k=1}^{\infty} (-1)^{k-1} \theta(k,s;d),\ \theta(k,s;d) = card\,(\mathcal{C}_k(\mathcal{C}_s(\{1, \cdots, d\}))) \tag{8}$$

Remark 8

(a) In fact the sum in (8) is finite: if $\theta(k,s;d) > 0$, then there exists a decomposition $\{1,\cdots,d\} = \bigcup_{j=1}^{j=k} I_j$, with $card(I_j) = s$, $I_j \neq I_h$ if $j \neq h$; this implies that $sk \geq d$, $k \leq C_s^d$. Therefore, $\theta(k,s;d) = 0$ if $k < \frac{d}{s}$ or $k > C_s^d$.
(b) In particular, if $s > d$, then $\Theta(s;d) = 0$.
(c) If $d = s = 1$, the corresponding term on the sum (7) must be intrepreted as $P(\{\xi_0 > u\})\, card\,(B_n \cap A)$.
(d) Indeed, we can also write down

$$P(\{N_n^*(B) \geq s\}) = \sum_{d=s}^{\infty} \Theta(s;d) \sum_{\vec{r} \in \mathbb{N}_L^{d-1}} P(\{\xi(\vec{r}) > u\})\, card\,(T_n(\vec{r}; B \cap A))\,,$$

because the extra terms (with respect to (7)) are null.

Proof of Lemma 1:

First, observe that with our notation the elementary inclusion-exclusion formula is:

$$P\left(\bigcup_{\gamma\in\Gamma} A_\gamma\right) = \sum_{k=1}^{k=card(\Gamma)} (-1)^{k-1} \sum_{C\in\mathcal{C}_k(\Gamma)} P\left(\bigcap_{\gamma\in C} A_\gamma\right),$$

for any finite Γ.
Therefore, we have:

$$\begin{aligned} P(\{N_n^*(B) \geq s\}) &= P\left(\bigcup_{I\in\mathcal{C}_s(B_n\cap A)} \{\xi(I) > u\}\right) \\ &= \sum_{k=1}^{k=card(\mathcal{C}_s(B_n\cap A))} (-1)^{k-1} \sum_{C\in\mathcal{C}_k(\mathcal{C}_s(B_n\cap A))} P\left(\bigcap_{I\in C} \{\xi(I) > u\}\right) \\ &= \sum_{k=1}^{k=card(\mathcal{C}_s(B_n\cap A))} (-1)^{k-1} \sum_{C\in\mathcal{C}_k(\mathcal{C}_s(B_n\cap A))} P\left(\{\xi(\bigcup_{I\in C} I) > u\}\right) \end{aligned}$$

$$= \sum_{k=1}^{k=card(\mathcal{C}_s(B_n\cap A))} (-1)^{k-1} \sum_{d=1}^{d=\infty} \sum_{H\in\mathcal{C}_d(B_n\cap A)} P(\{\xi(H) > u\}) \times card\left(\{C \in \mathcal{C}_k(\mathcal{C}_s(B_n \cap A)) \bigcup_{I\in C} I = H\}\right)$$

But, if $H \in \mathcal{C}_d(B_n \cap A)$, then

$$card\left(\{C \in \mathcal{C}_k(\mathcal{C}_s(B_n \cap A)) : \bigcup_{I\in C} I = H\}\right) = \theta(k,s;d)$$

and, on the other hand, by the stationarity of ξ, we get :

$$P(\{\xi(H) > u\}) = P(\{\xi(\vec{r}(H)) > u_n\}),$$

where , if $H = \{h_1, \cdots, h_d\}$ such that $h_i < h_{i+1}\ \forall i$, then :

$$\vec{r}(H) := (h_2 - h_1, h_3 - h_1, \cdots, h_d - h_1) \in [1, card(B_n \cap A)]_L^{d-1}.$$

It follows that :

$$P(\{N_n^*(B) \geq s\})$$

$$= \sum_{k=1}^{k=card(\mathcal{C}_s(B_n \cap A))} (-1)^{k-1} \sum_{d=1}^{d=\infty} \theta(k,s;d) \sum_{\vec{r} \in [1, card(B_n \cap A)]_L^{d-1}} P(\{\xi(\vec{r}) > u\})\, card(\{H \in \mathcal{C}_d(B_n \cap A) : \vec{r}(H) = \vec{r}\})$$

but an elementary argument shows that :

$$card(\{H \in \mathcal{C}_d(B_n \cap A)\,; \vec{r}(H) = \vec{r}\}) = card(T_n(\vec{r}; B \cap A))$$

and, using Remark 2, the Lemma follows $\diamond$

Next lemma will be used to simplify some formulae.

Lemma 2 *If* $s \leq d$, $\Theta(s;d) = (-1)^{d-1} \sum_{j=0}^{j=s-1} C_j^d (-1)^j$.

Proof of Lemma 2:

Pick $\rho \in (0,1)$ arbitrary and take $B = A = \mathbb{N}$, ξ *iid* and $u > 0$ such that $P(\{\xi_0 > u\}) = \rho$.

Applying Lemma 1, we obtain:

$$P(\{N_n^*(\mathbb{N}) \geq s\}) = \sum_{d=s}^{d=n} \Theta(s;d) \sum_{\vec{r} \in [1,n]_L^{d-1}} \rho^d card(T_n(\vec{r}; \mathbb{N})) \tag{9}$$

On the other hand, using that $N_n^*(\mathbb{N}) \sim Bin(n, \rho)$ and the binomial expansion for $(1-\rho)^k$, we have that

$$P(\{N_n^*(\mathbb{N}) \geq s\}) = \sum_{m=s}^{m=n} C_m^n \rho^m (1-\rho)^{n-m} = \sum_{m=s}^{m=n} C_m^n (-1)^m \left(\sum_{j=0}^{j=n-m} C_j^{n-m} (-\rho)^j \right) \rho^m \tag{10}$$

Since (9) and (10) give two polynomial expansions of the same function for $\rho \in (0,1)$, both polynomials are identical, and hence, we can equate its coefficients. We obtain:

$$\Theta(s;d) = \left(\sum_{\vec{r} \in [1,n]_L^{d-1}} card(T_n(\vec{r}; \mathbb{N})) \right)^{-1} C_d^n (-1)^d \left(\sum_{m=0}^{m=s-1} C_m^d (-1)^m \right) \tag{11}$$

Since the left-side term of (11) does not depend on n, we take the limit of the right-side for n tending to infinity, and, after an elementary computation, using that $\sum_{j=0}^{j=d} C_j^d(-1)^j = 0$, we obtain:

$$\begin{aligned}\Theta(s;d) &= (-1)^d \sum_{j=s}^{j=d} C_j^d(-1)^j \\ &= (-1)^{d-1} \sum_{j=0}^{j=s-1} C_j^d(-1)^j.\end{aligned}$$

and Lemma 2 follows $\diamond$.

Remark 9

Lemma 2 and a trivial computation show that, for s fix and d tending to infinity,

$$\Theta(s;d) \approx (-1)^{d+s} \frac{d^{s-1}}{(s-1)!}.$$

and that for any s, d

$$|\Theta(s;d)| \leq \frac{d^{s-1}}{(s-2)!}$$

We will introduce now a weak-dependence hypothesis:

(H1) There exist two non-decreasing sequences $(p_n)_{n\in\mathbb{N}}$, $(q_n)_{n\in\mathbb{N}}$, such that $\lim_{n\to\infty} p_n = \lim_{n\to\infty} q_n = \infty$, $\lim_{n\to\infty} \frac{p_n}{n} = \lim_{n\to\infty} \frac{q_n}{p_n} = 0$ and that satisfy:

if $B_n^i = [(i-1)(p_n+q_n), (i-1)(p_n+q_n)+p_n)$ then for any $B_1, \cdots, B_k \subset \mathbb{N}$ and $k \in \mathbb{N}$, $(N_n^*(B_1), \cdots, N_n^*(B_k))$ has the same asymptotic distribution as $(\hat{N}_n(B_1), \cdots, \hat{N}_n(B_k))$, where $\hat{N}_n(B) = \sum_{i=1}^{i=k_n} Z_n^i(B)$, with $k_n = int\left(\frac{n}{p_n+q_n}\right)$ and $(Z_n^i(B))_{1\leq i\leq k_n}$ are independent copies of $(N_n^*(B_n^i \cap B))_{1\leq i\leq k_n}$.

Remark 10

(a)It is very easy to check that mixing assumptions guarantee assumption **(H1)**. More precisely, let $s, t \in \mathbb{N}$ and

$$\Sigma_n(s,t) = \sigma(\{\{\xi_i > u_n\} : s \leq i \leq t\})$$

$$\alpha_{n,l} = sup\{|P(A\cap B) - P(A)P(B)| : A \in \Sigma_n(h, s+h), B \in \Sigma_n(s+h+l, n), h \geq 0, s+l+h < n\}$$

We will say that $N^* = \{N_n^*\}_{n\in\mathbb{N}}$ is strongly mixing if there exists a non-decreasing sequence $(q_n)_{n\in\mathbb{N}}$ such that $\lim_n q_n = \infty$ and $\lim_n \alpha_{n,q_n} = \lim \frac{q_n}{n} = 0$.

Then, it is straightforward to check that we can choose $(k_n)_{n\in\mathbb{N}}$ such that $\lim_n k_n = \lim \frac{n}{k_n+q_n} = \infty$ and $\lim_n k_n \alpha_{n,q_n} = 0$. Then, taking $p_n = \frac{n}{k_n}$, it follows that **(H1)** is satisfied (see [Hsing, Hüsler & Leadbetter (1988)], Lemma 2.2 for a detailed proof or [Leadbetter & Rootzén (1993)], Lemma 2.1; see [Bradley (1986)], [Doukhan, P. (1995)] for mixing conditions, exemples and covariance inequalities).

(b) We say that a random vector $(X_1, \cdots, X_n)$ is **positively associated** (or that it satisfies the FKG inequality) if, for any coordinatewise non-decreasing functions $g, h : \mathbb{R}^n \to \mathbb{R}$ such that $E\left(g(X_1, \cdots, X_n)^2\right) < \infty$, $E\left(h(X_1, \cdots, X_n)^2\right) < \infty$, we have that

$$Cov(g(X_1, \cdots, X_n), h(X_1, \cdots, X_n)) \geq 0$$

If for any nonempty subset B of $\{1, \cdots, n\}$ and any pair of coordinatewise non-decreasing functions, $g : \mathbb{R}^B \to \mathbb{R}$, $h : \mathbb{R}^{B^c} \to \mathbb{R}$ such that $E\left(g(X_t : t \in B)^2\right) < \infty$, $E\left(h(X_t : t \in B^c)^2\right) < \infty$, we have that

$$Cov(g(X_t : t \in B), h(X_t : t \in B^c)) \leq 0$$

we say that $(X_1, \cdots, X_t)$ is **negatively associated**.

A random process is said to be positively (negatively) associated if all its finite-dimensional projections are positively (negatively) associated. We will say that a random process is **associated** if it is either positively associated or negatively associated. Association is preserved by non-decreasing functions and restrictions of associated process are associated, This implies that if ξ is associated, then X, given by (4), is also associated.

If $(X_1, \cdots, X_n)$ is an associated vector whose coordinates have finite second moments, we have the following inequality:

$$|E\left(exp\{is\sum_{t=1}^{t=n} X_t\}\right) - \prod_{t=1}^{t=n} E\left(isX_t\right)| \leq \frac{s^2}{2} \sum_{1\leq k\neq t\leq n} |Cov(X_k, X_t)|$$

(For the proof see [Newman (1980)], Theorem 1, for positive association, and [Roussas (1994)], Proposition 3.1, for negative association).

Turning back to our process defined by (6), it is easy to check that if X is associated, the vector $(N_n^*(B_n^i))_{1\leq i\leq k_n}$ is associated.

Therefore, we deduce that **(H1)** holds if ξ is associated and

$$\lim_n \sum_{1\leq k\neq j\leq k_n} |Cov(N_n^*(B_n^k), N_n^*(B_n^j))| = 0.$$

But

$$\begin{aligned}\sum_{1\leq k\neq j\leq k_n} |Cov(N_n^*(B_n^k), N_n^*(B_n^j))| &= \sum_{1\leq k\neq j\leq_n} | \sum_{t\in B_n^k\cap A, s\in B_n^j\cap A} Cov(1\{\xi_0 > u_n\}, 1\{\xi_{s-t} > u_n\})| \\ &\leq \sum_{m=q_n}^{\infty} n|Cov(1\{\xi_0 > u_n\}, 1\{\xi_m > u_n\})|\end{aligned}$$

Hence, if ξ is associated and

$$\lim_n \sum_{m=q_n}^{\infty} n|Cov(1\{\xi_0 > u_n\}, 1\{\xi_m > u_n\})| = 0$$

then **(H1)** holds.

Similar conditions can be obtained for other weak-dependence structures. The reader will find in [Doukhan & Louichi (1996)] a summary of different weak-dependence conditions where **(H1)** can be proved in a similar way.

In the sequel, we will assume that :

$$\textbf{(H2)} \quad \limsup_n \sum_{m=1}^{\infty} n|Cov(1\{\xi_0 > u_n\}, 1\{\xi_m > u_n\})| < \infty$$

Proposition 1 *Let X be as in (4), and assume that (5),* **(H1)** *and* **(H2)** *hold. For any $s \in \mathbb{N}$, define*

$$Q_n(s) = \sum_{d=s}^{d=p_n} \Theta(s;d) \sum_{\vec{r} \in [1,p_n]_L^{d-1}} P\left(\{\xi(\vec{r}) > u_n\}/\{\xi_0 > u_n\}\right) F_n(\vec{r}; A),$$

where

$$F_n(\vec{r}; A) = \frac{1}{n} \sum_{1 \le i \le k_n : card(B_n^i \cap A) \ge d} card\left(T_n(\vec{r}, B_n^i \cap A)\right)$$

Assume further that, for any $s \in \mathbb{N}$,

$$\lim_n Q_n(s) \quad = \quad Q(s) \tag{12}$$

Then N_n converges in law (as a process) to a $CP(\nu)$ process, where

$$\nu_s = \lambda(Q(s) - Q(s+1)) \; \forall s \in \mathbb{N}.$$

Proof of Proposition 1:

By Theorem 4.2 of [Kallenberg (1983)], it suffices to show that for any k-tuple of semiclosed intervals $I_1, \cdots, I_k$, the random vector $(N_n(I_1), \cdots, N_n(I_k))$ converges in law to $(N(I_1), \cdots, N(I_k))$, where N is a $CP(\nu)$ process. Without loss of generallity, we may assume that $I_1, \cdots, I_k$ are disjoint. In this case, by **(H1)**, the coordinates of the random vector $(N_n(I_1), \cdots, N_n(I_k))$ are asymptotically independent, and therefore, it suffices to show that for any semiclosed interval I we have that $N_n(I)$ converges in law to $N(I)$. But an elementary argument shows that it is enough to consider the case $I = (0, a]$, $0 < a < 1$. Finally, this can be reduced, by a scale change, to check that

$$N_n((0,1]) \xrightarrow[n]{w} N((0,1]),$$

where N is a random variable whith Laplace transform

$$L(s) = exp\left(\sum_{j=1}^{\infty} \nu_j (e^{-sj} - 1)\right)$$

For the rest of this proof, $N_n((0,1])$ will be simply denoted by N_n.

By **(H1)** N_n is asymptotically equivalent to $\hat{N}_n = \sum_{i=1}^{i=k_n} Z_n^i$, with $(Z_n^i)_{1\leq i\leq k_n}$ independent copies of $(N_n^*(B_n^i))_{1\leq i\leq k_n}$.

We will first prove that $(\hat{N}_n)_{n\in\mathbb{N}}$ is tight. To prove that, we will show that $E(\hat{N}_n)$ and $Var(\hat{N}_n)$ are bounded, what implies the uniform integrability of $(\hat{N}_n)_{n\in\mathbb{N}}$, hence its tightness. $E(\hat{N}_n)$ is obviously bounded by the convergent sequence $nP(\{\xi_0 > u_n\})$; therefore $E(\hat{N}_n)$ is bounded. For the variance, we have:

$$Var(\hat{N}_n) = \sum_{i=1}^{i=k_n} Var\left(N_n^*(B_n^i)\right) = \sum_{i=1}^{i=k_n} \sum_{s,t\in B_n^i\cap A} Cov\left(1_{\{\xi_0>u_n\}}, 1_{\{\xi_{t-s}>u_n\}}\right)$$

$$= \sum_{i=1}^{i=k_n} card\left(B_n^i \cap A\right) P\left(\{\xi_0 > u_n\}\right)\left(1 - P\left(\{\xi_0 > u_n\}\right)\right)$$

$$+2\sum_{i=1}^{i=k_n}\sum_{m=1}^{m=p_n} Cov\left(1_{\{\xi_0>u_n\}}, 1_{\{\xi_m>u_n\}}\right) card\left(B_n^i \cap A \cap (B_n^i \cap A - m)\right)$$

The first term on the last expression is bounded by $nP(\{\xi_0 > u_n\})$ again, so it is bounded. The last term is also bounded, by **(H2)** and the fact that $card(B_n^i) = p_n$. Therefore, the variance is also bounded and tightness follows.

Given any subsequence $(\hat{N}_{n_h})_{h\in\mathbb{N}}$ pick a sub-subsequence that is weakly convergent to some random variable W. To simplify the notation, we will still call $(\hat{N}_{n_h})_{h\in\mathbb{N}}$ to this second subsequence. Since the law of W must be infinitely divisible and concentrated over $\mathbb{N}$, its Laplace transform is

$$H(s) = exp\left(\sum_{j=1}^{\infty} \pi_j(e^{-sj} - 1)\right)$$

Thus, it suffices to show that $\pi_j = \nu_j\ \forall j \in \mathbb{N}$. Observe that

$$P\left(\{N_n^*(B_n^i) \neq 0\}\right) \leq \sum_{k\in A\cap B_n^i} P\left(\{\xi_k > u_n\}\right) \leq p_n P\left(\{\xi_0 > u_n\}\right)$$

and therefore,

$$\lim_n P\left(\{N_n^*(B_n^i) \neq 0\}\right) = 0.$$

But, from Theorem 3 of [Rényi (1951)] (see also [Dziubdziela (1995)]) we deduce that

$$\pi_j = \lim_h \sum_{i=1}^{i=k_{n_h}} P\left(\{N_{n_h}^*(B_{n_h}^i) = j\}\right)\ \forall j$$

But (12) and Lemma 1 imply that

$$\nu_j = \lim_n \sum_{i=1}^{i=k_n} P\left(\{N_n^*(B_n^i) = j\}\right)$$

Therefore $\nu_j = \pi_j\ \forall j$ and the Lemma is proved $\diamond$

Corollary 1 *Assume that A is an APS and that X is as in Proposition 1. Assume further:*

(H3) $\forall d \in \mathbb{N},\ \forall \vec{r} \in \mathbb{N}_L^{d-1}, \lim_n P\left(\{\xi(\vec{r}) > u_n\}/\{\xi_0 > u_n\}\right) = a(\vec{r})$,

(H4) $\forall d \in \mathbb{N},\ \forall \vec{r} \in \mathbb{N}_L^{d-1}, P\left(\{\xi(\vec{r}) > u_n\}/\{\xi_0 > u_n\}\right) = a_n(\vec{r}) + b_n(\vec{r})$
where :

$$a_n(\vec{r}) \leq c(\vec{r})\ \forall n \in \mathbb{N}, \vec{r} \in \mathbb{N}_L^{d-1}$$

$$\sum_{d=s}^{\infty} |\Theta(s;d)| \sum_{\vec{r} \in \mathbb{N}_L^{d-1}} c(\vec{r}) < \infty$$

$$\lim_n \sum_{d=s}^{d=p_n} |\Theta(s;d)| \sum_{\vec{r} \in [1,p_n]_L^{d-1}} |b_n(\vec{r})| = 0.$$

Then N_n converges in law (as a process) to a $CP(\nu)$ process, where :

$$\nu_s = \lambda \sum_{d=s}^{\infty} (-1)^{s+d} C_s^d \sum_{\vec{r} \in \mathbb{N}_L^{d-1}} a(\vec{r}) F(\vec{r}; A)$$

Proof of Corollary 1:
We have

$$Q_n(s) = \sum_{d=s}^{d=p_n} \Theta(s;d) \sum_{\vec{r} \in [1,p_n]_L^{d-1}} a_n(\vec{r}) F_n(\vec{r}; A) + \sum_{d=s}^{d=p_n} \Theta(s;d) \sum_{\vec{r} \in [1,p_n]_L^{d-1}} b_n(\vec{r}) F_n(\vec{r}, A)$$

Since F_n is bounded by 1, the second term in the right-side of the last expression, converges, by **(H4)** to zero. Therefore, it suffices to show that

$$\lim_n \sum_{d=s}^{d=p_n} \Theta(s;d) \sum_{\vec{r} \in [1,p_n]_L^{d-1}} a_n(\vec{r}) F_n(\vec{r}; A) \quad = \quad \sum_{d=s}^{\infty} \Theta(s;d) \sum_{\vec{r} \in \mathbb{N}_L^{d-1}} a(\vec{r}) F(\vec{r}; A) \tag{13}$$

and that

$$\sum_{d=s}^{\infty} \Theta(s;d) \sum_{\vec{r} \in \mathbb{N}_L^{d-1}} a(\vec{r}) F(\vec{r}; A) - \sum_{d=s+1}^{\infty} \Theta(s+1;d) \sum_{\vec{r} \in \mathbb{N}_L^{d-1}} a(\vec{r}) F(\vec{r}; A) = \sum_{d=s}^{\infty} (-1)^{s+d} C_s^d \sum_{\vec{r} \in \mathbb{N}_L^{d-1}} a(\vec{r}) F(\vec{r}; A) \tag{14}$$

Let su first consider (13): assume that we have shown that

$$\lim_n F_n(\vec{r}; A) \quad = \quad F(\vec{r}, A) \forall \vec{r} \in \mathbb{N}^m, \forall m \tag{15}$$

Then, by **(H3)**, **(H4)** and Dominated Convergence, (13) follows. We will then prove (15).
Fix $d \in \mathbb{N}$ and $\vec{r} \in \mathbb{N}^{d-1}$ and let n be big enough such that $q_n > 2\|\vec{r}\|$.
Then, if $i \neq h$, $(B_n^i - r_j) \cap (B_n^h - r_k) = \emptyset$, $\forall j, k$. Thus,

$$card\left(T_n(\vec{r}; A)\right) \quad = \quad \sum_{i=1}^{i=k_n} card\left(T_n(\vec{r}; B_n^i \cap A)\right) + \Delta_n(\vec{r}) \tag{16}$$

where $\Delta_n(\vec{r})$ is the sum of the cardinals of all the sets of the form

$$\bigcap_{i=1}^{i=d-1} (K_i - r_i) \cap K_0 \tag{17}$$

where eack K_i is either:

- $B_n^j \cap A$ for some j, and if for some i, $K_i = B_n^j \cap A$, then there is not an h such that $K_h = B_n^k \cap A$, with $j \neq k$.
- $H_n^j \cap A$ for some j, where H_n^j is one of the "holes" $[(j-1)(p_n+q_n)+p_n, j(p_n+q_n) \wedge n)$. Furthermore, for at least one i this choice is taken.

Since the cardinal of the "holes" is at most q_n, the cardinal of a set of the form (17) is bounded by q_n. Taking into account that $\Delta_n(\vec{r})$ is the sum of at most $(k_n+1)(d-1)$ cardinals of sets of the form (17), we conclude that

$$\max_{\|\vec{r}\|<q_n/2} \Delta_n(\vec{r}) \quad \leq \quad q_n(k_n+1)(d-1) \tag{18}$$

Applying (18) and Definition 1 in (16) we obtain

$$\lim_n \frac{1}{n} \sum_{i=1}^{i=k_n} card\left(T_n(\vec{r}; B_n^i \cap A)\right) = F(\vec{r}; A).$$

But

$$\left|\frac{1}{n} \sum_{i=1}^{i=k_n} card\left(T_n(\vec{r}; B_n^i \cap A)\right) - F_n(\vec{r}; A)\right| \leq \frac{1}{n} \sum_{1 \leq i \leq k_n : card(B_n^i \cap A) \leq d} card\left(T_n(\vec{r}, B_n^i \cap A)\right)$$

$$\leq \frac{dk_n}{n} \underset{n}{\to} 0$$

and (15) follows.

Let us now turn to (14); we have that

$$\sum_{d=s}^{\infty} \Theta(s;d) \sum_{\vec{r} \in \mathbb{N}_L^{d-1}} a(\vec{r}) F(\vec{r}; A) - \sum_{d=s+1}^{\infty} \Theta(s+1;d) \sum_{\vec{r} \in \mathbb{N}_L^{d-1}} a(\vec{r}) F(\vec{r}; A) =$$

$$\Theta(s,s) \sum_{\vec{r} \in \mathbb{N}_L^{s-1}} a(\vec{r}) F(\vec{r}; A) + \sum_{d=s+1}^{\infty} (\Theta(s;d) - \Theta(s+1;d)) \sum_{\vec{r} \in \mathbb{N}_L^{d-1}} a(\vec{r}) F(\vec{r}; A)$$

But, by Lemma 2

$$\Theta(s;d) - \Theta(s+1;d) = \sum_{j=0}^{j=s-1} C_j^d (-1)^{j+d-1} - \sum_{j=0}^{j=s} C_j^d (-1)^{j+d-1} = -C_s^d (-1)^{s+d-1} = C_s^d (-1)^{s+d}$$

and, taking into account that $\sum_{j=0}^{j=s} C_j^s (-1) j = 0$, we get

$$\Theta(s,s) = \sum_{j=0}^{j=s-1} C_j^s (-1)^{j+s-1} = (-1)^{s+s} C_s^s = 1$$

we deduce (14) $\diamond$

Remark 11

$a(\vec{r})$ describes the probability of a cluster of exceedances, located at points whose distances are given by $\vec{r}$. In particular, $\sum_{\vec{r}\in\mathbb{N}_L^{d-1}} a(\vec{r})$ represents the probability of a cluster of size d; for $A = \mathbb{N}$, by Corollary 1, the result depends only on this "size cluster distribution", as proved in [Hsing, Hüsler & Leadbetter (1988)].

Remark 12

It seems reasonable to conjecture that if the set A is not an APS, then there exist an m-dependent random process X such that the CPLT does not hold, but authors do not know a proof for this statement.

4 Examples

4.1 m-dependent processes

Assume that ξ is a stationary m-dependent process and assume that (5) holds. Then ξ satisfies **(H1)** for any $(p_n)_{n\in\mathbb{N}}$, $(q_n)_{n\in\mathbb{N}}$, such that $\lim_n p_n = \lim_n q_n = \infty$, $\lim_n \frac{p_n}{n} = \lim_n \frac{q_n}{p_n} = 0$. In particular, we will choose $p_n = \sqrt{logn}$. Cauchy-Schwarz inequality gives **(H2)**. Let A be an APS and take X as in (4). Assume that **(H3)** holds. We will prove that **(H4)** holds, and therefore, Corollary 1 applies.

Assume that $\vec{r} \in \mathbb{N}_L^{d-1}$ is such that $\|\vec{r}\| > m$. Then $r_{d-1} > m$ and

$$P\left(\{\xi(\vec{r}) > u_n\}/\{\xi_0 > u_n\}\right) \leq P\left(\{\xi_0 > u_n, \xi_{r_{d-1}} > u_n\}/\{\xi_0 > u_n\}\right) = P\left(\{\xi_0 > u_n\}\right)$$

that goes to zero with n.

Define then :

$$\begin{aligned} a_n(\vec{r}) &= P\left(\{\xi(\vec{r}) > u_n\}/\{\xi_0 > u_n\}\right) 1_{\{\|\vec{r}\|\leq m\}} \\ b_n(\vec{r}) &= P\left(\{\xi(\vec{r}) > u_n\}/\{\xi_0 > u_n\}\right) 1_{\{\|\vec{r}\|> m\}} \\ c(\vec{r}) &= 1_{\{\|\vec{r}\|\leq m\}}. \end{aligned}$$

Observe first that if $\vec{r} \in \mathbb{N}_L^{d-1}$ and $d-1 > m$, then $\|\vec{r}\| > m$, thus :

$$\sum_{d=s}^{\infty} \Theta(s;d) \sum_{\vec{r}\in\mathbb{N}_L^{d-1}} c(\vec{r}) = \sum_{d=s}^{d=m+1} \Theta(s;d) card\left([1,m]_L^{d-1}\right) < \infty.$$

On the other hand, we have, by Remark 3 and (5) :

$$\sum_{d=s}^{d=p_n} |\Theta(s;d)| \sum_{\vec{r}\in[1,p_n]_L^{d-1}} b_n(\vec{r}) \leq C \sum_{d=s}^{d=p_n} d^{s-1} card\left([1,p_n]_L^{d-1}\right) P\left(\{\xi_0 > u_n\}\right)$$

$$= \frac{C}{n} \sum_{d=s}^{d=p_n} d^{s-1} \frac{p_n^{d-1}}{(d-1)!} \leq \frac{C}{n} p_n^{s-1} exp(p_n) \underset{n}{\to} 0$$

Hence, if ξ is stationary and m-dependent, A an APS and (4), (5), **(H3)** hold then the CPLT holds, and the limit is described by :

$$\nu_s = \sum_{d=s}^{d=m+1} (-1)^{s+d} C_s^d \sum_{\vec{r}\in[1,m]L^{d-1}} a(\vec{r}) F(\vec{r};A).$$

We will now give some specific exemples of this result.

Example 3

Consider a stationary, standardized, m-dependent gaussian process ξ, A any APS and assume (4) and (5). We will prove that **(H3)** holds, that $a(\vec{r}) = 0\ \forall \vec{r} \in \mathbb{N}_n^{d-1}$, $d > 1$ and therefore, that the limit process is simply a Poisson process of intensity λ.

Take $d > 1$ and fix $\vec{r} \in \mathbb{N}_L^{d-1}$. Pick $\delta > 0$; we get:

$$P\left(\{\xi(\vec{r}) > u_n\}/\{\xi_0 > u_n\}\right) \leq P\left(\{\xi_0 > u_n, \xi_{r_{d-1}} > u_n\}/\{\xi_0 > u_n\}\right)$$

$$= P\left(\{\xi_0 > u_n + \delta, \xi_{r_{d-1}} > u_n\}/\{\xi_0 > u_n\}\right) + P\left(\{u_n < \xi_0 \leq u_n + \delta, \xi_{r_{d-1}} > u_n\}/\{\xi_0 > u_n\}\right)$$

For the first term of the right-hand side we have the bound

$$\frac{P\left(\{\xi_0 > u_n + \delta\}\right)}{P\left(\{\xi_0 > u_n\}\right)} \underset{n}{\to} 0$$

For the second term, let ρ be the correlation coefficient between ξ_0 and $\xi_{r_{d-1}}$. Since ξ from its definition is stationary and m-dependent, $|\rho| < 1$. Then, since $\frac{1}{\sqrt{1-\rho^2}}[\xi_{r_{d-1}} - \rho\xi_0]$ is a standard gaussian variable independent of ξ_0, we obtain :

$$P\left(\{u_n < \xi_0 \leq u_n + \delta, \xi_{r_{d-1}} > u_n\}/\{\xi_0 > u_n\}\right) \leq$$

$$P\left(\{u_n < \xi_0 \leq u_n + \delta, \frac{1}{\sqrt{1-\rho^2}}[\xi_{r_{d-1}} - \rho\xi_0] > \frac{1}{\sqrt{1-\rho^2}}[u_n - \rho(u_n + \delta)]\}/\{\xi_0 > u_n\}\right)$$

$$= \frac{P\left(\{u_n < \xi_0 \leq u_n + \delta\}\right)}{P\left(\{\xi_0 > u_n\}\right)} P\left(\{Z > \frac{1}{\sqrt{1-\rho^2}}[u_n(1-\rho) - \rho\delta]\}\right)$$

The first term on the last expression goes to one and the second to zero, as n goes to infinity, so $a(\vec{r}) = 0$.

Example 4

In this case we will present a very simple example where the limit is a Compound Poisson process, but not Poisson. Let A be an APS such that $F(1;A) > 0$. Consider an *iid* sequence $(\zeta_k)_{k\in\mathbb{N}}$ that follows the Cauchy distribution. Set $\xi_k = \zeta_k + \zeta_{k+1}$. It is clear that ξ is stationary and 1-dependent. In addition, ξ_0 has the same distribution as $2C$ where C stands for a Cauchy variable. Indeed, as we have seen at the beginning of this section, we only need to show **(H3)** for $||\vec{r}|| \leq m$. Setting X as in (4) and u_n as in (5), it suffices to show that

$$\lim_n P\left\{\xi_1 > u_n\}/\{\xi_0 > u_n\}\right) = a(1) \in \left[0, \frac{1}{2}\right] \tag{19}$$

Indeed, if (19) holds, then the limit process satisfies :

$$\begin{aligned} \nu_1 &= \lambda\left(F(0;A) - 2a(1)F(1;A)\right) > 0 \\ \nu_2 &= \lambda a(1)F(1;A) > 0 \\ \nu_s &= 0\ \forall s \geq 3 \end{aligned}$$

Let us now prove (19). An elementary computation shows that :

$$\frac{d}{du}P(\{\xi_0 > u\}) \quad = \quad \frac{-2}{\pi(4+u^2)} := g(u) \tag{20}$$

$$\frac{d}{du}P(\{\xi_0 > u, \xi_1 > u\}) = (-1)\left(\int_{-\infty}^{\infty} f(x)f(u-x)q(x)dx + \int_{-\infty}^{\infty} f(x)\int_{u-x}^{\infty} f(y)f(u-y)dydx\right) := G(u),$$

where $f(x) = \frac{1}{\pi(1+x^2)}$ and $q(x) = \int_x^{\infty} f(y)dy$.
Using the fact that if $x \geq \frac{u}{2}$, then $(4+u^2)f(u) \leq \frac{4}{\pi}$ and Dominated Convergence, it follows that :

$$\lim_{u\to\infty} \frac{G(u)}{g(u)} = \frac{1}{2}$$

This implies (19) for $a(1) = \frac{1}{2}$.

4.2 ϕ-mixing processes.

Consider now a stationary process ξ, an APS A, and assume that (4), (5) and **(H3)** hold. With the notation of Remark 4, (a), define :

$$\phi(l) = sup\{|P(A_i/A_{1-i}) - P(A_i)| \, i = 0,1;\ A_0 \in \Sigma_n(h, s+h), A_1 \in \Sigma_n(s+h+l, n), h \geq 0, s+l+h < n, n \geq 1\}$$

Assume that :

(H5) $\sum_{l=1}^{\infty} \phi(l) = \Phi < 1$

For any $d-1 > 1$ and $\vec{r} \in \mathbb{N}^{d-1}$, set $r_0 = 0$ and $\pi_i(\vec{r}) = (r_0, r_1, \cdots, r_i)$ $i = 0, 1, \cdots, d-1$. Under **(H5)**, **(H1)** is obvious and **(H2)** follows from standard covariance inequalities for mixing processes (see [Bradley (1986)], [Doukhan, P. (1995)]). Condition **(H5)** is very strong, but it can be checked, for instance, for some Markov chains (see [Davydov (1973)], [Doukhan, P. (1995)]). We will also assume :

(H6) $\lim_n P(\{\xi(\pi_i(\vec{r})) > u_n\}/\{\xi(\pi_{i-1}(\vec{r})) > u_n\}) = a_i(\vec{r})$ $\forall i, \vec{r}, d-1$

Then we will show that **(H3)** and **(H4)** hold, and hence, the Corollary 1 applies.

Write down:

$$P(\{\xi(\vec{r}) > u_n\}/\{\xi_0 > u_n\}) = \prod_{i=1}^{i=d-1} P(\{\xi(\pi_i(\vec{r})) > u_n\}/\{\xi(\pi_{i-1}(\vec{r})) > u_n\});$$

it follows that **(H3)** holds for

$$a(\vec{r}) = \prod_{i=1}^{i=d-1} a_i(\vec{r})$$

On the other hand, set :

$$P(\{\xi(\pi_i(\vec{r})) > u_n\}/\{\xi(\pi_{i-1}(\vec{r})) > u_n\}) - P(\{\xi_0 > u_n\}) := D_{i,n}$$

and observe that :

$$|D_{i,n}| \leq \phi(r_i - r_{i-1})$$

Then :

$$P(\{\xi(\vec{r}) > u_n\}/\{\xi_0 > u_n\}) = P(\{\xi_0 > u_n\})^{d-1} + \sum_{k=1}^{k=d-1} \sum_{I \in \mathcal{C}_k(\{1,\cdots,d-1\})} \prod_{i \in I} D_{i,n} P(\{\xi_0 > u_n\})^{d-1-k}$$

Define :

$$b_n(\vec{r}) = P(\{\xi_0 > u_n\})^{d-1} + \sum_{k=1}^{k=d-2} \sum_{I \in \mathcal{C}_k(\{1,\cdots,d-1\})} \prod_{i \in I} D_{i,n} P(\{\xi_0 > u_n\})^{d-1-k},$$

$$a_n(\vec{r}) = \prod_{i=1}^{i=d-1} D_{i,n},$$

$$c(\vec{r}) = \prod_{i=1}^{i=d-1} \phi(r_i - r_{i-1})$$

Then, we get that :

$$\sum_{\vec{r} \in \mathbb{N}_L^{d-1}} b(\vec{r}) \leq \Phi^{d-1}$$

and, by Lemma 2 and the fact that $\Phi < 1$, we deduce that :

$$\sum_{d=s}^{\infty} \Theta(s;d) \sum_{\vec{r} \in \mathbb{N}_L^{d-1}} b(\vec{r}) \leq C \sum_{d=s}^{\infty} d^{s-1} \Phi^{d-1} < \infty$$

In a similar way, we get that :

$$\sum_{d=s}^{d=p_n} \Theta(s;d) \sum_{\vec{r} \in [1,p_n]_L^{d-1}} |b_n(\vec{r})| \leq \sum_{d=s}^{d=p_n} \Theta(s;d)[p_n P(\{\xi_0 > u_n\})]^{d-1} + \sum_{k=1}^{k=d-2} C_k^{d-1} \Phi^k P(\{\xi_0 > u_n\})^{d-1-k}$$

$$= \sum_{d=s}^{d=p_n} \Theta(s;d)[(p_n P(\{\xi_0 > u_n\}) + \Phi)^{d-1} - \Phi^{d-1}],$$

and then,

$$\lim_n \sum_{d=s}^{d=p_n} \Theta(s;d) \sum_{\vec{r} \in [1,p_n]_L^{d-1}} |b_n(\vec{r})| = 0$$

and **(H4)** holds.

Remark 13

a) Consider now a stationary associated process ξ and assume that :

$$\sum_{m=1}^{\infty} sup_n n |Cov(1\{\xi_0 > u_n\}, 1\{\xi_m > u_n\})| < \infty$$

Then **(H2)** holds, and, by Remark 4 (b), **(H1)** holds.
If A is an APS, and we assume (4), (5), **(H3)**, and **(H4)**, the CPLT holds.
b) In general, our hypotheses concerning ξ are restrictive, but, by our motivation (ozone peaks), we are interested in the case where ξ may be well be a "nice" process, but Y presents a more complicated structure.

5 CPLT for regression models and proof of the main result

First of all, we will present a very simple extension of Proposition 1 and Corollary 1. Consider an $\mathbb{R}^h$ -valued stationary process $\vec{\xi} = (\vec{\xi^1}, \cdots, \vec{\xi^h})$ and an APC $A^1, \cdots, A^h$.
Define :

$$X_t = \sum_{j=1}^{j=h} \xi_t^j 1_{A^j}(t) \tag{21}$$

Introduce the functions :

$$\begin{aligned} J : \mathbb{N} &\rightarrow \{1, \cdots, h\} \\ n &\longmapsto J(n) \text{ such that } n \in A^{J(n)} \end{aligned}$$

and,

$$\begin{aligned} \vec{J} : \mathbb{N}^d &\rightarrow \{1, \cdots, h\} \\ \vec{r} &\longmapsto \vec{J}(\vec{r}) = (J(r_1), \cdots, J(r_d)) \text{ such that } \vec{r} \in \left(A^{J(r_1)}, \cdots, A^{J(r_d)}\right) \end{aligned}$$

Then we can rewrite (21) as :

$$X_t = \xi_t^{J(t)} \tag{22}$$

We will set :

$$\{\xi^{\vec{j}}(\vec{r}) > u\} := \bigcap_{i=1}^{i=d} \{\xi_{r_i}^{j_i} > u\}, \forall \vec{r} \in \mathbb{R}^d, \ \forall \vec{j} \in \{1, \cdots, h\}^d.$$

We will assume now :

$$\begin{aligned} \lim_n \lambda_n(j) &= \lambda(j) \ \forall \ j = 1, \cdots, h \\ \text{where} & \\ \lambda_n(j) &= nP\left(\{\xi_0^j > u_n\}\right) \end{aligned} \tag{23}$$

and

(H7) $\limsup_n \sum_{m=1}^{\infty} n|Cov(1\{\xi_0^j > u_n\}, 1\{\xi_m^h > u_n\})| < \infty \ \forall j, h.$

Lemma 3 *Let B be a subset of $\mathbb{N}$, X as in (21), N_n^* as in (6), and $u > 0$, then*

$$P\left(\{N_n^*(B) \geq s\}\right) = \sum_{d=s}^{\infty} \Theta(s;d) \sum_{\vec{r} \in [1, card(B_n)]_L^{d-1}} \sum_{\vec{j} \in \{1, \cdots, h\}^d} P\left(\{\xi^{\vec{j}}(\vec{r}) > u\}\right) card\left(T_n(\vec{r}; B \cap \mathcal{A}(\vec{j}))\right),$$

where

$$\mathcal{A}(\vec{j}) := A^{j_0}, \cdots, A^{j_{d-1}}, \forall \vec{j} \in \mathbb{N}^d$$

and

$$\mathcal{A}(\vec{j}) \cap B := A^{j_0} \cap B, \cdots, A^{j_{d-1}} \cap B, \forall \vec{j} \in \mathbb{N}^d$$

Proof of Lemma 3 : Just mimic the proof of Lemma 2 $\diamond$

Proposition 2 *Let X be as in (21) and assume that (23),* **(H1)** *and* **(H7)** *hold. For any $s \in \mathbb{N}$, define :*

$$Q_n(s;j) = \sum_{d=s}^{d=p_n} \Theta(s;d) \sum_{\vec{r}\in[1,p_n]_L^{d-1}} \sum_{\{\vec{j}\in\{1,\cdots,h\}^d : J(0)=j\}} P\left(\{\xi^{\vec{j}}(\vec{r}) > u_n\}/\{\xi_0^j > u_n\}\right) F_n(\vec{r}; \mathcal{A}(\vec{j})),$$

where

$F_n(\vec{r}; \mathcal{A}(\vec{j})) = \frac{1}{n} \sum_{1\le i\le k_n : card(B_n^i)\ge d} card\left(T_n(\vec{r}, B_n^i \cap \mathcal{A}(\vec{j}))\right)$
$\mathrm{A}(\vec{j}) := \left(A^{j_0}, \cdots, A^{j_{d-1}}\right), \forall \vec{j} \in \mathbb{N}^d$
$\mathrm{A}(\vec{j}) \cap B := \left(A^{j_0} \cap B, \cdots, A^{j_{d-1}} \cap B\right), \forall \vec{j} \in \mathbb{N}^d$
Assume further that, for any $s \in \mathbb{N}$,

$$\lim_n Q_n(s;j) = Q(s;j) \ \forall j \tag{23}$$

Then N_n converges in law (as a process) to a $CP(\nu)$ process, where :

$$\nu_s = \sum_{j=1}^{j=k} \lambda(j)(Q(s;j) - Q(s+1;j)) \ \forall s \in \mathbb{N}.$$

Proof of Proposition 2:
Despite the greater complexity of notation, using the preceeding Lemma, this proof is obtained by exactly the same arguments as Proposition 1 $\diamond$.

We also have :

Corollary 2 *Assume that $A^1, \cdots, A^h$ is an APC and that X is as in Proposition 2. Assume further:*

(H8) $\forall d \in \mathbb{N}$, $\forall \vec{r} \in \mathbb{N}_L^{d-1}, \vec{j} \in \{1, \cdots, h\}^d \lim_n P\left(\{\xi^{\vec{j}}(\vec{r}) > u_n\}/\{\xi_0^{j_0} > u_n\}\right) = a(\vec{r}, \vec{j})$,
(H9) $\forall d \in \mathbb{N}$, $\forall \vec{r} \in \mathbb{N}_L^{d-1}$, $\vec{j} \in \{1, \cdots, h\}^d P\left(\{\xi^{\vec{j}}(\vec{r}) > u_n\}/\{\xi_0^{j_0} > u_n\}\right) = a_n(\vec{r}, \vec{j}) + b_n(\vec{r}, \vec{j})$
where :

$$a_n(\vec{r}, \vec{j}) \le c(\vec{r}, \vec{j}) \ \forall n,$$

$$\sum_{d=s}^{\infty} \Theta(s;d) \sum_{\vec{r}\in\mathbb{N}_L^{d-1}} c\left(\vec{r}, \vec{J}(\vec{r})\right) < \infty$$

$$\lim_n \sum_{d=s}^{d=p_n} \Theta(s;d) \sum_{\vec{r}\in[1,p_n]L^{d-1}} |b_n\left(\vec{r}, \vec{J}(\vec{r})\right)| = 0$$

Then N_n converges in law (as a process) to a $CP(\nu)$ process, where :

$$\nu_s = \sum_{j=1}^{j=h} \lambda(j)(Q(s;j) - Q(s+1;j)) \ \forall s \in \mathbb{N}$$

and

$$Q(s;j) = \sum_{d=s}^{\infty} (-1)^{s+d} C_s^d \sum_{\vec{r} \in \mathbb{N}_L^{d-1}} \sum_{\{\vec{j} \in \{1,\cdots,h\}^d : j_0 = j\}} a(\vec{r},\vec{j}) F(\vec{r}; A(\vec{j})).$$

Proof of Corollary 2:

Just follow the proof of Corollary 1 $\diamond$

Example 5

If ξ is an m-dependent stationary process such that :

$$\forall y \in \mathbb{R},\ \lim_n nP\left(\{\varphi(\xi_0, y) > u_n\}\right) = \lambda(y)$$

and

$$\forall d \in \mathbb{N},\ \vec{y} \in \mathbb{R}^d, \vec{r} \in \mathbb{N}_L^{d-1}, \lim_n P\left(\{\varphi\left(\xi(\vec{r}), \vec{y}\right) > u_n\} / \{\varphi(\xi_0, y_0) > u_n\}\right) = a(\vec{r}, \vec{y})$$

then ξ is a $\varphi - noise$ (see section 3.1).

If we replace "m-dependent "by assumption **(H5)**, we obtain the same result (see section 3.2).

We will present now the proof of the main result of this paper.

Proof of Theorem 1:

Consider first the case whre Y is regular. Using once again Theorem 4.2 of [Kallenberg (1983)] , we have to prove that for any $I_1, \cdots, I_k$ disjoint semiclosed intervals we have that $(N_n(I_1), \cdots, N_n(I_k))$ converges in law to $(N(I_1), \cdots, N(I_k))$, where N is a $CP(\nu)$ process. For the sake of simplicity, we will present here the case $k = 1$, but the general case is obtained by very similar arguments. Further, without loss of generality, we will set $I_1 = (0, 1]$ and hence, we will prove indeed that

$$N_n((0,1]) \xrightarrow[n]{w} N((0,1]),$$

where N is a random variable whith Laplace transform

$$L(s) = exp\left(\sum_{j=1}^{\infty} \nu_j (e^{-sj} - 1)\right)$$

Assume first that Y takes values on a finite set $y_1, \cdots, y_k$. Restrict the probability space to a set of total probability where the almost sure convergence of Definition 3 holds. Conditioning to Y, since ξ and Y are independent, the law of the exceedances point process for X $(N_n)$is the same as that of the point process $\bar{N}_n$ corresponding to (21) for $\xi^i = \varphi(\xi_0, y_i), A^i(\omega) = \{t \in \mathbb{N} : Y_t(\omega) = y_i\}$.Then, by Definition 4, the result follows.

Assume now that the result holds for Y bounded. Consider an unbounded Y let Y^K be the truncation of Y by K, $Y_n^K = c_K(Y_n)$, where $c_K(x) = x$ for $|x| < K$, $c_K(x) = Ksgn(x)$ for $|x| > K$; it is obvious that Y^K itself is ponderable. Call now N_n^K to the exceeding point process for $X^K = \varphi(\xi, Y^K)$, since Y^K is finite-valued, N_n^K converges to N^K, whose Laplace transform will be denoted by L^K.

But,

$$
\begin{aligned}
E(|N_n - N_n^K|) &\leq \sum_{i=1}^{i=n} P\left(\{\varphi(\xi_i, Y_i) > u_n\} \nabla \{\varphi(\xi_i, Y_i^K) > u_n\}\right) \\
&\leq \sum_{i=1}^{i=n} P\left([\{\varphi(\xi_i, Y_i) > u_n\} \nabla \{\varphi(\xi_i, Y_i^K) > u_n\}] \cap \{|Y_i| > K\}\right)
\end{aligned}
$$

(conditioning with respect to Y)

$$
\begin{aligned}
&\leq \sum_{i=1}^{i=n} \int_{|y|>K} P\left([\{\varphi(\xi_i, y) > u_n\} \nabla \{\varphi(\xi_i, sgn(y)K) > u_n\}] dP^{Y_i}(y)\right. \\
&\leq n 2 \sup_n \sup_{|y|>K} nP\left(\{\varphi(\xi_0, y) > u_n\}\right) \frac{1}{n} \sum_{i=1}^{i=n} P\left(|Y_i| > K\right)
\end{aligned}
$$

Then, by condition (a) of Definition 5, we get that :

$$
\lim_K \limsup_n E(|N_n - N_n^K|) \leq C \lim_K \mu^0([-K, K]^c) \quad = \quad 0 \tag{24}
$$

On the other hand,

$$
L^K(x) = exp\left(\sum_{j=1}^{\infty} \nu_j^K (e^{-xj} - 1)\right),
$$

with

$$
\nu_s^K = \tau(s)^K - \tau(s+1)^K \ \forall s \in \mathbb{N}
$$

$$
\tau(s)^K = \sum_{d=s}^{\infty} \Theta(s; d) \int_{\mathbb{R}^{d-1}} \sum_{\vec{r} \in \mathbb{N}_L^{d-1}} a(\vec{r}, \vec{y}) \lambda(y_0) \mu^{\vec{r}} \circ c_K^{-1}(dy)
$$

after some elementary computations, we obtain from Definition 5 and Dominated Convergence that :

$$
\lim_{K \to \infty} L^K \quad = \quad L \tag{25}
$$

From (24) and (25) we obtain the theorem for Y unbounded.

It suffices now to show the result for Y bounded. Without loss of generality, we may assume that Y takes values on [0,1). For $H \in \mathbb{N}$, define $d_H(x) = \sum_{i=1}^{i=H} \frac{i}{H} 1_{[\frac{i-1}{H}, \frac{i}{H})}(x)$ and set $Y_n(H) = d_H(Y_n)$. Then Y^H is also ponderable and takes values on a finite set. So,the result applies to $X_t(H) = \varphi(\xi_n, Y_n(H))$ On the other hand, from Definition 5 and Dominated Convergence we obtain analogous results to (24), (25), what shows that the result applies for Y bounded. Since we have already seen that the result applies for finite-valued Y, the proof of this part is finished .

If we turn now to the case where Y is not regular,we can easily see that for Y finite, the limit of N_n/Y can be obtained in the same way, but since the limit depends on Y, the limit of N_n is a mixture of Compound Poisson processes. The same approximation arguments extends the result to the general case. $\diamond$

Acknowledgements: Authors want to express their gratitude to Didier Dacunha-Castelle, who introduced them to the problem and to all the Probability and Statistics team of Orsay for their warming support and to an anonymous referee by its insightfull comments.

References

[Alpuim, Catkan & Hüsler (1995)] Extremes and clustering of nonstationary max-AR(1) sequences. *Stoch. Proc. Appl.* **56**, 171-184.

[Barbour, Holst & Janson (1992)] *Poisson approximation.* Oxford Studies in Probability **2**. Oxford University Press.

[Berman (1992)] *Sojourns and Extremes of Stochastic Processes.* Wadsworth & Brooks.

[Bradley (1986)] Basic properties of strong mixing conditions. *Dependence in Probability and Statistics: A Survey of Recent Results* (editors: Ebberlein,E. and Taqqu,M). Birhaüser, 165-192.

[Brown & Xia (1995)] On Stein-Chein factors for Poisson approximation. *Statist. Probab. Letters* **23**, 327-332.

[Bryc & Smolenski(1993)] Moment Conditions for almost sure convergence of weakly correlated random variables. *Proc. Am. Math. Soc.* **119**, No. 2, 355-373.

[Cohen (1989)] On the Compound Poisson Limit Theorem for High Level Exceedances. *J. Appl. Prob.* **26**, 458-465.

[Cox & Isham (1992)] *Point Processes.* Monographs on Apllied Probability and Statistics **12**. Chapman & Hall.

[Davydov (1973)] Mixing conditions for Markov Chains. *Theory Probab. Appl.* **18**, No.2, 312-328.

[Doukhan, P. (1995)] *Mixing:Properties and Examples.* Lectures Notes in Statistics **85**, Springer Verlag.

[Doukhan & Louichi (1996)] Weak dependence and moment inequalities. *Prepublications Mathématiques d'Orsay* **97.08**.

[Dziubdziela (1988)] A Compound Poisson Limit Theorem for Stationary Mixing Sequences. *Rev. Roumaine Math. Pures Appl.* **33**, 1-2, 39-45.

[Dziubdziela (1995)] On the limit distribution of the sums of mixing Bernoulli random variables *Statist. Probab. Letters* **23**, 179-182.

[Falk, Hüsler & Reiss (1994)] *Laws of Small Numbers: Extremes and Rare Events.* DMV Seminar **23**, Birkhäuser-Verlag.

[Ferreira (1993)] Joint exceedances of high levels under a local dependence condition. *J. Appl. Prob.* **30** 112-120.

[Guyon (1995)] *Random Fields on a Network. Modeling, Statistics and Applications.* Probability and its Applications, Springer-Verlag.

[Hudson, Tucker & Veeh (1989)] Limit distributions of sums of m-dependent Bernoulli random variables. *Probab. Theory Related Fields* **82**, 9-17.

[Hüsler (1993)] A Note on Exceedances and rare events of non-stationary sequences. *J. Appl. Prob.* **30**, 877-888.

[Hsiau (1997)] Compound Poisson Limit Theorems for Markov Chains. *J. Appl. Prob.* **34**, 24-34.

[Hsing, Hüsler & Leadbetter (1988)] On the Exceedance Point Process for a stationary sequence. *Probab. Th. Rel. Fields* **78**, 97-112.

[Kallenberg (1983)] *Random Measures*, 3rd. edition. Academic Press.

[Leadbetter, Lindgren & Rootzén (1983)] *Extremes and Related Properties of Random Sequences and Processes.* Spinger Series in Statistics, Springer-Verlag.

[Leadbetter & Rootzén (1988)] Extremal Theory for Stochastic Processes. *Ann. Probab.* **16**, No. 2, 431-478.

[Leadbetter & Nandagopalan (1989)] On Exceedance Point Processes for Stationary Sequences under Mild Oscilation Restrictions. *Extreme Value Theory, Proc. Conf. Oberwolfach, 1987.* Lectures Notes in Statistics **51**, 69-80.

[Leadbetter & Hsing (1990)] Limit theorems for strongly mixing stationary random measures. *Stoch. Proc. Appl.* **36**, 231-243.

[Leadbetter (1991)] On a basis for a "Peak over Treshold" modelling. *Statist. Probab. Letters* **12**, 357-362.

[Leadbetter & Rootzén (1993)] On Central Limit Theory for Families of Strongly Mixing Additive Random Functions. *Stochastic Processes: A Festschrift in Honour of Gopinath Kallianpur* (editors: S. Cambanis, J.K. Ghosh, R.L. Karandikar, P.K. Sen), Springer-Verlag.

[Leadbetter (1995)] On high level exceedance modeling and tail inference. *J. Statist. Plann. Inference* **45**, 247-260.

[Newman (1980)] Normal Fluctuations and the FKG inequalities. *Commun. Math. Phys.* **74**, 119-128.

[Perera (1994)] Spatial statistics, central limit theorems for mixing random fields and the geometry of $\mathbb{Z}^d$ *C.R. Acad. Sci. Paris* t.**319**, Série I, 1083-1088.

[Perera (1997)a] Geometry of $\mathbb{Z}^d$ and the Central Limit Theorem for weakly dependent random fields. *J. Theoret. Probab.* Vol.10, No. 3, 581-603.

[Perera (1997)b] Applications of Central Limit Theorems over asymptotically measurable sets: regression models. *C. R. Acad. Sci. Paris* t.**324**, Série I, p. 1275-1280.

[Rényi (1951)] On composed Poisson distribution II. *Acta Math. Acad. Sci. Hungar.* **2**, 83-98.

[Resnick (1987)] *Extreme Values, Regular Variation, and Point Processes.* Applied Probability Series **4**, Springer-Verlag.

[Roussas (1994)] Asymptotic Normality of Random Fields of Positively of Negatively Associated Processes. *J. Multivariate Anal.* **50**, 152-173.

[Smith & Shively (1994)] A Point Process Approach to Modelling trends in Tropospheric Ozone Based On Exceedances of a High Treshold. Technical Report **16**, National Institute of Statistical Sciences.

[Volkonskii & Rozanov (1959)] Some limit theorems for random functions, I. *Theory Probab. Appl.* **4**, 178-197.

[Volkonskii & Rozanov (1961)] Some limit theorems for random functions, I. *Theory Probab. Appl.* **6**, 186-198.

[Wschebor (1985)] *Surfaces aléatoires: mesure géométrique des ensembles de niveau.* Lecture Notes in Mathematics **1147**, Springer-Verlag.

6 Appendix 1 : Example of an asymptotically mesurable set that is not an asymptotically ponderable set

This example is based on elementary computations, but we will present it here in detail.
Let us consider $\mathbf{p} = (p_{(i,j,k,r)})_{(i,j,k,r)\in\{0,1\}^4}$ such that

$$p_{(i,j,k,r)} \geq 0 \ \forall i,j,k$$
$$\sum_{(i,j,k,r)\in\{0,1\}^4} p_{(i,j,k,r)} = 1$$

Consider (U_0, U_1, U_2, U_3) a vector of Bernoulli random variables such that

$$p_{ijkr} = P(U_0 = i, U_1 = j, U_2 = k, U_3 = r) \ \forall (i,j,k,r)$$

First, let us just compute the laws of couples (U_s, U_t) If $s \neq t \in \{0,1,2,3\}$ and if $\sigma^{i,j}_{s,t}(k,r)$ denotes the permutation of (i,j,k,r) that in the coordinate s takes the value i, in the coordinate t takes the value j and in the other two coordinates takes the values (k,r) (for instance $\sigma^{1,1}_{1,3}(0,0) = (1,0,1,0)$, $\sigma^{1,0}_{1,3}(1,0) = (1,1,0,0)$) then

$$P(U_s = i, U_t = j) = \sum_{(k,j)\in\{0,1\}^2} p_{\sigma^{i,j}_{s,t}(k,r)} \ \forall (i,j) \in \{0,1\}^2 \tag{26}$$

Fix now $p_{(i,j,k,r)} = \frac{1}{16} \ \forall (i,j,k,r)$; by taking indepent copies of (U_0, U_1, U_2, U_3), extend this vector to a random process $U = (U_t)_{t\in\mathbb{N}}$ such that

$(U_{4m}, U_{4m+1}, U_{4m+2}, U_{4m+3}) \sim (U_0, U_1, U_2, U_3) \ \forall m \in \mathbb{N}$
$((U_{4m}, U_{4m+1}, U_{4m+2}, U_{4m+3})_{m\in\mathbb{N}}$ are independent

Observe that, by (26) and construction, we have

$$P(U_{4m+t} = i, U_{4m+s} = j) = \frac{1}{4} \ \forall (i,j) \in \{0,1\}^2, \ \forall s \neq t \in \{0,1,2,3\}$$
$$P(U_t = i, U_s = j, U_r = k) = \frac{1}{8} \ \forall (i,j,k) \in \{0,1\}^3, \ \forall s \neq t \neq r \neq s \in \{0,1,2,3\} \tag{27}$$

We will now construct another process $V = (V_t)_{t\in\mathbb{N}}$ using the same construction as before, but for a different choice of $\mathbf{p}$ (to avoid confussions, we will denote $\mathbf{q}$ to this second chice). More precisely, V satisfies that

$$P(V_0 = i, V_1 = j, V_2 = k, V_3 = r) = q_{(i,j,k,r)} \ \forall \{i,j,k,r\} \in \{0,1\}^4$$
$$(V_{4m}, V_{4m+1}, V_{4m+2}, V_{4m+3}) \sim (V_0, V_1, V_2, V_3) \ \forall m \in \mathbb{N}$$
$$((V_{4m}, V_{4m+1}, V_{4m+2}, V_{4m+3}))_{m\in\mathbb{N}} \text{ are independent}$$

with

$$q_{(i,j,k,r)} = \frac{1}{24} \ \forall (i,j,k,r) \in \{(0,0,0,0),(0,0,1,1),(0,1,1,0),(0,1,0,1),(1,0,0,0),(1,0,1,1),(1,1,1,0),(1,1,0,1)\}$$
$$q_{(i,j,k,r)} = \frac{1}{12} \text{ in any other case} \tag{28}$$

Using again (26) and the construction we deduce that

$\mathrm{P}(V_{4m+t} = i, V_{4m+s} = j) = \frac{1}{4}\ \forall (i,j) \in \{0,1\}^2,\ \forall s \neq t \in \{0,1,2,3\}$
$P(V_{4m} = i, V_{4m+1} = j, V_{4m+2} = k) = P(V_{4m} = i, V_{4m+1} = j, V_{4m+3} = k)$
$= P(V_{4m} = i, V_{4m+2} = j, V_{4m+3} = k) \frac{1}{8}\ \forall (i,j,k) \in \{0,1\}^3$
$P(V_{4m+1} = i, V_{4m+2} = j, V_{4m+3} = k) = \frac{1}{6}\ \forall (i,j,k) \in \{(0,0,1),(0,1,0),(1,0,0),(1,1,1)\}$;
$P(V_{4m+1} = i, V_{4m+2} = j, V_{4m+3} = k) = \frac{1}{12}$ in any other case

Observe that we got two 3-dependent binary processes U and V which have identical two-dimensional laws ($(U_t, U_s) \sim (V_t, V_s)\ \forall s \neq t$) but with some three-dimensional laws that differ ($(U_{4m+1}, U_{4m+2}, U_{4m+3})$ and $(V_{4m+1}, V_{4m+2}, V_{4m+3}), m \in \mathbb{N}$ have different laws). It is clear that we can also assume that U and V are independent. (It will not be used here, but observe that in fact U is *iid* and V is 3-dependent, pairwise-independent, with $V_{4m}, V_{4m+1}, V_{4m+2}$ independent, $V_{4m}, V_{4m+1}, V_{4m+3}$ independent, $V_{4m}, V_{4m+2}, V_{4m+3}$ independent for any m, but with $V_{4m+1}, V_{4m+2}, V_{4m+3}$ dependent.)

For any $n \in \mathbb{N}$ define $I(n) = [100^{2^{n-1}}, 100^{2^n})$ and set

$$B = \bigcup_{r=0}^{\infty} I(2r)$$

and finally, define

$$A(\omega) = \{n \in B \setminus U_n(\omega) = 1\} \bigcup \{n \in B^c \setminus V_n(\omega) = 1\}$$

We have then

$$\frac{card(A_n)}{n} = \frac{1}{n}[\sum_{j=1}^{j=n} (X_j - E(X_j)) + \sum_{j=1}^{j=n} E(X_j)] \tag{29}$$

Where

$$X_n = \begin{cases} U_n & \text{if } n \in B \\ V_n & \text{if } n \in B^c \end{cases}$$

wich is clearly a 3-dependent process of Bernoulli variables.
Applying a Borel-Cantelli argument to the first term on the right side of (29) we deduce that it converges almost surely to zero and we get that

$$\frac{card(A_n)}{n} \text{ has the same asymptotics as } \frac{1}{n}\sum_{j=1}^{j=n} E(X_j)$$

but we deduce from their construction that

$$E(U_j) = E(V_j) = \frac{1}{2}\ \forall j, \text{ hence } E(X_j) = \frac{1}{2}\ \forall j$$

and

$$\lim_n \frac{card(A_n)}{n} = \frac{1}{2}\ a.s. \tag{30}$$

Consider now $r \geq 1$

$$\frac{card\,(A_n \cap (A_n - r))}{n} = \frac{1}{n}[\sum_{j=1}^{j=n} (\ X_j X_{j+r} - E(X_j X_{j+r})) + \sum_{j=1}^{j=n} E(X_j X_{j+r})]$$

Using again a Borel-Cantelli on the first term of the right side of the previous expression and the fact that U and V have identical two-dimensional laws, we get that :

$$\frac{card\,(A_n \cap (A_n - r))}{n} \text{ has the same asymptotic as } \lim_n \frac{1}{n} \sum_{j=1}^{j=n} E(U_j U_{j+r})$$

But $E(U_j U_{j+r}) = \frac{1}{4}\ \forall j$ and then

$$\lim_{n \to +\infty} \frac{card\,(A_n \cap (A_n - r))}{n} = \frac{1}{4}\ a.s.$$

what implies that A is, with probability one, an asymptotically measurable set in the sense of [Perera (1994)], [Perera (1997)a], with

$$F(r; A) = \frac{1}{4}\ \forall r \geq 1$$

We will now prove that, with probability one, A **is not an APS**.
It suffices to show that for $\vec{r} = (0, 1, 2) \in \mathbb{N}^3$ the following statement holds

(D) $\limsup_n \frac{card\{T_n(\vec{r}; A)\}}{n} > \liminf_n \frac{card\{T_n(\vec{r}; A)\}}{n}\ a.s.$
But

$$\frac{card\{T_n(\vec{r}; A)\}}{n} = \frac{card\,(A_n \cap (A_n - 1) \cap (A_n - 2))}{n} = \frac{1}{n}[\sum_{j=1}^{j=n} (\ X_j X_{j+1} X_{j+2} - E(X_j X_{j+1} X_{j+2})) + \sum_{j=1}^{j=n} E(X_j X_{j+1} X_{j+}$$

Using once again Borel-Cantelli on the first term of the right side of the last equation, we deduce that we only have to study the limit behaviour of

$$d_n = \frac{1}{n} \sum_{j=1}^{j=n} E(X_j X_{j+1} X_{j+2})]$$

Consider now a subsequence of the form

$$n_k = 100^{2^{2k}},\ k \text{ tending to infinity}$$

Since

$$0 \leq X_j \leq 1\ \forall j$$

we have that

$$d_{n_k} \leq 100^{-2^{2k}} \left(100^{2^{2k-1}} + \sum_{j=100^{2^{2k-1}}}^{j=100^{2^{2k}}} E(X_j X_{j+1} X_{j+2}) \right) = c_{n_k}$$

by its definition, we are now only considering $j \in B$, and since $0 \leq X_j \leq 1$ we can neglect in the summation any finite number of indexes; therefore, we can assume that we are considering

$$j, j+1, j+2 \in B$$

and we deduce that c_{n_k} has the same asymptotics as

$$100^{-2^{2k}} \left(100^{2^{2k-1}} + \sum_{j=100^{2^{2k-1}}}^{100^{2^{2k}}} E(U_j U_{j+1} U_{j+2}) \right) = 100^{-2^{2k}} \{100^{2^{2k-1}} + (100^{2^{2k}} - 100^{2^{2k-1}} + 1)\frac{1}{8}\}$$

since $100^{2^{2k-1}} = \sqrt{100^{2^{2k}}}$ we finally get that

$$\lim_k c_{n_k} = \frac{1}{8} \tag{31}$$

and hence

$$\liminf_n \frac{card\{T_n(\vec{r}; A)\}}{n} \leq \frac{1}{8} \; a.s. \tag{32}$$

Consider now a second subsequence (we still use the same notation for it for the sake of simplicity)

$$n_k = 100^{2^{2k+1}}, \; k \text{ tending to infinity}$$

We have now that, since $0 \leq X_j \leq 1$, $j \in \mathbb{N}$

$$d_{n_k} \geq 100^{-2^{2k+1}} \sum_{j=100^{2^{2k}}}^{j=100^{2^{2k+1}}} E(X_j X_{j+1} X_{j+2}) = s_{n_k}$$

we can now assume. without affecting the asymptotics, that in the summation we are only considering

$$j, j+1, j+2 \in B^c$$

and hence

$$E(X_j X_{j+1} X_{j+2}) = E(V_j V_{j+1} V_{j+2}) = \begin{cases} \frac{1}{6} & \text{if } j = 4m+1 \text{ for some } m \in \mathbb{N} \\ \frac{1}{8} & \text{in any other case} \end{cases}$$

(Note that we have used here that, for instance, by the of independence of V_{4m+4} with respect to $(V_t)_{t \leq 4m+3}$, $E(V_{4m+2} V_{4m+3} V_{4m+4}) = E(V_{4m+2} V_{4m+3}) E(V_{4m+4}) = \frac{1}{4}\frac{1}{2}$)
we deduce that

$$\lim_k s_{n_k} = \frac{1}{4}\frac{1}{6} + \frac{3}{4}\frac{1}{8} = \frac{13}{96}$$

and hence

$$\limsup_n \frac{card\{T_n(\vec{r}; A)\}}{n} \geq \frac{13}{96} \; a.s. \tag{33}$$

Since $\frac{13}{96} > \frac{1}{8}$, by (32) and (33), we obtain that, with probability one, $\frac{card\{T_n(\vec{r};A)\}}{n}$ does not converge and therefore A is not an APS.

7 Appendix 2: Brief discussion concerning ozone data.

In this appendix we will present some very rough analysis of real ozone-data and discuss whether a Compound-Poisson asymptotic approach for this sort of data could be reasonable. We also refer to [Smith & Shively (1994)] for an example of real modelling concerning this type of data.

In this case, we have considered 765 summer days where ozone levels were registered Azusa (next to Los Angeles, USA), and we have taken $u = 400$ as the "high-level"for the exceedances. We have subdivided the sample into 63 intervals containing 12 days. Denote $N_1, ..., N_{63}$ to the number of exceedances observed in each period of 12 days. Usual hypothesis testing to detect departures from the *iid* assumption (runs up-dowm, Spearman's correlation test) indicates that there is no significative evidence to reject that $N_1, ..., N_{63}$ is *iid*. This is compatible with the situation described in this paper where the point process of high-level excedances over a large period of time is close to a Compound Poisson process and hence the number of exceedances over a fixed number of disjoint intervals of the same length should be close to an *iid* sample. At

second, this N_t should have a distribution that is close to a Compound Poisson-law, what is also compatible with this set of data after a standard goodness-of-fit checking (chi-square, Kolmogorov-Smirnov). But we think that the most controversial point is the fact if the process of exceedances is close to a process with stationary and independent increments, because this point is related to the "degree "of non-stationarity and the range of dependence of the available data. Despite the lack of sensitivity of standard test for departures from the *iid* assumption, we may argue that this concrete example may not be representative of the degree of complexity usually found in real modelling, and perhaps non-stationarity and dependence are "local "enough in this case to make it not significative for after a 12-values grouping. It seems reasonable to suspect that it will often require larger samples (or larger groups) to observe the same phenomena for other data, but we think that an approach as the one presented could be applied for to describe other similar data.

The basic point is that regular I-decomposable models appears to be natural when one can expect that "averaging ", make the non-stationary influence of some explicative variables dissappear and let the weakly dependent and stationary component control the asymptotic behaviour. Of course, as in the case of non-regular I-decomposable models, the explicative variables may have a long-range dependence structure that does not allow the stationary weakly dependent component to produce a limit with stationary and independent increments: we have in the limit a random mixture of Compound-Poisson processes and the explicative variables have still a significative effect in the limit, wich may have increments that are nor independent neither stationary. And if "averaging "is not at all possible for the explicative variables (even with a random limit for averages), then we are out of the context of this paper. To know whether averaging on the explicative variables is possible, and whether it leads to non-random limit averages appears as a key point for modelling using this type of approach. We hope this brief discussion will illustrate in more intuitive terms, what are the essential points in the type of models presented in this paper.

Conclusion

En conclusion à cette étude, l'analyse des données de pollution atmosphérique permet l'application et le développement de méthodes statistiques les plus variées. En pratique, il existe de nombreuses difficultés, principalement d'ordre métrologique, évoquées dans les chapitres précédents.

Dans la deuxième partie sur l'influence du transport en région parisienne, le faible nombre d'années dont nous disposions (trois années, 1994 à 1996) et le nombre important de données manquantes, ne nous ont pas permis d'envisager l'utilisation de techniques statistiques différentes. Il serait donc naturel d'essayer d'affiner les résultats obtenus, en reprenant cette étude quand le corpus de données sera plus important. Il serait alors possible de prendre en compte "l'espace "par des techniques statistiques plus spécifiques.

Dans la troisième partie, nous nous sommes limités aux valeurs élevées d'O_3. Les modèles mis en œuvre, en utilisant les résultats de la théorie des valeurs extrêmes, nous ont permis de mettre en évidence en région parisienne, une tendance sur la période 1988-1997 par station. La tendance estimée sur chacun des quatre sites de mesure est relativement complexe puisqu'elle dépend de la température. Ces tendances peuvent aussi traduire des changements réels locaux dans les niveaux d'o_3 enregistrés. En effet, l'o_3 est consommé par le NO émis par les automobiles: si la station de mesure est située à proximité d'une route très passante, les niveaux d'o_3 enregistrés y seront plus faibles que ceux mesurés à proximité d'une rue calme. Par conséquent, si pendant la période d'étude, le trafic automobile aux abords d'un site a évolué, l'estimation de la tendance obtenue traduira principalement ce phénomène local. Il serait donc intéressant d'obtenir un modèle non plus par station, mais un modèle global où l'effet station serait contrôlé: il permettrait d'estimer une tendance régionale et de conserver les spécificités locales.

Enfin, dans la quatrième partie, il faudrait envisager l'application du théorème limite Poisson Composé obtenu aux données numériques de pollution dont nous disposons et vérifier si sa réciproque est vraie.

Partie V

Annexes

Annexe A

Annexes Partie II

A.1 Boîtes à moustaches par station des valeurs du maximum d'ozone en fonction de la direction du vent

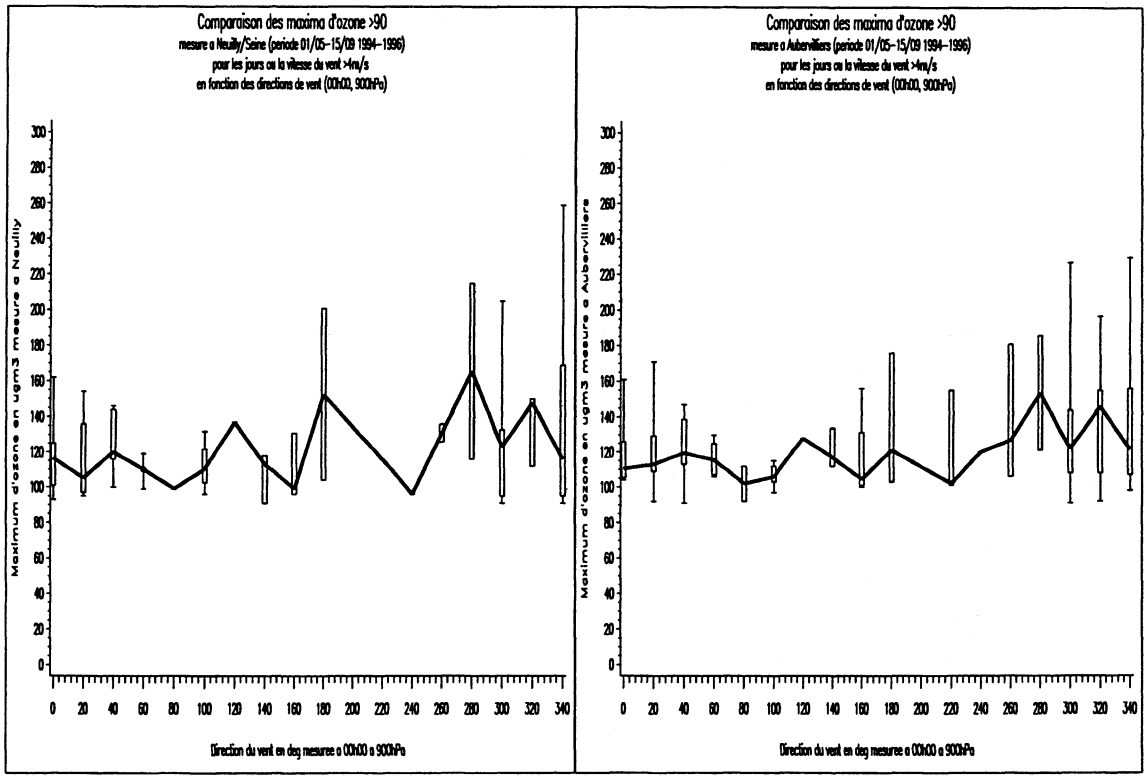


Figure A.1: Comparaison des maxima d'ozone en fonction de la direction de vent mesurée à 850 hPa

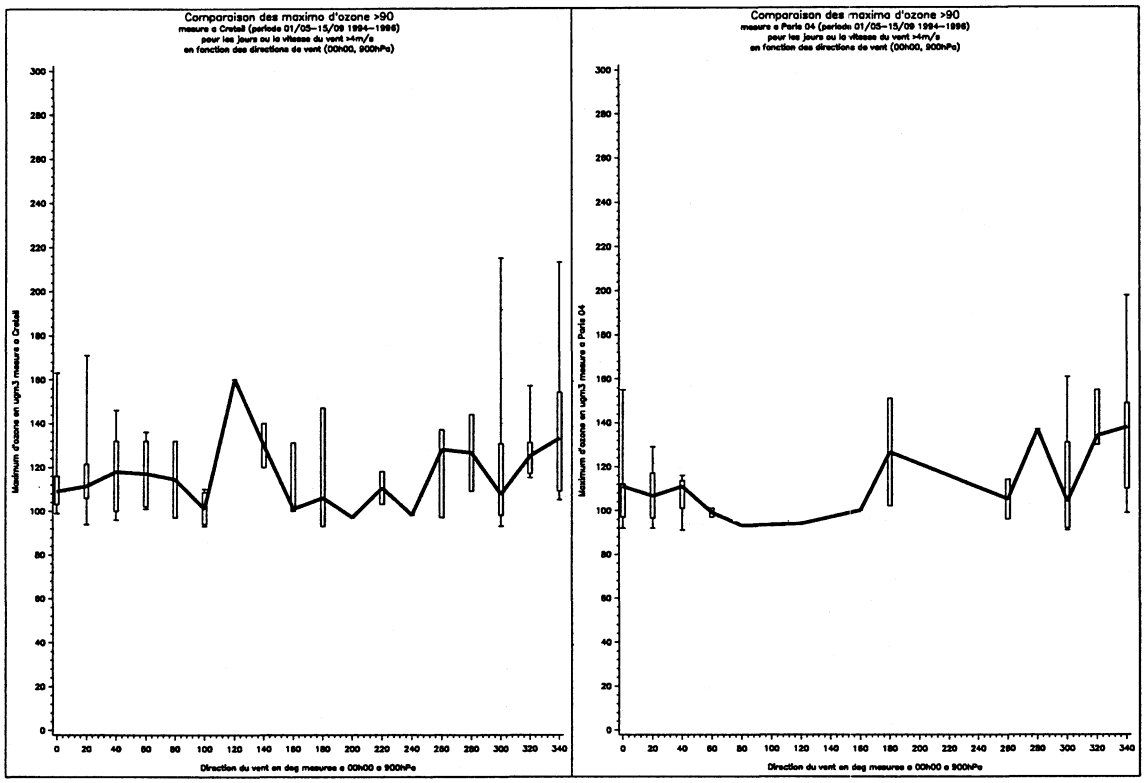


Figure A.2: Comparaison des maxima d'ozone en fonction de la direction de vent mesurée à 850 hPa

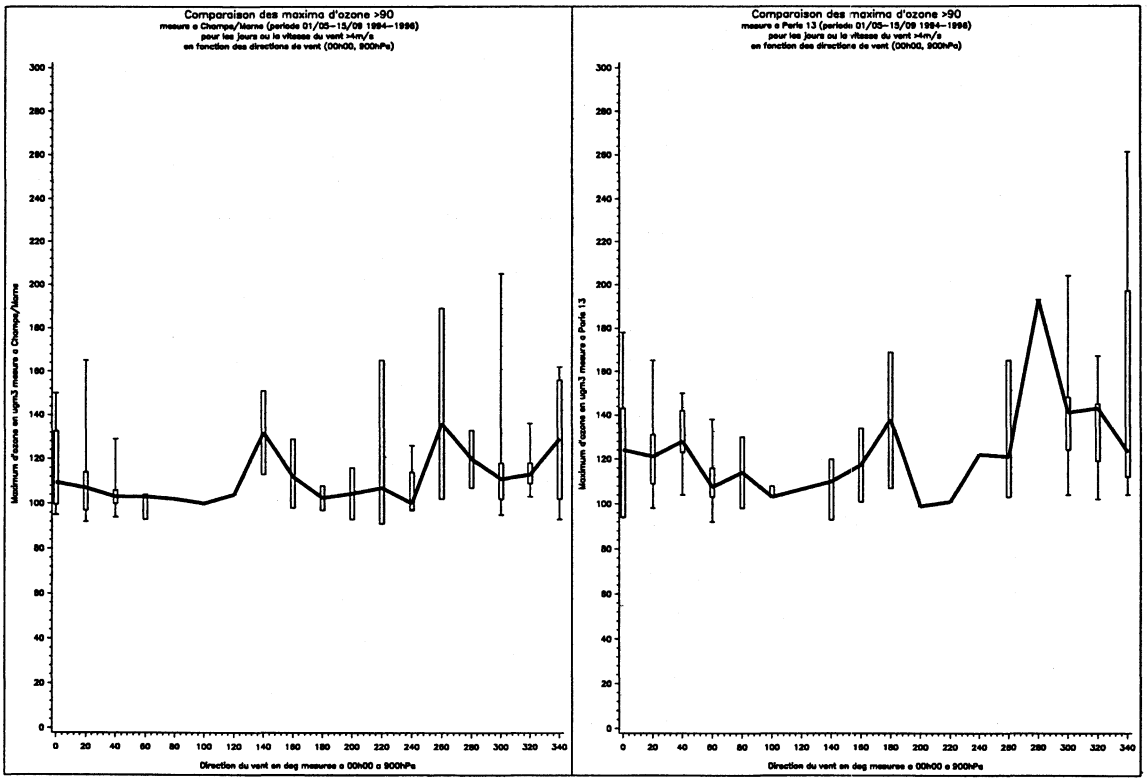


Figure A.3: Comparaison des maxima d'ozone en fonction de la direction de vent mesurée à 850 hPa

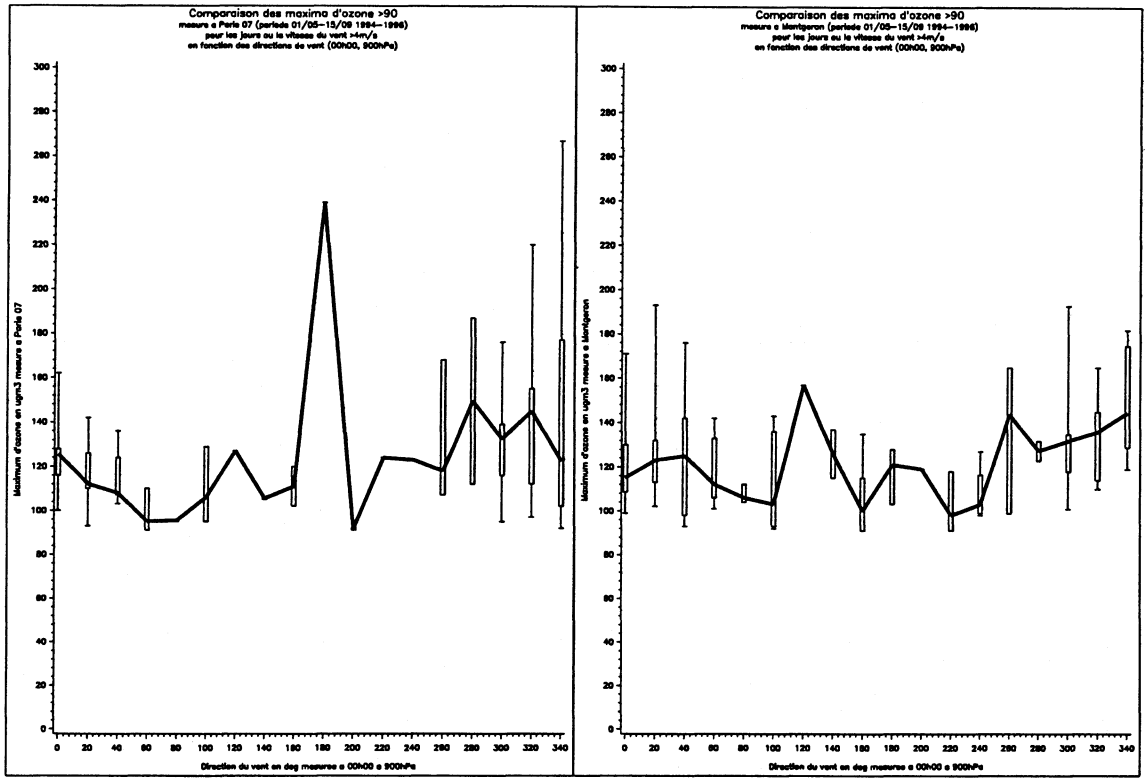


Figure A.4: Comparaison des maxima d'ozone en fonction de la direction de vent mesurée à 850 hPa

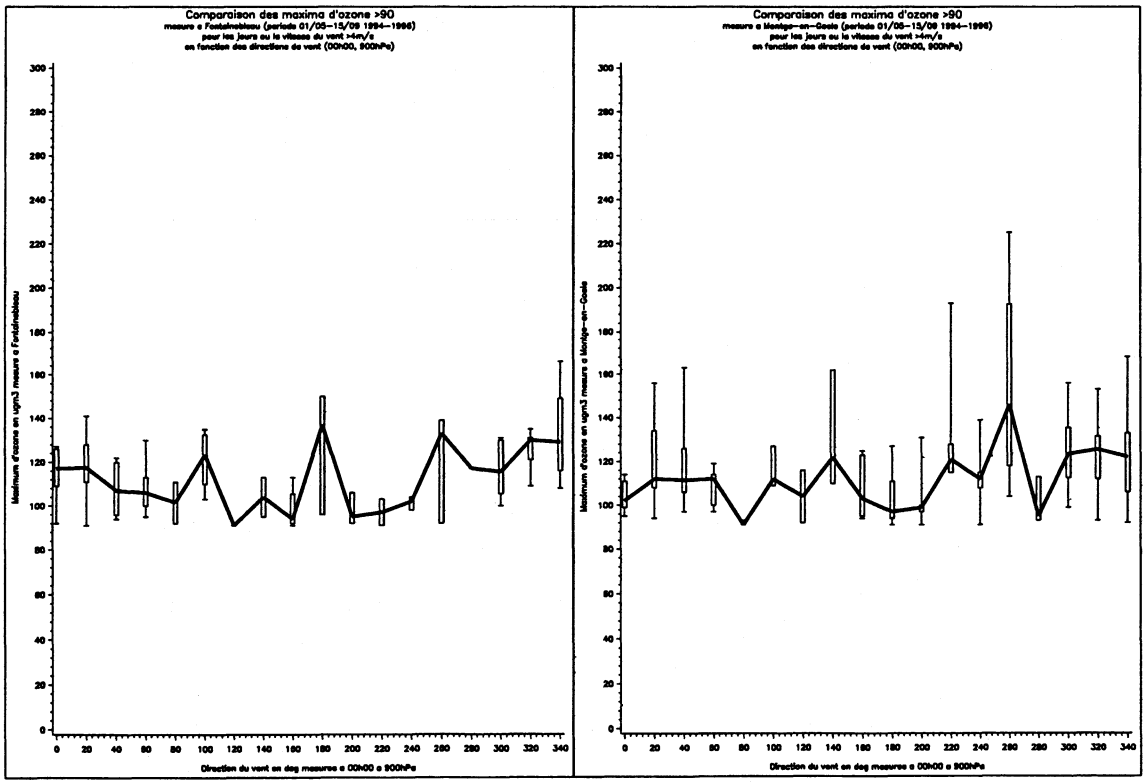


Figure A.5: Comparaison des maxima d'ozone en fonction de la direction de vent mesurée à 850 hPa

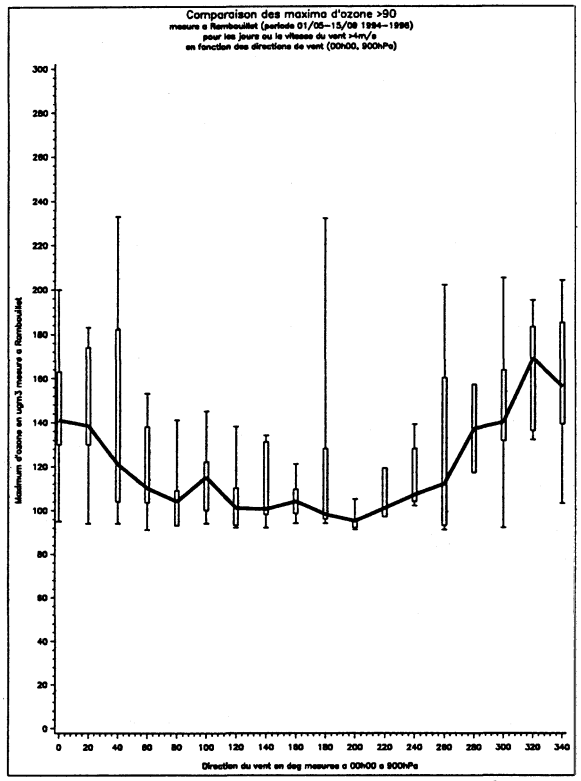


Figure A.6: Comparaison des maxima d'ozone en fonction de la direction de vent mesurée à 850 hPa

A.2 Boîtes à moustaches par classe de directions de vent mesuré à 00h00 à Trappes

A.2.1 TC

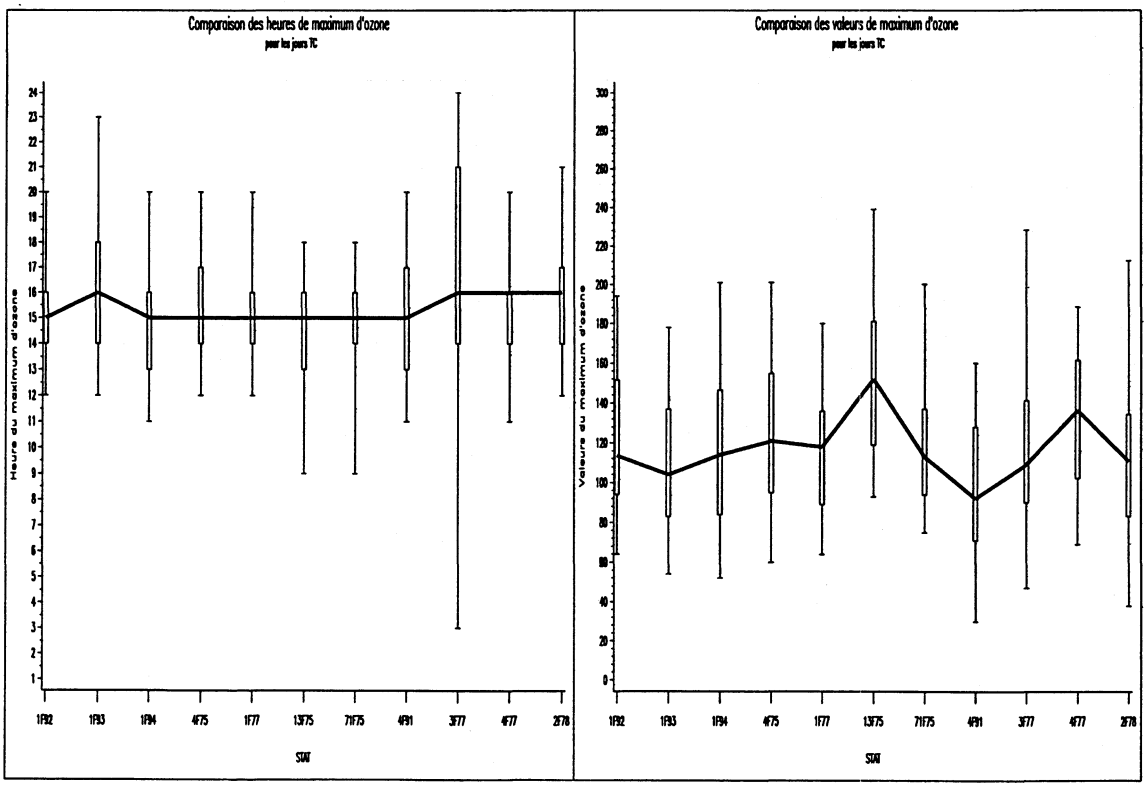


Figure A.7: Comparaison des maxima d'ozone et des heures du maximum pour la classe TC à 00h00

A.2.2 OCEAN

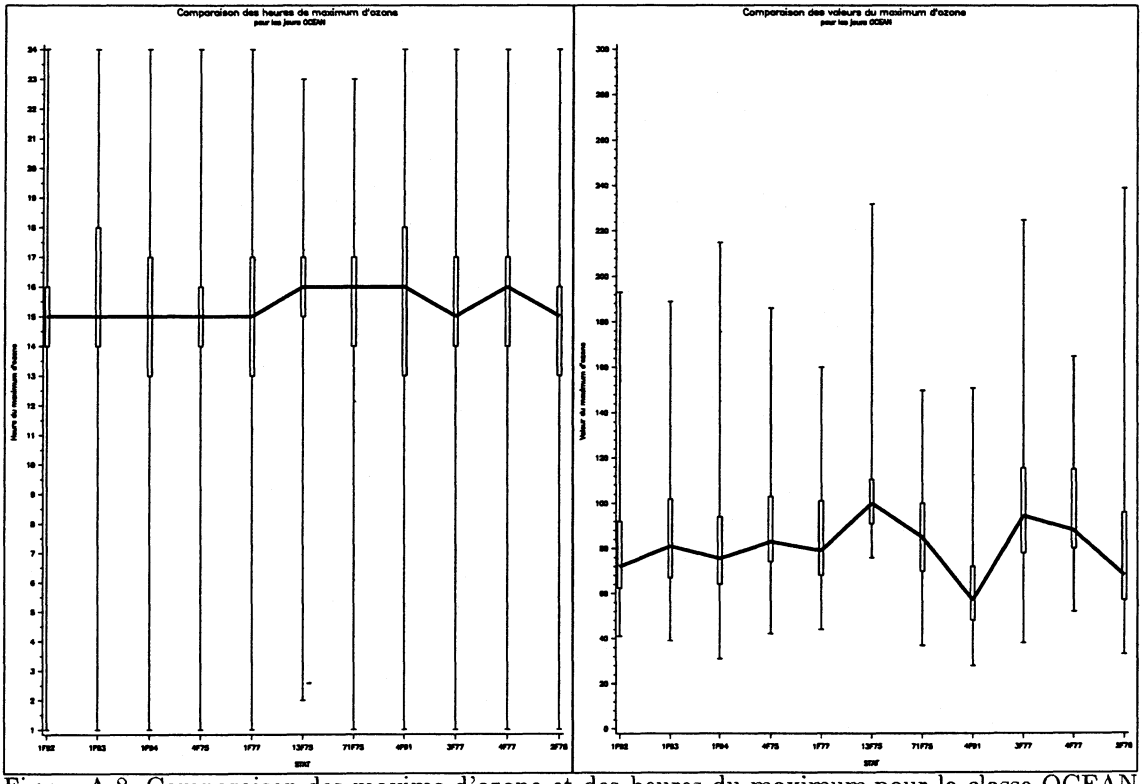


Figure A.8: Comparaison des maxima d'ozone et des heures du maximum pour la classe OCEAN à 00h00

A.2.3 CONTI

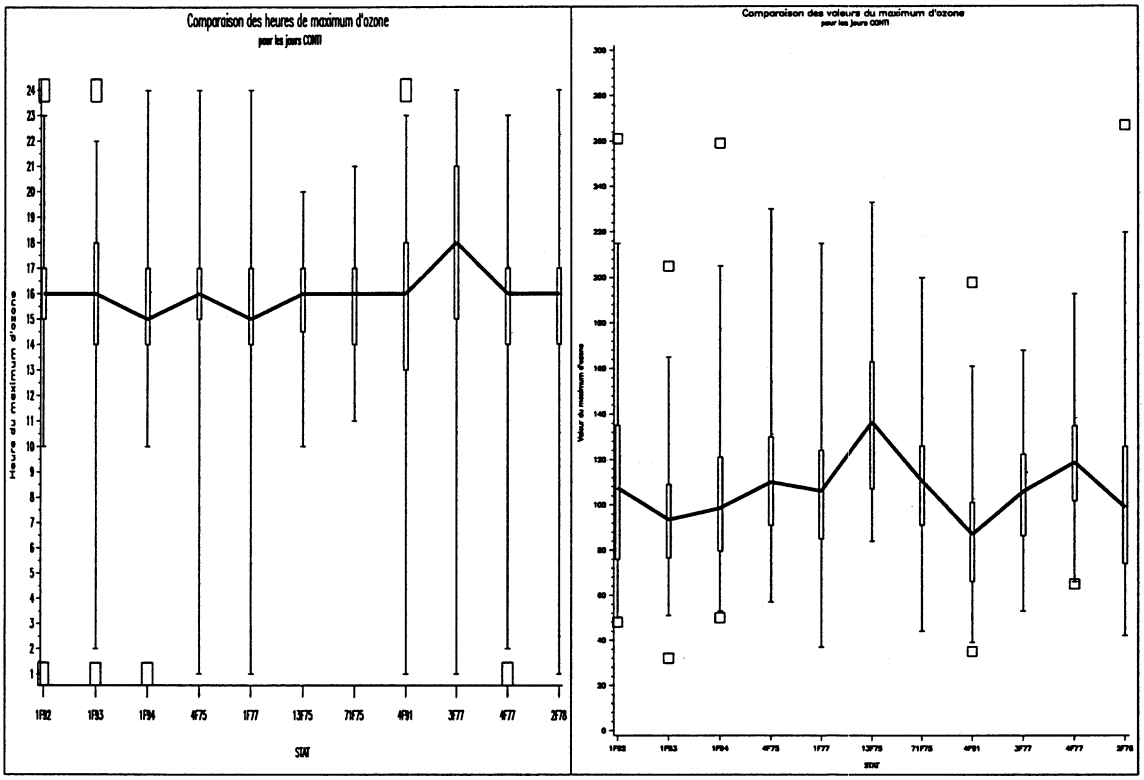


Figure A.9: Comparaison des maxima d'ozone et des heures du maximum pour la classe CONTI à 00h00

A.3 Test de Flury : comparaison d'axes principaux

Ce test permet de vérifier l'hypothèse que les axes principaux de G populations peuvent être considérés comme identiques. Notons :

- $X_g, 1 \leq g \leq G$ la matrice de l'échantillon de taille n_g de la population g ($(x_{ij})_g$ représente la valeur de la variable x_j pour l'observation i de la population g, $1 \leq j \leq p$ et $1 \leq i \leq n_g$).

- $Y_g, 1 \leq g \leq G$, la matrice des données centrées :

$$Y_g = (I - \frac{\mathbf{1}\mathbf{1}'}{n_g})X_g$$

où I est la matrice identité $n_g \times n_g$ et $\mathbf{1}$ est le vecteur unité de taille n_g.

- S_g la matrice de variance-covariance de la population g, symétrique définie positive, définie par

$$S_g = \frac{1}{n_g - 1} Y_g' Y_g$$

- $U_g = [u_1, \cdots, u_p]_g$ la matrice orthogonale des vecteurs propres de S_g, correspondant d'après les résultats de l'ACP aux facteurs principaux, et L_g la matrice diagonale dont les p valeurs propres l_j sont celles de S_g (on supposera $l_1 > l_2 > \cdots > l_p$) . On a :

$$U_g' S_g U_g = L_g$$

Supposant que chaque échantillon provient d'une population normale de vecteur moyenne μ_g et de matrice de variance-covariance Σ_g, respectivement. Notant Φ la matrice commune des facteurs principaux et Δ_g la matrice diagonale des valeurs propres du groupe g. Le test d'hypothèses est le suivant :

$$H_0 : \Phi' \Sigma_g \Phi = \Delta_g$$

En utilisant la procédure de calcul, décrite dans [24], consistant à calculer les vecteurs propres de la matrice de covariance-variance moyenne $S = \frac{\sum_{g=1}^{G}(n_g-1)S_g}{\sum_{g=1}^{G}(n_g-1)}$, on obtient une estimation notée F de la matrice Φ . On peut alors estimer les composantes principales de chacunes des g populations par $Z_g = X_g F$ et poser $F_g = F' S_g F$. Dans [30], les auteurs montrent qu'alors la statistique du rapport des log-vraisemblances s'écrit :

$$X^2 = \sum_{g=1}^{G}(n_g - 1) \log \frac{|\text{diag } F_g|}{|F_g|}$$

qui, en appliquant la théorie générale des tests de rapport de vraisemblance, suit asymptotiquement une loi du χ^2 à $(g-1)p(p-1)/2$ degrés de liberté sous H_0.

A.4 Analyse procustéenne : rappel

Notations :

- $X = [X_1\ X_2\ X_3] \in M_{11*9}$, où $X_i \in M_{11*3}\ \forall i \in \{1, \cdots, 3\}$ représente le triplet de facteurs principaux i correspondant à une direction donnée.

- on choisit comme mesure de proximité de X_i et X_j, $i \neq j$ la quantité :

$$d^2(X_i, X_j) = \sum_{k=11}^{k=11} \| X_{ki} - X_{kj} \|^2 = tr\left((X_i - X_j)(X_i - X_j)'\right)$$

$$\forall (i,j) \in \{1,2,3\}^2$$

- $\tilde{X}_i$ matrice centrée correspondant à la matrice X_i. On supposera de plus que, $tr(\tilde{X}_i \tilde{X}_i{}') = 1,\ \forall i \in \{1,2,3\}$ de façon à réduire la variabilité relative à chaque matrice.

Le but de cette analyse est de trouver la matrice orthogonale T_{ij} qui minimise $d^2(\tilde{X}_i, \tilde{X}_j T_{ij})$, ou qui maximise $tr(T_{ij} \tilde{X}_i{}' \tilde{X}_j)$, puisque :

$$\begin{aligned} d^2(\tilde{X}_i, \tilde{X}_j T_{ij}) &= tr\left(\left(\tilde{X}_i - \tilde{X}_j T_{ij}\right)\left(\tilde{X}_i - \tilde{X}_j T_{ij}\right)'\right) \\ &= tr(\tilde{X}_i \tilde{X}_i{}') + tr(\tilde{X}_j \tilde{X}_j{}') - 2tr(T_{ij} \tilde{X}_i{}' \tilde{X}_j) \end{aligned}$$

En utilisant la décomposition en valeurs singulières de $\tilde{X}_i{}' \tilde{X}_j = U \Theta V'$, on peut montrer que l'estimateur de T_{ij} est :

$$\hat{T}_{ij} = VU' = \left(\tilde{X}_j{}' \tilde{X}_i \tilde{X}_i{}' \tilde{X}_j\right)^{-\frac{1}{2}} \left(\tilde{X}_j{}' \tilde{X}_i\right)$$

La mesure de proximité $d^2(\tilde{X}_i, \tilde{X}_j \hat{T}_{ij})$, ainsi obtenue est une mesure de la distance entre les deux structures factorielles X_i et X_j.

A.5 Positionnement multidimensionnel : rappel

On trouvera de plus amples détails dans (cf. [31])
Soit D une matrice n*n de **dissemblances**, c'est-à-dire telle que :

$$\begin{aligned} d_{rr} &= 0 \\ d_{rs} &\geq 0 \; \forall r \neq s \\ d_{rs} &= d_{sr} \; \forall r \neq s \end{aligned}$$

Le but de la méthode du positionnement multidimensionnel (Multidimensional Scaling : MDS en anglais), est de trouver les points $P_1, P_2, \cdots, P_n$ en dimension k tels que si $\hat{d_{rs}}$ représente la distance Euclidienne entre P_r et P_s, alors $\hat{D}$ est "similaire" dans un certain sens à D. La dimension k est en général inconnue.

A.5.1 Quelques résultats théoriques

Définition 9 *Une matrice de dissemblances D est dite* **Euclidienne**, *s'il existe une configuration de n points dans un certain espace Euclidien dont les distances entre points sont données par D ; c'est-à-dire, si pour un certain p, il existe des points $x_1, x_2, \cdots, x_n \in \mathbb{R}^p$ tels que :*

$$d_{rs}^2 = (x_r - x_s)' (x_r - x_s)$$

Pour toute matrice de dissemblances D, notons :

- $A = (a_{rs}) = -\frac{1}{2} d_{rs}^2$
- $B = HAH$ où $H = I - n^{-1} II'$ est la matrice $n \times n$ de centrage.

Le théorème suivant permet de déterminer si D est Euclidienne, et, si oui, comment trouver la configuration de points correspondants :

Théorème 20 *Soit D une matrice de dissemblances et B comme définie plus haut, alors D est Euclidienne si et seulement si, B est définie semi-positive. En particulier :*

- *(a) Si D est la matrice Euclidienne des distances inter-points pour une configuration $Z = (z_1, z_2, \cdots, z_n)'$, alors*

  $$b_{rs} = (z_r - \bar{z})' (z_r - \bar{z}) \;\; (r; s) \in \{1, \cdots, n\}^2$$

  ou sous forme matricielle,

  $$B = (HZ)' (HZ)$$

  donc $B \geq 0$.

- *(b) Inversement, si B est définie semi-positive de rang p, alors la configuration correspondant à B peut-être construite comme ci-suit :*
  Soient $\lambda_1 > \lambda_2 > \cdots > \lambda_p$, les valeurs propres positives de B, de vecteurs propres respectifs $X_{(i)}$ normalisés par $X'_{(i)} X_{(i)} = \lambda_i \; i \in \{1, \cdots, p\}$ alors, les points P_r dans $\mathbb{R}^p$ de coordonnées $x_r = (x_{r1}, \cdots, x_{rp})'$ (où x_r est donc la $r^{ième}$ ligne de X) ont des distances inter-points données par D.

Remarque :
La $r^{ième}$ ligne de X contient les coordonnées du $r^{ième}$ point, alors que la $i^{ième}$ colonne de X contient le vecteur propre correspondant à λ_i.

A.5.2 Algorithme pratique

Un choix possible de configuration en dimension k est suggéré par le théorème précédent : choisir la configuration dans $\mathbb{R}^k$ dont les coordonnées sont déterminées par les k premiers vecteurs propres de B. Si les k premières valeurs propres de B sont "grandes" et positives et si les autres valeurs propres sont proches de 0 (positives ou "faiblement" négatives), alors les distances entre les points de cette configuration seront une approximation de D. Cette configuration est appelée solution classique du problème de positionnement multidimensionnel.

Résumé des calculs nécessaires :

- (a) A partir de D, construire $A = \left(-\frac{1}{2}d_{rs}^2\right)$
- (b) Obtenir B telle que : $b_{rs} = a_{rs} - \bar{a_{r.}} - \bar{a_{.s}} - \bar{a_{..}}$
- (c) Trouver les k plus grandes valeurs propres $\lambda_1 > \lambda_2 > \cdots > \lambda_k$ de B, de vecteurs propres respectifs $X = \left(X_{(1)}, \cdots, X_{(k)}\right)$ normalisés par $X_{(i)}^{'} X_{(i)} = \lambda_i \; i \in \{1, \cdots, p\}$.
- (d) Les coordonnées des points P_r sont alors $x_r = \left(x_{r1}, \cdots, x_{rp}\right)^{'} \; \forall r \in \{1, \cdots, n\}$ lignes de X.

Remarque :
En pratique, on limite souvent la dimension à k=1, 2 ou 3, ceci pour faciliter l'interprétation de la solution.

A.6 Histogrammes pour les jours où au moins une des stations de mesure dépasse $90\mu g/m3$

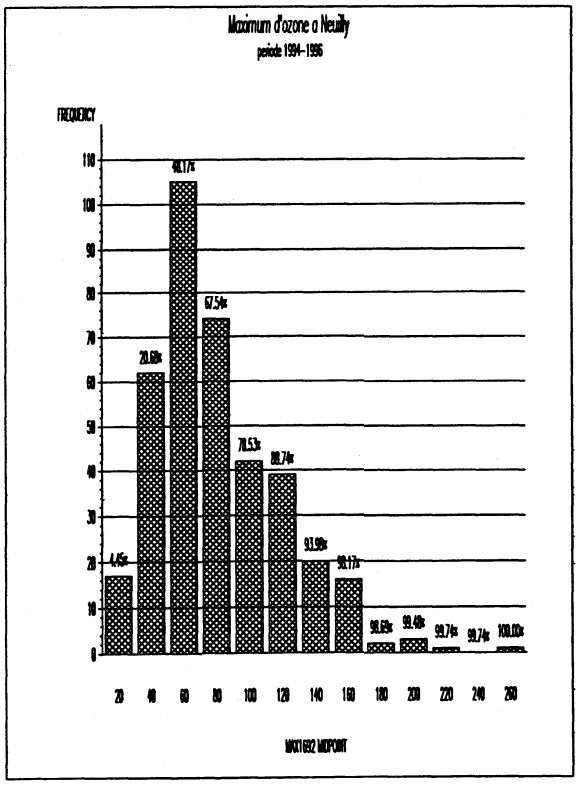


Figure A.10: Histogramme des valeurs du maximum d'ozone mesurées à Neuilly/Seine

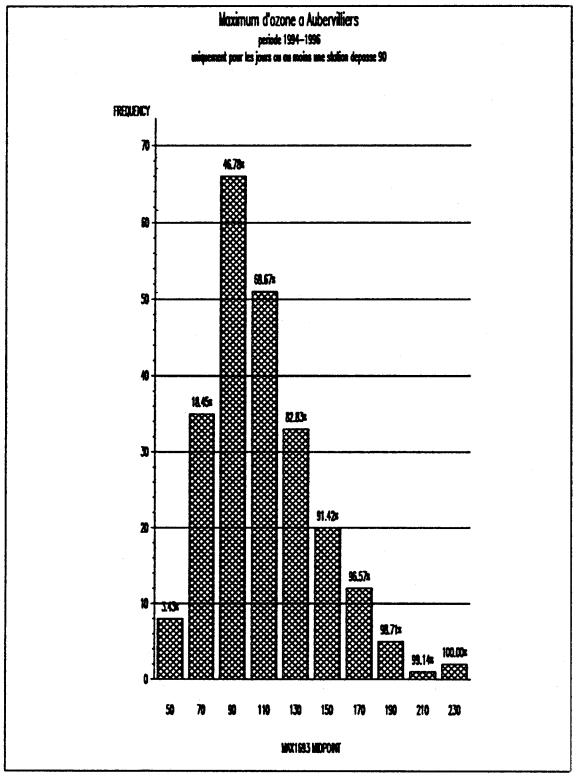


Figure A.11: Histogramme des valeurs du maximum d'ozone mesurées à Aubervilliers

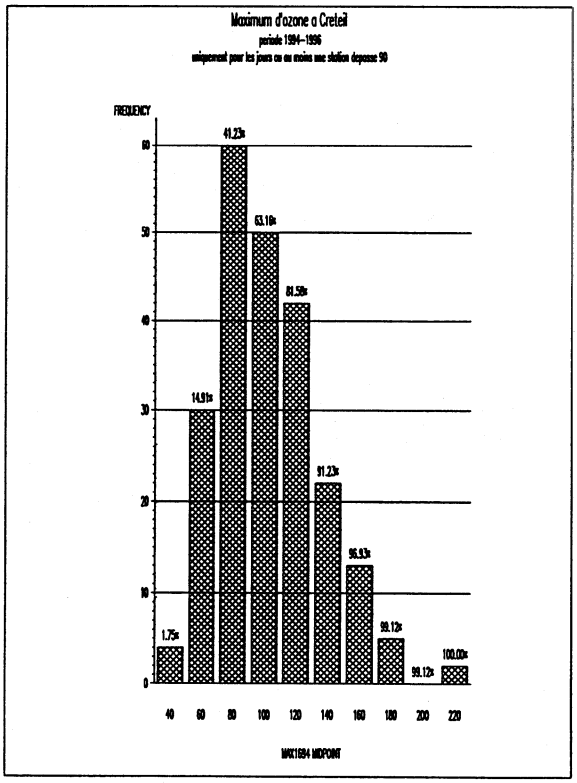


Figure A.12: Histogramme des valeurs du maximum d'ozone mesurées à Créteil

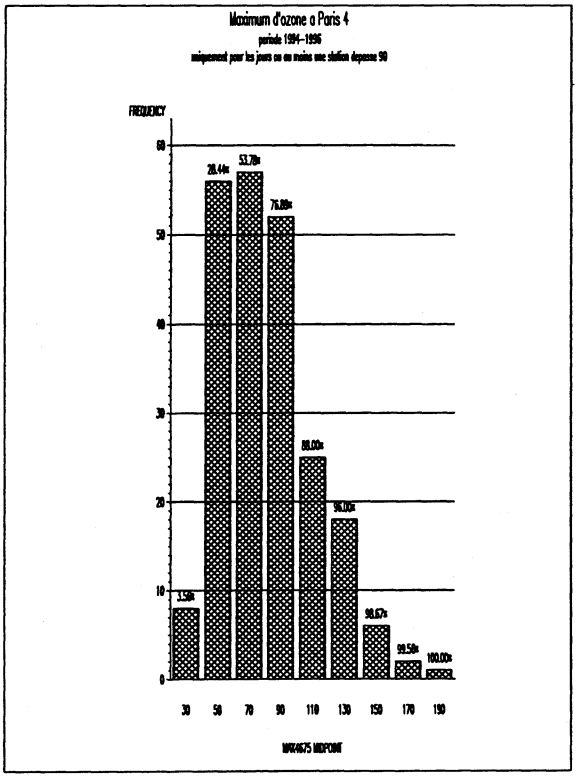


Figure A.13: Histogramme des valeurs du maximum d'ozone mesurées à Paris 04

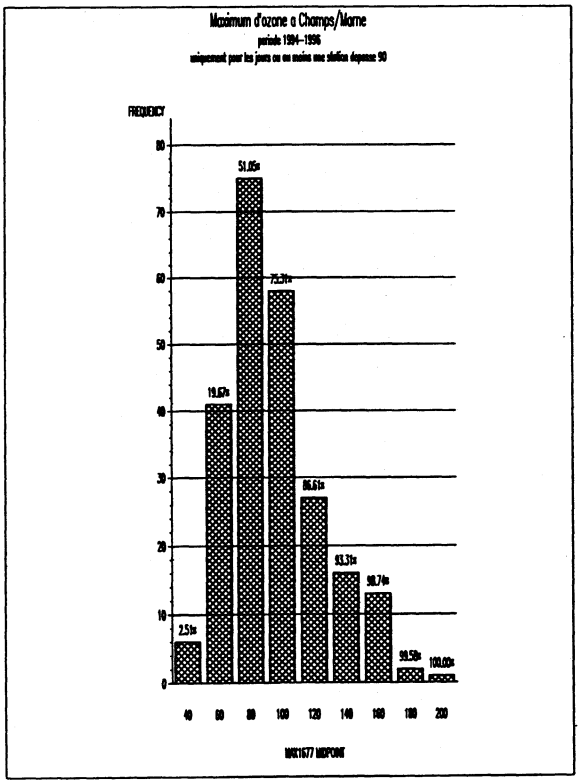


Figure A.14: Histogramme des valeurs du maximum d'ozone mesurées à Champs sur Marne

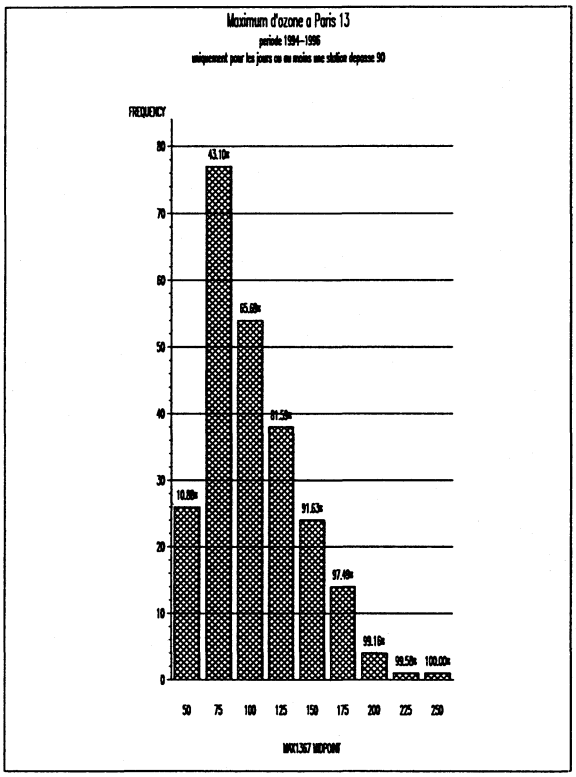


Figure A.15: Histogramme des valeurs du maximum d'ozone mesurées à Paris 13

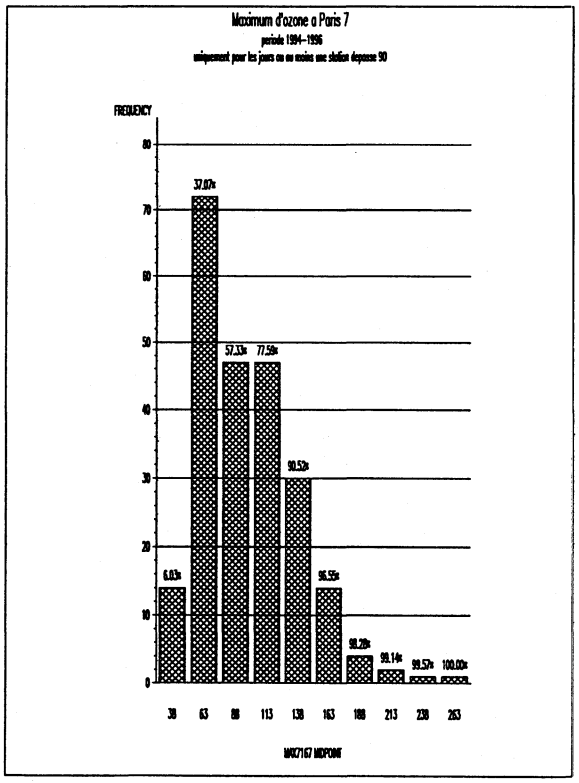


Figure A.16: Histogramme des valeurs du maximum d'ozone mesurées à Paris 07

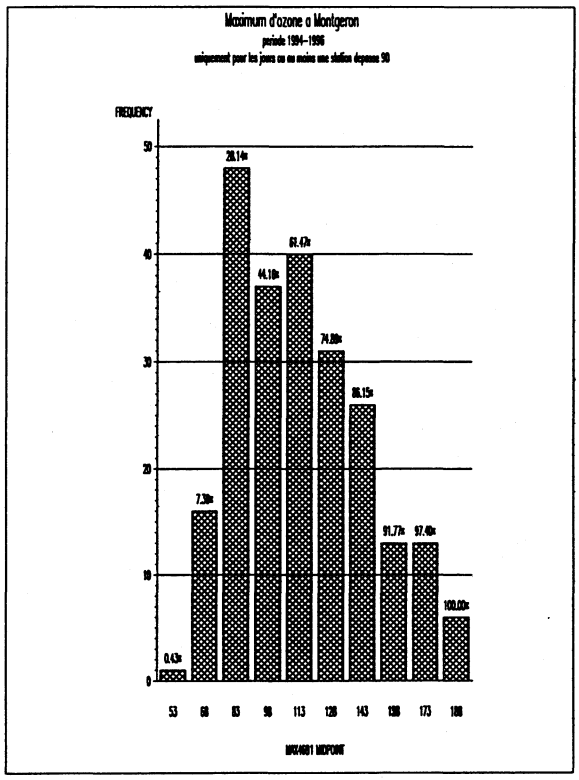


Figure A.17: Histogramme des valeurs du maximum d'ozone mesurées à Montgeron

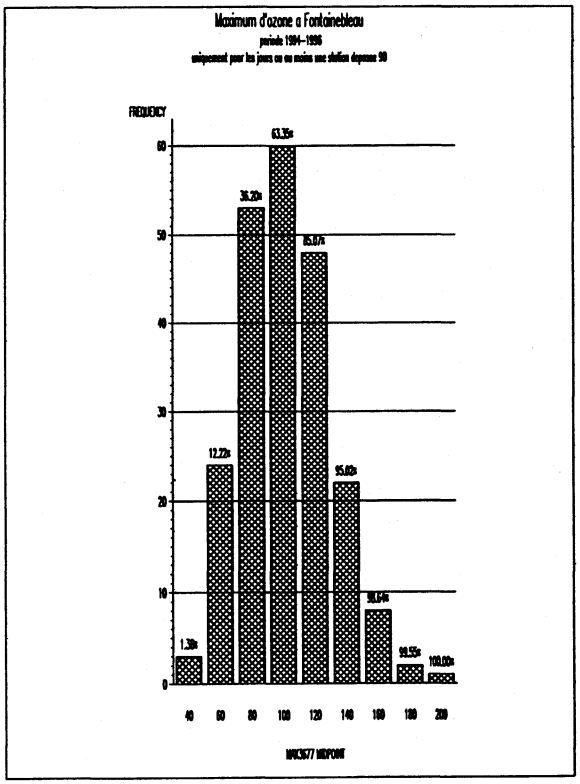


Figure A.18: Histogramme des valeurs du maximum d'ozone mesurées à Fontainebleau

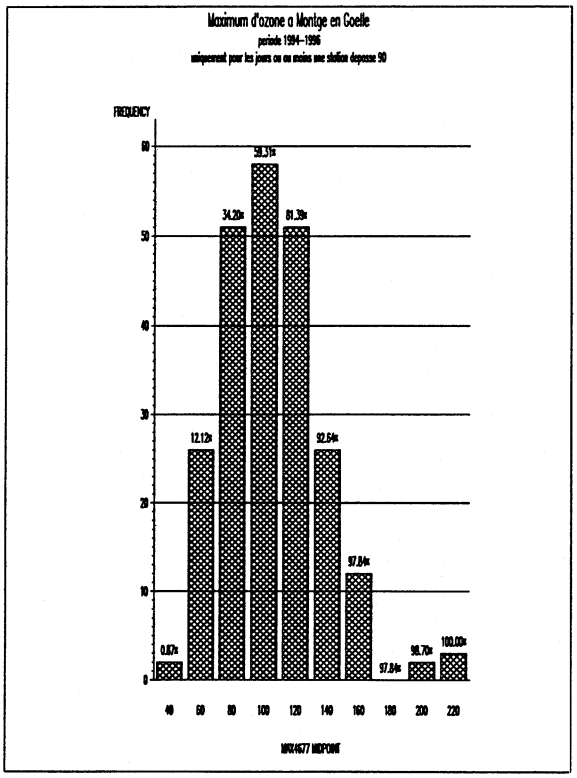


Figure A.19: Histogramme des valeurs du maximum d'ozone mesurées à Montgé-en-Goële

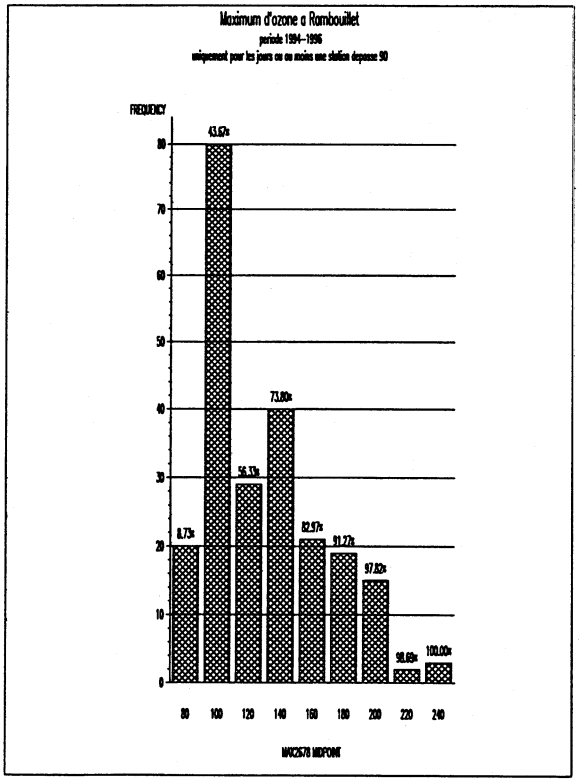


Figure A.20: Histogramme des valeurs du maximum d'ozone mesurées à Rambouillet

A.7 Programmes utilisés

A.7.1 Analyse en Composantes principales

Le logiciel SAS ([41]), ([42])) a été utilisé pour programmer les différents modules d'analyse.

```
data o3941692;
filename fich1 'd:\o3\o3941692.don';
infile fich1 lrecl=1500;
input ind sta jour mois an p1-p24;

data o3951692;
filename fich1 'd:\o3\o3951692.don';
infile fich1 lrecl=1500;
input ind sta jour mois an p1-p24;

data o3961692;
filename fich1 'd:\o3\o3961692.don';
infile fich1 lrecl=1500;
input ind sta jour mois an p1-p24;
if an=1996 then an=96;

data o31692;
  set o3941692  o3951692  o3961692 ;
if ((p1^=-1) & (p2^=-1) & (p3^=-1) & (p4^=-1) & (p5^=-1) &
   (p6^=-1) & (p7^=-1) & (p8^=-1) & (p9^=-1) & (p10^=-1) &
   (p11^=-1) & (p12^=-1) & (p13^=-1) & (p14^=-1) & (p15^=-1) &
   (p16^=-1) & (p17^=-1) & (p18^=-1) & (p19^=-1) & (p20^=-1) &
   (p21^=-1) & (p22^=-1) & (p23^=-1) & (p24^=-1)) then
  do;
   max1692=max(of p1-p24);
  end;
else
  do;
    max1692=.;
   end;

if max1692=p1 then tm1692=1;
if max1692=p2 then tm1692=2;
if max1692=p3 then tm1692=3;
if max1692=p4 then tm1692=4;
if max1692=p5 then tm1692=5;
if max1692=p6 then tm1692=6;
if max1692=p7 then tm1692=7;
if max1692=p8 then tm1692=8;
if max1692=p9 then tm1692=9;
if max1692=p10 then tm1692=10;
if max1692=p11 then tm1692=11;
if max1692=p12 then tm1692=12;
if max1692=p13 then tm1692=13;
if max1692=p14 then tm1692=14;
if max1692=p15 then tm1692=15;
if max1692=p16 then tm1692=16;
if max1692=p17 then tm1692=17;
if max1692=p18 then tm1692=18;
if max1692=p19 then tm1692=19;
if max1692=p20 then tm1692=20;
if max1692=p21 then tm1692=21;
if max1692=p22 then tm1692=22;
if max1692=p23 then tm1692=23;
if max1692=p24 then tm1692=24;
if max1692=. then tm1692=.;
keep jour mois an max1692 tm1692;

..... (idem pour les autres stations)
```

```
data o3;
  merge  o31692 o31693 o31694 o34675 o31677 o31367 o37167 o34691
         o33677 o34677 o32678;
  by an mois jour;
  if  (mois>4) & (mois<10);
  if ((mois=9) & (jour>15)) then delete;
  if (max1692>90) | (max1693>90) | (max1694>90) |
    (max4675>90) | (max1677>90) | (max1367>90) | (max7167>90) |
   (max4691>90) | (max3677>90) | (max4677>90) | (max2678>90);
proc sort data=o3;
by an mois jour;
run;

data vent;
  filename fich 'd:\vent\trappes\lise00h.don';
  infile fich lrecl=1750;
  input jour mois an fv1000 d1000 s $;
  if an<94 then delete;
run;

 data decal;
   merge o3(in=a) vent;
   by an mois jour;
   if a;
if (mois=5) | (mois=6) | (mois=7) | (mois=8) | ((mois=9) & (jour<16));
run;

 proc sort data=decal;
   by s;
run;

proc princomp data=decal cov n=3
      outstat=acpheur;
title1 'Etude de l''heure du max., periode 1994-1996 : ACP';
title2 'reseau d''alerte o3, en fonction de la direction du vent a 1000m.';
title3 'mesuree a TRAPPES a 00h00';
title4 'suppression des jours ou aucune station ne depasse 90.';
var tm1692 tm1693 tm1694 tm4675 tm1677 tm1367 tm7167 tm4691
    tm3677 tm4677 tm2678;
  by s;
 run;

/* creation d''un fichier par hauteur 1000  et heure (00h00)
   contenant les facteurs des differentes directions du vent
 _NAME_  FOOCEAN1 FOOCEAN2 FOCONTI1 FOCONTI2 FOFRONT1 FOFRONT2 FOTC1 FOTC2;

*/

 data sest;
set acpheur;
if ((s='OCEAN') & (_TYPE_='SCORE') & ((_NAME_='PRIN1') | (_NAME_='PRIN2')
     | (_NAME_='PRIN3')));
keep s tm1692 tm1693 tm1694 tm4675 tm1677 tm1367 tm7167 tm4691
     tm3677 tm4677 tm2678
;
run;
proc transpose data=sest
               out=sest1
               prefix=FOOCEAN;
var
 tm1692 tm1693 tm1694 tm4675 tm1677 tm1367 tm7167 tm4691
 tm3677 tm4677 tm2678;
  by s;
run;
data sest1;
  set sest1;
drop s;

 data norou;
set acpheur;
```

```
if ((s='CONTI') & (_TYPE_='SCORE') & ((_NAME_='PRIN1') | (_NAME_='PRIN2')
   | (_NAME_='PRIN3')));
keep s tm1692 tm1693 tm1694 tm4675 tm1677 tm1367 tm7167 tm4691
     tm3677 tm4677 tm2678;
run;
proc transpose data=norou
               out=norou1
               prefix=FOCONTI;
var
 tm1692 tm1693 tm1694 tm4675 tm1677 tm1367 tm7167 tm4691
 tm3677 tm4677 tm2678;
  by s;
run;
data norou1;
  set norou1;
drop s;

 data tempc;
set acpheur;
if ((s='TC') & (_TYPE_='SCORE') & ((_NAME_='PRIN1') | (_NAME_='PRIN2')
   | (_NAME_='PRIN3')));
keep s tm1692 tm1693 tm1694 tm4675 tm1677 tm1367 tm7167 tm4691
     tm3677 tm4677 tm2678;
proc transpose data=tempc
               out=tempc1
               prefix=FOTC;
var
 tm1692 tm1693 tm1694 tm4675 tm1677 tm1367 tm7167 tm4691
 tm3677 tm4677 tm2678;
  by s;
run;
data tempc1;
  set tempc1;
drop s;

data F1000;
merge norou1 sest1 tempc1;
run;

data ffi;
  set F1000;

/**** creation du fichier final ac00bis.don
     (fichier des var vent a 00h00 et directions vent par classes
    : variables sont...
*****/

  filename fich 'd:\lise\decal\acpruseu\ac00bis.don';
  file fich lrecl=1750;
  put _NAME_  FOOCEAN1 FOOCEAN2 FOOCEAN3 FOCONTI1 FOCONTI2 FOCONTI3
              FOTC1 FOTC2 FOTC3;
run;
```

A.7.2 Standardisation des triplets de facteurs principaux

```
/*****************************************************************
 ACP : STANDARDISATION DES VARIABLES :
 pour appliquer l analyse procusteenne,
- on centre les facteurs principaux
- somme des carres de chaque facteur principaux est egale a 1

creation du fichier permanent d:\lise\decal\acpruseu\acpstd.don
*****************************************************************/

data factt;
  filename fich 'd:\lise\decal\acpruseu\ac00bis.don';
  infile fich lrecl=1750;
 input sta $ FOOCEAN1 FOOCEAN2 FOOCEAN3 FOCONTI1 FOCONTI2 FOCONTI3
             FOTC1 FOTC2 FOTC3;
run;
```

```
proc standard data=factt
              out=ffa
              mean=0
              std=1
              ;
  var
 FOOCEAN1 FOOCEAN2 FOOCEAN3 FOCONTI1 FOCONTI2 FOCONTI3 FOTC1 FOTC2 FOTC3;
run;

/** reduction tq la somme des carres de chacun des facteurs principaux soit egale a 1 **/
data factor;
set ffa;
a=sqrt(10);

        FOCONTI1=(FOCONTI1)/a;
        FOCONTI2=(FOCONTI2)/a;
        FOCONTI3=(FOCONTI3)/a;
        FOOCEAN1=(FOOCEAN1)/a;
        FOOCEAN2=(FOOCEAN2)/a;
        FOOCEAN3=(FOOCEAN3)/a;
        FOTC1=(FOTC1)/a;
        FOTC2=(FOTC2)/a;
        FOTC3=(FOTC3)/a;


/***************** fichier permanent ***************/

  data dda;
        set factor;
        drop a;
  filename fich 'd:\lise\decal\acpruseu\acstd00b.don';
  file fich lrecl=1750;
  put sta $
   FOOCEAN1 FOOCEAN2 FOOCEAN3 FOCONTI1 FOCONTI2 FOCONTI3 FOTC1 FOTC2 FOTC3;
run;
 proc print;
title1 'Etude du decalage entre tm pour la periode 1994-1996 01/05-15/09';
title2 'facteurs principaux (standardises)';
title3 ' heure(00h00) et direction de vent.';
run;
```

A.7.3 Analyse procustéenne

```
/************************************************
nom du programme : d:\lise\decal\acpruseu\procus0b.sas

  calcul des distances entre triplets de facteurs a 00h00 1000
        ANALYSE PROCUSTEENNE
            ex: (FOOCEAN1,FOOCEAN2,FOOCEAN3) et (FOCONTI1,FOCONTI2,FOCONTI3)
Remarque :
  filename fich 'd:\lise\decal\acprural\procus0b.don';
************************************************/


  data fact;
  filename fich 'd:\lise\decal\acpruseu\acstd00b.don';
  infile fich lrecl=1750;
  input sta $
   FOOCEAN1 FOOCEAN2 FOOCEAN3 FOCONTI1 FOCONTI2 FOCONTI3 FOTC1 FOTC2 FOTC3;
run;

data factor;
set fact;
drop sta;
run;

proc iml;
use factor;
read all var _ALL_ into cc;
```

```
close factor;

ddeux=j(3,3,0);
a=j(2,2,1); T=j(3,3,1); YT=j(11,3,1); dif=YT; dif2=dif;
 I3=j(3,1,1); I11=j(1,11,1);
i=0; j=0;

       do i=1 to 8 by 3;
           X=cc(|,i:i+2|);
              do j=4 to 8 by 3;
                  if j>i then do;
                        Y=cc(|,j:j+2|);
                        call svd(u,q,v,t(X)*Y);
                        T=v*t(u);
                        ddeux(|(i+2)/3,(j+2)/3|)=
                        sqrt(trace(X*t(X))+trace(Y*t(Y))-2*trace(T*t(X)*Y));
                        ddeux(|(j+2)/3,(i+2)/3|)=ddeux(|(i+2)/3,(j+2)/3|);
                        YT=Y*T;
                       dif=X-YT;
                       dif2=(X-YT)##2;
                       d2=I11*dif2*I3;
                       print i j dif dif2 d2;
             /*          print Y YT;  */

                     free T; free u; free q; free v;
                     end;
                     free Y;

                end;


                 free X;
       end;



bb={'dd11' 'dd12' 'dd13' };
create dist from ddeux(|colname=bb|);
append from ddeux;
close dist;
quit;

proc print data=dist;
title1 'matrice des distances entre couples de facteurs:';
title2 'tq d(X,Y)=sqrt(tr((X-YT)(X-YT)''))';
title3 'ou T= matrice carre orthogonale qui minimise (d(X,Y))**2';
title4 'Analyse procusteenne.';

data dd;
set dist;
  filename fich 'd:\lise\decal\acpruseu\procus0b.don';
  file fich lrecl=1750;
  put dd11 dd12 dd13;

run;
```

A.7.4 Positionnement multidimensionnel

```
/*****************************************************
        Representation par carte des structures factorielles
        par classe de directions de vent
        MULTIVARIATE SCALING
(cf chap 14 Multivariate Analysis MARDIA, KENT et BIBBY)
*****************************************************/

       data fact;
  filename fich 'd:\lise\decal\acpruseu\procus0b.don';
  infile fich lrecl=1750;
  input dd11 dd12 dd13 ;
```

```
run;


proc iml;
use fact;
read all var _ALL_ into cc;
close fact;
A=j(3,3,1);
I3=I(3);
id=j(3,1,1);
        A=(-0.5)#(cc##2);
        C=I3-(id*t(id))#(1/3);
        B=C*A*C;
call eigen(m,e,B);
print 'Les valeurs propres de B sont  ' m;
print 'Les vecteurs propres de B sont  ' e;
m1=m(|1|); m2=m(|2|);
x1=sqrt(m1)#e(|,1|);
x2=sqrt(m2)#e(|,2|);
P=x1 || x2 ;
/* print 'Les coordonnees des points Pr sont les lignes de la matrice P ';
print P;
*/
bb={'x1' 'x2'};
create uu from P(|colname=bb|);
append from P;
close P;
quit;

data dirr;
  input direc $;
   cards;
OCEAN
CONTI
TC
 data vect12;
merge dirr uu;
run;
proc print data=vect12;
title1 'coordonnees des points Pr';
title2 'correspondant a une direction donnee (OCEAN, CONTI et TC)';
title3 'a  H0 (00h00 TU)';
title4 'a une altitude donnee B (1000m)';
run;

proc plot;
  plot  x2*x1=direc;
run;
```

Annexe B

Annexes Partie III

B.1 Résolution d'un système d'équations simultanées par la méthode de Newton-Raphson

Soit $\hat{C} = (\hat{\alpha_1}, \cdots, \hat{\alpha_p})$ l'estimateur du maximum de vraisemblance de C.
Notons :

$$\begin{aligned} C_i &= p_i(C) \\ t\,[q(C)] &= t\left(\frac{\partial l_1}{\partial C_1}, \cdots, \frac{\partial l_1}{\partial C_p}\right) \\ H_{ij}(C) &= \frac{\partial {l_1}^2}{\partial C_i C_k} = \frac{\partial {l_1}^2}{\partial C_k C_i} \; \forall (i,j) \in \{1, \cdots, p\} \end{aligned}$$

Nous devons donc considérer les situations dans lesquelles le vecteur $\hat{C}$ est un maximum local de la log-vraisemblance l_1 et satisfait le système d'équations de vraisemblance. Supposons que C_0 soit un premier voisin de $\hat{C}$, écrivons alors le développement en série de Taylor de $q(C)$ au premier ordre au voisinage de C_0 :

$$q(C) = q(C_0) + H(C_0)(C - C_0)$$

Comme $\hat{C}$ est solution du système, on a $q(\hat{C}) = 0$, donc

$$0 = q(C_0) + H(C_0)(\hat{C} - C_0)$$

D'où l'approximation de $\hat{C}$:

$$\hat{C} = C_0 - [H(C_0)]^{-1}\, q(C_0) \tag{B.1}$$

En pratique, (B.1) permet de définir le processus itératif permettant d'obtenir $\hat{C}$:

$$C^{(s+1)} = C^{(s)} - \left[H(C^{(s)})\right]^{-1} q(c^{(s)})$$

en supposant $H(C^{(s)})$ non singulière. La condition d'arrêt de l'itération est :

$$\begin{cases} |C^{(s+1)} - C^{(s)}| \leq \varepsilon \; \varepsilon \text{ petit et fixe} \\ q(C^{(s+1)}) = 0 \end{cases}$$

On en déduit alors que $C^{(s+1)}$ est une bonne approximation de $\hat{C}$.

B.2 Test de Kolmogorov-Smirnov

Il s'agit d'un test non paramétrique d'ajustement à une distribution entièrement spécifiée de fonction de répartition $F(x)$. Supposons que l'on ait un échantillon aléatoire de taille n pour une

variable aléatoire X de distribution continue inconnue $F_X(x)$. Soit $F_0(x)$ la distribution complètement spécifiée. On considère les tests d'hypothèse suivants :

$$H_0 : \ F_X = F_0$$

contre trois alternatives différentes :

- (a) $H_1^{(a)} : \ F_X = F_0$
- (b) $H_1^{(b)} : \ F_X \geq F_0$ et $F_X \neq F_0$
- (c) $H_1^{(c)} : \ F_X \leq F_0$ et $F_X \neq F_0$

Les statistiques utilisées pour tester ces hypothèses sont basées sur les mesures de la distance entre $F_0(x)$ et la distribution empirique $F_n(x)$. Elles ont la propriété d'être indépendantes de la distribution (ie non paramétriques) : sous l'hypothèse nulle leurs distributions ne dépendent pas de $F_0(x) = F_X(x)$. Ceci est une conséquence de la transformation intégrale des probabilités : si la variable aléatoire possède la distribution $F_X(x)$, alors la variable aléatoire $U = F_X(x)$ est uniformément distribuée sur $(0; 1]$.
Donc, si on a n variables alátoires $X_1, X_2, \cdots, X_n$ de distribution $F_X(x)$ et les statistiques d'ordre $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ alors $U_1 = F_X(X_1), U_2 = F_X(X_2), \cdots, U_n = F_X(X_n)$ sont des variables aléatoires uniformément distribuées sur $(0; 1]$, de statistiques d'ordre $U_{(1)} = F_X(X_{(1)}), \cdots, U_{(n)} = F_X(X_{(n)})$. Cette transformation permet de réduire le problème à tester si les points observés $u_1 = F_0(x_1), \cdots, u_n = F_0(x_n)$ suivent une loi uniforme sur $(0; 1]$, ce qui peut se faire par le test de Kolmogorov.
Les statistiques de Kolmogorov-Smirnov sont :

$$D_n^+ = \sup_{-\infty<x<+\infty} \{F_n(x) - F_0(x)\} = \max_{1\leq i\leq n} \left\{ \frac{i}{n} - F_0(X_{(i)}) \right\}$$

$$D_n^- = \sup_{-\infty<x<+\infty} \{F_0(x) - F_n(x)\} = \max_{1\leq i<n} \left\{ F_0(X_{(i)}) - \frac{(i-1)}{n} \right\}$$

$$D_n = \sup_{-\infty<x<+\infty} |F_n(x) - F_0(x)| = \max(D_n^+, D_n^-)$$

De nombreux travaux ont été effectués sur les distributions de ces statistiques (Darling 1957). Les distributions ont été tabulées par Birnbaum (1952), Miller (1956) et Owen (1962). Elles ne sont pas asymptotiquement normales. Par exemple, D_n^+ et D_n^- admettent la distribution asymptotique suivante :

$$\lim_{n\to+\infty} P\{D_n^+\sqrt{n} \leq d\} = 1 - \exp(-2d^2) \ (0 \leq d \leq +\infty)$$

D'où si $d_{n,\alpha}$ est une valeur telle que :

$$P\left\{D_n\sqrt{n} > d_{n,\alpha}\right\} = \alpha$$

alors le test utilisant la statistique D_n rejette l'hypothèse nulle au niveau α, si la valeur observée de D_n est plus grande que $\frac{d_{n,\alpha}}{\sqrt{n}}$.
Tomassone, Dervin et Masson, ont programmé dans [52] un algorithme permettant de calculer les valeurs critiques de la statistique de Kolmogorov-Smirnov à deux décimales pour $n > 35$, ce qui paraît suffisant dans notre cas :

Application pratique :
Au seuil $\alpha = 0.05$ et si $n > 35$, la région critique est $D_\alpha > \frac{1.36}{\sqrt{n}}$.

	niveau α				
$n > 35$	0.01	0.05	0.10	0.15	0.20
D_α	1.63	1.36	1.22	1.14	1.07

Tableau B.1: Valeurs critiques de la statistique de Kolmogorov-Smirnov

B.3 Diagrammes en Boîtes à "moustaches"

voir ([40]) :
Le diagramme en boîte représente schématiquement les principales caractéristiques d'une variable numérique en utilisant les quartiles Q_1, Q_2, Q_3, définis par $F(Q_1) = 0.25$ $F(Q_2) = 0.5$ $F(Q_3) = 0.75$. Dans la version courante, la partie centrale de la distribution est représentée par une boîte de largeur arbitraire et dont la longueur correspond à l'intervalle interquartile. On trace à l'intérieur la position de la médiane. La boîte est alors complétée par des "moustaches" correspondant aux valeurs adjacentes :

- adjacente supérieure : plus grande valeur inférieure à $Q_3 + 1.5(Q_3 - Q_1)$;
- adjacente inférieure : plus grande valeur inférieure à $Q_3 - 1.5(Q_3 - Q_1)$.

Les valeurs "extérieures" représentées par des carrés sont celles qui sortent des "moustaches".

B.4 Estimation du coefficient de corrélation entre intervalles de temps adjacents par la méthode Bootstrap

On utilisera les notations suivantes dans toute la suite :

- N nombre d'estimations du coefficient de corrélation, nous avons choisi $N = 600$
- cormoy : estimation de la valeur moyenne du coefficient de corrélation
- corstd : estimation de l'écart-type du coefficient de corrélation
- corinf : estimation de la valeur minimale du coefficient de corrélation
- comax : estimation de la valeur maximale du coefficient de corrélation
- cormed : estimation de la valeur médiane du coefficient de corrélation
- q0025 : estimation de la valeur du quantile de 2.5 % du coefficient de corrélation
- q0975 : estimation de la valeur du quantile de 97.5 % du coefficient de corrélation

Remarques :

- Nous pouvons cependant nous interroger sur le sens de cette corrélation (même problème dans [50]) : que signifie la relation entre la dernière valeur d'une année et la première de la suivante? Pour vérifier si les variables S_i et S_{i-1} étaient liées à la variable aléatoire année t, nous avons calculé le coefficient de corrélation partielle entre S_i et S_{i-1}, mais les résultats de la méthode Bootstrap appliquée au coefficient de corrélation partiel étant sensiblement identiques á ceux obtenus pour le coefficient de corrélation nous n'avons retenu ici que les résultats concernant le coefficient de corrélation.
- Il semble difficile d'appliquer des tests plus sophistiqués à l'indépendance des intervalles de temps entre deux jours de dépassement du seuil fixé, à cause de la complexité du modèle. Ces difficultés proviennent du fait que :
  (i) les intervalles ne sont pas stationnaires
  (ii) les intervalles ne sont pas normalement distribués
  (iii) le processus stochastique utilisé pour approcher le processus discret des jours de dépassement est continu.

B.5 Quelques rappels sur la régression logistique

cf. [22]

B.5.1 Différences avec la régression linéaire

- La première différence concerne la nature de la relation entre la variable et les covariables. Dans tout problème de régression la quantité clef est l'espérance conditionnelle $E[Z/s(t), w_1(t), \cdots, w_p(t)]$ où Z représente la variable et $s(t)$ et $w_j(t), 1 \leq j \leq p$ les valeurs des covariables. Pour la régression linéaire, on suppose que cette espérance peut s'exprimer sous la forme :

$$E[Z/s(t), w_1(t), \cdots, w_p(t)] = \alpha_0 + \alpha_1 s(t) + \sum_{j=2}^{p} \alpha_j w_j(t)$$

Cette expression implique que $E[Z/s(t), w_1(t), \cdots, w_p(t)]$ peut prendre toutes les valeurs possibles entre $-\infty$ et $+\infty$. Pour la régression logistique, Z est une variable aléatoire dichotomique et donc son espérance conditionnelle doit être comprise entre 0 et 1. Par conséquent, La forme spécifique du modèle de régression logistique est la suivante :

$$\pi(t) = \frac{e^{\alpha_0 + \alpha_1 s(t) + \sum_{j=2}^{p} \alpha_j w_j(t)}}{1 + e^{\alpha_0 + \alpha_1 s(t) + \sum_{j=2}^{p} \alpha_j w_j(t)}}$$

La transformation logit centrale dans le modèle logistique, est définie par :

$$g(s(t); w_j(t) : j = 1, \cdots p) = ln\left[\frac{\pi(t)}{1 + \pi(t)}\right] = \alpha_0 + \alpha_1 s(t) + \sum_{j=2}^{p} \alpha_j w_j(t)$$

- La deuxième différence importante entre la régression linéaire et la régression logistique concerne la distribution conditionnelle de la variable aléatoire Z. Dans le cas du modèle de régression linéaire, on suppose qu'une observation de la variable Z peut être exprimée sous la forme $y = E[Z/s(t), w_1(t), \cdots, w_p(t)]$. La quantité ε appelée l'erreur est supposée suivre une loi normale d'espérance 0 et de variance constante quel que soit le niveau de la covariable. Ainsi, la distribution conditionnelle de Z sachant $s(t), w_j(t)\ \forall j = 1, \cdots, p$ sera normale d'espérance $E[Z/s(t), w_1(t), \cdots, w_p(t)]$ et de variance constante.
Dans le cas où Z est dichotomique, la situation est différente : on écrit $Z = \pi(t) + \varepsilon$. La quantité ε ne peut prendre que deux valeurs : si $y = 1$, alors $\varepsilon = 1 - \pi(t)$ avec la probabilité $\pi(t)$, et si $y = 0$ alors $\varepsilon = -\pi(t)$ avec la probabilité $1 - \pi(t)$. ε possède donc une distribution d'espérance nulle et de variance égale à $\pi(t)\,(1 - \pi(t))$. Par conséquent, la distribution conditionnelle de Z sachant $s(t), w_j(t)\ \forall j = 1, \cdots, p$ est une distribution de Bernoulli de probabilité donnée par l'espérance conditionnelle, $\pi(t)$.

B.5.2 Interprétation des coefficients du modèle de régression logistique

Cas simple

Le **odds ratio** ou **rapport des chances**, ψ, est un paramètre intéressant dans une régression logistique car son interprétation est simple. Il permet de quantifier par exemple dans l'étude de la tendance l'évolution du nombre de dépassements d'un seuil u élevé sur la période d'étude. On définit l'estimation du log-odds pour l'effet année $s(t)$ sur l'ensemble de la période d'étude (1988-1997) des données de pollution mesurées en région parisienne, par :

$$ln(\hat{\psi}(s(t) = 1, s(t) = 10)) = \hat{g}(s(t) = 10; w_j(t), j = 1, \cdots, p) - \hat{g}(s(t) = 1; w_j(t), j = 1, \cdots, p) = \hat{\alpha}_1 \times 9$$

où $\hat{\psi}$ est l'estimation du **rapport des chances**, donc

$$\hat{\psi}(s(t)=1, s(t)=10) = \frac{\hat{\pi}(s(t)=10)\,(1-\hat{\pi}(s(t)=10))}{\hat{\pi}(s(t)=1)\,(1-\hat{\pi}(s(t)=1))} = \exp\left[\hat{\alpha}_1 \times 9\right]$$

L'estimation du rapport des chances sur 10 ans $\hat{\psi}(s(t)=1, s(t)=10) = x$ indique que le risque d'observer un dépassement en 1997 est x fois celui de 1988. De plus, en théorie, pour des échantillons de taille suffisamment grande, la distribution de $\hat{\psi}$ est normale. Malheureusement, cette taille d'échantillon s'avère dans de nombreux cas pratiques trop grande. Par conséquent, les inférences sont souvent basées sur la distribution de $\hat{\alpha_1}$ qui tend vers une loi normale pour des échantillons de taille beaucoup plus petite. Un intervalle de confiance à $100(1-\alpha)$ % estimé pour le rapport des chances est obtenu en calculant les bornes de l'intervalle de confiance de $\hat{\alpha_1}$, puis prenant l'exponentielle des ces valeurs. En général, les bornes sont données par (ici pour $\alpha = 5\%$) :

$$\exp\left[\hat{\alpha}_1 \times 9 \pm 1.96 \times 9 \times \hat{\sigma}(\hat{\alpha}_1)\right]$$

Estimation des rapports des chances dans le cas d'un modèle logistique avec interaction

Considérons le modèle contenant les covariables $s(t), w_j(t) : j = 1, \cdots, p$ et $w_2(t) \times s(t)$. La fonction Logit pour ce modèle s'écrit :

$$g(s(t); w_j(t) : j = 1, \cdots, p; s(t)w_2(t)) = \alpha_0 + \alpha_1 s(t) + \sum_{j=2}^{p} \alpha_j w_j(t) + \alpha_{12} s(t) \times w_2(t)$$

le log-odds pour $s(t) = 1$ contre $s(t) = 10$ avec $w_2(t) = tmax$ est

$$\begin{aligned} ln\left[\psi(s(t)=1, s(t)=10, w_2(t)=tmax)\right] &= g(1, tmax) - g(10, tmax) && \text{(B.2)} \\ &= \alpha_1(10-1) + \alpha_{12} tmax \times (10-1) && \text{(B.3)} \end{aligned}$$

L'estimation du log odds est obtenu en remplaçant les paramètres dans l'équation B.2 par leurs estimations.
L'estimation de la variance de l'expression dans B.2 est :

$$\hat{var}\{ln\left[\hat{\psi}(s(t)=1, s(t)=10, w_2(t)=tmax)\right]\} = \quad \text{(B.4)}$$

$$\hat{var}(\hat{\alpha}_1)(10-1)^2 + \hat{var}(\hat{\alpha}_{12})[tmax(10-1)]^2 + 2\hat{cov}(\hat{\alpha}_1, \hat{\alpha}_{12}) tmax[10-1]^2 \quad \text{(B.5)}$$

et les bornes de l'intervalle de confiance à 95 % de $\psi(s(t)=1, s(t)=10, w_2(t)=tmax)$ sont :

$$\exp\left(\left[\hat{\alpha}_1 \times 9 + \hat{\alpha}_{12} tmax \times 9\right] \pm 1.96 \left(\hat{var}\{ln\left[\hat{\psi}(s(t)=1, s(t)=10, w_2(t)=tmax)\right]\}\right)^{1/2}\right).$$

B.6 Description de la procédure LOGISTIC sous SAS

La procédure LOGISTIC estime les paramètres de modèles de régression logistique linéaire pour des variables réponse binaires ou ordinales par la méthode du maximum de vraisemblance. Des sous-ensembles de variables explicatives peuvent-être choisis par des méthodes de sélection de modèles variées. Dans notre étude, nous avons utilisé cette procédure pour la variable réponse binaire dépassement/non-dépassement.

B.6.1 Modèles utilisés par la procédure LOGISTIC pour une variable réponse binaire

La réponse,Y, d'une unité expérimentale ou d'un individu peut prendre une des deux valeurs notées 1 et 2 (par exemple $Y = 1$ si une maladie est présente; sinon $Y = 2$). Supposons que

X soit un vecteur de variables explicatives et que $p = Pr(Y = 1/X)$ soit la probabilité à modéliser, le modèle linéaire logistique a la forme:

$$logit\,(p) = \log(p/(1-p)) = \alpha + \beta' X$$

où α est le paramètre constant, et β est le vecteur des paramètres de tendance. Le modèle logistique possède une forme commune avec une classe plus générale de modèles linéaires dans laquelle une fonction $g = g(\mu)$ (appelée fonction de lien) de la moyenne de la variable réponse est supposée être liée linéairement aux variables explicatives. Puisque la moyenne μ dépend implicitement de la forme stochastique de la réponse, et que les variables explicatives sont supposées fixées, la fonction g fournit le lien entre la composante aléatoire et la composante déterministe de la variable réponse Y. Les fonctions de liens couramment utilisées en pratique sont les fonctions logit, normit et log-log.Cette classe de modèles s'écrit:

$$g(p) = \alpha + \beta' X$$

Dans notre cas, la fonction de lien utilisée est la fonction logit $g(p) = \log(p/(1-p))$ qui est l'inverse de la distribution logistique $F(x) = 1/(1+\exp(-x))$.

B.6.2 Description statistique de la procédure

Observations déterminantes pour la vraisemblance

Supposons que la variable réponse puisse prendre les valeurs ordonnées 1,...,k,k+1, où k est un entier ≥ 1. La probabilité que la j-ième observation ait la réponse i est donnée par:

$$Pr(Y_j = i/X_j) = \begin{cases} F(\alpha_1 + \beta' X_j) & i = 1 \\ F(\alpha_i + \beta' X_j) - F(\alpha_{i-1} + \beta' X_j) & 1 < i \leq k \\ 1 - F(\alpha_k + \beta' X_j) & i = k+1 \end{cases}$$

où Y_j est la variable réponse correspondant au vecteur $X_j{}'$ des variables explicatives; $\alpha_1, \cdots, \alpha_k$ sont les paramètres constants et β le vecteur des autres paramètres.

PROC LOGISTIC s'utilise de façon similaire aux autres procédures de régression existantes sous SAS. Avant d'entamer l'étape d'estimation, la procédure LOGISTIC calcule le score statistique global pour tester la signification commune à toutes les variables explicatives du modèle.
Les estimateurs du maximum de vraisemblance des paramètres de régression sont calculés en utilisant l'algorithme des moindres carrés avec repondération itérative (Iteratively Reweighted Least Squares (IRLS)).

IRLS

Considérons la variable multinomiale $Z_j = (Z_{1j}, \cdots, Z_{(k+1)j})'$ telle que:

$$Z_{ij} = \begin{cases} 1 \text{ si } Y_j = i \\ 0 \text{ sinon.} \end{cases}$$

Avec p_{ij} représentant la probabilité que la j-ième observation ait la réponse i, la valeur estimée de Z_j est $p_j = (p_{1j}, \cdots, p_{(k+1)j})'$. La matrice de covariance de Z_j est V_j. Soient γ le vecteur des paramètres de régression, c'est-à-dire, $\gamma' = (\alpha_1, \cdots, \alpha_k, \beta')$ et D_j la matrice des dérivées partielles de p_j par rapport à γ. L'équation à résoudre pour estimer les paramètres de régression est:

$$\Sigma_j D_j{}' W_j (Z_j - p_j) = 0$$

où $W_j = w_j V_j^-$, w_j est le poids de la j-ième observation et V_j^- est l'inverse généralisée de V_j. La procédure LOGISTIC choisit V_j^- comme l'inverse de la matrice diagonale formée par les

p_j.
Les estimateurs sont obtenus de façon itérative :

$$\hat{\gamma_{m+1}} = \hat{\gamma_m} + \left(\Sigma_j \hat{D_j}' \hat{W_j} \hat{D_j}\right)^{-1} \Sigma_j \hat{D_j}' \hat{W_j} \left(Z_j - \hat{p_j}\right)$$

où $\hat{D_j}$, $\hat{W_j}$ et $\hat{p_j}$ sont respectivement D_j, W_j et p_j évalués en $\hat{\gamma_m}$. L'expression après le signe plus correspond à la taille du pas. Si la vraisemblance évaluée en $\hat{\gamma}_{m+1}$ est plus petite que celle évaluée en $\hat{\gamma}_m$, alors $\hat{\gamma}_{m+1}$ est recalculé en utilisant la moitié de la taille du pas.
La matrice de covariance estimée de $\hat{\gamma}_{m+1}$ est alors :

$$cov\left(\hat{\gamma}_{m+1}\right) = \left(\Sigma_j \hat{D_j}' \hat{W_j} \hat{D_j}\right)^{-1}$$

Critères d'évaluation du modèle

Supposons que le modèle comprenne s variables explicatives. Soit y_j la valeur réponse de la j-ième observation. L'estimation $\hat{p_j} = Pr\left(Y_j = y_j / X_j\right)$ est obtenue en remplaçant les coefficients de régression par leurs estimations. Les trois critères (scores) suivants sont calculés :

- $-2\log$-vraisemblance
$$-2\log L = -2\Sigma_j w_j \log\left(\hat{p_j}\right)$$
où w_j est le poids de la j-ième observation.
- Critère d'information d'Akaike
$$AIC = -2\log L + 2\left(k + s\right)$$
où k est le nombre de valeurs ordonnées moins une pour la variable réponse, et s le nombre de variables explicatives.
- Critère de Schwartz
$$SC = -2\log L + \left(k + s\right)\log\left(N\right)$$
où N est le nombre total d'observations.

La statistique $-2\log L$ admettant la distribution du χ^2 sous l'hypothèse nulle (toutes les variables explicatives dans le modèle sont nulles) et la procédure affiche la p-valeur de cette statistique. Les statistiques AIC et SC fournissent deux ajustements de la statistique $-2\log L$ tenant compte du nombre de termes dans le modèle et du nombre d'observations utilisées ; ainsi elles pénalisent les modèles surparamétrés.

Calcul du χ-deux résiduel

Pour comprendre la forme générale des statistiques de score, soit $U\left(\gamma\right)$ le vecteur des dérivées partielles de la log-vraisemblance par rapport au vecteur de paramètres γ et $-I\left(\gamma\right)$ la matrice des dérivées partielles secondes de la log-vraisemblance par rapport au vecteur de paramètres γ. Sous l'hypothèse $\gamma = \gamma_0$, la statistique du χ-deux résiduel définie par :

$$U\left(\gamma_0\right)' I\left(\gamma_0\right)^{-1} U\left(\gamma_0\right)$$

admet une distribution asymptotique du χ-deux à r degrés de liberté, où r est la dimension de γ.
Si on utilise l'option SELECTION=FORWARD, BACKWARD ou STEPWISE, la procédure calcule une statistique de χ-deux résiduel et affiche la statistique, ses degrés de liberté et la valeur de la probabilité sous H_0 (appelée p-valeur dans toute la suite).

Supposons que le modèle comprenne s variables explicatives. Le modèle complet possède le vecteur paramètre :

$$\gamma = (\alpha_1, \cdots, \alpha_k, \beta_1, \cdots, \beta_s)'$$

Soient $\hat{\alpha_1}, \cdots, \hat{\alpha_k}$ et $\hat{\beta}_1, \cdots, \hat{\beta}_t$ les estimateurs du maximum de vraisemblance d'un modèle réduit à t variables explicatives ($t < s$). Le χ-deux résiduel est la «χ-Squared score »statistique évaluée en γ_0 donné par :

$$\gamma_0 = \left(\hat{\alpha_1}, \cdots, \hat{\alpha_k}, \hat{\beta}_1, \cdots, \hat{\beta}_t, 0, \cdots, 0\right)'$$

Le χ-deux résiduel admet une distribution asymptotique de χ-deux à s-t degrés de liberté.

Sélection des variables

Après quelques essais, nous avons opté pour une sélection descendante (SELECTION=BACKWARD). La table contenue dans la sortie de la procédure, intitulée «Summary of Backward Elimination Procedure »fournit l'approximation de la statistique du χ-deux pour la variable éliminée et la p-valeur correspondant à une distribution du χ-deux à 1 degré de liberté.

Prédicteur linéaire, probabilité estimée et intervalles de confiance

Pour un vecteur X de variables explicatives, le prédicteur linéaire $\eta_i = g\left(Pr\left(Y \leq i/X\right)\right) = \alpha_i + \beta' X$, où $1 \leq i \leq k$, est estimé par $\hat{\eta}_i = \hat{\alpha_i} + \hat{\beta}' X$. L'écart-type estimé de η_i est $\hat{\sigma}\left(\hat{\eta}_i\right)$ qui correspond à la racine carrée de la forme quadratique $(1, X')\, V_b (1, X')'$ où V_b est la matrice de covariance estimée des paramètres estimés. L'intervalle de confiance asymptotique à $100\,(1-\alpha)\,\%$ pour η_i est donné par :

$$\hat{\eta}_i \pm z_{\alpha/2} \hat{\sigma}\left(\hat{\eta}_i\right)$$

où $z_{\alpha/2}$ est le percentile $100\,(1-\alpha)$ d'une distribution normale $N(0,1)$. Ainsi, dans le cas de la fonction de lien logit, la valeur prédite de $Pr\left(Y \leq i/X\right)$ est $1/\left(1+\exp\left(-\hat{\eta}_i\right)\right)$, et les bornes inférieure et supérieure de l'intervalle de confiance sont respectivement $1/\left(1+\exp\left(-\left(\hat{\eta}_i - z_{\alpha/2}\hat{\sigma}\left(\hat{\eta}_i\right)\right)\right)\right)$ et $1/\left(1+\exp\left(-\left(\hat{\eta}_i + z_{\alpha/2}\hat{\sigma}\left(\hat{\eta}_i\right)\right)\right)\right)$.

Analyse des estimateurs du maximum de vraisemblance

La sortie de la procédure LOGISTIC comporte une table contenant :

- l'estimation du maximum de vraisemblance de chaque paramètre.
- L'écart-type estimé de chaque paramètre estimé, égal à la racine carrée de l'élément diagonal (correspondant au paramètre) de la matrice de covariance estimée .
- La statistique de Wald du χ-deux égale au carré du paramètre estimé divisé par son écart-type estimé.
- La p-valeur de la statistique du χ-deux de Wald obtenue à partir de la distribution du χ-deux à un degré de liberté.
- L'estimation standardisée de chaque paramètre β_j égale à au paramètre β_j estimé divisé par le ratio de l'écart-type de la distribution sous-jacente (inverse de la fonction de lien) et de l'écart-type de la variable explicative calculé sur l'échantillon de données.

Corrélation de rang entre les réponses observées et les probabilités estimées

Pour étudier la qualité de l'ajustement aux données, il faut regarder la relation entre la réponse et la probabilité estimée. Pour deux réponses différentes, on dira que cette paire d'observations est concordante, si la probabilité estimée et la réponse observée concordent. Inversement, on dira que cette paire est discordante, si la probabilité estimée et la réponse observée sont différentes. Une paire d'observations sera dite ex aequo (tied en anglais), si ces deux probabilités sont égales. Notant :

- N, le nombre total d'observations
- t, le nombre total de paires avec des réponses différentes
- nc, le nombre de paires concordantes
- nd, le nombre de paires discordantes

On observe donc $t - nc - nd$ paires ex aequo. La procédure LOGISTIC calcule les quatre indices de corrélation de rang permettant d'évaluer la capacité prédictive du modèle :

$$c = \frac{nc + 0.5(t - nc - nd)}{t}$$

$$\text{Somers'D} = \frac{nc - nd}{t}$$

$$\text{Goodman-Kruskal Gamma} = \frac{nc - nd}{nc + nd}$$

$$\text{Kendall's Tau-a} = \frac{nc - nd}{0.5N(N-1)}$$

B.7 Résultats détaillés des modèles sans interaction par station pour la région parisienne

B.7.1 Neuilly/Seine

- $u = 120\ \mu g/m3$
  Dans ce premier modèle, nous avons utilisé un seuil u de 120 $\mu g/m3$ et déterminé parmi les variables s, $TMAX$, $TRANGE$, $WSAVG$ et $WSRANGE$ les variables pertinentes pour modéliser la fréquence et la taille des dépassements.
  * *Fréquence des dépassements*

    Nous avons posé $dep120 = 1$ si la valeur du maximum d'ozone est strictement supérieure à 120 $\mu g/m3$ et $dep120 = 2$ sinon. Puis, nous avons utilisé la procédure LOGISTIC de SAS (cf. annexe (B.6)) avec sélection progressive descendante des variables :

```
The LOGISTIC Procedure

Data Set: WORK.HLEVEL
Response Variable: DEP120
Response Levels: 2
Number of Observations: 856
Link Function: Logit

Response Profile

Ordered
Value     DEP120  Count
    1          1     95
    2          2    761
```

```
WARNING: 524 observation(s) were deleted due to missing values for the
response or explanatory variables.
```

Model Fitting Information and Testing Global Null Hypothesis BETA=0

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	598.738	349.704	.
SC	603.491	363.961	.
-2 LOG L	596.738	343.704	253.034 with 2 DF (p=0.0001)
Score	.	.	210.670 with 2 DF (p=0.0001)

Residual Chi-Square = 4.3129 with 4 DF (p=0.3653)

Summary of Backward Elimination Procedure

Step	Variable Removed	Number In	Wald Chi-Square	Pr > Chi-Square
1	TRANGE	5	0.1318	0.7165
1	WSRANGE	4	0.4748	0.4908
1	T	3	2.1580	0.1418
1	T92	2	1.4886	0.2224

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	-11.7548	1.2392	89.9828	0.0001	.	.
TMAX	1	0.4474	0.0436	105.4438	0.0001	1.221772	1.564
WSAVG	1	-0.5076	0.1311	14.9893	0.0001	-0.417758	0.602

Association of Predicted Probabilities and Observed Responses

Concordant = 91.8%	Somers' D = 0.839		
Discordant = 7.9%	Gamma = 0.842		
Tied = 0.3%	Tau-a = 0.166		
(72295 pairs)	c = 0.920		

Estimated Correlation Matrix

Variable	INTERCPT	TMAX	WSAVG
INTERCPT	1.00000	-0.94185	-0.27035
TMAX	-0.94185	1.00000	-0.04889
WSAVG	-0.27035	-0.04889	1.00000

Le modèle pour $\alpha(t)$ est :

$$\hat{\alpha}(t) = \hat{\alpha_1} + \hat{\alpha_3} * \omega_3(t) + \hat{\alpha_5} * \omega_5(t)$$

où les différents paramètres prennent les valeurs indiquées ci-dessus.
Les coefficients associés aux variables météorologiques sont tels que nous pouvions le prévoir :

- Le coefficient positif 0.4474 associé à Tmax implique que plus la température maximale est élevée durant la journée, plus la probabilité d'avoir un dépassement est grande.
- Le coefficient négatif -0.5076 associé à Wmoy implique que plus la vitesse moyenne du vent est élevée durant la journée, plus la probabilité d'avoir un dépassement est petite.

On observe de plus que l'effet moyen et l'effet Tmax sont très corrélés (-0.94185). On effectue ensuite un test de Kolmogorov-Smirnov, comme décrit dans l'annexe (B.2). On obtient les résultats résumés dans le tableau (B.2) et le graphe (B.1). On

n	D_n^+	D_n^-	$d_{n,0.05}$
109	0.118	0.084	0.130

Tableau B.2: Test sur la distribution des intervalles ($u = 120\ \mu g/m3$)

accepte donc l'hypothèse nulle H_0 : $F_S = F_0$ au niveau 0.05, puisque la valeur observée de D_n est strictement inférieure à celle de $d_{109,0.05}$. Le processus de Poisson non homogène utilisé pour modéliser la fréquence des jours où le maximum d'ozone dépasse 120 $\mu g m^{-3}$ est une approximation acceptable.

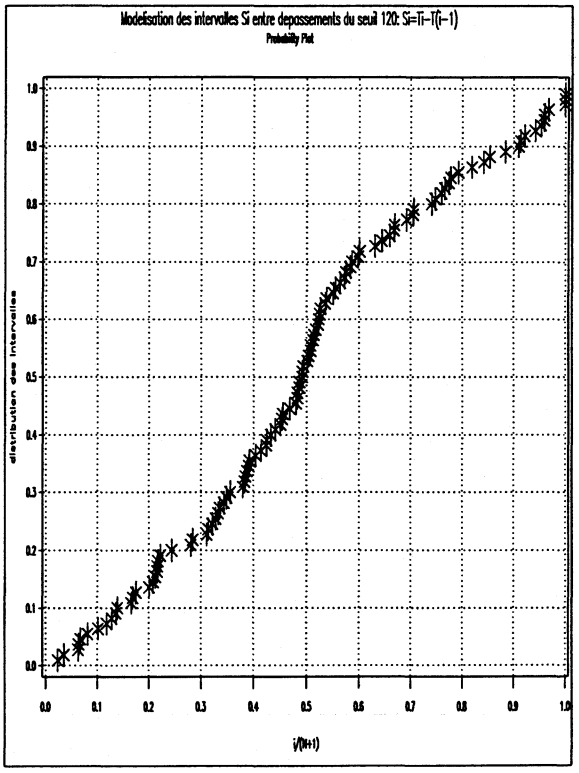


Figure B.1: Probability plot ($u = 120\ \mu g/m3$)

* *Tailles des dépassements*

Le modèle pour $\beta(t)$ est :

$$\hat{\beta}(t) = \hat{\beta}_1 + \hat{\beta}_3 * \omega_3(t) + \hat{\beta}_5 * \omega_5(t)$$

où les différents paramètres prennent les valeurs indiquées dans le tableau (B.3). L'observation majeure qui en découle est le fait que l'effet année n'est pas significatif. Il semblerait donc qu'à conditions météorologiques constantes, la taille des dépassements n'ait pas évolué durant la période d'étude. De plus, comme l'estimation de la taille du dépassement le jour t est $1/\beta(t)$, on en déduit que :

- Le coefficient négatif -0.0024 associé à Tmax implique que plus la température maximale est élevée durant la journée, plus l'estimation de la taille du dépassement est grande.

- Le coefficient positif 0.0137 associé à WSAVG implique que plus la vitesse moyenne du vent est élevée durant la journée, plus l'estimation de la taille du dépassement est petite.

Variable	Paramètre	Estimation	Ecart-type	T-ratio
c^{te}	β_1	0.0692	0.0297	2.3339
Tmax	β_3	-0.0024	0.0009	-2.6932
WSAVG	β_5	0.0137	0.0039	3.5578

Tableau B.3: Taille des dépassements ($u = 120\ \mu g/m3$)

On obtient la matrice de corrélation entre paramètres estimés suivante :

$$\text{Cor} = \begin{pmatrix} 1.00000 & -0.95367 & -0.38832 \\ -0.95367 & 1.00000 & 0.11184 \\ -0.38832 & 0.11184 & 1.00000 \end{pmatrix}$$

On observe que l'effet moyen et l'effet de Tmax sont très corrélés (-0.95367).
On effectue ensuite un test de Kolmogorov-Smirnov, comme décrit dans l'annexe (B.2). On obtient les résultats résumés dans le tableau (B.4) et le graphe (B.2). On

n	D_n^+	D_n^-	$d_{n,0.05}$
95	0.053	0.046	0.140

Tableau B.4: Test sur la distribution des tailles de dépassement ($u = 120\ \mu g/m3$)

accepte donc l'hypothèse nulle H_0 : $F_S = F_0$ au niveau 0.05, puisque la valeur observée de D_n est inférieure à celle de $d_{95,0.05}$.

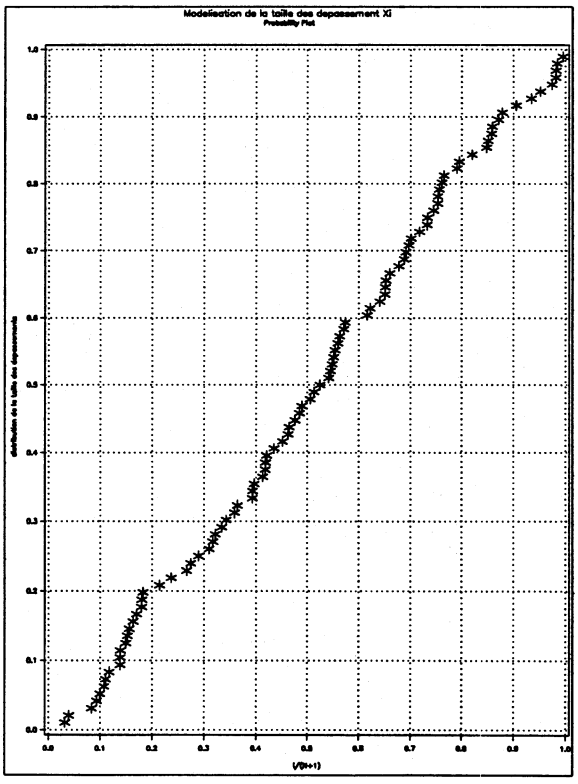


Figure B.2: Probability plot ($u = 120\ \mu g/m3$)

La densité exponentielle est donc un modèle approprié aux tailles de dépassement X_i.

Remarque :
Le nombre d'observations sur lequel est effectué le test de Kolmogorov-Smirnov pour la distribution des intervalles S_i et celui pour la distribution des tailles des dépassements n'est pas le même. En effet, dans le cas de la taille des dépassements, seuls les jours de dépassements où toutes les variables météorologiques ont pu être mesurées sont comptabilisés. Alors que, dans le cas des intervalles S_i entre deux dépassements, la distribution des intervalles doit vérifier (7.30), nous avons conservé tous les jours où un dépassement avait eu lieu et supposé que l'absence de données météorologiques le jour d'un dépassement avait lieu aléatoirement, de façon à conserver un maximum d'intervalles S_i pour effectuer le test.

Conclusions et analyse critique du modèle

Nous avions observé dans une étude préliminaire que le seuil u de 110 μgm^{-3} était trop petit pour que le processus du nombre de dépassements de ce seuil sur le site de Neuilly/Seine soit modélisé par un processus de Poisson non-homogène, alors que la distribution de la taille des dépassements du seuil 110 μgm^{-3} était bien approchée par une distribution exponentielle.
Nous avons alors dans ce travail changé de seuil et déterminé que $u = 120$ μgm^{-3} permettait de modéliser le processus du nombre de dépassements de ce seuil par un processus de Poisson non-homogène. Mais cette modélisation ne permet malheureusement pas de dégager de tendance à long terme dans les valeurs du maximum d'ozone troposphérique mesurées à Neuilly-sur-Seine et dépassant 120 μgm^{-3}.
Il nous est alors apparu intéressant de reprendre cette étude en prenant le seuil 130 μgm^{-3}.

- $u = 130$ $\mu g/m3$

Nous présentons ici uniquement les résultats pour le seuil $u = 130$ $\mu g/m3$, seuil à partir duquel une tendance se dégage.

* *Fréquence des dépassements*

Nous avons donc posé $dep130 = 1$ si la valeur du maximum d'ozone est strictement supérieure à 130 $\mu g/m3$ et $dep130 = 2$ sinon.

```
          The LOGISTIC Procedure

    Data Set: WORK.HLEVEL
    Response Variable: DEP130
    Response Levels: 2
    Number of Observations: 856
    Link Function: Logit

          Response Profile

          Ordered
          Value     DEP130   Count
              1          1      62
              2          2     794

WARNING: 524 observation(s) were deleted due to missing values for the
response or explanatory variables.


        Model Fitting Information and Testing Global Null Hypothesis BETA=0

                        Intercept
        Intercept          and
```

Criterion	Only	Covariates	Chi-Square for Covariates
AIC	446.914	256.457	.
SC	451.666	280.218	.
-2 LOG L	444.914	246.457	198.457 with 4 DF (p=0.0001)
Score	.	.	164.342 with 4 DF (p=0.0001)

Residual Chi-Square = 0.0580 with 2 DF (p=0.9714)

Summary of Backward Elimination Procedure

Step	Variable Removed	Number In	Wald Chi-Square	Pr > Chi-Square
1	WSRANGE	5	0.0112	0.9158
1	TRANGE	4	0.0468	0.8287

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	-9.7515	1.5369	40.2577	0.0001	.	.
T	1	0.3756	0.1276	8.6715	0.0032	0.562694	1.456
T92	1	-2.3780	0.6808	12.2007	0.0005	-0.608362	0.093
TMAX	1	0.4522	0.0535	71.4476	0.0001	1.235001	1.572
WSAVG	1	-0.8894	0.1994	19.8934	0.0001	-0.732015	0.411

Association of Predicted Probabilities and Observed Responses

Concordant	= 94.0%	Somers' D	= 0.881
Discordant	= 5.9%	Gamma	= 0.882
Tied	= 0.1%	Tau-a	= 0.119
(49228 pairs)		c	= 0.941

Estimated Correlation Matrix

Variable	INTERCPT	T	T92	TMAX	WSAVG
INTERCPT	1.00000	0.20710	-0.26279	-0.82989	-0.23005
T	0.20710	1.00000	-0.86572	0.09094	-0.39283
T92	-0.26279	-0.86572	1.00000	-0.17181	0.33669
TMAX	-0.82989	0.09094	-0.17181	1.00000	-0.19661
WSAVG	-0.23005	-0.39283	0.33669	-0.19661	1.00000

Le modèle pour $\alpha(t)$ est :

$$\hat{\alpha}(t) = \hat{\alpha_0} + \hat{\alpha_1} * s(t) + \hat{\alpha_2} * t92(t) + \hat{\alpha_3} * \omega_3(t) + \hat{\alpha_5} * \omega_5(t)$$

où les différents paramètres prennent les valeurs indiquées page suivante :

- Le coefficient positif 0.4522 associé à Tmax implique que plus la température maximale est élevée durant la journée, plus la probabilité d'avoir un dépassement est grande.
- Le coefficient négatif -0.8894 associé à Vmoy implique que plus la vitesse moyenne du vent est élevée durant la journée, plus la probabilité d'avoir un dépassement est petite.
- Le coefficient positif 0.3756 associé à $s(t)$ implique qu'à conditions météorologiques constantes, il y a une forte tendance globale à la hausse dans la fréquence des dépassements.
- Le coefficient négatif -2.3780 associé à $t92$ implique qu'en tenant compte des conditions météorologiques, il y a une tendance à la baisse dans la fréquence des dépassements entre la période 1988-1991 et la période 1992-1997.

On observe que l'effet de tendance global s et l'effet de tendance $t92$ entre les période 88-91 et 92-97 sont très corrélés négativement (-0.86572). En effectuant un test de Kolmogorov-Smirnov, on obtient les résultats résumés dans le tableau (B.5) et le graphe (B.3). On accepte donc l'hypothèse nulle H_0 : $F_S = F_0$ au niveau 0.05,

n	D_n^+	D_n^-	$d_{n,0.05}$
74	0.118	0.032	0.158

Tableau B.5: Test sur la distribution des intervalles ($u = 130\ \mu g/m3$ et introduction de $t92$)

puisque la valeur observée de D_n est strictement inférieure à celle de $d_{74,0.05}$. Le processus de Poisson non-homogène utilisé pour modéliser la fréquence des jours où le maximum d'ozone dépasse 130 $\mu g m^{-3}$ est une bonne approximation.

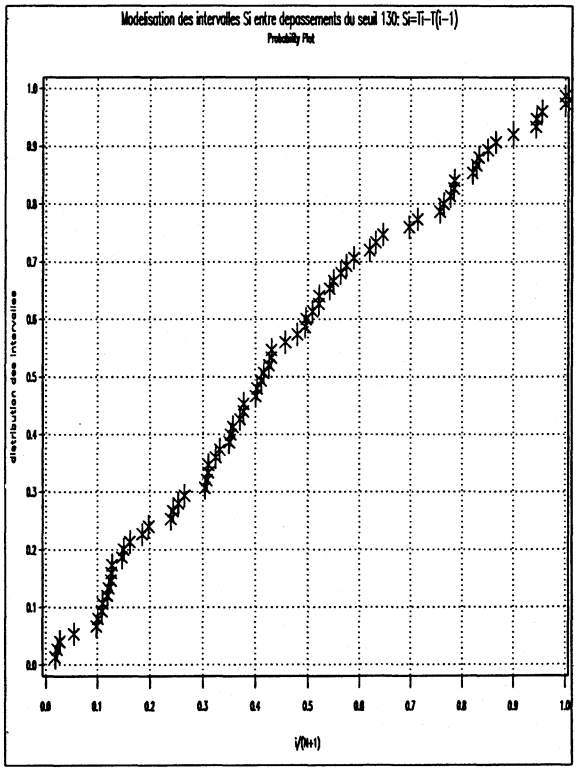


Figure B.3: Probability plot ($u = 130\ \mu g/m3$ et introduction de $t92$)

* *Tailles des dépassements*

Le modèle pour $\beta(t)$ est :

$$\hat{\beta}(t) = \hat{\beta}_1 * s(t) + \hat{\beta}_2 * t92(t) + \hat{\beta}_3 * \omega_3(t) + \hat{\beta}_5 * \omega_5(t)$$

où les différents paramètres prennent les valeurs indiquées dans le tableau (B.6). L'observation majeure qui découle est le fait que l'effet année "global" et l'effet de "période" sont significatifs. Il semblerait donc qu'à conditions météorologiques constantes, la taille des dépassements ait évolué durant la période d'étude. De plus, comme l'estimation de la taille du dépassement le jour t est $1/\beta(t)$, on en déduit que :

- Le coefficient négatif -0.0060 associé à s implique qu'en tenant compte des effets des conditions météorologiques, il y a une tendance à la hausse dans la taille des dépassements durant la période totale d'étude.

- Le coefficient positif 0.0358 associé à $t92$ implique qu'en tenant compte des effets des conditions météorologiques, il y a une tendance à la baisse dans la taille des dépassements entre la période 1988-1991 et la période 1992-1997.

- Le coefficient négatif -0.0012 associé à Tmax implique que plus la température maximale est élevée durant la journée, plus l'estimation de la taille du dépassement est grande.

- Le coefficient positif 0.0188 associé à WSAVG implique que plus la vitesse moyenne du vent est élevée durant la journée, plus l'estimation de la taille du dépassement est petite.

Variable	Paramètre	Estimation	Ecart-type	T-ratio
s	β_1	-0.0060	0.0030	-2.0301
$t92$	β_2	0.0358	0.0145	2.4695
Tmax	β_3	-0.0012	0.0004	-2.9457
WSAVG	β_5	0.0188	0.0046	4.0538

Tableau B.6: Taille des dépassements ($u = 130\ \mu g/m3$ et introduction de $t92$)

On obtient la matrice de corrélation entre paramètres estimés suivante :

$$\text{Cor} = \begin{pmatrix} 1.00000 & -0.89956 & 0.46218 & -0.26930 \\ -0.89956 & 1.00000 & -0.64082 & 0.15222 \\ 0.46218 & -0.64082 & 1 & -0.65847 \\ -0.26930 & 0.15222 & -0.65847 & 1.00000 \end{pmatrix}$$

On observe que l'effet de tendance "global" et l'effet de tendance "période" sont très corrélés (-0.89956).
En effectuant un test de Kolmogorov-Smirnov, on obtient les résultats résumés dans le tableau (B.7) et le graphe (B.4). On accepte donc l'hypothèse nulle H_0 : $F_S = F_0$

n	D_n^+	D_n^-	$d_{n,0.05}$
62	0.066	0.028	0.173

Tableau B.7: Test sur la distribution des tailles de dépassement ($u = 130\ \mu g/m3$ et introduction de $t92$)

au niveau 0.05, puisque la valeur observée de D_n est inférieure à celle de $d_{62,0.05}$.

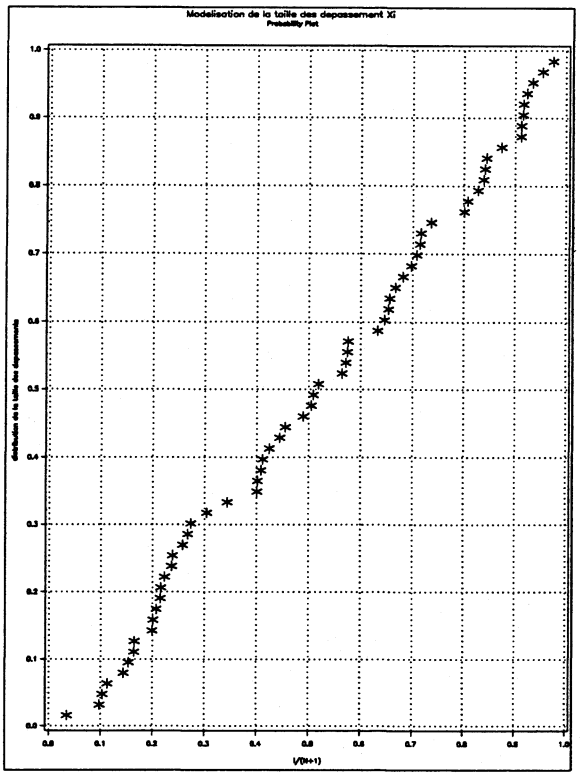


Figure B.4: Probability plot ($u = 130\ \mu g/m3$ et introduction de $t92$)

La densité exponentielle est donc un modèle acceptable pour modéliser les tailles de dépassement X_i.

Conclusions

La modélisation de la fréquence des dépassements du seuil $u = 130\ \mu g/m3$ et de leur taille par un processus de Poisson non homogène bi-dimensionnel permet d'observer qu'à conditions météorologiques constantes :

- les valeurs des excédents d'ozone ont augmenté.

- Le rapport des chances estimé pour une augmentation de 10 années est $\exp(10 * 0.3756) = 42.8$, par conséquent le risque d'avoir un dépassement de seuil $u = 130\ \mu g/m3$ en 1997 est 42.8 fois supérieur à celui de 1988 !
- Le rapport des chances estimé pour le passage de la période 1988-1991 à 1992-1997 est $\exp(-2.3780) = 0.093$, par conséquent le risque d'avoir un dépassement de seuil $u = 130\ \mu g/m3$ sur 1988-1991 est 0.093 inférieur à celui de la deuxième sous-période.
- les valeurs des excédents ont diminué entre les périodes 1988-1991 et 1992-1997.
- Une explication plausible de cet effet bloc (division significative de la période en deux sous périodes) pourrait être la meilleure précision des mesures dans la deuxième période, due non pas au changement de l'analyseur (effectué en Janvier 1993), mais à un changement dans le suivi et l'entretien de l'analyseur.

– COMPARAISONS DES DEUX MODÈLES ET CONCLUSIONS

A conditions météorologiques constantes, la fréquence et la taille des dépassements sur le site de Neuilly/Seine n'ont pas évolué durant la période, pour le seuil $u = 120\ \mu g/m3$; alors que, le risque d'avoir un dépassement dépassements du seuil $u = 130\ \mu g/m3$ en 1997 est environ 43 fois supérieur à celui de 1988 et a diminué entre les périodes 1988-1991 et 1992-1997. Par conséquent, cette modélisation permet de mettre en évidence

une évolution croissante du nombre et de la taille des dépassements du seuil u fixé à 130 $\mu g/m3$.

B.7.2 Champs/Marne

- $u = 120\ \mu g/m3$

  * *Fréquence des dépassements*
  Le modèle pour $\alpha(t)$ est :

$$\hat{\alpha}(t) = \hat{\alpha_0} + \hat{\alpha_1} * s(t) + \hat{\alpha_3} * \omega_3(t) + \hat{\alpha_5} * \omega_5(t)$$

où les différents paramètres prennent les valeurs indiquées ci-dessous.

The LOGISTIC Procedure

Data Set: WORK.HLEVEL
Response Variable: DEP120
Response Levels: 2
Number of Observations: 985
Link Function: Logit

Response Profile

Ordered Value	DEP120	Count
1	1	43
2	2	942

WARNING: 395 observation(s) were deleted due to missing values for the response or explanatory variables.

Model Fitting Information and Testing Global Null Hypothesis BETA=0

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	355.399	180.226	.
SC	360.292	199.796	.
-2 LOG L	353.399	172.226	181.173 with 3 DF (p=0.0001)
Score	.	.	153.204 with 3 DF (p=0.0001)

Residual Chi-Square = 1.1140 with 3 DF (p=0.7737)

Summary of Backward Elimination Procedure

Step	Variable Removed	Number In	Wald Chi-Square	Pr > Chi-Square
1	WSRANGE	5	0.0101	0.9200
1	TRANGE	4	0.5221	0.4699
1	T93	3	0.5731	0.4490

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	-18.9664	2.2830	69.0195	0.0001	.	.
T	1	0.7147	0.1242	33.1237	0.0001	1.081283	2.043
TMAX	1	0.4889	0.0636	59.1909	0.0001	1.325551	1.631
WSAVG	1	-0.5175	0.1856	7.7755	0.0053	-0.427218	0.596

Association of Predicted Probabilities and Observed Responses

Concordant	= 96.3%	Somers' D	= 0.928
Discordant	= 3.6%	Gamma	= 0.928
Tied	= 0.1%	Tau-a	= 0.078
(40506 pairs)		c	= 0.964

Estimated Correlation Matrix

Variable	INTERCPT	T	TMAX	WSAVG
INTERCPT	1.00000	-0.59282	-0.93379	-0.04117
T	-0.59282	1.00000	0.38093	-0.43290
TMAX	-0.93379	0.38093	1.00000	-0.05737
WSAVG	-0.04117	-0.43290	-0.05737	1.00000

On obtient l'ajustement graphique B.5.

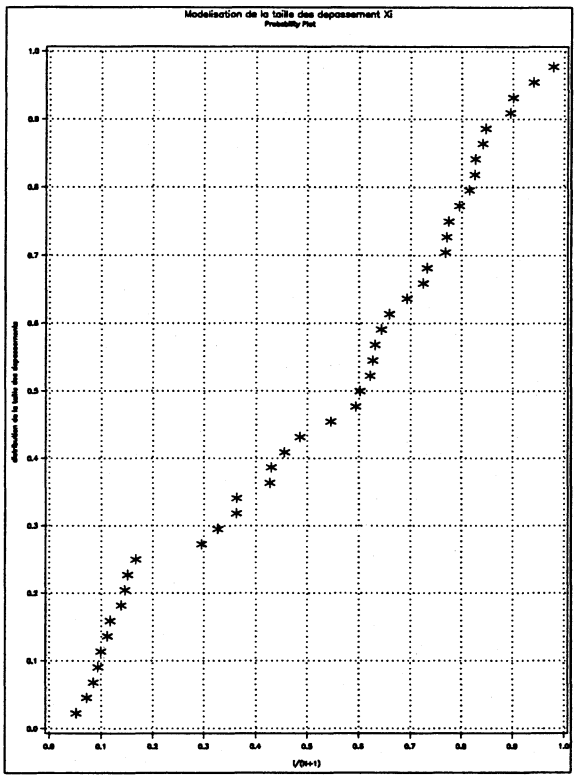


Figure B.5: Probability plot ($u = 120\ \mu g/m3$)

* *Taille des dépassements*
Le modèle pour $\beta(t)$ est :

$$\hat{\beta}(t) = \hat{\beta}_4 * \omega_4(t)$$

où le paramètre prend la valeur indiquée dans le tableau (B.8)
On obtient l'ajustement graphique B.6.

Variable	Paramètre	Estimation	Ecart-type	T-ratio
Trange	β_4	0.0032	0.0004	7.9838

Tableau B.8: Taille des dépassements ($u = 120\ \mu g/m3$)

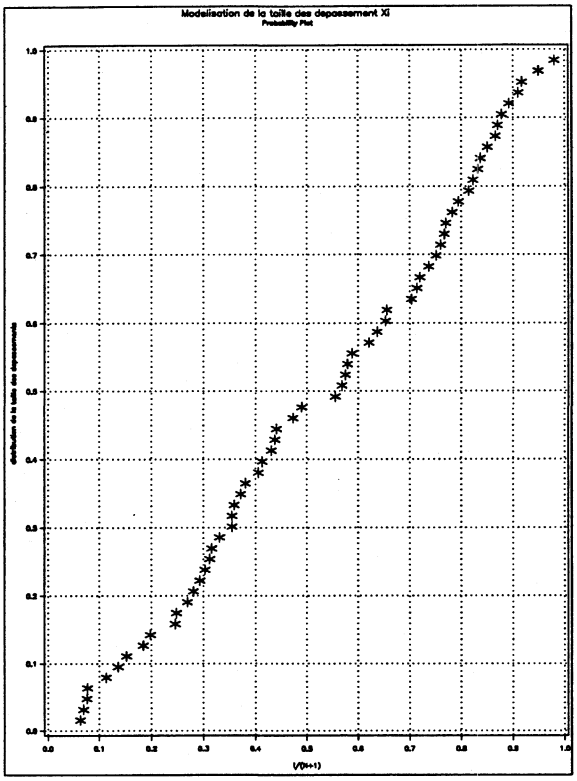


Figure B.6: Probability plot ($u = 120\ \mu g/m3$)

- $u = 130\ \mu g/m3$

  * *Fréquence des dépassements*
    Le modèle pour $\alpha(t)$ est :

$$\hat{\alpha}(t) = \hat{\alpha_0} + \hat{\alpha_1} * s(t) + \hat{\alpha_3} * \omega_3(t) + \hat{\alpha_5} * \omega_5(t)$$

    où les différents paramètres prennent les valeurs indiquées ci-dessous :

```
The LOGISTIC Procedure

        Data Set: WORK.HLEVEL
        Response Variable: DEP130
        Response Levels: 2
        Number of Observations: 985
        Link Function: Logit


          Response Profile
        Ordered
          Value  DEP130   Count
              1       1      32
              2       2     953

WARNING: 395 observation(s) were deleted due to missing values for the
response or explanatory variables.

        Model Fitting Information and Testing Global Null Hypothesis BETA=0

                          Intercept
                 Intercept     and
    Criterion        Only    Covariates   Chi-Square for Covariates
    AIC           284.271      138.360
```

SC	289.164	157.931	.
-2 LOG L	282.271	130.360	151.911 with 3 DF (p=0.0001)
Score	.	.	123.162 with 3 DF (p=0.0001)

Residual Chi-Square = 0.6395 with 3 DF (p=0.8873)

Summary of Backward Elimination Procedure

Step	Variable Removed	Number In	Wald Chi-Square	Pr > Chi-Square
1	WSRANGE	5	0.000921	0.9758
1	TRANGE	4	0.00142	0.9699
1	T93	3	0.6322	0.4266

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	-18.7919	2.6551	50.0929	0.0001	.	.
T	1	0.7998	0.1579	25.6714	0.0001	1.210136	2.225
TMAX	1	0.5030	0.0758	44.0609	0.0001	1.363639	1.654
WSAVG	1	-1.0858	0.2850	14.5164	0.0001	-0.896426	0.338

Association of Predicted Probabilities and Observed Responses

Concordant = 96.7%		Somers' D = 0.936	
Discordant = 3.1%		Gamma = 0.939	
Tied = 0.2%		Tau-a = 0.059	
(30496 pairs)		c = 0.968	

Estimated Correlation Matrix

Variable	INTERCPT	T	TMAX	WSAVG
INTERCPT	1.00000	-0.60958	-0.92831	0.14572
T	-0.60958	1.00000	0.40764	-0.53826
TMAX	-0.92831	0.40764	1.00000	-0.26205
WSAVG	0.14572	-0.53826	-0.26205	1.00000

On obtient l'ajustement graphique B.7.

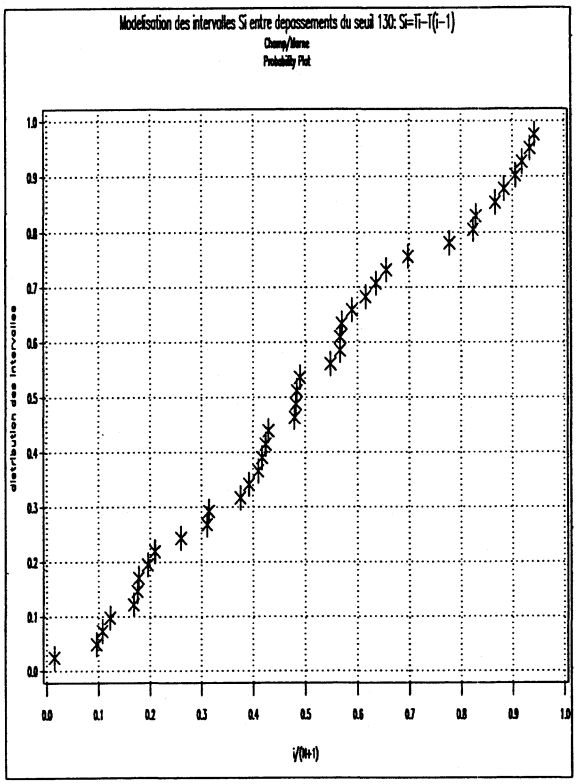


Figure B.7: Probability plot ($u = 130\ \mu g/m3$)

* *Taille des dépassements*
Le modèle pour $\beta(t)$ est :

$$\hat{\beta}(t) = \hat{\beta}_4 * \omega_4(t)$$

où le paramètre prend la valeur indiquée dans le tableau (B.9).

Variable	Paramètre	Estimation	Ecart-type	T-ratio
Trange	β_4	0.0036	0.0005	6.7418

Tableau B.9: Taille des dépassements ($u = 130\ \mu g/m3$)

On obtient l'ajustement graphique B.8.

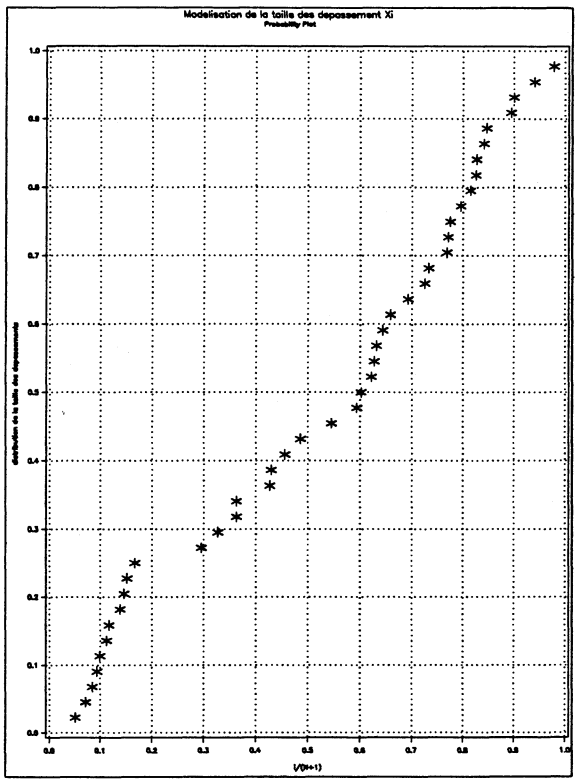


Figure B.8: Probability plot ($u = 130\ \mu g/m3$)

B.7.3 Aubervilliers, $u = 130\ \mu g/m3$

Fréquence des dépassements

Le modèle pour $\alpha(t)$ est :

$$\hat{\alpha}(t) = \hat{\alpha_0} + \hat{\alpha_1} * t93(t) + \hat{\alpha_3} * \omega_3(t) + \hat{\alpha_5} * \omega_5(t)$$

où les différents paramètres prennent les valeurs indiquées ci-dessous :

```
The LOGISTIC Procedure

        Data Set: WORK.HLEVEL
        Response Variable: DEP130
        Response Levels: 2
        Number of Observations: 857
        Link Function: Logit


              Response Profile

           Ordered
             Value DEP130  Count
                 1      1     48
                 2      2    809

WARNING: 523 observation(s) were deleted due to missing values for the
response or explanatory variables.

        Model Fitting Information and Testing Global Null Hypothesis BETA=0
```

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	371.955	163.965	.
SC	376.708	182.979	.
-2 LOG L	369.955	155.965	213.990 with 3 DF (p=0.0001)
Score	.	.	173.945 with 3 DF (p=0.0001)

Residual Chi-Square = 5.2354 with 3 DF (p=0.1553)

Summary of Backward Elimination Procedure

Step	Variable Removed	Number In	Wald Chi-Square	Pr > Chi-Square
1	T	5	1.1712	0.2791
1	TRANGE	4	2.0699	0.1502
1	WSRANGE	3	2.1380	0.1437

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	-23.0979	2.8276	66.7291	0.0001	.	.
T93	1	2.4179	0.5215	21.4969	0.0001	0.665214	11.222
TMAX	1	0.6588	0.0815	65.3783	0.0001	1.802999	1.933
WSAVG	1	-0.4614	0.2015	5.2428	0.0220	-0.381474	0.630

Association of Predicted Probabilities and Observed Responses

Concordant	= 97.3%	Somers' D	= 0.946
Discordant	= 2.7%	Gamma	= 0.947
Tied	= 0.1%	Tau-a	= 0.100
(38832 pairs)		c	= 0.973

Estimated Correlation Matrix

Variable	INTERCPT	T93	TMAX	WSAVG
INTERCPT	1.00000	-0.62503	-0.93606	-0.03364
T93	-0.62503	1.00000	0.40173	-0.15551
TMAX	-0.93606	0.40173	1.00000	-0.15585
WSAVG	-0.03364	-0.15551	-0.15585	1.00000

On obtient l'ajustement graphique B.9.

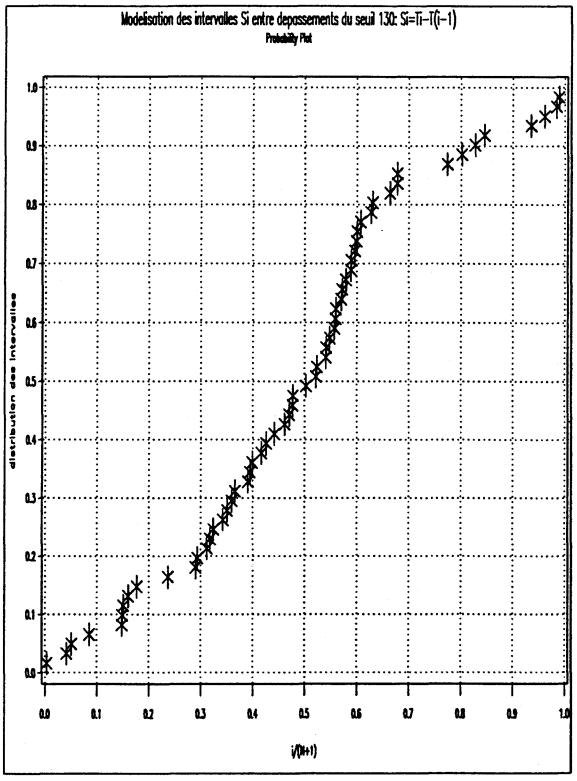


Figure B.9: Probability plot ($u = 130\ \mu g/m3$)

Taille des dépassements

Le modèle pour $\beta(t)$ est :

$$\hat{\beta}(t) = \hat{\beta}_0 + \hat{\beta}_3 * \omega_3(t) + \hat{\beta}_5 * \omega_5(t)$$

où le paramètre prend la valeur indiquée dans le tableau (B.10)

Variable	Paramètre	Estimation	Ecart-type	T-ratio
c^{te}	β_0	0.1413	0.0623	2.2663
Tmax	β_3	-0.0044	0.0018	-2.4778
Wsavg	β_5	0.0143	0.0064	2.2210

Tableau B.10: Taille des dépassements ($u = 130\ \mu g/m3$)

On obtient la matrice de corrélation entre paramètres estimés suivante :

$$\text{Cor} = \begin{pmatrix} 1.00000 & -0.97020 & -0.39374 \\ -0.97020 & 1.00000 & 0.17409 \\ -0.39374 & 0.17409 & 1.00000 \end{pmatrix}$$

On obtient l'ajustement graphique B.10.

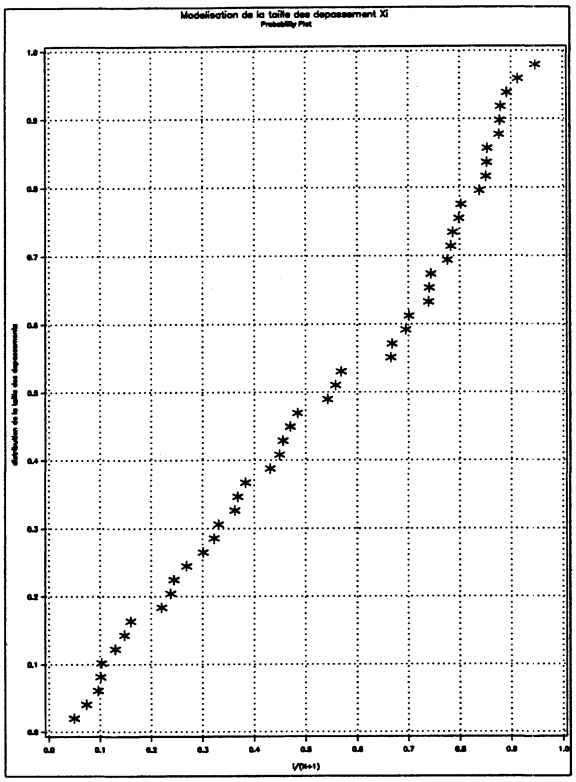


Figure B.10: Probability plot ($u = 130\ \mu g/m3$)

B.7.4 Créteil, $u = 130\ \mu g/m3$

Fréquence des dépassements

Le modèle pour $\alpha(t)$ est :

$$\hat{\alpha}(t) = \hat{\alpha_0} + \hat{\alpha_1} * s(t) + \hat{\alpha_3} * \omega_3(t) + \hat{\alpha_5} * \omega_5(t)$$

où les différents paramètres prennent les valeurs indiquées ci-dessous :

```
The LOGISTIC Procedure

        Data Set: WORK.HLEVEL
        Response Variable: DEP130
        Response Levels: 2
        Number of Observations: 855
        Link Function: Logit

              Response Profile

            Ordered
              Value DEP130  Count
                  1      1     50
                  2      2    805

WARNING: 525 observation(s) were deleted due to missing values for the
 response or explanatory variables.

        Model Fitting Information and Testing Global Null Hypothesis BETA=0

                              Intercept
                  Intercept       and
```

Criterion	Only	Covariates	Chi-Square for Covariates
AIC	382.925	248.042	.
SC	387.676	267.047	.
-2 LOG L	380.925	240.042	140.883 with 3 DF (p=0.0001)
Score	.	.	125.231 with 3 DF (p=0.0001)

Residual Chi-Square = 0.7562 with 3 DF (p=0.8599)

Summary of Backward Elimination Procedure

Step	Variable Removed	Number In	Wald Chi-Square	Pr > Chi-Square
1	WSRANGE	5	0.1209	0.7281
1	T93	4	0.3266	0.5677
1	TRANGE	3	0.3066	0.5797

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	-11.5328	1.4633	62.1169	0.0001	.	.
T	1	0.1504	0.0749	4.0354	0.0446	0.214693	1.162
TMAX	1	0.3800	0.0476	63.7152	0.0001	1.042516	1.462
WSAVG	1	-0.5737	0.1734	10.9422	0.0009	-0.467641	0.563

Association of Predicted Probabilities and Observed Responses

Concordant = 92.1%		Somers' D	= 0.844
Discordant = 7.7%		Gamma	= 0.846
Tied	= 0.3%	Tau-a	= 0.093
(40250 pairs)		c	= 0.922

Estimated Correlation Matrix

Variable	INTERCPT	T	TMAX	WSAVG
INTERCPT	1.00000	-0.22764	-0.90530	-0.24790
T	-0.22764	1.00000	-0.00417	-0.24284
TMAX	-0.90530	-0.00417	1.00000	-0.02460
WSAVG	-0.24790	-0.24284	-0.02460	1.00000

On obtient l'ajustement graphique B.11.

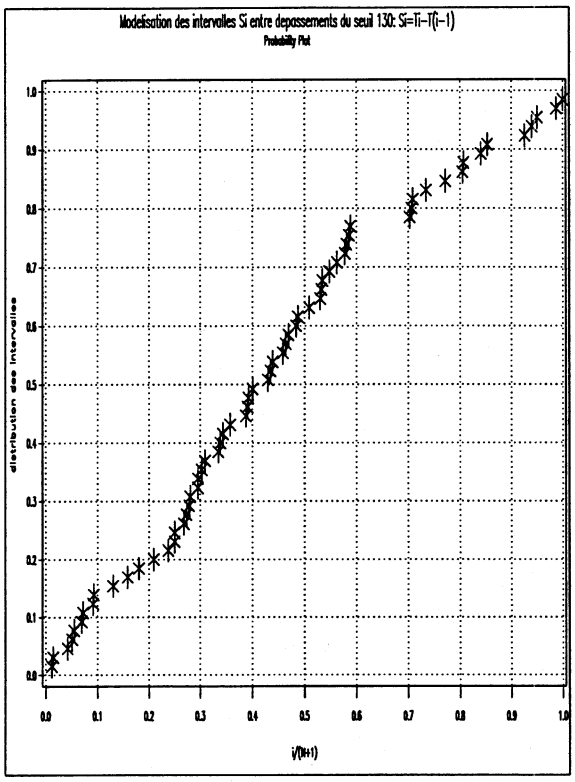


Figure B.11: Probability plot ($u = 130\ \mu g/m3$)

Taille des dépassements

Le modèle pour $\beta(t)$ est :

$$\hat{\beta}(t) = \hat{\beta}_0 + \hat{\beta}_3 * \omega_3(t) + \hat{\beta}_5 * \omega_5(t)$$

où le paramètre prend la valeur indiquée dans le tableau (B.11). L'observation majeure qui en découle est le fait que l'effet année n'est pas significatif. Il semblerait donc qu'à conditions météorologiques constantes, la taille des dépassements n'ait pas évolué durant la période d'étude.

Variable	Paramètre	Estimation	Ecart-type	T-ratio
c^{te}	β_0	0.1840	0.0780	2.3604
Tmax	β_3	-0.0063	0.0024	-2.5753
Wsavg	β_5	0.0237	0.0074	3.2047

Tableau B.11: Taille des dépassements ($u = 130\ \mu g/m3$)

On obtient la matrice de corrélation entre paramètres estimés suivante :

$$\text{Cor} = \begin{pmatrix} 1.00000 & -0.98031 & -0.20579 \\ -0.98031 & 1.00000 & 0.02386 \\ -0.20579 & 0.02386 & 1.00000 \end{pmatrix}$$

On obtient l'ajustement graphique B.12.

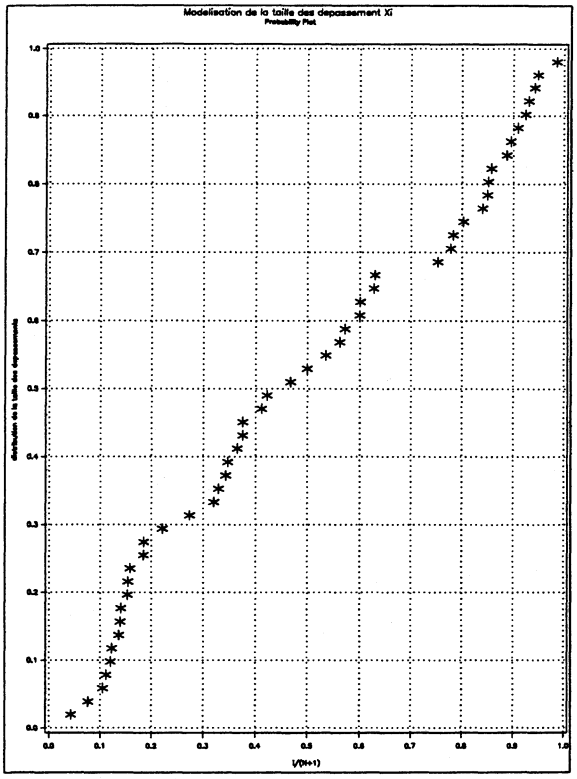


Figure B.12: Probability plot ($u = 130\ \mu g/m3$)

B.8 Résultats détaillés des modèles avec interaction associés à la fréquence des dépassements,pour la région parisienne

B.8.1 Neuilly/Seine

- JOURS SUPPRIMÉS

OBS	JOUR	MOIS	AN	PMAX1692	PMAX1693	PMAX1677	PMAX1694	TMAX	TRANGE	WSAVG	WSRANGE
1	9	9	88	26	.	72	.	22.1	7.2	4.92	2
2	17	6	89	129	.	85	131	25.5	9.8	4.00	4
3	11	5	91	128	59	57	56	16.5	12.4	4.23	4
4	9	8	91	179	73	41	.	24.5	11.8	3.69	1
5	18	8	91	146	63	33	.	22.1	10.2	2.62	2
6	16	6	92	121	106	95	117	24.2	7.1	5.46	3
7	15	6	94	135	150	105	.	23.1	10.0	2.85	3
8	15	8	94	125	124	85	112	23.9	12.4	2.00	2
9	10	8	95	111	131	127	133	27.8	13.8	6.23	5
10	23	7	96	126	127	136	128	28.6	7.3	8.08	2
11	3	5	97	111	123	132	124	25.8	17.3	4.62	4
12	26	5	97	123	124	112	116	22.2	14.3	2.46	1
13	24	8	97	116	136	131	132	32.6	14.6	3.23	6

- RÉSULTATS DE LA RÉGRESSION LOGISTIQUE

Response Profile

Ordered

Value	DEP110	DEP120	DEP130	DEP140	DEP150	DEP160	DEP170	DEP180	DEP190	DEP200
1	125	85	59	41	29	24	16	14	12	10
2	718	758	784	802	814	819	827	829	831	833

Model Fitting Information and Testing Global Null Hypothesis BETA=0

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
DEP110			
AIC	709.638	357.703	.
SC	714.375	376.651	.
-2 LOG L	707.638	349.703	357.934 with 3 DF (p=0.0001)
Score	.	.	243.757 with 3 DF (p=0.0001)
DEP120			
AIC	553.160	269.551	.
SC	557.897	288.499	.
-2LOG L	551.160	261.551	289.609 with 3 DF (p=0.0001)
Score	.	.	203.280 with 3 DF (p=0.0001)
DEP130			
AIC	429.584	238.658	.
SC	434.320	257.606	.
-2LOG L	427.584	230.658	196.925 with 3 DF (p=0.0001)
Score	.	.	146.070 with 3 DF (p=0.0001)
DEP140			
AIC	329.891	186.365	.
SC	334.628	205.313	.
-2LOG L	327.891	178.365	149.526 with 3 DF (p=0.0001)
Score	.	.	115.919 with 3 DF (p=0.0001)
DEP150			
AIC	254.432	144.267	.
SC	259.169	163.214	.
-2LOG L	252.432	136.267	116.165 with 3 DF (p=0.0001)
Score	.	.	80.901 with 3 DF (p=0.0001)
DEP160			
AIC	220.138	128.516	.
SC	224.875	147.464	.
-2LOG L	218.138	120.516	97.622 with 3 DF (p=0.0001)
Score	.	.	67.552 with 3 DF (p=0.0001)
DEP170			
AIC	160.554	99.520	.
SC	165.291	118.468	.
-2LOG L	158.554	91.520	67.035 with 3 DF (p=0.0001)
Score	.	.	41.797 with 3 DF (p=0.0001)
DEP180			
AIC	144.508	91.431	.
SC	149.245	110.378	.
-2LOG L	142.508	83.431	59.077 with 3 DF (p=0.0001)
Score	.	.	37.233 with 3 DF (p=0.0001)
DEP190			
AIC	127.878	75.656	.
SC	132.615	94.603	.
-2LOG L	125.878	67.656	58.222 with 3 DF (p=0.0001)
Score	.	.	33.854 with 3 DF (p=0.0001)
DEP200			
AIC	110.569	62.385	.
SC	115.306	81.333	.
-2LOG L	108.569	54.385	54.183 with 3 DF (p=0.0001)
Score	.	.	30.151 with 3 DF (p=0.0001)

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
Dep110 26.9C							
INTERCPT	1	1.6290	0.4781	11.6106	0.0007	.	.
T	1	-2.6715	0.2883	85.8764	0.0001	-3.994213	0.069
WSAVG	1	-0.8253	0.1374	36.0676	0.0001	-0.677926	0.438
TTMAX	1	0.0994	0.0101	96.6595	0.0001	3.625435	1.104
Dep120 27.5C							
INTERCPT	1	1.2292	0.5675	4.6921	0.0303	.	.

```
     T          1   -3.0360   0.3766   64.9926      0.0001   -4.539308   0.048
     WSAVG      1   -0.8786   0.1733   25.7067      0.0001   -0.721782   0.415
     TTMAX      1    0.1103   0.0127   75.0484      0.0001    4.022650   1.117
   DEP130  28C
     INTERCPT   1    0.6057   0.6199    0.9549      0.3285    .          .
     T          1   -2.6044   0.3713   49.1940      0.0001   -3.893951   0.074
     WSAVG      1   -0.8132   0.1902   18.2914      0.0001   -0.668060   0.443
     TTMAX      1    0.0931   0.0121   59.6064      0.0001    3.394435   1.098
   DEP140  27.7C
     INTERCPT   1    0.4366   0.7322    0.3556      0.5510    .    .
     T          1   -2.4748   0.4135   35.8184      0.0001   -3.700212   0.084
     WSAVG      1   -1.0074   0.2440   17.0448      0.0001   -0.827536   0.365
     TTMAX      1    0.0892   0.0132   45.8406      0.0001    3.251929   1.093
   DEP150  29.1C
     INTERCPT   1    1.0369   0.8639    1.4407      0.2300    .          .
     T          1   -2.8404   0.5335   28.3428      0.0001   -4.246763   0.058
     WSAVG      1   -1.2246   0.3039   16.2318      0.0001   -1.005947   0.294
     TTMAX      1    0.0977   0.0166   34.5413      0.0001    3.564194   1.103
   DEP160    28.8C
     INTERCPT   1    1.0794   0.9301    1.3467      0.2459    .          .
     T          1   -2.6373   0.5408   23.7851      0.0001   -3.943178   0.072
     WSAVG      1   -1.3881   0.3435   16.3295      0.0001   -1.140312   0.250
     TTMAX      1    0.0915   0.0168   29.5615      0.0001    3.337486   1.096
   DEP170    29.8C
     INTERCPT   1    1.3060   1.0871    1.4432      0.2296    .          .
     T          1   -2.6540   0.6423   17.0711      0.0001   -3.968037   0.070
     WSAVG      1   -1.5358   0.4146   13.7232      0.0002   -1.261595   0.215
     TTMAX      1    0.0892   0.0197   20.4344      0.0001    3.253529   1.093
   DEP180    29.3C
     INTERCPT   1    1.2259   1.1522    1.1319      0.2874    .          .
     T          1   -2.4556   0.6457   14.4650      0.0001   -3.671520   0.086
     WSAVG      1   -1.6461   0.4514   13.2999      0.0003   -1.352235   0.193
     TTMAX      1    0.0837   0.0199   17.6886      0.0001    3.053106   1.087
DEP190    29.2C
     INTERCPT   1    2.1772   1.3414    2.6344      0.1046    .          .
     T          1   -2.5880   0.7328   12.4715      0.0004   -?.869490   0.075
     WSAVG      1   -2.2030   0.5769   14.5833      0.0001   -1.809745   0.110
     TTMAX      1    0.0885   0.0228   15.0636      0.0001    3.228543   1.093
DEP200    28.4C
     INTERCPT   1    2.8992   1.6034    3.2694      0.0706    .          .
     T          1   -2.4122   0.7819    9.5189      0.0020   -3.606632   0.090
     WSAVG      1   -2.8442   0.7526   14.2811      0.0002   -2.336440   0.058
     TTMAX      1    0.0848   0.0247   11.8036      0.0006    3.093859   1.089

      Association of Predicted Probabilities and Observed Responses

     DEP110
      Concordant = 94.4%   Somers' D = 0.890
      Discordant =  5.5%   Gamma     = 0.891
      Tied       =  0.1%   Tau-a     = 0.225
      (89750 pairs)        c         = 0.945
     DEP120
       Concordant = 95.4%  Somers'D  =  0.908
       Discordant = 4.5%   Gamma  =  0.909
       Tied = 0.1%         Tau-a = 0.165
       (64430  pairs)      c  =  0.954
     DEP130
       Concordant = 94.7%  Somers'D = 0.895
       Discordant = 5.2%   Gamma = 0.896
       Tied = 0.1%         Tau-a = 0.117
       (46256  pairs)      c = 0.947
     DEP140
       Concordant = 95.1%  Somers'D = 0.904
       Discordant = 4.8%   Gamma = 0.905
       Tied = 0.1%         Tau-a = 0.084
       (32882  pairs)      c = 0.952
     DEP150
       Concordant = 95.6%  Somers'D = 0.912
       Discordant =  4.4%  Gamma = 0.913
       Tied =  0.1%        Tau-a = 0.061
```

```
    (23606  pairs)      c = 0.956
  DEP160
    Concordant = 95.2%  Somers'D = 0.905
    Discordant = 4.7%   Gamma = 0.906
    Tied = 0.1%         Tau-a = 0.050
    (19656  pairs)      c = 0.952
  DEP170
    Concordant = 94.9%  Somers'D = 0.900
    Discordant = 4.9%   Gamma = 0.902
    Tied = 0.2%         Tau-a = 0.034
    (13232  pairs)      c = 0.950
  DEP180
    Concordant = 94.2%  Somers'D  = 0.889
    Discordant = 5.4%   Gamma = 0.892
    Tied = 0.4%         Tau-a = 0.029
    (11606  pairs)      c = 0.944
  DEP190
    Concordant = 95.1%  Somers'D  = 0.908
    Discordant = 4.3%   Gamma = 0.913
    Tied = 0.6%         Tau-a = 0.026
    (9972  pairs)       c = 0.954
  DEP200
    Concordant = 96.6%  Somers'D = 0.936
    Discordant = 3.0%   Gamma = 0.939
    Tied = 0.3%         Tau-a = 0.022
    (8330  pairs)       c = 0.968

      Estimated Correlation Matrix

DEP110
   Variable   INTERCPT    T        WSAVG     TTMAX
   INTERCPT    1.00000  -0.33809 -0.81018   0.26714
   T          -0.33809   1.00000  0.15467  -0.98334
   WSAVG      -0.81018   0.15467  1.00000  -0.17590
   TTMAX       0.26714  -0.98334 -0.17590   1.00000
DEP120
    Variable  INTERCPT    T        WSAVG     TTMAX
    INTERCPT  1.00000  -0.37769  -0.80627   0.32043
    T        -0.37769   1.00000   0.19369  -0.98647
    WSAVG    -0.80627   0.19369   1.00000  -0.22178
    TTMAX     0.32043  -0.98647  -0.22178   1.00000
DEP130
    Variable  INTERCPT    T         WSAVG    TTMAX
    INTERCPT  1.00000  -0.37417  -0.79186   0.31013
    T        -0.37417   1.00000   0.14680  -0.98387
    WSAVG    -0.79186   0.14680   1.00000  -0.17865
    TTMAX     0.31013  -0.98387  -0.17865   1.00000
DEP140
    Variable  INTERCPT    T        WSAVG     TTMAX
    INTERCPT  1.00000  -0.41298  -0.77171   0.35438
    T        -0.41298   1.00000   0.18014  -0.98142
    WSAVG    -0.77171   0.18014   1.00000  -0.23388
    TTMAX     0.35438  -0.98142  -0.23388   1.00000
DEP150
    INTERCPT  1.00000  -0.47079  -0.79293   0.42934
    T        -0.47079   1.00000   0.23906  -0.98529
    WSAVG    -0.79293   0.23906   1.00000  -0.29161
    TTMAX     0.42934  -0.98529  -0.29161   1.00000
DEP160
    Variable  INTERCPT    T        WSAVG     TTMAX
    INTERCPT  1.00000  -0.47717  -0.78403   0.43340
    T        -0.47717   1.00000   0.24963  -0.98255
    WSAVG    -0.78403   0.24963   1.00000  -0.31240
    TTMAX     0.43340  -0.98255  -0.31240   1.00000
DEP170
    Variable  INTERCPT    T        WSAVG     TTMAX
    INTERCPT  1.00000  -0.46598  -0.79771   0.41943
    T        -0.46598   1.00000   0.23876  -0.98310
    WSAVG    -0.79771   0.23876   1.00000  -0.29168
    TTMAX     0.41943  -0.98310  -0.29168   1.00000
```

DEP180

Variable	INTERCPT	T	WSAVG	TTMAX
INTERCPT	1.00000	-0.45463	-0.78956	0.40125
T	-0.45463	1.00000	0.23054	-0.98063
WSAVG	-0.78956	0.23054	1.00000	-0.28719
TTMAX	0.40125	-0.98063	-0.28719	1.00000

DEP190

Variable	INTERCPT	T	WSAVG	TTMAX
INTERCPT	1.00000	-0.47434	-0.81124	0.42434
T	-0.47434	1.00000	0.28163	-0.98044
WSAVG	-0.81124	0.28163	1.00000	-0.33769
TTMAX	0.42434	-0.98044	-0.33769	1.00000

DEP200

Variable	INTERCPT	T	WSAVG	TTMAX
INTERCPT	1.00000	-0.46330	-0.82647	0.40683
T	-0.46330	1.00000	0.29461	-0.97692
WSAVG	-0.82647	0.29461	1.00000	-0.35043
TTMAX	0.40683	-0.97692	-0.35043	1.00000

Estimated Covariance Matrix

Pour u=110

Variable	INTERCPT	T	WSAVG	TTMAX
INTERCPT	0.2285465699	-0.046594786	-0.053222419	0.001291076
T	-0.046594786	0.0831045748	0.0061269794	-0.002865751
WSAVG	-0.053222419	0.0061269794	0.0188823475	-0.000244352
TTMAX	0.001291076	-0.002865751	-0.000244352	0.0001021987

Pour u=120

Variable	INTERCPT	T	WSAVG	TTMAX
INTERCPT	0.3220127269	-0.080714486	-0.079286954	0.0023147056
T	-0.080714486	0.1418246279	0.0126406347	-0.004729198
WSAVG	-0.079286954	0.0126406347	0.0300312194	-0.000489254
TTMAX	0.0023147056	-0.004729198	-0.000489254	0.0001620514

pour u=130

Variable	INTERCPT	T	WSAVG	TTMAX
INTERCPT	0.3842418217	-0.086123907	-0.093336181	0.0023171291
T	-0.086123907	0.1378812906	0.0103648098	-0.004403501
WSAVG	-0.093336181	0.0103648098	0.0361571106	-0.000409465
TTMAX	0.0023171291	-0.004403501	-0.000409465	0.000145282

pour u=140

Variable	INTERCPT	T	WSAVG	TTMAX
INTERCPT	0.5361125894	-0.125038804	-0.137871628	0.0034166189
T	-0.125038804	0.1709950466	0.0181755883	-0.005343794
WSAVG	-0.137871628	0.0181755883	0.0595374467	-0.000751434
TTMAX	0.0034166189	-0.005343794	-0.000751434	0.0001733811

pour u=150

Variable	INTERCPT	T	WSAVG	TTMAX
INTERCPT	0.7463283227	-0.216992041	-0.208206179	0.0061665154
T	-0.216992041	0.2846487127	0.0387669816	-0.00873966
WSAVG	-0.208206179	0.0387669816	0.0923829296	-0.001473602
TTMAX	0.0061665154	-0.00873966	-0.001473602	0.0002764098

Pour u=160

Variable	INTERCPT	T	WSAVG	TTMAX
INTERCPT	0.865159456	-0.240011047	-0.250508348	0.0067839441
T	-0.240011047	0.2924309394	0.0463719042	-0.008941403
WSAVG	-0.250508348	0.0463719042	0.1180003925	-0.001805888
TTMAX	0.0067839441	-0.008941403	-0.001805888	0.0002831924

B.8.2 Champs/Marne

– JOURS SUPPRIMÉS

OBS	JOUR	MOIS	AN	PMAX1692	PMAX1693	PMAX1677	PMAX1694	TMAX	TRANGE	WSAVG	WSRANGE
1	9	9	88	126	.	72	.	22.1	7.2	4.92	2
2	17	6	89	129	.	85	131	25.5	9.8	4.00	4
3	11	5	91	128	59	57	56	16.5	12.4	4.23	4
4	9	8	91	179	73	41	.	24.5	11.8	3.69	1
5	18	8	91	146	63	33	.	22.1	10.2	2.62	2

6	16	6	92	121	106	95	117	24.2	7.1	5.46	3
7	15	6	94	135	150	105	.	23.1	10.0	2.85	3
8	20	7	94	111	130	71	132	26.4	11.3	2.77	1
9	15	8	94	125	124	85	112	23.9	12.4	2.00	2
10	1	7	95	142	169	178	124	29.1	10.1	3.38	5
11	10	8	95	111	131	127	133	27.8	13.8	6.23	5
12	23	7	96	126	127	136	128	28.6	7.3	8.08	2
13	3	5	97	112	123	132	124	25.8	17.3	4.62	4
14	26	5	97	123	124	112	116	22.2	14.3	2.46	1
15	24	8	97	116	136	131	132	32.6	14.6	3.23	6

– RÉSULTATS DE LA RÉGRESSION LOGISTIQUE

Response Profile
Ordered

Value	DEP110	DEP120	DEP130	DEP140	DEP150	DEP160	DEP170	DEP180	DEP190	DEP200
1	59	38	28	19	13	9	3	2		
2	911	932	942	951	957	961	967	968		

Model Fitting Information and Testing Global Null Hypothesis BETA=0

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
DEP110			
AIC	446.708	194.080	.
SC	451.585	213.589	.
-2 LOG L	444.708	186.080	258.628 with 3 DF (p=0.0001)
Score	.	.	274.733 with 3 DF (p=0.0001)
DEP120			
AIC	322.709	150.856	.
SC	327.587	170.365	.
-2LOG L	320.709	142.856	177.854 with 3 DF (p=0.0001)
Score	.	.	204.361 with 3 DF (p=0.0001)
DEP130			
AIC	255.709	104.417	.
SC	260.586	123.926	.
-2LOG L	253.709	96.417	157.292 with 3 DF (p=0.0001)
Score	.	.	159.425 with 3 DF (p=0.0001)
DEP140			
AIC	189.074	98.293	.
SC	193.951	117.802	.
-2LOG L	187.074	90.293	96.781 with 3 DF (p=0.0001)
Score	.	.	99.328 with 3 DF (p=0.0001)
DEP150			
AIC	139.946	77.128	.
SC	144.823	96.637	.
-2LOG L	137.946	69.128	68.818 with 3 DF (p=0.0001)
Score	.	.	76.211 with 3 DF (p=0.0001)
DEP160			
AIC	104.158	58.543	.
SC	109.035	78.053	.
-2LOG L	102.158	50.543	51.614 with 3 DF (p=0.0001)
Score	.	.	56.772 with 3 DF (p=0.0001)
DEP170			
AIC	42.663	26.902	.
SC	47.540	46.411	.
-2LOG L	40.663	18.902	21.761 with 3 DF (p=0.0001)
Score	.	.	25.212 with 3 DF (p=0.0001)
DEP180			
AIC	30.732	13.780	.
SC	35.610	33.289	.
-2LOG L	28.732	5.780	22.952 with 3 DF (p=0.0001)
Score	.	.	27.137 with 3 DF (p=0.0001)

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
DEP110	20.7C						

```
  INTERCPT  1  -3.1936   0.6918   21.3103  0.0001     .          .
  T         1  -1.8345   0.3085   35.3642  0.0001  -2.771304  0.160
  WSAVG     1  -0.7914   0.1940   16.6366  0.0001  -0.652784  0.453
  TTMAX     1   0.0887   0.0109   65.7494  0.0001   3.200734  1.093
DEP120  18.7C
  INTERCPT  1  -4.1897   0.9081   21.2863  0.0001     .          .
  T         1  -1.4302   0.3288   18.9185  0.0001  -2.160473  0.239
  WSAVG     1  -0.8808   0.2446   12.9649  0.0003  -0.726600  0.414
  TTMAX     1   0.0763   0.0111   46.9618  0.0001   2.752647  1.079
DEP130  19.1C
  INTERCPT  1  -2.7078   1.1698    5.3585  0.0206     .          .
  T         1  -1.7890   0.4656   14.7623  0.0001  -2.702585  0.167
  WSAVG     1  -2.0213   0.4623   19.1209  0.0001  -1.667379  0.132
  TTMAX     1   0.0935   0.0168   31.0107  0.0001   3.373750  1.098
DEP140  20.3C
  INTERCPT  1  -2.2943   1.1974    3.6712  0.0554     .          .
  T         1  -1.5291   0.4710   10.5420  0.0012  -2.309985  0.217
  WSAVG     1  -1.6854   0.4332   15.1393  0.0001  -1.390312  0.185
  TTMAX     1   0.0752   0.0155   23.5969  0.0001   2.713388  1.078
DEP150  17.8C
  INTERCPT  1  -3.9556   1.5171    6.7986  0.0091     .          .
  T         1  -1.1948   0.5099    5.4906  0.0191  -1.804930  0.303
  WSAVG     1  -1.5298   0.4659   10.7802  0.0010  -1.261899  0.217
  TTMAX     1   0.0673   0.0162   17.2301  0.0001   2.428467  1.070
DEP160  20.6C
  INTERCPT  1  -3.7064   1.7657    4.4061  0.0358     .          .
  T         1  -1.6117   0.6889    5.4732  0.0193  -2.434714  0.200
  WSAVG     1  -1.5686   0.5440    8.3137  0.0039  -1.293913  0.208
  TTMAX     1   0.0781   0.0216   13.0620  0.0003   2.817880  1.081
DEP170  24.4C
  INTERCPT  1  -5.1834   3.1643    2.6834  0.1014     .          .
  T         1  -2.7667   1.5203    3.3121  0.0688  -4.179536  0.063
  WSAVG     1  -1.4498   0.8561    2.8679  0.0904  -1.195925  0.235
  TTMAX     1   0.1136   0.0484    5.5042  0.0190   4.102469  1.120
DEP180
  INTERCPT  1  -29.3254  32.9339   0.7929  0.3732     .          .
  T         1  -10.2434  13.6674   0.5617  0.4536  -15.474140  0.000
  WSAVG     1  -4.1126    5.5810   0.5430  0.4612   -3.392515  0.016
  TTMAX     1   0.4526    0.5787   0.6117  0.4341   16.339589  1.572

Association of Predicted Probabilities and Observed Responses

DEP110
  Concordant = 96.9%  Somers' D = 0.940
  Discordant =  2.9%  Gamma     = 0.942
  Tied       =  0.2%  Tau-a     = 0.107
  (53749 pairs)       c         = 0.970
DEP120
  Concordant = 97.4%  Somers' D = 0.948
  Discordant = 2.6%   Gamma = 0.949
  Tied = 0.1%         Tau-a = 0.071
  (35416  pairs)      c = 0.974
DEP130
  Concordant = 98.4%  Somers' D = 0.969
  Discordant = 1.5%   Gamma = 0.969
  Tied = 0.0%         Tau-a = 0.054
  (26376  pairs)      c = 0.985
DEP140
  Concordant = 97.9%  Somers' D = 0.958
  Discordant = 2.1%   Gamma = 0.958
  Tied = 0.0%         Tau-a = 0.037
  (18069 pairs)       c = 0.979
DEP150
  Concordant = 97.9%  Somers' D = 0.959
  Discordant = 2.0%   Gamma = 0.960
  Tied = 0.1%         Tau-a = 0.025
  (12441 pairs)       c = 0.980
DEP160
  Concordant = 98.1%  Somers' D = 0.963
  Discordant = 1.8%   Gamma = 0.964
```

```
  Tied = 0.1%          Tau-a = 0.018
  (8649 pairs)         c = 0.981
DEP170
  Concordant = 98.9%   Somers' D = 0.979
  Discordant = 1.0%    Gamma = 0.979
  Tied = 0.1%          Tau-a = 0.006
  (2901 pairs)         c = 0.989
DEP180
  Concordant = 99.9%   Somers' D = 0.998
  Discordant = 0.1%    Gamma = 0.998
  Tied = 0.0%          Tau-a = 0.004
  (1936 pairs)         c = 0.999

Estimated  Correlation  Matrix

DEP110
  Variable  INTERCPT     T       WSAVG    TTMAX
  INTERCPT  1.00000  -0.18021  -0.39169 -0.02651
  T        -0.18021   1.00000   0.00586 -0.94386
  WSAVG    -0.39169   0.00586   1.00000 -0.15223
  TTMAX    -0.02651  -0.94386  -0.15223  1.00000
DEP120
  Variable  INTERCPT    T        WSAVG     TTMAX
  INTERCPT  1.00000  -0.26454  -0.25887  -0.02482
  T        -0.26454   1.00000  -0.03989  -0.91051
  WSAVG    -0.25887  -0.03989   1.00000  -0.16809
  TTMAX    -0.02482  -0.91051  -0.16809   1.00000
DEP130
  Variable  INTERCPT    T        WSAVG     TTMAX
  INTERCPT  1.00000  -0.39724  -0.27548   0.14702
  T        -0.39724   1.00000   0.30946  -0.91527
  WSAVG    -0.27548   0.30946   1.00000  -0.54157
  TTMAX     0.14702  -0.91527  -0.54157   1.00000
DEP140
  Variable  INTERCPT    T        WSAVG      TTMAX
  INTERCPT  1.00000  -0.45872  -0.41212   0.25350
  T        -0.45872   1.00000   0.21584  -0.92802
  WSAVG    -0.41212   0.21584   1.00000  -0.41071
  TTMAX     0.25350  -0.92802  -0.41071   1.00000
DEP150
  INTERCPT  1.00000  -0.44306  -0.25342   0.14280
  T        -0.44306   1.00000   0.09412  -0.89792
  WSAVG    -0.25342   0.09412   1.00000  -0.32994
  TTMAX     0.14280  -0.89792  -0.32994   1.00000
DEP160
  Variable  INTERCPT    T        WSAVG      TTMAX
  INTERCPT  1.00000  -0.44471  -0.30189   0.19860
  T        -0.44471   1.00000   0.16472  -0.92911
  WSAVG    -0.30189   0.16472   1.00000  -0.34926
  TTMAX     0.19860  -0.92911  -0.34926   1.00000
DEP170
  Variable  INTERCPT    T        WSAVG      TTMAX
  INTERCPT  1.00000  -0.29770  -0.25738   0.08242
  T        -0.29770   1.00000   0.32711  -0.95978
  WSAVG    -0.25738   0.32711   1.00000  -0.44094
  TTMAX     0.08242  -0.95978  -0.44094   1.00000
DEP180
  Variable  INTERCPT    T        WSAVG      TTMAX
  INTERCPT  1.00000   0.95856   0.91366  -0.97146
  T         0.95856   1.00000   0.96700  -0.99838
  WSAVG     0.91366   0.96700   1.00000  -0.96936
  TTMAX    -0.97146  -0.99838  -0.96936   1.00000

Estimated Covariance Matrix

Pour u=110
    Variable  INTERCPT          T           WSAVG         TTMAX
    INTERCPT  0.4785871189  -0.038458589  -0.052572013  -0.000200529
    T        -0.038458589    0.0951659372  0.0003505718 -0.003183907
    WSAVG    -0.052572013    0.0003505718  0.0376419533 -0.000322961
```

```
    TTMAX    -0.000200529   -0.003183907   -0.000322961    0.0001195696
Pour u=120
    Variable  INTERCPT          T           WSAVG         TTMAX
    INTERCPT  0.8246387296  -0.078988779   -0.057507398   -0.000250844
    T        -0.078988779    0.1081153715  -0.003208834   -0.003331278
    WSAVG    -0.057507398   -0.003208834    0.0598442299  -0.000457544
    TTMAX    -0.000250844   -0.003331278   -0.000457544    0.0001238139
Pour u=130
    Variable  INTERCPT          T           WSAVG         TTMAX
    INTERCPT  1.3683392318  -0.216366627  -0.14896102     0.0028863415
    T        -0.216366627    0.2168111495  0.0666084122  -0.007152397
    WSAVG    -0.14896102     0.0666084122  0.2136780157  -0.004201472
    TTMAX     0.0028863415  -0.007152397  -0.004201472    0.0002816614
Pour u=140
    Variable  INTERCPT          T           WSAVG         TTMAX
    INTERCPT  1.4337921136  -0.258686769  -0.213757847    0.0046968376
    T        -0.258686769    0.2218053414  0.0440332937  -0.006762907
    WSAVG    -0.213757847    0.0440332937  0.187636062   -0.002752854
    TTMAX     0.0046968376  -0.006762907  -0.002752854    0.0002394322
Pour u=150
    Variable  INTERCPT          T           WSAVG         TTMAX
    INTERCPT  2.3015100628  -0.342733469  -0.179124134    0.003511047
    T        -0.342733469    0.2600031901  0.0223602348  -0.007420343
    WSAVG    -0.179124134    0.0223602348  0.217080507   -0.002491421
    TTMAX     0.003511047   -0.007420343  -0.002491421    0.0002626583
```

B.8.3 Aubervilliers

– JOURS SUPPRIMÉS

OBS	JOUR	MOIS	AN	PMAX1692	PMAX1693	PMAX1677	PMAX1694	TMAX	TRANGE	WSAVG	WSRANGE
1	9	9	88	126	.	72	.	22.1	7.2	4.92	2
2	17	6	89	129	.	85	131	25.5	9.8	4.00	4
3	11	5	91	128	59	57	56	16.5	12.4	4.23	4
4	9	8	91	179	73	41	.	24.5	11.8	3.69	1
5	18	8	91	146	63	33	.	22.1	10.2	2.62	2
6	16	6	92	121	106	95	117	24.2	7.1	5.46	3
7	15	6	94	135	150	105	.	23.1	10.0	2.85	3
8	9	8	94	.	133	106	131	28.6	11.7	6.15	4
9	15	8	94	125	124	85	112	23.9	12.4	2.00	2
10	10	8	95	111	131	127	133	27.8	13.8	6.23	5
11	23	7	96	126	127	136	128	28.6	7.3	8.08	2
12	3	5	97	112	123	132	124	25.8	17.3	4.62	4
13	26	5	97	123	124	112	116	22.2	14.3	2.46	1
14	17	8	97	194	177	143	181	28.9	14.2	.	.
15	24	8	97	116	136	131	132	32.6	14.6	3.23	6

– RÉSULTATS DE LA RÉGRESSION LOGISTIQUE

```
Response Profile
 Ordered
 Value  DEP110  DEP120  DEP130  DEP140  DEP150  DEP160  DEP170  DEP180  DEP190
 1        100      70      44      31      22      18       9       7       3
 2        745     775     801     814     823     827     836     838     842

Model Fitting Information and Testing Global Null Hypothesis BETA=0

                  Intercept
          Intercept  and
 Criterion  Only  Covariates  Chi-Square  for  Covariates
DEP110
 AIC       616.502  322.464     .
 SC        621.242  341.421     .
 -2 LOG L  614.502  314.464  300.038 with 3 DF (p=0.0001)
 Score        .        .     271.368 with 3 DF (p=0.0001)
DEP120
 AIC       484.752  217.565     .
 SC        489.491  236.522     .
 -2LOG L   482.752  209.565  273.187  with  3  DF  (p=0.0001)
```

```
  Score          .          .      250.064  with  3  DF  (p=0.0001)
DEP130
  AIC        347.721    160.843       .
  SC         352.460    179.800       .
  -2LOG L    345.721    152.843    192.878  with  3  DF  (p=0.0001)
  Score          .          .      178.238  with  3  DF  (p=0.0001)
DEP140
  AIC        267.780    127.400       .
  SC         272.520    146.358       .
  -2LOG L    265.780    119.400    146.380  with  3  DF  (p=0.0001)
  Score          .          .      137.963  with  3  DF  (p=0.0001)
DEP150
  AIC        205.947     96.666       .
  SC         210.686    115.623       .
  -2LOG L    203.947     88.666    115.281  with  3  DF  (p=0.0001)
  Score          .          .      111.215  with  3  DF  (p=0.0001)
DEP160
  AIC        176.177     80.933       .
  SC         180.916     99.890       .
  -2LOG L    174.177     72.933    101.244  with  3  DF  (p=0.0001)
  Score          .          .       94.265  with  3  DF  (p=0.0001)
DEP170
  AIC        101.662     46.337       .
  SC         106.401     65.294       .
  -2LOG L     99.662     38.337     61.325  with  3  DF  (p=0.0001)
  Score          .          .       42.030  with  3  DF  (p=0.0001)
DEP180
  AIC         83.050     40.482       .
  SC          87.789     59.440       .
  -2LOG L     81.050     32.482     48.567  with  3  DF  (p=0.0001)
  Score          .          .       30.129  with  3  DF  (p=0.0001)
DEP190
  AIC         41.834     19.769       .
  SC          46.573     38.727       .
  -2LOG L     39.834     11.769     28.064  with  3  DF  (p=0.0001)
  Score          .          .       11.627  with  3  DF  (p=0.0088)

Analysis of Maximum Likelihood Estimates

                    Parameter  Standard    Wald      Pr >   Standardized  Odds
  Variable      DF  Estimate   Error   Chi-Square  Chi-Square  Estimate   Ratio

DEP110  22.5C
  INTERCPT      1   -1.6464   0.5166    10.1558     0.0014        .         .
  T             1   -1.8147   0.2238    65.7258     0.0001     -2.337686  0.163
  WSAVG         1   -0.4838   0.1402    11.9111     0.0006     -0.398567  0.616
  TTMAX         1    0.0805   0.00810   98.7333     0.0001      2.626866  1.084
DEP120  23.2C
  INTERCPT      1   -1.6709   0.6484     6.6417     0.0100        .         .
  T             1   -2.2208   0.3126    50.4762     0.0001     -2.860840  0.109
  WSAVG         1   -0.7671   0.1978    15.0343     0.0001     -0.632007  0.464
  TTMAX         1    0.0957   0.0111    74.8679     0.0001      3.123542  1.100
DEP130  25.3C
  INTERCPT      1   -1.3062   0.8000     2.6657     0.1025        .         .
  T             1   -2.5623   0.4149    38.1388     0.0001     -3.300726  0.077
  WSAVG         1   -0.8443   0.2551    10.9566     0.0009     -0.695552  0.430
  TTMAX         1    0.1011   0.0138    53.6874     0.0001      3.300845  1.106
DEP140  26.3C
  INTERCPT      1   -1.4452   0.9361     2.3838     0.1226        .         .
  T             1   -2.7271   0.4987    29.9019     0.0001     -3.513039  0.065
  WSAVG         1   -0.8473   0.3035     7.7942     0.0052     -0.698079  0.429
  TTMAX         1    0.1038   0.0161    41.7926     0.0001      3.388797  1.109
DEP150  25.2C
  INTERCPT      1   -1.5705   1.1355     1.9130     0.1666        .         .
  T             1   -2.6098   0.5664    21.2335     0.0001     -3.361899  0.074
  WSAVG         1   -1.2929   0.4241     9.2922     0.0023     -1.065141  0.274
  TTMAX         1    0.1036   0.0184    31.6190     0.0001      3.381066  1.109
DEP160  25.7C
  INTERCPT      1   -1.1901   1.2557     0.8982     0.3433        .         .
  T             1   -2.8759   0.6614    18.9097     0.0001     -3.704774  0.056
```

```
  WSAVG      1  -1.5543   0.5049    9.4788   0.0021   -1.280537    0.211
  TTMAX      1   0.1121   0.0216   26.9402   0.0001    3.658530    1.119
DEP170  29.3C
  INTERCPT   1   1.2670   1.7262    0.5388   0.4629           .        .
  T          1  -4.4096   1.1653   14.3191   0.0002   -5.680461    0.012
  WSAVG      1  -2.1735   0.7436    8.5440   0.0035   -1.790629    0.114
  TTMAX      1   0.1503   0.0366   16.8779   0.0001    4.906775    1.162
DEP180  30.6C
  INTERCPT   1   1.5543   1.8906    0.6758   0.4110           .        .
  T          1  -4.7410   1.3525   12.2879   0.0005   -6.107397    0.009
  WSAVG      1  -2.0247   0.7704    6.9061   0.0086   -1.668055    0.132
  TTMAX      1   0.1549   0.0411   14.1824   0.0002    5.056978    1.168
DEP190  32.3C
  INTERCPT   1   9.1520   5.9028    2.4039   0.1210           .        .
  T          1  -6.4860   3.1712    4.1831   0.0408   -8.355292    0.002
  WSAVG      1  -6.0111   2.9619    4.1187   0.0424   -4.952261    0.002
  TTMAX      1   0.2005   0.0944    4.5098   0.0337    6.544540    1.222

Association of Predicted Probabilities and Observed Responses

DEP110
  Concordant = 93.9%  Somers' D = 0.881
  Discordant = 5.9%   Gamma  = 0.882
  Tied  = 0.2%        Tau-a  = 0.184
  (74500 pairs)       c    = 0.940
DEP120
  Concordant = 96.6%  Somers' D = 0.932
  Discordant = 3.3%   Gamma = 0.933
  Tied = 0.1%         Tau-a = 0.142
  (54250 pairs)       c = 0.966
DEP130
  Concordant = 97.2%  Somers' D = 0.944
  Discordant = 2.8%   Gamma = 0.945
  Tied = 0.1%         Tau-a = 0.093
  (35244 pairs)       c = 0.972
DEP140
  Concordant = 97.3%  Somers' D = 0.947
  Discordant = 2.6%   Gamma = 0.947
  Tied = 0.1%         Tau-a = 0.067
  (25234 pairs)       c = 0.973
DEP150
  Concordant = 97.7%  Somers' D = 0.955
  Discordant = 2.2%   Gamma = 0.956
  Tied = 0.1%         Tau-a = 0.049
  (18106 pairs)       c = 0.978
DEP160
  Concordant = 98.4%  Somers' D = 0.968
  Discordant = 1.6%   Gamma = 0.969
  Tied = 0.1%         Tau-a = 0.040
  (14886 pairs)       c = 0.984
DEP170
  Concordant = 98.8%  Somers' D = 0.976
  Discordant = 1.2%   Gamma = 0.977
  Tied = 0.1%         Tau-a = 0.021
  (7524 pairs)        c = 0.988
DEP180
  Concordant = 98.6%  Somers' D = 0.973
  Discordant = 1.3%   Gamma = 0.974
  Tied = 0.2%         Tau-a = 0.016
  (5866 pairs)        c = 0.986
DEP190
  Concordant = 99.7%  Somers' D = 0.994
  Discordant = 0.3%   Gamma = 0.994
  Tied = 0.0%         Tau-a = 0.007
  (2526 pairs)        c = 0.997


Estimated Correlation Matrix

DEP110
```

Variable	INTERCPT	T	WSAVG	TTMAX
INTERCPT	1.00000	-0.14755	-0.58388	-0.02625
T	-0.14755	1.00000	-0.08662	-0.95130
WSAVG	-0.58388	-0.08662	1.00000	-0.00214
TTMAX	-0.02625	-0.95130	-0.00214	1.00000

DEP120

Variable	INTERCPT	T	WSAVG	TTMAX
INTERCPT	1.00000	-0.23228	-0.54176	0.07987
T	-0.23228	1.00000	0.06957	-0.95780
WSAVG	-0.54176	0.06957	1.00000	-0.18043
TTMAX	0.07987	-0.95780	-0.18043	1.00000

DEP130

Variable	INTERCPT	T	WSAVG	TTMAX
INTERCPT	1.00000	-0.32825	-0.59175	0.20699
T	-0.32825	1.00000	0.13824	-0.96686
WSAVG	-0.59175	0.13824	1.00000	-0.23644
TTMAX	0.20699	-0.96686	-0.23644	1.00000

DEP140

Variable	INTERCPT	T	WSAVG	TTMAX
INTERCPT	1.00000	-0.34979	-0.60048	0.24029
T	-0.34979	1.00000	0.13939	-0.96979
WSAVG	-0.60048	0.13939	1.00000	-0.23707
TTMAX	0.24029	-0.96979	-0.23707	1.00000

DEP150

Variable	INTERCPT	T	WSAVG	TTMAX
INTERCPT	1.00000	-0.40360	-0.51542	0.27876
T	-0.40360	1.00000	0.22358	-0.95797
WSAVG	-0.51542	0.22358	1.00000	-0.37453
TTMAX	0.27876	-0.95797	-0.37453	1.00000

DEP160

Variable	INTERCPT	T	WSAVG	TTMAX
INTERCPT	1.00000	-0.42878	-0.51300	0.31408
T	-0.42878	1.00000	0.28457	-0.96020
WSAVG	-0.51300	0.28457	1.00000	-0.43897
TTMAX	0.31408	-0.96020	-0.43897	1.00000

DEP170

Variable	INTERCPT	T	WSAVG	TTMAX
INTERCPT	1.00000	-0.49555	-0.66813	0.43293
T	-0.49555	1.00000	0.41597	-0.98070
WSAVG	-0.66813	0.41597	1.00000	-0.50540
TTMAX	0.43293	-0.98070	-0.50540	1.00000

DEP180

	INTERCPT	T	WSAVG	TTMAX
INTERCPT	1.00000	-0.49680	-0.69279	0.43562
T	-0.49680	1.00000	0.37815	-0.98340
WSAVG	-0.69279	0.37815	1.00000	-0.45167
TTMAX	0.43562	-0.98340	-0.45167	1.00000

DEP190

Variable	INTERCPT	T	WSAVG	TTMAX
INTERCPT	1.00000	-0.75229	-0.93728	0.70921
T	-0.75229	1.00000	0.72716	-0.99159
WSAVG	-0.93728	0.72716	1.00000	-0.72452
TTMAX	0.70921	-0.99159	-0.72452	1.00000

Estimated Covariance Matrix

Pour u=110

Variable	INTERCPT	T	WSAVG	TTMAX
INTERCPT	0.2669197063	-0.017063037	-0.042285439	-0.000109825
T	-0.017063037	0.0501033862	-0.002717918	-0.001724415
WSAVG	-0.042285439	-0.002717918	0.0196495325	-2.425163E-6
TTMAX	-0.000109825	-0.001724415	-2.425163E-6	0.000065581

Pour u=120

Variable	INTERCPT	T	WSAVG	TTMAX
INTERCPT	0.4203860312	-0.04707582	-0.0694964	0.0005726653
T	-0.04707582	0.0977080988	0.0043023687	-0.003310714
WSAVG	-0.0694964	0.0043023687	0.0391434411	-0.000394756
TTMAX	0.0005726653	-0.003310714	-0.000394756	0.0001222826

Pour u=130

Variable	INTERCPT	T	WSAVG	TTMAX

	INTERCPT	0.6400696735	-0.108958733	-0.120751699	0.0022852226
	T	-0.108958733	0.1721401832	0.0146289454	-0.005535737
	WSAVG	-0.120751699	0.0146289454	0.0650552594	-0.000832208
	TTMAX	0.0022852226	-0.005535737	-0.000832208	0.0001904337

Pour u=140

Variable	INTERCPT	T	WSAVG	TTMAX
INTERCPT	0.8762016495	-0.163291036	-0.170597074	0.0036118125
T	-0.163291036	0.2487122977	0.0210978875	-0.007766107
WSAVG	-0.170597074	0.0210978875	0.0921168792	-0.00115539
TTMAX	0.0036118125	-0.007766107	-0.00115539	0.0002578443

Pour u=150

Variable	INTERCPT	T	WSAVG	TTMAX
INTERCPT	1.2892961931	-0.259546635	-0.248217261	0.0058300156
T	-0.259546635	0.3207589669	0.053705103	-0.00999316
WSAVG	-0.248217261	0.053705103	0.1798857646	-0.002925856
TTMAX	0.0058300156	-0.00999316	-0.002925856	0.0003392541

B.8.4 Créteil

– JOURS SUPPRIMÉS

OBS	JOUR	MOIS	AN	PMAX1692	PMAX1693	PMAX1677	PMAX1694	TMAX	TRANGE	WSAVG	WSRANGE
1	13	5	88	.	.	86	131	23.0	10.3	3.77	3
2	14	5	88	.	.	67	123	24.2	13.3	4.08	3
3	9	9	88	126	.	72	.	22.1	7.2	4.92	2
4	15	6	89	134	40	85	153	27.2	11.2	2.92	2
5	17	6	89	129	.	85	131	25.5	9.8	4.00	4
6	11	5	91	128	59	57	56	16.5	12.4	4.23	4
7	9	8	91	179	73	41	.	24.5	11.8	3.69	1
8	18	8	91	146	63	33	.	22.1	10.2	2.62	2
9	4	9	91	109	.	65	132	29.7	14.2	3.62	1
10	5	9	91	112	.	53	138	27.0	11.5	4.23	3
11	15	5	92	141	124	45	162	28.6	15.1	3.23	2
12	16	6	92	121	106	95	117	24.2	7.1	5.46	3
13	27	6	92	124	111	87	137	25.4	12.1	2.77	1
14	1	8	92	114	119	118	136	32.0	11.0	4.46	3
15	15	6	94	135	150	105	.	23.1	10.0	2.85	3
16	9	8	94	.	133	106	131	28.6	11.7	6.15	4
17	14	8	94	99	112	102	132	22.5	10.9	1.85	1
18	15	8	94	125	124	85	112	23.9	12.4	2.00	2
19	10	8	95	111	131	127	133	27.8	13.8	6.23	5
20	23	7	96	126	127	136	128	28.6	7.3	8.08	2
21	3	5	97	112	123	132	124	25.8	17.3	4.62	4
22	26	5	97	123	124	112	116	22.2	14.3	2.46	1
23	17	8	97	194	177	143	181	28.9	14.2	.	.
24	24	8	97	116	136	131	132	32.6	14.6	3.23	6

– RÉSULTATS DE LA RÉGRESSION LOGISTIQUE

```
Response Profile
Ordered
Value DEP110  DEP120 DEP130 DEP140 DEP150 DEP160 DEP170 DEP180 DEP190 DEP200
1        92     58     38     24     18     12      5      3
2       744    778    798    812    818    824    831    833

Model Fitting Information and Testing Global Null Hypothesis BETA=0

                 Intercept
          Intercept and
Criterion Only Covariates Chi-Square for Covariates

DEP110
 AIC       581.541 326.197     .
 SC        586.270 345.111     .
 -2 LOG L 579.541 318.197 261.344 with 3 DF (p=0.0001)
 Score        .       .    231.215 with 3 DF (p=0.0001)
DEP120
 AIC      423.389 215.471     .
 SC       428.118 234.385     .
```

```
 -2LOG L 421.389 207.471 213.918 with 3 DF (p=0.0001)
 Score      .       .    194.332 with 3 DF (p=0.0001)
DEP130
 AIC     311.165 158.903     .
 SC      315.894 177.817     .
 -2LOG L 309.165 150.903 158.262 with 3 DF (p=0.0001)
 Score      .       .     141.673 with 3 DF (p=0.0001)
DEP140
 AIC     219.732 108.443     .
 SC      224.461 127.357     .
 -2LOG L 217.732 100.443 117.289 with 3 DF (p=0.0001)
 Score      .       .    101.701 with 3 DF (p=0.0001)
DEP150
 AIC     175.787 103.863     .
 SC      180.516 122.778     .
 -2LOG L 173.787  95.863  77.924 with 3 DF (p=0.0001)
 Score      .       .     72.525 with 3 DF (p=0.0001)
DEP160
 AIC     127.676 79.809      .
 SC      132.405 98.724      .
 -2LOG L 125.676 71.809   53.867 with 3 DF (p=0.0001)
 Score      .       .     53.928 with 3 DF (p=0.0001)
DEP170
 AIC     63.162 39.147       .
 SC      67.891 58.061       .
 -2LOG L 61.162 31.147    30.015 with 3 DF (p=0.0001)
 Score      .       .     26.448 with 3 DF (p=0.0001)
DEP180
 AIC     41.769 21.724       .
 SC      46.498 40.639       .
 -2LOG L 39.769 13.724    26.045 with 3 DF (p=0.0001)
 Score      .       .     19.334 with 3 DF (p=0.0002)
```

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
DEP110 24.0C							
INTERCPT	1	-0.5080	0.5047	1.0129	0.3142	.	.
T	1	-1.7667	0.2261	61.0713	0.0001	-2.515120	0.171
WSAVG	1	-0.6379	0.1426	20.0189	0.0001	-0.519480	0.528
TTMAX	1	0.0737	0.00784	88.2533	0.0001	2.591586	1.076
DEP120 24.3C							
INTERCPT	1	-0.5658	0.6587	0.7378	0.3904	.	.
T	1	-2.0064	0.3118	41.4034	0.0001	-2.856248	0.134
WSAVG	1	-0.9404	0.2112	19.8288	0.0001	-0.765853	0.390
TTMAX	1	0.0825	0.0104	62.6171	0.0001	2.900724	1.086
DEP130 23.5C							
INTERCPT	1	-0.2081	0.8341	0.0623	0.8029	.	.
T	1	-1.8458	0.3649	25.5889	0.0001	-2.627705	0.158
WSAVG	1	-1.4742	0.3172	21.6074	0.0001	-1.200588	0.229
TTMAX	1	0.0784	0.0120	42.9675	0.0001	2.758680	1.082
DEP140 21.0C							
INTERCPT	1	-0.4230	1.0931	0.1497	0.6988	.	.
T	1	-1.6308	0.4432	13.5377	0.0002	-2.321541	0.196
WSAVG	1	-2.2119	0.4867	20.6555	0.0001	-1.801315	0.109
TTMAX	1	0.0778	0.0150	26.8904	0.0001	2.736548	1.081
DEP150 23.5C							
INTERCPT	1	-1.0976	1.1050	0.9867	0.3205	.	.
T	1	-1.6614	0.4776	12.1023	0.0005	-2.365120	0.190
WSAVG	1	-1.4673	0.4056	13.0841	0.0003	-1.194942	0.231
TTMAX	1	0.0706	0.0150	22.2666	0.0001	2.484995	1.073
DEP160 22.2C							
INTERCPT	1	-2.2530	1.3551	2.7641	0.0964	.	.
T	1	-1.4749	0.5524	7.1300	0.0076	-2.099648	0.229
WSAVG	1	-1.3899	0.4579	9.2141	0.0024	-1.131887	0.249
TTMAX	1	0.0665	0.0170	15.3704	0.0001	2.338821	1.069
DEP170 22.0C							
INTERCPT	1	-2.3985	2.2417	1.1449	0.2846	.	.

```
 T         1  -1.7216   1.0148   2.8783    0.0898    -2.450929    0.179
 WSAVG     1  -2.1347   0.7511   8.0785    0.0045    -1.738484    0.118
 TTMAX     1   0.0784   0.0313   6.2935    0.0121     2.757507    1.082
DEP180 23.5C
 INTERCPT  1  -1.9933   3.5069   0.3231    0.5698            .        .
 T         1  -2.8679   2.0860   1.8901    0.1692    -4.082747    0.057
 WSAVG     1  -3.3980   1.2908   6.9299    0.0085    -2.767289    0.033
 TTMAX     1   0.1222   0.0669   3.3377    0.0677     4.296856    1.130

Association of Predicted Probabilities and Observed Responses

DEP110
 Concordant = 93.3%  Somers' D = 0.869
 Discordant = 6.5%   Gamma = 0.870
 Tied = 0.2%         Tau-a = 0.170
 (68448 pairs)       c = 0.934
DEP120
 Concordant = 95.9%  Somers' D = 0.919
 Discordant = 4.0%   Gamma = 0.920
 Tied = 0.1%         Tau-a = 0.119
 (45124 pairs)       c = 0.959
DEP130
 Concordant = 96.4%  Somers' D = 0.928
 Discordant = 3.6%   Gamma = 0.929
 Tied = 0.0%         Tau-a = 0.081
 (30324 pairs)       c = 0.964
DEP140
 Concordant = 97.5%  Somers' D = 0.950
 Discordant = 2.5%   Gamma = 0.950
 Tied = 0.0%         Tau-a = 0.053
 (19488 pairs)       c = 0.975
DEP150
 Concordant = 96.8%  Somers' D = 0.937
 Discordant = 3.1%   Gamma = 0.938
 Tied = 0.1%         Tau-a = 0.040
 (14724 pairs)       c = 0.969
DEP160
 Concordant = 96.7%  Somers' D = 0.935
 Discordant = 3.2%   Gamma = 0.936
 Tied = 0.1%         Tau-a = 0.026
 (9888 pairs)        c = 0.967
DEP170
 Concordant = 97.8%  Somers' D = 0.958
 Discordant = 2.0%   Gamma = 0.959
 Tied = 0.1%         Tau-a = 0.011
 (4155 pairs)        c = 0.979
DEP180
 Concordant = 99.4%  Somers' D = 0.988
 Discordant = 0.6%   Gamma = 0.988
 Tied = 0.0%         Tau-a = 0.007
 (2499 pairs)        c = 0.994

Estimated Correlation Matrix

DEP110
 Variable INTERCPT     T       WSAVG      TTMAX
 INTERCPT 1.00000  -0.23731  -0.68743   0.10739
 T       -0.23731   1.00000  -0.02395  -0.96254
 WSAVG   -0.68743  -0.02395   1.00000  -0.03712
 TTMAX    0.10739  -0.96254  -0.03712   1.00000
DEP120
 Variable INTERCPT     T       WSAVG      TTMAX
 INTERCPT 1.00000  -0.31381  -0.64489   0.19816
 T       -0.31381   1.00000   0.06432  -0.96378
 WSAVG   -0.64489   0.06432   1.00000  -0.15704
 TTMAX    0.19816  -0.96378  -0.15704   1.00000
DEP130
 Variable INTERCPT     T       WSAVG      TTMAX
 INTERCPT 1.00000  -0.42087  -0.60455   0.31027
 T       -0.42087   1.00000   0.15817  -0.95174
```

```
 WSAVG   -0.60455   0.15817    1.00000  -0.30874
 TTMAX    0.31027  -0.95174   -0.30874   1.00000
DEP140
 Variable INTERCPT     T       WSAVG      TTMAX
 INTERCPT 1.00000  -0.49739   -0.47305   0.34082
 T       -0.49739   1.00000    0.25576  -0.92484
 WSAVG   -0.47305   0.25576    1.00000  -0.48500
 TTMAX    0.34082  -0.92484   -0.48500   1.00000
DEP150
 Variable INTERCPT     T       WSAVG      TTMAX
 INTERCPT 1.00000  -0.45036   -0.53663   0.31432
 T       -0.45036   1.00000    0.13317  -0.94867
 WSAVG   -0.53663   0.13317    1.00000  -0.28881
 TTMAX    0.31432  -0.94867   -0.28881   1.00000
DEP160
          INTERCPT     T       WSAVG     TTMAX
 INTERCPT 1.00000  -0.45185   -0.43544   0.27034
 T       -0.45185   1.00000    0.08262  -0.93924
 WSAVG   -0.43544   0.08262    1.00000  -0.25255
 TTMAX    0.27034  -0.93924   -0.25255   1.00000
DEP170
          INTERCPT     T       WSAVG     TTMAX
 INTERCPT 1.00000  -0.46899   -0.33166   0.26524
 T       -0.46899   1.00000    0.16553  -0.94894
 WSAVG   -0.33166   0.16553    1.00000  -0.31501
 TTMAX    0.26524  -0.94894   -0.31501   1.00000
DEP180
 Variable INTERCPT     T       WSAVG      TTMAX
 INTERCPT 1.00000  -0.39911   -0.30356   0.22716
 T       -0.39911   1.00000    0.37802  -0.97203
 WSAVG   -0.30356   0.37802    1.00000  -0.47342
 TTMAX    0.22716  -0.97203   -0.47342   1.00000
```

Estimated Covariance Matrix

Pour u=110

Variable	INTERCPT	T	WSAVG	TTMAX
INTERCPT	0.254735308	-0.027077541	-0.049464241	0.0004251137
T	-0.027077541	0.051110112	-0.000771808	-0.001706685
WSAVG	-0.049464241	-0.000771808	0.020325574	-0.000041503
TTMAX	0.0004251137	-0.001706685	-0.000041503	0.0000615125

Pour u=120

Variable	INTERCPT	T	WSAVG	TTMAX
INTERCPT	0.4339276998	-0.064455824	-0.089715421	0.00136043
T	-0.064455824	0.097225988	0.0042358293	-0.003131922
WSAVG	-0.089715421	0.0042358293	0.0446005998	-0.000345643
TTMAX	0.00136043	-0.003131922	-0.000345643	0.0001086132

Pour u=130

Variable	INTERCPT	T	WSAVG	TTMAX
INTERCPT	0.695692	-0.12809315	-0.15992185	0.0030964085
T	-0.12809315	0.1331462411	0.0183040692	-0.004155238
WSAVG	-0.15992185	0.0183040692	0.1005848009	-0.001171568
TTMAX	0.0030964085	-0.004155238	-0.001171568	0.0001431614

Pour u=140

Variable	INTERCPT	T	WSAVG	TTMAX
INTERCPT	1.1949132288	-0.240980867	-0.251661269	0.0055896496
T	-0.240980867	0.1964419604	0.055169981	-0.006149899
WSAVG	-0.251661269	0.055169981	0.2368594818	-0.003541377
TTMAX	0.0055896496	-0.006149899	-0.003541377	0.0002250978

Pour u=150

Variable	INTERCPT	T	WSAVG	TTMAX
INTERCPT	1.2209287915	-0.237650575	-0.240529084	0.005199982
T	-0.237650575	0.2280681336	0.025798144	-0.006783126
WSAVG	-0.240529084	0.025798144	0.1645488092	-0.001754033
TTMAX	0.005199982	-0.006783126	-0.001754033	0.0002241611

B.9 Résultats détaillés pour Los Angeles

B.9.1 Liste des covariables météorologiques disponibles

Le fichier de données ozone et météorologie mesurées à Los Angeles (datafr2.don), transmis par J. Casmassi, possède la structure suivante:

```
# Variable (en colonne) :                                    codage
1 year                                                         an
2 month                                                        mois
3 day                                                          jour
4 1200 UTC 500 mb pattern (1-low pressure,                     pre1
 2-trough approaching, 3-zonal flow,
 4-high pressure ridge, 5-cut off low)
5 1200 UTC 500 mb height at Vandenberg AFB (dam)               pre2
6 1500 UTC lax-wjf surface pressure gradient mb                pre3
7 1500 UTC Summation pressure gradient mb:                     pre4
 [(lgb-dag)+(san-las)+(riv-dag)]
8 1500 UTC lax-sfo surface pressure gradient mb                pre5
9 1500 UTC san-las surface pressure gradient mb                pre6
10 1300 UTC lax surface temperature (C)                        t1
11 1300 UTC lax 1000 mb temperature (C)                        t2
12 1300 UTC lax 950 mb temperature (C)                         t3
13 1300 UTC lax 900 mb temperature (C)                         t4
14 1300 UTC lax 850 mb temperature (C)                         t5
15 1300 UTC lax inversion base temperature (C)                 inv1
 (if no inversion = 0.0)
16 1300 UTC lax inversion top temperature (C)                  inv2
 (if no inversion = 0.0)
17 1300 UTC lax inversion base height (ft)                     inv3
 (surface=100.0, no inversion=9999.)
18 1300 UTC lax inversion top height (ft)                      inv4
 (no inversion = 9999.0)
19 daily 1-hr max o3 at azusa (ug/m3)                          oz1
20 daily 1-hr max o3 at burbank (ug/m3)                        oz2
21 daily 1-hr max o3 at long beach (ug/m3)                     oz3
22 daily 1-hr max o3 at reseda (ug/m3)                         oz4
23 daily 1-hr max o3 at pomona (ug/m3)                         oz5
24 daily 1-hr max o3 at whittier (ug/m3)                       oz6
25 daily 1-hr max o3 at lynwood (ug/m3)                        oz7
26 daily 1-hr max o3 at pico rivera (ug/m3)                    oz8
27 daily 1-hr max o3 at downtown los angeles (ug/m3)           oz9
28 daily 1-hr max o3 at pasadena (ug/m3)                       oz10
29 daily 1-hr max o3 at santa clarita (ug/m3)                  oz11
30 daily 1-hr max o3 at west los angeles (ug/m3)               oz12
31 daily 1-hr max o3 at hawthorne/lennox (ug/m3)               oz13
32 daily 1-hr max o3 at anaheim (ug/m3)                        oz14
33 daily 1-hr max o3 at la habra (ug/m3)                       oz15
34 daily 1-hr max o3 at el toro (ug/m3)                        oz16
35 daily 1-hr max o3 at los alamitos (ug/m3)                   oz17
36 daily 1-hr max o3 at costa mesa (ug/m3)                     oz18
37 daily 1-hr max o3 at palm springs (ug/m3)                   oz19
38 daily 1-hr max o3 at riverside (ug/m3)                      oz20
39 daily 1-hr max o3 at perris (ug/m3)                         oz21
40 daily 1-hr max o3 at banning (ug/m3)                        oz22
41 daily 1-hr max o3 at norco (ug/m3)                          oz23
42 daily 1-hr max o3 at upland (ug/m3)                         oz24
43 daily 1-hr max o3 at crestline (ug/m3)                      oz25
44 daily 1-hr max o3 at fontana (ug/m3)                        oz26
45 daily 1-hr max o3 at san bernardino (ug/m3)                 oz27
46 daily 1-hr max o3 at redlands (ug/m3)                       oz28
47 daily 1-hr max o3 at glendora (ug/m3)                       oz29
48 max 1-hr average temperature at                             tmaxl
 downtown los angeles (F)
49 24-hr average temperature at                                tmoyl
 downtown los angeles (F)
50 max 1-hr average temperature at azusa (F)                   tmaxa
51 24-hr average temperature at azusa (F)                      tmoya
52 24-hour average wind speed at                               vvl
```

```
 downtown los angeles (mph)
53 24-hour resultant wind direction at                          dvl
 downtown los angeles
 (degrees: eg. 90 from the east...180 from the south)
```

Remarques :

- Ce fichier contient tous les jours de la période 01/01/1981-31/12/1996 (ne conserver que les périodes " été " 01/05-15/09 de 1981-1996 pour l'étude de l'ozone).
- Les variables tmaxl, tmoyl, tmaxa et tmoya ne sont mesurées que depuis 1989.

B.9.2 Le site de mesure d'Azusa

Jours dépassant le seuil 500 μgm^{-3}

OBS	AN	MOIS	JOUR	OZ1
1	81	5	9	520
2	81	6	4	540
3	81	6	17	700
4	81	6	18	540
5	81	6	19	560
6	81	6	20	560
7	81	6	24	600
8	81	6	25	560
9	81	6	26	560
10	81	7	14	560
11	81	7	21	520
12	81	7	23	520
13	81	8	1	540
14	81	8	5	540
15	81	8	6	640
16	81	8	7	600
17	81	8	20	540
18	81	8	21	520
19	81	8	26	600
20	81	8	27	540
21	81	8	28	700
22	82	7	20	520
23	82	8	4	520
24	82	8	5	560
25	82	8	10	560
26	82	8	17	640
27	82	8	18	540
28	82	8	19	540
29	82	9	1	620
30	82	9	2	720
31	82	9	3	640
32	82	9	7	580
33	83	5	25	540
34	83	5	27	660
35	83	5	28	560
36	83	6	6	520
37	83	6	13	520
38	83	6	14	620
39	83	7	11	780
40	83	7	12	660
41	83	7	13	620
42	83	7	20	540
43	83	7	23	540
44	83	7	30	520
45	83	7	31	540
46	83	8	1	520
47	83	8	2	620
48	83	8	3	560
49	83	8	4	540
50	83	8	5	720
51	83	8	6	540

52	83	8	26	580
53	83	8	27	600
54	83	8	29	520
55	83	9	2	540
56	83	9	6	520
57	83	9	11	740
58	83	9	14	540
59	84	5	13	560
60	84	7	1	520
61	84	7	6	620
62	84	7	11	540
63	84	7	13	520
64	84	8	3	520
65	84	8	6	580
66	84	8	7	540
67	84	8	8	560
68	84	8	29	580
69	84	8	30	560
70	84	9	14	540
71	85	5	23	560
72	85	6	6	640
73	85	6	7	660
74	85	6	8	540
75	85	7	3	560
76	85	7	8	560
77	85	8	23	560
78	85	8	24	720
79	85	8	25	520
80	85	8	28	580
81	85	8	30	540
82	86	6	26	620
83	86	6	27	560
84	86	9	6	620
85	87	7	31	600
86	87	8	1	520
87	87	9	1	540
88	88	5	11	560
89	88	7	16	600
90	88	7	17	580
91	88	7	20	540
92	88	8	25	540
93	88	9	2	540
94	88	9	3	540
95	89	6	13	520
96	89	6	14	620
97	89	6	18	520
98	89	6	19	520
99	89	7	19	660
100	89	7	20	600
101	89	9	13	520
102	91	8	9	560
103	92	8	1	520

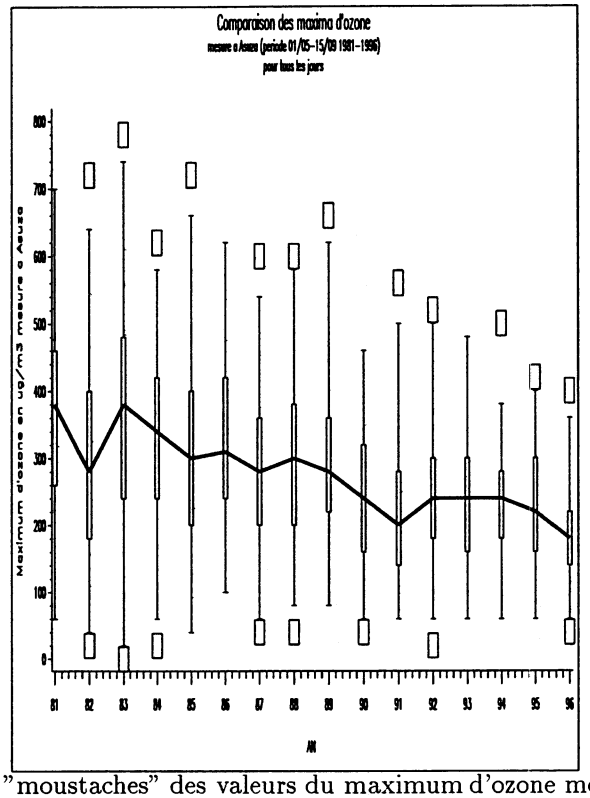


Figure B.13: Boîtes à "moustaches" des valeurs du maximum d'ozone mesurées à Azusa

B.9.3 Le site de mesure de Long Beach

fréquence des dépassements

– JOURS SUPPRIMÉS

OBS	JOUR	MOIS	AN	OZ3	PRE5	PRE6	T2	T4
1	9	9	81	100	-2.9	-1.1	18.0	22.1
2	29	7	82	60	-3.8	1.7	18.7	24.8
3	30	7	82	20	-3.2	1.0	20.0	25.6
4	30	5	84	40	-2.2	2.2	18.8	27.5
5	16	6	84	220	0.0	3.2	16.0	14.6
6	26	6	84	120	-2.4	2.6	16.4	25.3
7	26	8	85	80	-4.1	-0.1	19.2	25.0
8	27	6	90	60	-7.9	1.0	21.2	29.0
9	22	8	93	200	1.0	1.6	16.0	20.0

– RÉSULTATS DE LA RÉGRESSION LOGISTIQUE

```
Response Profile

                Ordered
                 Value  DEP180     DEP200
                     1     168        115
                     2     577        710

Model Fitting Information and Testing Global Null Hypothesis BETA=0

                                  Intercept
                       Intercept      and
          Criterion         Only Covariates    Chi-Square for Covariates

DEP180
          AIC            797.341    179.432          .
          SC             801.954    207.112          .
          -2 LOG L       795.341    167.432    627.909 with 5 DF (p=0.0001)
```

	Score	.	.	463.655 with 5 DF (p=0.0001)
DEP200				
	AIC	668.372	145.598	.
	SC	673.087	173.890	.
	-2 LOG L	666.372	133.598	532.774 with 5 DF (p=0.0001)
	Score	.	.	426.219 with 5 DF (p=0.0001)

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
DEP180							
INTERCPT	1	-5.8528	1.6555	12.4985	0.0004	.	.
T	1	-0.3668	0.0487	56.8183	0.0001	-0.976413	0.693
PRE5	1	-0.5523	0.1101	25.1568	0.0001	-0.700584	0.576
PRE6	1	-0.3206	0.0847	14.3254	0.0002	-0.612310	0.726
T2	1	-0.3102	0.1054	8.6546	0.0033	-0.486854	0.733
T4	1	0.6314	0.0774	66.4880	0.0001	2.461032	1.880
DEP200							
INTERCPT	1	-7.8707	1.9353	16.5397	0.0001	.	.
T	1	-0.4160	0.0576	52.2219	0.0001	-1.080277	0.660
PRE5	1	-0.6463	0.1283	25.3874	0.0001	-0.796506	0.524
PRE6	1	-0.3259	0.0865	14.2071	0.0002	-0.599561	0.722
T2	1	-0.3982	0.1227	10.5366	0.0012	-0.608207	0.672
T4	1	0.7453	0.1046	50.7585	0.0001	2.757125	2.107

Association of Predicted Probabilities and Observed Responses

DEP180		
Concordant = 98.9%	Somers' D = 0.978	
Discordant = 1.1%	Gamma = 0.978	
Tied = 0.0%	Tau-a = 0.342	
(96936 pairs)	c = 0.989	
DEP200		
Concordant = 99.1%	Somers' D = 0.983	
Discordant = 0.8%	Gamma = 0.983	
Tied = 0.0%	Tau-a = 0.236	
(81650 pairs)	c = 0.992	

Estimated Correlation Matrix

Variable	INTERCPT	T	PRE5	PRE6	T2	T4
DEP180						
INTERCPT	1.00000	0.14559	0.34818	-0.16524	-0.57515	-0.33036
T	0.14559	1.00000	0.25545	0.06487	0.13606	-0.51911
PRE5	0.34818	0.25545	1.00000	0.05507	0.07012	-0.35844
PRE6	-0.16524	0.06487	0.05507	1.00000	0.11606	-0.06605
T2	-0.57515	0.13606	0.07012	0.11606	1.00000	-0.54238
T4	-0.33036	-0.51911	-0.35844	-0.06605	-0.54238	1.00000
DEP200						
Variable	INTERCPT	T	PRE5	PRE6	T2	T4
INTERCPT	1.00000	0.19864	0.40213	0.01521	-0.36670	-0.46749
T	0.19864	1.00000	0.32555	0.15215	0.22410	-0.53005
PRE5	0.40213	0.32555	1.00000	0.14849	0.20467	-0.46631
PRE6	0.01521	0.15215	0.14849	1.00000	0.08832	-0.15002
T2	-0.36670	0.22410	0.20467	0.08832	1.00000	-0.62471
T4	-0.46749	-0.53005	-0.46631	-0.15002	-0.62471	1.00000

Estimated Covariance Matrix

Variable	INTERCPT	T	PRE5	PRE6	T2	T4
DEP180						
INTERCPT	2.7407750223	0.0117304403	0.0634736716	-0.023174092	-0.100387922	-0.042353771
T	0.0117304403	0.0023684498	0.0013689637	0.0002674405	0.0006980983	-0.001956387
PRE5	0.0634736716	0.0013689637	0.0121255381	0.0005136669	0.0008140348	-0.003056547
PRE6	-0.023174092	0.0002674405	0.0005136669	0.007176213	0.001036568	-0.000433311
T2	-0.100387922	0.0006980983	0.0008140348	0.001036568	0.0111155434	-0.004428276
T4	-0.042353771	-0.001956387	-0.003056547	-0.000433311	-0.004428276	0.0059968751

Variable	INTERCPT	T	PRE5	PRE6	T2	T4
DEP200						

INTERCPT	3.7454023584	0.022129646	0.0998240753	0.0025447972	-0.087059039	-0.094642828
T	0.022129646	0.0033136712	0.0024037672	0.0007573719	0.0015825174	-0.003191826
PRE5	0.0998240753	0.0024037672	0.0164523873	0.001646921	0.0032205189	-0.006256814
PRE6	0.0025447972	0.0007573719	0.001646921	0.0074773402	0.0009369034	-0.001357
T2	-0.087059039	0.0015825174	0.0032205189	0.0009369034	0.0150489795	-0.00801676
T4	-0.094642828	-0.003191826	-0.006256814	-0.001357	-0.00801676	0.0109430051

B.10 Programmes utilisés

Le logiciel SAS ([41]), ([42]) et MATLAB ([33]), ([32]) ont été utilisés pour programmer nos différents modules de calcul.

B.10.1 graphiques

Histogrammes

```
/************************************************************
Histogramme des differentes variables
************************************************************/
data sds;
filename fich'd:\lise\highleve\donnees\dontrend.don';
 infile fich lrecl=1500;
 input jour mois an pmax1692 pmax1693 pmax1677 pmax1694
       tmax trange wsavg wsrange ;
 x1692=pmax1692-130;
  if x1692>0;
run;

goptions reset=global gunit=border
      ftext=simplex htitle=0.35cm htext=0.25cm;
libname c 'd:\lise\highleve\graph\';

proc gchart data=sds gout=c.c1;
  vbar pmax1692 / cpercent autoref;
title1 'Valeurs du maximum d''ozone journalier mesurees a Neuilly/Seine
        periode 1988-1997';
title2 'pour les jours ou la valeur du maximum d''ozone depasse 130.';
run;

proc gchart data=sds gout=c.c1;
  vbar tmax /cpercent autoref;
title1 'Valeurs de la temperature maximale mesurees a Saclay
        periode 1988-1997';
title2 'pour les jours ou la valeur du maximum d''ozone depasse 130.';
run;

proc gchart data=sds gout=c.c1;
  vbar trange /cpercent autoref;
title1 'Valeurs de la difference entre Tmax et Tmin mesurees a Saclay
        periode 1988-1997';
title2 'pour les jours ou la valeur du maximum d''ozone depasse 130.';
run;

proc gchart data=sds gout=c.c1;
  vbar wsavg /cpercent autoref;
title1 'Valeurs de la vitesse du vent moyenne journaliere mesurees
        a 58m a Saclay periode 1988-1997';
title2 'pour les jours ou la valeur du maximum d''ozone depasse 130.';
run;

proc gchart data=sds gout=c.c1;
  vbar wsrange / cpercent autoref;
title1 'Valeurs de la difference entre Vmax et Vmin mesurees a 58m
         a Saclay periode 1988-1997';
title2 'pour les jours ou la valeur du maximum d''ozone depasse 130.';
```

```
run;

/*********************************************************
proc greplay gout=c.c1 igout=c.c1;
goptions gaccess=gsasfile device=psepsf gsfmode=append;
filename gsasfile "d:\lise\highleve\graph\c1.eps";
*********************************************************/
run;
```

Boîtes à "moustaches"

```
/*********************************************************
 Programme SAS Boite a "moustaches"
*********************************************************/
data tot;
 filename fich'd:\lise\highleve\donnees\dontrend.don';
 infile fich lrecl=1500;
 input jour mois an pmax1692 pmax1693 pmax1677 pmax1694
       tmax trange wsavg wsrange ;
x1692=pmax1692-130;
if x1692>0;
run;
goptions reset=global gunit= border
        ftext=simplex htitle=0.35cm htext=0.25cm;
libname c 'd:\lise\highleve\graph\';

symbol interpol=boxjt01
       width=3
       value=square
       height=2;

proc gplot data=tot gout=c.c2;
     plot pmax1692*an / haxis=axis1
                        vaxis=axis2;
axis1 value=('88' '89' '90' '91' '92' '93' '94' '95' '96' '97');
axis2 label=(a=90 'Maximum d''ozone en gm3 mesure a Neuilly/Seine')
      order=(0 to 300 by 100);
title1 'Comparaison des maxima d''ozone';
title2 'mesure a Creteil entre 6h00 et 18h00
                  (periode 01/05-15/09 1988-1997)';
title3 'pour les jours ou le maximum d''ozone depasse 130.';

/*******
proc greplay gout=c.c2 igout=c.c2;
goptions gaccess=gsasfile device=psepsf gsfmode=append;
filename gsasfile "d:\lise\highleve\graph\boxcra.eps";
*******/
run;
```

Graphes comparatifs

```
/*********************************************************
Graphiques comparatifs
 ozone >130 /temperature....
*********************************************************/
data sds;
filename fich'd:\lise\highleve\donnees\dontrend.don';
 infile fich lrecl=1500;
 input jour mois an pmax1692 pmax1693 pmax1677 pmax1694
       tmax trange wsavg wsrange ;
x1692=pmax1692-130;
 if x1692>0;

goptions reset=global gunit= border
        ftext=simplex htitle=0.35cm htext=0.25cm;
libname c 'd:\lise\highleve\graph\';

proc gplot data=sds gout=c.c3;
```

```
     plot pmax*tmax / grid
                  haxis=axis1
                  vaxis=axis2;
axis1 label=('Temperature maximale mesuree a Saclay')
      width=1;
axis2 label=(a=90 'Max d''ozone journalier a Neuilly/Seine 1998-1997') ;
title1 'Comparaison des maxima d''ozone>130 avec tmax';
title2 'pour les jours ou le maximum d''ozone est > 130.';

proc gplot data=sds gout=c.c3;
     plot pmax*trange / grid
                  haxis=axis1
                  vaxis=axis2;
axis1 label=('Difference entre tmax et tmin mesurees a Saclay')
      width=1;
axis2 label=(a=90 'Max d''ozone journalier a Neuilly/Seine 1998-1997') ;
title1 'Comparaison des maxima d''ozone>130 avec trange';
title2 'pour les jours ou le maximum d''ozone est > 130.';
run ;
proc gplot data=sds gout=c.c3;
     plot pmax*wsavg / grid
                  haxis=axis1
                  vaxis=axis2;
axis1 label=('Vitesse moyenne du vent mesuree a Saclay')
      width=1;
axis2 label=(a=90 'Max d''ozone journalier >130 a Neuilly/Seine 1998-1997') ;
title1 'Comparaison des maxima d''ozone avec vmoy';
title2 'pour les jours ou le maximum d''ozone est > 130.';

proc gplot data=sds gout=c.c3;
     plot pmax*wsrange / grid
                  haxis=axis1
                  vaxis=axis2;
axis1 label=('Difference entre vmax et vmin')
      width=1;
axis2 label=(a=90 'Max d''ozone journalier>130 a Neuilly/Seine 1998-1997') ;
title1 'Comparaison des maxima d''ozone avec wsrange';
title2 'pour les jours ou le maximum d''ozone est > 130.';

proc gplot data=sds gout=c.c3;
     plot tmax*wsavg / grid
                  haxis=axis1
                  vaxis=axis2;
axis1 label=('Vitesse moyenne du vent mesuree a Saclay')
      width=1;
axis2 label=(a=90 'Temperature max mesuree a Saclay 1998-1997') ;
title1 'Comparaison tmax avec vitesse du vent moyenne';
title2 'pour les jours ou le maximum d''ozone est > 130.';

proc gplot data=sds gout=c.c3;
     plot tmax*wsrange / grid
                  haxis=axis1
                  vaxis=axis2;
axis1 label=('Difference entre vmax et vmin')
      width=1;
axis2 label=(a=90 'Temperature max mesuree a Saclay 1998-1997') ;
title1 'Comparaison tmax avec wsrange';
title2 'pour les jours ou le maximum d''ozone est > 130.';
run;

/*****
proc greplay gout=c.c3 igout=c.c3;
goptions gaccess=gsasfile device=psepsf gsfmode=append;
filename gsasfile "d:\lise\highleve\graph\comp1a.eps";
******/
run;
```

Choix des seuils u raisonnables

- INDÉPENDANCE MUTUELLE ENTRE INTERVALLES DE TEMPS SUCCESSIFS
  Ce programme, utilisant le logiciel Matlab, permet d'estimer le coefficient de corrélation entre intervalles de temps successifs par la méthode Bootstrap.

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%Estimation de la correlation serielle
%des intervalles de temps entre deux valeurs du maximum d'ozone
%mesurees sur le site de Neuilly/Seine et depassant le seuil de 130
%periode 1988-1997
%par la methode Bootstrap
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

load  d:\lise\highlevel\donnees\inneu120.don -ascii
x=inneu120;
r=corrcoef(x);
rxy=r(1,2)
 [n,p]= size(x);
c=1000;
 alpha=0.025;
 y=ones(c,1)*0;

 for j= 1:c
    k=round(n*rand(n,1)+0.5);
for i= 1: n
xb(i,:)=x(k(i),:);
end
rb=corrcoef(xb);
 y(j, 1)=rb(1,2);

 end
 n
 ymoy=mean(y)
 ystd=std(y)
 yinf=min(y)
 ymax=max(y)
 ymed=median(y)
 yt=sort(y);
 q0025=yt(c*alpha)
 q975=yt(c*(1 -alpha))
 hist(yt)
title('Histogramme des differentes valeurs de correlations serielles obtenues')
print -f1 -deps pomax1.ps
```

- INDÉPENDANCE MUTUELLE ENTRE INTERVALLES DE TEMPS SUCCESSIFS
  même programme.

B.10.2 Estimation des paramètres

- FRÉQUENCE DES DÉPASSEMENTS

```
 /*************************************************************
 Regression logistique pour la station Neuilly
*************************************************************/
data tot;
 filename fich'd:\lise\highleve\donnees\dontrend.don';
 infile fich lrecl=1500;
 input jour mois an pmax1692 pmax1693 pmax1677 pmax1694
       tmax trange wsavg wsrange ;
run;

data hlevel;
   set tot;
t=an-87;
if an<92 then  do;
               t92=1;
                end;
          else do;
               t92=2;
```

```
              end;
if pmax1692>120 then dep120=1;
if ((pmax1692<=120) & (pmax1692^=.)) then dep120=2;
if pmax1692=. then dep120=.;

if pmax1692>130 then dep130=1;
if ((pmax1692<=130) & (pmax1692^=.)) then dep130=2;
if pmax1692=. then dep130=.;

 proc logistic ;
   model dep130=t t92 tmax trange wsavg wsrange
             / selection=backward
               fast
               slstay=0.05
               corrb ;
      output out=pred p=pp;
 run;
```

– TAILLES DES DÉPASSEMENTS

```
/*************************************************************
Modelisation de la taille des depassements de niveau u=130
Station Neuilly/Seine, periode 1988-1997
ajout de la variable t92
suppression de l'effet moyen
*************************************************************/
data level;
 filename fich'd:\lise\highleve\donnees\dontrend.don';
 infile fich lrecl=1500;
 input jour mois an pmax1692 pmax1693 pmax1677 pmax1694
       tmax trange wsavg wsrange ;
t=an-87;
if an<92 then  do;
                t92=1;
               end;
          else do;
                t92=2;
               end;
 x=pmax1692-130;
 if x>0;
 obs=_N_;
keep obs jour mois  t t92 tmax trange wsavg wsrange x;

proc iml;
  use level;
  read all var {obs t t92 tmax wsavg trange wsrange x}
  into cc;
  close level;
  nc=ncol(cc); nl=nrow(cc);

/*** nombre d iterations totales ***/
nn=40;
c={-0.0006046,0.035789,-00.001244,0.0187940};
er=1E-4; k=1; s=er;

/*************************************************************
calcul de la valeur de la vraisemblance
pour l estimation du vecteur de parametre c
*************************************************************/

beta=j(nl,1,1); aa=beta; bb=beta;
  do i=1 to nl;
      beta(|i|)=c(|1|)*cc(|i,2|)+c(|2|)*cc(|i,3|)+c(|3|)*cc(|i,4|)
          +c(|4|)*cc(|i,5|) ;
      aa(|i|)=-beta(|i|)*cc(|i,8|);
      bb(|i|)=log(beta(|i|));
  end;
a=aa(|+|); b=bb(|+|); f= a+b;
print k c f;
free aa bb a b;
vf=f; vf1=f; ll=1;
```

```
do until (((s<er) & (abs(f-vf1)<0.1)) | (k>nn));

/***********************************************************
calcul des derivees premieres de la log-vraisemblance
***********************************************************/
der=j(4,1,1); df=j(nl,4,1);
do i=1 to nl;
  do j=1 to 4;
     df(|i,j|)=(cc(|i,j+1|)/beta(|i|))-cc(|i,j+1|)*cc(|i,8|);
  end;
end;
/********* derivees premieres **********/
do j=1 to 4;
der(|j|)=df(|+,j|);
end;
free df;


/***************************************
calcul des derivees secondes de la log-vraisemblance
***************************************/
second=j(4,4,1);
h1=j(nl,4,0); h2=j(nl,4,0); h3=j(nl,4,0); h4=j(nl,4,0);

do i=1 to nl;
   h1(|i,1|)=cc(|i,2|)**2/((beta(|i|))**2);
     do j=2 to 4;
      h1(|i,j|)=cc(|i,2|)*cc(|i,j+1|)/((beta(|i|))**2);
     end;
  h2(|i,2|)=cc(|i,3|)**2/((beta(|i|))**2);
     do j=3 to 4;
      h2(|i,j|)=cc(|i,3|)*cc(|i,j+1|)/((beta(|i|))**2);
     end;
  h3(|i,3|)=cc(|i,4|)**2/((beta(|i|))**2);
  h3(|i,4|)=cc(|i,4|)*cc(|i,5|)/((beta(|i|))**2);
  h4(|i,4|)=cc(|i,5|)**2/((beta(|i|))**2);
end;

do j=1 to 4;
   second(|1,j|)=-h1(|+,j|);
end;
do j=2 to 4;
   second(|2,j|)=-h2(|+,j|);
end;
do j=3 to 4;
   second(|3,j|)=-h3(|+,j|);
end;
second(|4,4|)=-h4(|+,4|);

do i=1 to 4;
  do j=1 to 4;
     if i>j then second(|i,j|)=second(|j,i|);
  end;
end;

free h1; free h2; free h3; free h4;
ca=c; ck=ll*ginv(second)*der; c=ca-ck;

/***********************************************************
calcul de la nouvelle valeur de la log-vraisemblance
avec valeur de c changee
***********************************************************/

beta=j(nl,1,1); aa=beta; bb=beta;
  do i=1 to nl;
     beta(|i|)=c(|1|)*cc(|i,2|)+c(|2|)*cc(|i,3|)+c(|3|)*cc(|i,4|)
          +c(|4|)*cc(|i,5|);
     aa(|i|)=-beta(|i|)*cc(|i,8|);
     bb(|i|)=log(beta(|i|));
```

```
  end;
a=aa(|+|); b=bb(|+|); f= a+b;
free aa bb a b;

if (f>vf) then do;
               if (ll<1) then ll=ll*10;
               end;
          else do;
               do until ((f>vf1) | (ll<(1E-20)));
                     ll=ll/10;
                     ck=ll*ginv(second)*der;
                     c=ca-ck;
                     vf=f;
beta=j(nl,1,1); aa=beta; bb=beta;
  do i=1 to nl;
     beta(|i|)=c(|1|)*cc(|i,2|)+c(|2|)*cc(|i,3|)+c(|3|)*cc(|i,4|)
          +c(|4|)*cc(|i,5|) ;
     aa(|i|)=-beta(|i|)*cc(|i,8|);
     bb(|i|)=log(beta(|i|));
  end;

  a=aa(|+|); b=bb(|+|); f= a+b;
free aa bb a b;

 end;
       end;
print k c f vf s;
vf1=vf; vf=f; s=sqrt(ssq(ck)); ll=1;
      k=k+1;
 end;

cor=second; cov=-ginv(second);
print cov;
  do i=1 to 4;
      do j=1 to 4;
          cor(|i,j|)=cov(|i,j|)/sqrt(cov(|i,i|)*cov(|j,j|))  ;
      end;
  end;
print cor;

if k>nn then do;
     print 'Interations insuffisantes';
     print c;
             end;

else do;
  cg=t(c); cov=-ginv(second); b=-eigval(second);
  print b;
  ectyp=c; tratio=c; var=j(4,3,0);
     do i=1 to 4;
      ectyp(|i|)=sqrt(cov(|i,i|));
      tratio(|i|)=c(|i|)/ectyp(|i|);
     end;
  var=c || ectyp || tratio;
  az={'estimation' 'ecart-type' 'T ratio'};
  bb={'beta1' 'beta2' 'beta3' 'beta4' };
  print var[rowname=bb colname=az];
     end;
quit;
```

B.10.3 Validation du modèle

– FRÉQUENCE DES DÉPASSEMENTS : VALIDATION DE LA DISTRIBUTION DES INTERVALLES S_i

```
/*************************************************************
Modelisation de la frequence des jours de depassement du seuil
u=130 pour Neuilly/Seine
```

```
Notations :
t=an-87
t92=1 si an<92 2 sinon
utilisation de la periode 1988-1997
effet t92 et t !!!
test de Kolmogorov-Smirnov
sur la distribution des intervalles Si entre les
depassements (processus de Poisson non homogene
************************************************************/

data level;
 filename fich'd:\lise\highleve\donnees\dontrend.don';
 infile fich lrecl=1500;
 input jour mois an pmax1692 pmax1693 pmax1677 pmax1694
       tmax trange wsavg wsrange ;

/****** etude u=130 ***************/
 x=pmax1692-130;
 t=an-87;
 if an<92 then  do;
              t92=1;
               end;
          else do;
               t92=2;
               end;
obs=_N_;
run;

data dep;
  set level;
 if x>0;

proc iml;
    use level;
    read all var {obs t t92 tmax wsavg x}
    into tot;
    close level;
    use dep;
    read all var {obs t t92 tmax wsavg x}
    into depa;
    close dep;
    nctot=ncol(tot);
    nltot=nrow(tot);
    ncdepa=ncol(depa);
    nldepa=nrow(depa);
v=j(nldepa,1,0);
s=j(nldepa,1,0);
  do j=2 to nldepa;
     s(|j|)=depa(|j,1|)-depa(|j-1,1|);
  end;
b=max(s);
print b;
 nnl=nldepa-1;
print nnl;
alpha=j(nnl,b,0); aalpha=alpha;  l=j(nnl,1,1); F=l;
     do i=1 to nnl;
        do j=(depa(|i,1|)+1) to depa(|i+1,1|);
      alpha(|i,j-depa(|i,1|)|)=-9.7515+0.3756*tot(|j,2|)
         -2.3780*tot(|j,3|)+0.4522*tot(|j,4|)-0.8894*tot(|j,5|);
         end;
     end;

/************************************************************
 proba que le seuil 130 soit depasse connaissant les cond meteo :
    regression logistique
************************************************************/

  do i=1 to nnl;
    do j=1 to b;
       aalpha(|i,j|)=exp(alpha(|i,j|))/(1+exp(alpha(|i,j|)));
```

```
      if aalpha(|i,j|)=0.5 then aalpha(|i,j|)=0;
    end;
  end;
  ff=aalpha(|,1:10|);

  do i=1 to nnl;
     l(|i|)=aalpha(|i,+|);
     if l(|i|)=0 then l(|i|)=.;
     F(|i|)=1-exp(-l(|i|));
  end;
print l F;
bb={'U'};
create unif from F(|colname=bb|);
append from F;
 close unif;
 quit;

proc sort data=unif;
  by u;
run;
proc print data=unif;
run;

/*************************************************************
 suppression des valeurs de unif manquantes
*************************************************************/

 proc iml;
    use unif;
    read all var {u}
    into un;
    close unif;
  nl=nrow(un); nl1=nl+1;  s=j(nl,1,0);

   do i=1 to nl;
       if  un(|i,1|)=. then s(|i|)=1;
   end;
   nmis=s(|+|); nmis1=nmis+1;
F=un(|nmis1:nl,|); nlf=nrow(F); nlf1=nlf+1;
print un F;
w=j(nlf,1,0);
  do i=1 to nlf;
     w(|i|)=i/nlf1;
  end;
/*************************************************************
 test de kolmogorov valeur de la statistique Dn
*************************************************************/

dplusn=w-F; dmoinsn=F-w; dpn=max(dplusn);
dmn=max(dmoinsn); dn=max(dpn,dmn);
test= F || w;
bb={'FF' 'N'};
create testu from test(|colname=bb|);
append from test;
close testu;
quit;
/*************************************************************
Probability plot
*************************************************************/

goptions reset=global gunit= border
         ftext=simplex htitle=0.35cm htext=0.25cm;
libname c 'd:\lise\highleve\graph\';

proc gplot data=testu gout=c.c4;
     plot N*FF=1  / overlay
                 grid
                 haxis=axis1
                 vaxis=axis2
                 hminor=0
```

```
                vminor=0;
goptions gaccess=gsasfile device=win gsfmode=append;
filename gsasfile "d:\lise\highleve\graph\frene130.eps";
axis1 order=(0 to 1 by 0.1)
      label=('i/(N+1)')
      width=1;
axis2 order=(0 to 1 by 0.1)
      label=(a=90 'distribution des intervalles') ;
title1 'Modelisation des intervalles Si entre depassements du seuil 130:
             Si=Ti-T(i-1)';
title2 'Probability Plot';
symbol1 c=black i=none value=star height=2;

/*********
proc greplay gout=c.c4 igout=c.c4;
goptions gaccess=gsasfile device=psepsf gsfmode=append;
filename gsasfile "d:\lise\highleve\graph\frene130.eps";
********/
run;
```

- TAILLE DES DÉPASSEMENTS

  - Test de Kolmogorov-Smirnov

```
/*************************************************************
Modelisation de la frequence des jours de depassement
du seuil u=130
test de Kolmogorov-Smirnov
sur la distribution de la taille des depassements

periode 1988-1997
ajout de la variable t92
*************************************************************/
data level;
 filename fich'd:\lise\highleve\donnees\dontrend.don';
 infile fich lrecl=1500;
 input jour mois an pmax1692 pmax1693 pmax1677 pmax1694
       tmax trange wsavg wsrange ;

/****** etude u=130 ***************/
 x=pmax1692-130;
 t=an-87;
 if an<92 then  do;
              t92=1;
               end;
          else do;
               t92=2;
               end;
obs=_N_;
run;

data dep;
  set level;
 if x>0;

 proc iml;
    use dep;
    read all var {obs t92 t wsavg tmax trange x}
 into depa;
    close dep;
    ncdepa=ncol(depa); nldepa=nrow(depa);
    beta=j(nldepa,1,0);  xx=beta; F=beta;
 do i=1 to nldepa;
    beta(|i|)=0.0357893*depa(|i,2|)-0.006046*depa(|i,3|)+0.0187945*depa(|i,4|)
      -0.001244*depa(|i,5|);
    xx(|i|)=-beta(|i|)*depa(|i,7|);
 end;
F=1-exp(xx);
print F;
```

```
bb={'U'};
create unif from F(|colname=bb|);
append from F;
 close unif;
 quit;

proc sort data=unif;
  by u;
run;
/*************************************************************
  supprimer des valeurs de unif manquantes
*************************************************************/

 proc iml;
    use unif;
    read all var {u}
    into un;
    close unif;
  nl=nrow(un); nl1=nl+1;  s=j(nl,1,0);

   do i=1 to nl;
      if  un(|i,1|)=. then s(|i|)=1;
   end;

nmis=s(|+|); nmis1=nmis+1; F=un(|nmis1:nl,|);
nlf=nrow(F); nlf1=nlf+1;
print un F;
w=j(nlf,1,0);
  do i=1 to nlf;
     w(|i|)=i/nlf1;
  end;

/*************************************************************
 test de kolmogorov valeur de la statistique Dn
*************************************************************/

dplusn=w-F; dmoinsn=F-w; dpn=max(dplusn);
dmn=max(dmoinsn); dn=max(dpn,dmn);
test= F || w;
nltest=nrow(test);
bb={'U' 'N'};
create testu from test(|colname=bb|);
append from test;
 close testu;
quit;

/*************************************************************
Probability plot
*************************************************************/

goptions reset=global gunit= border
         ftext=simplex htitle=0.35cm htext=0.25cm;
libname c 'd:\lise\highleve\graph\';

proc gplot data=testu gout=c.c4;
     plot N*U=1   / overlay
                 grid
                 haxis=axis1
                 vaxis=axis2
                 hminor=0
                 vminor=0;
axis1 order=(0 to 1 by 0.1)
      label=('i/(N+1)')
      width=1;
axis2 order=(0 to 1 by 0.1)
      label=(a=90 'distribution de la taille des depassements ') ;
title1 'Modelisation de la taille des depassement Xi ';
title2 'Probability Plot';
symbol1 c=black i=none value=star height=2;
```

```
/************************
proc greplay gout=c.c4 igout=c.c4;
goptions gaccess=gsasfile device=psepsf gsfmode=append;
filename gsasfile "d:\lise\highleve\graph\testaill.eps";
************************/
run;
```

– Contours de la vraisemblance

```
[x, out]=fmins('beta1', [0.1 , -0.001, 0.01]);
x
beta1(x)
out
Y=[
19 26.4 14.7 1.23077 1;
40 26.2 11.2 2.38462 4;
8 27.4 13.0 3.23077 4;
9 26.6 10.2 2.15385 3;
26 29.2 12.3 2.30769 3;
54 28.9 12.9 1.84615 3;
6 22.1 7.2 4.92308 2;
11 22.0 13.4 2.07692 2;
15 24.7 13.2 2.15385 2;
13 26.3 11.8 2.30769 3;
5 27.5 14.0 3.23077 3;
1 28.9 13.9 2.84615 2;
14 28.4 12.4 2.69231 3;
16 24.0 12.5 2.38462 1;
14 27.2 11.2 2.92308 2;
10 26.1 10.6 3.46154 2;
9 25.5 9.8 4.00000 4;
0 25.9 9.6 4.92308 2;
7 27.4 11.5 2.92308 3;
12 32.2 17.3 3.76923 3;
41 36.3 15.6 1.84615 2;
117 32.3 17.7 1.76923 2;
44 33.3 18.8 3.84615 4;
36 28.6 10.8 3.07692 4;
1 32.2 16.6 2.84615 2;
8 16.5 12.4 4.23077 4;
69 32.4 15.5 3.76923 1;
23 26.3 8.4 2.84615 2;
59 24.5 11.8 3.69231 1;
41 26.2 12.8 2.00000 0;
137 28.0 13.6 2.00000 2;
143 30.1 15.1 2.84615 4;
26 22.1 10.2 2.61538 2;
98 23.9 14.8 2.07692 2;
93 26.6 15.6 1.69231 2;
80 29.6 16.6 3.69231 3;
17 29.2 18.7 3.38462 2;
21 28.6 15.1 3.23077 2;
5 25.4 12.2 1.53846 1;
21 25.4 11.5 2.15385 2;
20 26.1 10.0 3.30769 1;
1 24.2 7.1 5.46154 3;
4 25.4 12.1 2.76923 1;
4 29.7 13.9 2.38462 1;
7 32.1 14.1 2.07692 2;
5 29.3 16.8 1.69231 2;
5 30.1 16.8 2.53846 3;
47 34.2 17.3 2.38462 2;
82 30.1 12.8 1.69231 2;
9 29.2 17.6 1.46154 1;
75 34.6 17.8 2.07692 2;
10 35.1 15.2 3.00000 3;
0 25.5 12.5 1.92308 2;
15 23.1 10.0 2.84615 3;
81 29.6 14.2 1.15385 1;
81 31.2 12.4 1.30769 1;
```

```
62 30.0 11.9 2.92308 3;
17 26.8 11.5 1.92308 1;
26 29.1 11.9 2.07692 2;
46 29.8 12.2 2.38462 1;
35 31.5 11.7 3.15385 2;
30 31.3 12.1 2.38462 3;
7 30.0 12.0 1.92308 2;
4 30.1 12.3 2.07692 2;
139 32.3 12.7 1.30769 1;
3 32.2 12.1 3.38462 3;
95 34.9 15.2 2.46154 4;
10 28.2 8.8 2.53846 1;
5 23.9 12.4 2.00000 2;
22 29.1 10.1 3.38462 5;
16 29.0 12.7 2.92308 2;
30 32.2 13.4 3.15385 4;
13 34.2 15.4 3.15385 2;
16 35.0 13.3 3.84615 4;
10 29.7 12.2 3.15385 2;
5 31.5 12.7 3.30769 4;
39 33.1 15.7 3.15385 3;
28 34.1 14.3 4.15385 4;
13 31.3 13.5 5.15385 4;
42 31.2 14.3 3.92308 4;
52 32.5 15.1 3.53846 3;
31 31.1 13.0 3.00000 3;
0 27.5 13.2 3.84615 3;
21 30.5 15.6 2.53846 3;
26 27.8 16.7 3.69231 3;
4 23.4 11.2 2.69231 3;
2 27.6 15.9 3.30769 2;
6 28.6 7.3 8.07692 2;
25 28.9 13.8 3.38462 2;
8 27.3 14.8 3.15385 3;
16 29.1 16.0 3.84615 2;
3 22.2 14.3 2.46154 1;
17 27.6 11.5 3.07692 3;
6 30.2 12.4 5.69231 5;
47 31.2 11.8 2.23077 3;
18 30.1 10.3 3.38462 4;
50 28.7 12.2 2.76923 2;
4 32.5 13.9 3.61538 3
];
d=Y(:,1);
Tmax= Y(:,2);
Trange= Y(:,3);
Wsavg= Y(:,4);
Wsrange= Y(:,5);
[n,p]=size(Y)
a=0.076
x1=-0.0036:0.0001:-0.0016;
y1=0.0041:0.001:0.0241;
[X1,Y1]=meshgrid(x1,y1);
[nx,px]=size(X1);
[ny,py]=size(Y1);
Z=zeros(nx,px);
for i=1:n
Z1=a*ones(nx,px)+Tmax(i)*X1+Wsavg(i)*Y1;
if Z1<=0
Z2=0
else
Z2=-log(Z1)+Z1.*ones(nx,px)*d(i);
end
Z=Z+Z2;
end
v=[420:5:450]
contour(x1,y1,Z,v)
%contour3(X1,Y1,Z, 20)
```

Bibliographie

[1] Académie des Sciences. *Ozone et propriétés oxydantes de la troposphère : Essai d'évaluation scientifique.* Technique & Documentation-Lavoisier, Paris, Octobre 1993. Rapport n 30.

[2] AIRPARIF. *Surveillance de la qualité de l'Air en Ile-de-France : Les résultats 1997.* 1996.

[3] Azais J.-M. Analyse de variance non orthogonale, l'exemple de SAS/GLM. *Revue de Statistique Appliquée*, XLII(2): 27–41, 1994.

[4] Bel L., Bellanger L., Bobbia M., Ciuperca G., Dacunha-Castelle D., Gilibert E., Jackubowicz P., Oppenheim G., Tomassone R. On forecasting ozone episodes in the Paris area. *Listy Biometryczne- Biometrical Letters*, 35(1), 1998.

[5] Bel L., Bellanger L., Bonneau V., Ciuperca G., Dacunha-Castelle D. , Deniau C., Ghattas B., Misiti M., Misiti Y., Oppenheim G., Poggi J.M., Tomassone R. Eléments de comparaison de prévisions statistiques des pics d'ozone. *Revue de Statistique Appliquée*, 1998. à paraître.

[6] Bel L., Bellanger L., Bonneau V., Ciuperca G., Dacunha-Castelle D., Deniau C., Ghattas B., Misiti M., Misiti Y., Oppenheim G., Poggi J-M., Tomassone R. Prévision des pointes de pollution dans la région parisienne. Technical report, Laboratoire Modélisation Stochastique et Statistique, Université de Paris-Sud, Orsay, 1997.

[7] Bellanger L., Tomassone R. Wind direction and maximum pollutants concentration : a case-study. In Blasco ed., editor, *Advances in Environmental and Ecological Modelling.* Kluwer, Amsterdam, 1998. à paraître.

[8] Bloomfield P., Royle A., Yang Q. 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, 1993.

[9] Bloomfield P., Royle A., Yang Q. 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, 1993.

[10] Brion D., Gilibert E. Forecasting atmospheric pollution peaks over the Ile-de-France. *à paraître*, 1996.

[11] Carlier P., Mouvier G. Initiation à la physico-chimie de la basse troposphère. *Pollution Atmosphérique*, Janvier-Mars: 12–24, 1988.

[12] Cox D. R., Lewis P. A. *The Statistical Analysis of Series of Events.* John Wiley, New York, 1966.

[13] Cox W. M., Shao-Hang C. Meteorologically adjusted ozone trends in urban areas: A probabilistic approach. *Atmospheric Environment*, 27B(4): 425–434, 1993.

[14] Crow L.H. Reliability analysis for complex repairable systems. In F.Proschan and R.J. Serfling (eds), editors, *Reliability and Biometry SIAM, pages 379–410. Philadelphia, PA, 1974.*

[15] Davidson A.C. Modelling excesses over high thresholds, with an application. In Statistical extremes and applications (ed. J.Tiago de Oliveira), *pages 621–638, 1984. Dordrecht: Reidel.*

[16] Davidson A.C., Smith R. L. Models for exceedances over high thresholds (with discussion). J.R. Statist. Soc., *52: 393–442, 1990.*

[17] Davis J.M., Eder B.K., Bloomfield P. Modeling Ozone in the Chicago Urban Area. In Cox L.H. Nychka D., Piegorsch W.W., editor, Case Studies in Environmental Statistics. *Springler Verlag, New York, 1998.*

[18] Eder B. E., Davis J. M., Bloomfield P. An automated classification scheme designed to better elucidate the dependence of ozone on meteorology. Journal of Applied Meteorology, *33, 1994.*

[19] Gao F., Sacks J., Welch W. J. Predicting the urban ozone levels and trends with semiparametric modelling. Technical Report 14, National Institute of Statistical Sciences, P.O. Box 14162, Research Triangle Park, N.C. 27709, 1994.

[20] Graf-Jaccottet M., Jaunin M.-H. Predictive models for ground ozone and nitrogen dioxyde time series. Environmetrics, *9: 393–406, 1998.*

[21] Hosking J.M.R., Wallis J.R. Parameter and quantile estimation for the generalized Pareto distribution. Technometrics, *29: 339–349, 1987.*

[22] Hosmer D. W., Lemeshow S. Applied Logistic Regression. *John Wiley & Sons, New York, 1989.*

[23] Johnson R. A., Wehrly T. Measures and models for angular correlation and angular-linear correlation. Journal of the Royal Statistical Society, *(Serie B) 39: 222–229, 1977.*

[24] Krzanowski, W. J. Principal component analysis in the presence of group structure. Applied Statistics, *33: 164–168, 1984.*

[25] Leadbetter M. R. On a basis for "Peaks over Threshold "modeling. Statistics and Probability Letters, *12: 357–362, 1991.*

[26] Leadbetter M. R. On high level exceedance modelling and tail inference. J. Stat. Plan. Inference, *45(1-2): 247–260, 1995.*

[27] Leadbetter M.R. On exceedance based environmental criteria. Technical Report 9, National Institute of Statistical Sciences, P.O. Box 14162, Research Triangle Park, N.C. 27709, 1993.

[28] Leadbetter M.R., Lindgren G., Rootzén H. Extremes and Related Properties of Random Sequences and Series. *Springer Verlag, New York, 1983.*

[29] Lee L. Testing adequacy of the Weibull and log linear rate models for a Poisson process. Technometrics, *22(2): 195–199, 1980.*

[30] Legay J.M., Tomassone R. La comparaison de régressions orthogonales. Revue de Statistique Appliquée, *1998. à paraître.*

[31] Mardia K. V., Kent J. T., Bibby J. M. Multivariate Analysis. *Academic Press Inc., London, 1982.*

[32] Math Works Inc. MATLAB Reference Guide,. Cary,The Math Works Inc., 1995.

[33] Math Works Inc. *MATLAB User's Guide,. Cary,The Math Works Inc., 1995.*

[34] Milionis A. E., Davies T. D. Regression and stochastic models for air pollution-I. Review, comments and suggestions. Atmospheric Environment, *28(17): 2801–2810, 1994.*

[35] Pickands J. The two-dimensional Poisson process and extremal processes. J. Appl. Probab., *8: 745–756, 1971.*

[36] Pickands J. Statistical inference using extreme order statistics. Ann. Statist., *3: 119–131, 1975.*

[37] Pugh D., Vassie J.M. Applications of the joint probability method for extreme sea level computations. Proc. Instn Civ. Engrs, Part 2, *69: 959–975, 1980.*

[38] Rootzén H., Leadbetter M. R., de Haan L. On the distribution of tail array sums for strongly mixing stationary sequences. 1994. en preparation.

[39] Ross G.J.S. Nonlinear Estimation. Springer-Verlag, London, 1990.

[40] Saporta G. *Probabilités Analyse des données et statistique.* Editions Technip, Paris, 1990.

[41] SAS Institute Inc. *SAS/IML software: Usage and reference, Version 6, First Edition. Cary,NC: SAS Institute Inc., 1989.*

[42] SAS Institute Inc. SAS/STAT User's Guide : Version 6, Fourth Edition, Vol. 1 and 2. Cary,NC: SAS Institute Inc., 1994.

[43] Shively T.S. An analysis of the long-term trend in ozone data from two Houston Texas monitoring sites. *Atmospheric Environment*, 24B: 293–301, 1990.

[44] Shively T.S. An analysis of the trend in ground-level ozone using nonhomogeneous Poisson processes. *Atmospheric Environment*, 25B: 387–396, 1991.

[45] Smith R.L. Threshold methods for sample extremes. *In Statistical extremes and applications (ed. J.Tiago de Oliveira)*, pages 621–638, 1984. Dordrecht : Reidel.

[46] Smith R.L. Maximum likelihood estimation in a class of nonregular cases. *Biometrika*, 72(1): 67–90, 1985.

[47] Smith R.L. Extreme value theory based on the r largest annual events. *J. Hydrology*, 86: 27–43, 1986.

[48] Smith R.L. Estimating tails of probability distributions. *The annals of Statistics*, 15(3): 1174–1207, 1987.

[49] Smith R.L. Extreme value analysis of environmental time series : An application to trend detection in ground-level ozone (with discussion). *Statistical Science*, 4: 367–393, 1989.

[50] Smith R.L., Shively T. S. Point process approach to modeling trends in tropospheric ozone based on exceedances of a high threshold. *Atmospheric Environment*, 29(23): 3489–3499, 1995.

[51] Somerville M. C., Mukerjee S., Fox D. Estimating the wind direction of maximum air pollutant concentration. *Environmetrics*, 7: 231–243, 1996.

[52] Tomassone R., Dervin C. et Masson J.P. *Biométrie : Modélisation de phénomènes biologiques.* Masson, Paris, 1995.

[53] Toupance G. L'ozone dans la basse troposphère. théorie et pratique. *Pollution Atmosphérique*, Janvier-Mars: 32–42, 1988.

[54] Toupance G., Perros P., Soedomo M. The rapid rotation of the wind direction and sharp ozone peak features. *Physicochemical Behavior of Atmospheric Pollutants*, 1986.

[55] Vaquera-Huerta H., Villasenor J.A., Hughes J. Statistical analysis of trends in urban ozone. *Statistics for the Environment 3 : Pollution Assessment and Control (edited by Barnett V. and Turkman K.F.)*, pages 175–183, 1997. John Wiley & Sons Ltd.

[56] Vautard R., Beekman M., Honoré C., Deleuze I. La pollution photochimique en région parisienne simulée par le modèle chimere et l'influence du transport régional d'ozone. Technical report, 1998.

[57] Weissman I. Estimation of parameters and large quantiles based on the k largest observations. *J. Amer. Statist. Assoc.*, 73: 812–815, 1978.

[58] Zeldin M.D., Cassmassi J.C. Development of improved methods for predicting air quality levels in the south coast air bassin. *Final Report to California Air Ressources Board, Technology Service Corp., Santa Monica, CA, Contract AG-192-30*, 1979.