Université Paris VI
et
Ecole Nationale du Génie Rural

Thèse

présentée par

Keltoum Chaouche

pour obtenir le grade de

Docteur de L'ENGREF

en

Hydrologie stochastique

# Approche multifractale de la modélisation stochastique en hydrologie

Soutenue publiquement le 4 Janvier 2001 devant le jury suivant:

M. **Jean-Noel BACRO**
M. **Philippe BOIS**
M. **Pierre HUBERT**
M. **Gabriel LANG**
M. **Christian ONOF**
M. **Daniel SCHERTZER**

Le mémoire de thèse de Keltoum Chaouche a été approuvé par:

---

Professeur Ph. Bois Date

---

Professeur C. Onof Date

Université Paris VI

Janvier 2001

## Résumé

La plupart des séries d'observations hydrologiques possèdent des caractéristiques peu communes (grande variabilité sur une large gamme d'échelles spatiales et temporelles, périodicité, corrélation temporelle à faible décroissance), difficiles à mesurer (séries tronquées et intégrées sur des pas de temps qui ne respectent pas la nature du phénomène) et compliquées à intégrer dans un modèle stochastique classique (ARMA, Markov). Le modélisateur doit aussi faire face aux problèmes liés à l'échelle : en hydrologie (et en météorologie) où les données sont issues de pas de temps très divers, il est particulièrement intéressant de disposer de modèles à la fois capable d'intégrer des données à pas de temps différents et de fournir des résultats à un pas de temps différent de celui des entrées (désagrégation ou agrégation de séries temporelles).

Ce travail de thèse explore les possibilités d'application d'un nouveau type de modèle, les modèles multifractals, qui traduisent le plus simplement possible des propriétés d'invariance de certains paramètres (invariance spatiale, temporelle ou spatio-temporelle). En particulier, dans les modèles de cascades multifractales de générateur algébrique, c'est le coefficient de décroissance algébrique qui est un invariant d'échelle. Des résultats issus de la théorie probabiliste des valeurs extrêmes sont dès lors très utiles pour estimer ce paramètre. L'élaboration d'un outil statistique capable de détecter un comportement algébrique et d'estimer le paramètre de décroissance algébrique constitue une étape préalable au développement de cette thèse. Il est aussi montré dans ce travail, que la forme de dépendance (longue ou courte) des modèles en cascade multifractale diffère selon le générateur de la cascade.

L'étude exploite une base de données de 232 séries annuelles de divers sites, de nombreuses séries de pluie à divers pas de temps (mois, jour, heure, minute) ainsi que quelques séries de débits. Elle conduit aux résultats suivants :

- Les lois de type algébrique sont adaptées à la modélisation des grandes périodes de retour des séries de pluie étudiées.

- Sur ces mêmes séries, le coefficient de décroissance algébrique est un paramètre invariant d'échelle (sur des gammes d'échelles supérieures à l'heure).

- L'estimation de ce coefficient en divers sites à travers le monde est très peu variable.

- La propriété de longue dépendance est décelable au sein de certaines séries de débits, notamment des séries de rivières sur craie.

Ces résultats incitent donc à l'emploi de cascades multifractales pour la modélisation des séries de pluie, bien qu'un travail concernant la détection et l'estimation de longue dépendance reste à accomplir pour que le choix du générateur respecte la forme de dépendance de la série.

# Table des matières

# Liste des Figures

# Liste des Tableaux

## Remerciements

J'adresse mes remerciements à tous ceux et à toutes celles qui m'ont aidée, par leur encadrement et par leurs conseils scientifiques, ou par leur appui matériel, à réaliser ce mémoire de thèse.

J'exprime ma gratitude à mes deux directeurs de thèse, M. Pierre Hubert et M. Gabriel Lang. P. Hubert, directeur de la formation doctorale Hydrologie et Hydrogéologie Quantitatives de l'Ecole des Mines de Paris, a donné l'impulsion de départ à ce travail et, par ses conseils, m'a permis de prospecter de nouvelles voies. G. Lang, chercheur au GRESE (laboratoire de Gestion du Risque en Sciences de l'Eau) a encadré cette thèse et n'a pas ménagé son temps pour me relire ou pour discuter mes approches.

Je remercie M. Bacro pour l'aide qu'il m'a apportée pour la modélisation des extrêmes.

Je remercie M. Bois, professeur d'Hydrologie a l'ENSHMG et chercheur au LTHE (Laboratoire d Etudes des Transferts en Hydrologie et Environnement) ainsi que M. Onof, professeur de l'Imperial College de Londres qui me font l'honneur d'être les rapporteurs de ma thèse.

Cette thèse n'aurait pas vu le jour sans Eric Parent, directeur du GRESE qui m'a accueillie et hébergée dans son laboratoire, sans M. de Marsily et MM. Degoutte et Miller, qui m'ont admise dans l'Ecole doctorale ; à tous ceux-là, j'exprime ma reconnaissance.

Je remercie enfin tous ceux qui ont fourni les données hydrologiques : MM. T. Lebel, A. Barcelo et A. Castéanu.

# Chapitre 1

# Introduction

## 1.1 Problème général

L'hydrologie est une science qui étudie le cycle de l'eau dans tous ses aspects, vaste sujet s'étendant de l'évaporation jusqu'à la formation et la propagation des écoulements, superficiels et souterrains. Cependant cette science est aussi chargée d'apporter des réponses concrètes aux nombreuses demandes des aménageurs et des responsables locaux. Ces derniers sont en particulier préoccupés par la prédétermination et la prévision des inondations ou des sécheresses, le dimensionnement des ouvrages de franchissement ou de barrage, l'exploitation de l'eau et la gestion des réserves, l'assainissement urbain ou le drainage des terres agricoles (Réméniéras, [70]).

Suite aux récentes catastrophes survenues en France, de nombreuses actions et réflexions ont été suscitées dans les milieux scientifiques. L'amélioration de la gestion du risque d'inondation est ainsi devenue une préoccupation majeure de la recherche hydrologique. Avant la réalisation de toute étude, l'hydrologue commence par la récolte de mesures et d'observations de certains phénomènes physiques ayant trait à l'eau, pour étudier le problème plus en amont, et tenter d'aboutir à une analyse de la dynamique qui les soutend. L'élaboration de modèles opérationnels de gestion des ressources hydriques, de modèles pluie débit ou de dimensionnement d'ouvrage ne constituent que la phase finale du travail de l'hydrologue, et il importe de modéliser la structure sous-jacente du phénomène pluvieux.

Cependant, on ne possède actuellement qu'une compréhension partielle des mécanismes physiques des champs de pluie (Cho, [18]). En effet, la pluie est, par sa constitution même en gouttes discrètes tombant de façon intermittente, un phénomène extrêmement variable et discontinu dans l'espace et dans le temps. La plupart des hydrologues s'orientent alors vers une approche stochastique de la modélisation, tentant d'exploiter au mieux les séries de données collectées.

Aux problèmes liés à l'incompréhension de la physique du phénomène pluvieux s'ajoute celui de sa mesure. Cette dernière est généralement effectuée en collectant des séries temporelles de variables hydrologiques telles que par exemple des cumuls de pluie ou des débits de rivières. Or ces variables résultent d'une intégration en temps ou en espace et ne rendent pas compte de la grande discontinuité du phénomène. De plus, la plupart de

des séries ne couvrent que quelques décennies, et rares sont celles qui s'étalent sur plus de 100 ans. Pour accéder à des informations plus anciennes concernant le climat ou les records météorologiques historiques, diverses voies ont été proposés telles que les méthodes de paléo-hydrologie. Cependant ces procédés ne permettent d'accéder qu'à des estimations grossières sur les fluctuations climatiques historiques, et ne fournissent jamais de séries temporelles hydrologiques fiables.

S'il se restreignait à une analyse fréquentiste, l'hydrologue serait obligé de se contenter des courtes séries "lissées" disponibles aujourd'hui et serait peu à même de répondre aux divers demandes évoquées précédemment. Il est donc nécessaire d'élaborer un modèle, c'est-à-dire un ensemble d'hypothèses ainsi qu'une formalisation mathématique permettant d'atteindre ces objectifs grâce aux données. Les connaissances que l'on possède à priori sur le phénomène pluvieux sont à faire figurer dans le modèle. Ces dernières peuvent être issues de la physique, mais aussi être de simples évidences liées à des symétries ou des invariances d'échelle. Cependant, l'élaboration de modèles de pluie sur de courtes séries soulève divers problèmes d'ordre statistique:

- Celui de la stationnarité ou non de séries par nature finies : La stationnarité d'une série de variables est une notion relative à la longueur de la série. En effet, lorsque l'on considère la série de la figure 1.1 avant troncature, on peut conclure à la présence d'une tendance ou d'un saut au lieu d'un cycle.
- Le problème de la qualité de l'estimation en général (des paramètres) et de la convergence des estimateurs.
- Ce dernier problème se conjugue à celui de la grande variabilité des séries.

Les récentes avancées des moyens de mesure des événements pluvieux (satellite, radar, nouveaux pluviomètres et pluviographes) orientent aujourd'hui l'hydrologue vers des modèles capables d'intégrer des données à différents pas de temps et d'espace. A ce niveau aussi, on peut d'ores et déjà énoncer les problèmes statistiques qui vont se poser :

- L'intégration de séries d'un même site issues de pas de temps différents dans un modèle pose des problèmes techniques difficiles (parfois insurmontables) dans la plupart des modèles statistiques utilisés en hydrologie.
- De même la réalisation de transfert d'information spatiale est peu aisée dans les modèles statistiques classiques.

Actuellement, la plupart des modèles hydrologiques sont développés en un site et à une certaine échelle de temps, en ajustant de façon subjective une distribution et/ou un processus stochastique à une série pluviométrique. La période de retour d'un événement est en effet fonction du modèle choisi. Les modèles exponentiels et algébriques, qui peuvent être équivalents pour l'estimation des événements fréquents, procurent des résultats fondamentalement différents pour l'estimation des grandes périodes de retour. En effet, dans un modèle exponentiel, on passe du débit de durée de retour $N$ années au débit de durée de retour $10 \times N$ années par l'addition d'une constante alors que dans un modèle algébrique, il

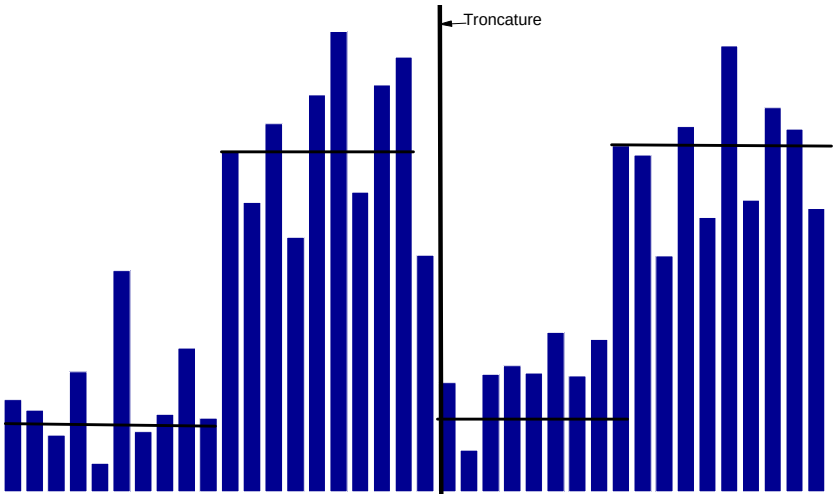



Figure 1.1: Effet de la troncature d'une série.

faut multiplier par une constante (Hubert, [43]). On a accès à l'erreur d'adéquation du modèle en comparant les périodes de retour estimées aux périodes empiriques. Les événements extrêmes recensés aujourd'hui montrent que les modèles de type exponentiel surestiment les grandes périodes de retour. On a représenté sur la figure 1.2 une comparaison des périodes de retour estimées par un modèle normal aux périodes observées sur 232 longues séries annuelles en divers endroits du globe (projet FRIEND-AMY).

Il apparaît une nette surestimation des périodes de retour, dans un rapport pouvant aller jusqu'à 40. Ces résultats sont corroborés par divers travaux dont celui de Bendjoudi et Hubert, [9].

Ces événements extrêmes (donc rares) font souvent l'objet d'études indépendantes (en chaque site et pour chaque pas de temps), alors qu'une modélisation unifiée permettrait d'intégrer plus d'information et d'extraire des caractéristiques globales qui réduiraient le nombre de paramètres de modélisation.

En fait, de récentes études montrent qu'en hydrologie (tout comme dans de nombreux domaines géophysiques d'ailleurs), les séries issues de divers régimes climatiques ou collectées selon des pas des temps différents présentent des caractéristiques invariantes. Sous un angle purement géométrique, la forme des nuages de pluie se révèle de nature fractale (Lovejoy, 83 [57]). Hubert et Carbonnel, 89 [45] ou Olsson et al., 92 [68] analysent l'occurrence temporelle de pluie et concluent à une structure mono-fractale de cette dernière, résultat qui peut être étendu à l'occurrence spatiale (Hubert et Carbonnel, 88 [44]). Mais des propriétés multifractales sont aussi décelées sur des données radar (Schertzer and Lovejoy, 87 [75]), ainsi que sur des données satellites (Lovejoy and Schertzer, 91 [58]). Plusieurs travaux concluent aussi à une invariance du paramètre de décroissance algébrique des queues de distribution sur diverses séries temporelles de cumuls pluviométriques à pas de temps

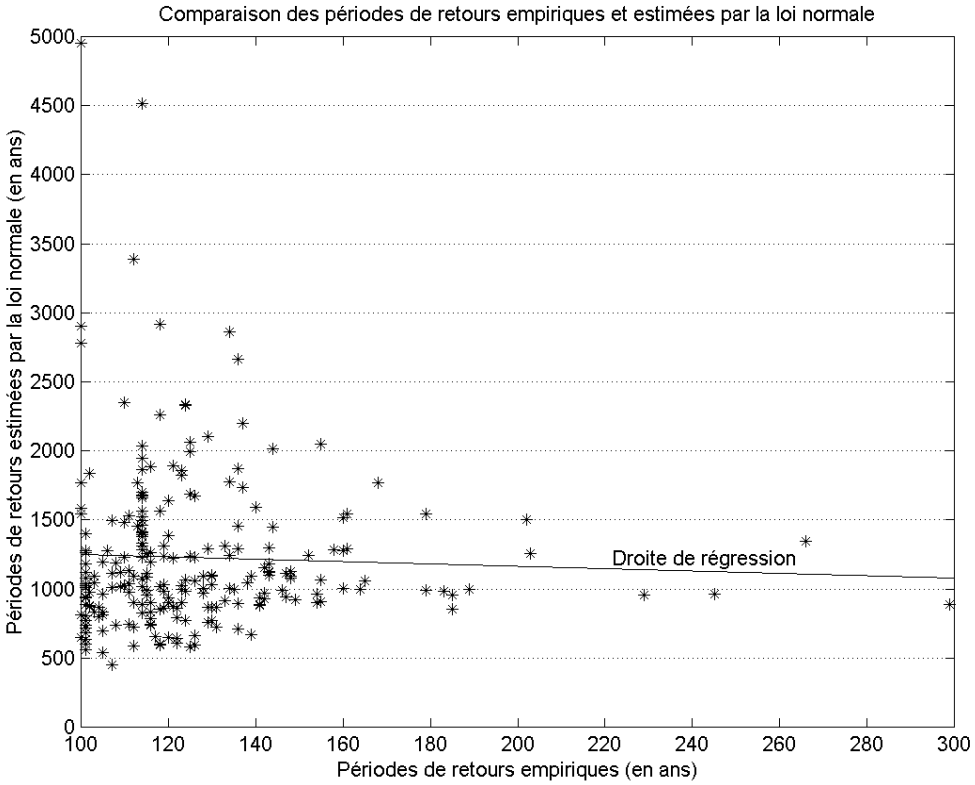



Figure 1.2: Sur-estimation des périodes de retour estimées par la loi normale.

différents (de Lima, 98 [25]).

## 1.2 Objectifs de la thèse

L'objectif de ce travail est de proposer un modèle stochastique de pluie qui intègre les connaissances physiques dont nous disposons a priori sur ce phénomène. En particulier, le modèle devra être multi-échelle et réagir le plus pertinemment possible par rapport aux artefacts de mesure précités. Les modèles en cascades multifractales constituent un bon moyen d'accéder directement aux caractéristiques constatées, tout en restant proche de la réalité physique du phénomène pluvieux. Cependant, l'estimation et la prévision au sein de séries multifractales est très délicate en raison de leur grande variabilité et de leur forme de dépendance.

L'originalité de cette thèse consiste à développer une nouvelle méthode statistique d'estimation de l'indice de décroissance algébrique, paramètre entrant en jeu dans des modèles en cascades multifractales, et à effectuer une classification des formes de dépendance théoriques de tels modèles.

Cette thèse vise plus précisément les objectifs suivants :

- Pour palier les problèmes liés à la mesure et à l'irrégularité des séries, il est nécessaire d'accentuer le travail d'élaboration de méthodes d'exploration et d'estimation statistique. Un outil statistique adapté à de telles séries est donc développé.
- La notion de longue dépendance est de plus en plus abordée dans la littérature hydrologique. La mise en relation de cette propriété statistique avec les modèles en cascades constitue la second objectif de ce travail. La détermination des modèles de cascades qui possèdent ou non cette propriété permet, d'une part d'approfondir l'état des connaissances de tels modèles et d'autre part d'affiner l'estimation des paramètres du modèle (grâce à des outils statistiques appropriés existant dans la littérature).
- Ce modèle, muni de ces outils statistiques, devra être confronté à des données hydrologiques réelles (pluies et débits) issues de divers sites et à divers pas de temps : les résultats obtenus seront comparés à ceux recensés dans la littérature.
- Enfin, ce travail dressera une liste de recommandations aux ingénieurs pour aboutir au développement de modèles hydrologiques opérationnels.

## 1.3 Démarche proposée

Ce travail s'organise en deux parties.

Dans la première partie, on commence par rappeler les aspects météorologiques des précipitations et les limites liées à la mesure des variables hydrologiques. Puis, après une description de l'état de l'art en matière de méthodes statistiques d'estimation spatiale, une revue des principaux modèles stochastiques utilisés en hydrologie est dressée. Enfin les fractals et les modèles en cascades multifractales sont présentés sous leur forme générale ainsi qu'une revue des principaux auteurs les ayant appliquées en hydrologie. Une

structure fractale de l'occurrence de pluie est décelée sur des séries Sahéliennes (données BADINAGE).

La deuxième partie est consacrée aux lois de type algébrique (qui apparaissent dans les modèles en cascades multifractales). Dans un premier temps, elles sont définies puis on présente le cadre probabiliste dans lequel elles se développent (le domaine d'attraction de la loi de Fréchet). Une revue bibliographique des divers outils statistiques adaptés aux extrêmes est dressée. Un outil statistique original d'exploration et d'estimation du paramètre de décroissance algébrique est élaboré. Il est confronté aux autres outils sur des simulations, puis appliqué à diverses séries de pluies (pas de temps différents et 232 séries annuelles à travers le monde). Cette sous-partie descriptive conclut à :

- Un comportement algébrique des séries des extrêmes pluviométriques
- Une invariance spatiale et temporelle du paramètre de décroissance algébrique (données du projet FRIEND-AMY dans le cadre de l'UNESCO et longues séries de cumuls de pluie à pas de temps allant de la journée à la minute).
- L'exploration de la persistance de ce comportement à faible pas de temps est effectuée à partir de séries à basculement d'auget (données EPSAT-Niger et série de l'Ile de la Réunion), suite à un travail préliminaire d'adaptation de l'outil aux séries de fréquence de basculement.

La troisième partie de ce travail analyse la forme de dépendance des séries hydrologiques. Les cascades multifractales sont des modèles simples qui retranscrivent l'invariance d'échelle du paramètre de décroissance algébrique constatée dans la deuxième partie. Mais rendent-ils compte d'autres propriétés statistiques constatées en pratique ? La propriété de longue dépendance est une autre caractéristique couramment abordée dans la littérature hydrologique. Il est donc intéressant d'analyser, sur un plan théorique, le type de dépendance (à court terme et à long terme) des cascades multifractales. Quelques cascades particulières sont étudiées plus en détail (log-Gamma, log-Normale ou log-Poisson). Des outils performants recensés dans la littérature statistique sont appliqués à nos séries pour l'estimation du paramètre de longue dépendance. Ils concluent à la présence de longue dépendance sur les séries de rivières sur craie.

Enfin, dans la partie "Conclusion", une synthèse des résultats est présentée, ainsi que leurs limites d'interprétation.

# Chapitre 2

# Champs de pluie et modélisation stochastique

Parmi les principaux objectifs de l'hydrologie figurent la génération des données synthétiques (par exemple pour tester la résistance d'un ouvrage face à divers scenarii) et la prévision de données futures (pluie ou débit). S'il veut exploiter le plus largement possible l'information dont il dispose et ne pas se contenter d'une analyse descriptive des données, l'hydrologue doit au préalable élaborer un modèle. En général, un modèle comporte deux volets : déterministe ou stochastique. Parfois le volet déterministe est absent et le modèle est purement stochastique, d'autres fois le volet stochastique est absent et le modèle est purement déterministe (ou mécaniste). Mais ces deux volets peuvent aussi se compléter : dans le modèle linéaire par exemple le volet stochastique est le réceptacle de l'erreur.

Les modèles purement stochastiques (ou boîtes noires) ne sont d'aucune utilité pour la modélisation des champs de pluie où l'on possède des connaissances a priori sur la causalité du phénomène (voir par exemple la hiérarchie de structures établie par Austin et Houze, 72 [3]). Les modèles purement déterministes quant à eux, sont surtout développés en météorologie et tentent de reproduire la dynamique des champs de pluie. Comme les équations exactes régissant le phénomène sont inconnues, ces modèles reposent sur des équations aux dérivées partielles de Navier-Stokes qui sont tronquées et approximées pour décrire la dynamique de l'atmosphère, puis intégrées numériquement (sous hypothèse d'homogénéité à chaque échelle). Malgré ces simplifications bien souvent éloignées de la réalité, les équations résultant de ces modèles restent complexes et les calculs numériques très lourds. De plus, les échelles sont étudiées indépendamment les unes des autres.

Les modèles à la fois stochastiques et déterministes rencontrés dans la littérature sont habituellement temporels et construits autour d'une seule échelle (de type série chronologique). Or les liens connus entre réalisation du phénomène à différentes échelles rendent inadéquats ce genre de modèles. Les modèles spatiaux et spatio-temporels intègrent plusieurs échelles en se basant généralement sur la hiérarchie des structures de champs de pluie de Austin et Houze [3]. Mais il n'est physiquement pas justifié qu'on doive privilégier certaines échelles dites caractéristiques. Dans un effort d'intégration du maximum d'information, le modélisateur se devra de mettre en jeu une gamme d'échelle aussi large que possible.

Dans cette partie, on présente brièvement l'état des connaissances que l'on possède a priori sur le phénomène pluvieux, les difficultés à le mesurer ainsi que des outils stochastiques pertinents pour la modélisation de ce dernier. Enfin , on présentera les modèles fractals et multifractals (Mandelbrot, 75 [59], Schertzer et Lovejoy, 93 [76]) qui ont connu un développement récent. Ils offrent un cadre formel prenant directement en compte une forme d'invariance d'échelle du phénomène, permettant ainsi une réduction du nombre de paramètres et une extrapolation de l'information entre les échelles.

## 2.1 Les champs de pluie

Avant de décrire les procédés de mesure de pluie, nous allons brièvement rappeler l'état des connaissances sur la structure des champs de pluie et les gammes d'échelle sur lesquelles il est possible de mesurer ce phénomène. Un tableau fournira ensuite les échelles pour lesquelles ces techniques de mesure sont susceptibles de conduire à des estimations fiables. Nous terminerons par quelques remarques générales sur la pertinence des hypothèses statistiques de continuité et de stationnarité des séries pluvieuses.

### 2.1.1 Mécanismes générateurs des champs de pluie

#### Les principaux types de précipitations

On distingue principalement trois types de précipitations qui peuvent survenir ensemble et se compléter :

- Précipitations liées au passage d'une perturbation frontale : Elles sont dues aux rencontres de masses d'air de températures différentes. La trace au sol de la frontière entre ces deux masses d'air est appelée front. Les fronts chauds (provoqués par un déplacement d'air chaud au dessus d'air froid dans la même direction) occasionnent des pluies de faible intensité et de longue durée. Les fronts froids (provoqués par l'avancée d'une masse d'air froid qui soulève de l'air chaud) sont marqués par de forts pics d'intensité.

- Précipitations liées à une convection locale : Un réchauffement des basses couches de l'atmosphère crée un équilibre thermique instable. L'ascendance de l'air chaud donne naissance à des cellules convectives (zone intertropicale et zone tempérée en été).

- Précipitations orographiques : Elle sont dues à la présence d'une barrière montagneuse qui, sous l'effet du vent, provoque l'élévation de masses d'air humide. Il en résulte un refroidissement de ces masses d'air. Il peut alors y avoir un équilibre thermique instable (instabilité convective) ou bien une instabilité en air saturé.

Signalons dès maintenant que selon le type de précipitation, les hypothèses d'un modèle de pluie seront différentes : l'hypothèse d'indépendance des épisodes pluvieux successifs semble plus plausible dans le cas de précipitations convectives que dans le cas de précipitations frontales.

| Echelle | Spatiale | Temporelle |
|---|---|---|
| synoptique | >10 000 $km^2$ | Plusieurs jours |
| Grande méso-échelle | 1 000 à 10 000 $km^2$ | Plusieurs heures |
| Petite méso-échelle | 100 à 400 $km^2$ | Environ une heure |
| Cellule de pluie | Environ 10 $km^2$ | 1 min à 1/2 heure |

Tableau 2.1: Organisation des champs de pluie

**Hiérarchie des structures de pluie**

A partir d'un suivi radar d'épisode en Nouvelle Angleterre, Austin et Houze ([3]) établissent une hiérarchie des champs précipitants en situation extra-tropicale (tableau 2.1). Ils distinguent des structures imbriquées les unes aux autres. Chaque structure est intrinsèquement liée à une échelle spatio-temporelle (comportant quatre niveaux) et possède des caractéristiques propres (durée de vie de la structure, intensité, taille).

Les deux dernières structures concernent les orages violents (par leur forte intensité qui peut atteindre jusqu'à 100 mm/h). Les cellules de pluie apparaissent isolées ou bien de façon organisée à l'intérieur d'un front.

La représentation météorologique de Austin et Houze permet donc, malgré un chaos apparent, de mettre en évidence une organisation des champs de pluie. L'inconvénient de cette analyse est qu'elle déconnecte les échelles les unes des autres, faisant apparaître artificiellement un carcan d'échelles caractéristiques. Or la pluie est un phénomène qui se développe sur une gamme d'échelles très étendue et ses grandeurs n'ont pas d'échelle caractéristique : de 1 $km^2$à 10 000 $km^2$en espace (averse, front pluvieux, ...), et de quelques minutes aux échelles géologiques en temps (voir tableau 2.1). Il n'est donc pas envisageable de privilégier une échelle plutôt qu'une autre, et tout effort de modélisation se doit d'établir une description unifiée des champs de pluie.

### 2.1.2 Mesure de la pluie

Les données de précipitations ont longtemps été produites par des mesures directes (pluviographiques ou pluviométriques). L'apparition récente des mesures indirectes (radars météorologiques et systèmes d'observation satellites) a permis de décrire la variabilité spatiale de la pluie qui jusque là n'avait été accessible que par des réseaux denses de pluviomètres ou de pluviographes.

- Mesure pluviométrique : Ce sont les mesures le plus couramment utilisées. L'eau de pluie est recueillie sur une surface de quelques $cm^2$et accumulée dans un réservoir. Son volume sur un pas de temps fixe (en général 24 heures, mais parfois de l'ordre de la saison) est quantifié par divers procédés (mesure de la hauteur d'eau dans le réservoir, mais aussi pesée du réservoir). La résolution temporelle est donc égale au pas de temps de mesure et la résolution spatiale est de quelques $cm^2$, mais ces mesures sont assimilées à des mesures ponctuelles en pratique. Les erreurs sont principalement

dues au vent, à des effets d'écran, à l'évaporation et parfois liées à l'obstruction ou au débordement de l'appareil.

- Mesure pluviographique : C'est une mesure un peu moins simple que la précédente. Le pluviographe classique enregistre continûment les cumuls de pluie dans le temps, mais il existe aussi des pluviographes à dépouillement manuel. Le pluviographe à basculement enregistre des dates de basculement, chaque basculement correspondant au remplissage d'un auget d'une contenance de 0.1 mm ou de 0.5 mm. La résolution spatiale est là aussi considérée comme ponctuelle, tandis que la résolution temporelle varie de 1 à 15 minutes. Les sources d'erreurs sont identiques aux précédentes mais elles peuvent aussi être dues à la saturation de l'auget qui impose une borne supérieure aux mesures.

- Mesure par spectro-pluviomètre optique : Son principe est basé sur l'atténuation produite par les gouttes de pluie lors de leur passage dans un faisceau lumineux. Le disdromètre, ou compteur de gouttes, donne aussi accès à la granulométrie des gouttes par des procédés optiques ou acoustiques

- Mesures radar : Le radar est un dispositif de mesure actif : Il émet un rayonnement incident pour mesurer la réflectivité radar $Z$. Celle-ci est ensuite convertie en intensité de pluie $R$ par la relation dite de Marshall Palmer $Z = A \times R^b$, où $A$ et $b$ sont des coefficients calculés par une calibration utilisant des données au sol issues d'un réseau de pluviomètres. Cette relation est fondée sur des hypothèses très fortes telles qu'une répartition homogène des gouttes, une forme de goutte sphérique et une intensité de pluie constante (Andrieux 86 [2], Blanchet 93 [12], Bourel 94 [14]).

  Plusieurs erreurs découlent de la relation $R-Z$. La relation de Marshall Palmer est non linéaire, ce qui implique que les coefficients de calage $A$ et $b$ varient par changement d'échelle. On recense dans la littérature plus d'une centaine de couples $(A, b)$.

  Le phénomène dit de bande brillante constitue aussi une limite à la validité de la relation de Marshall Palmer. Il se produit au dessus de l'isotherme 0°, et est dû au fait que la réflectivité des particules de glace est beaucoup plus grande que celle des particules d'eau.

  Ajoutons d'autres sources d'erreurs qui peuvent être dues à la mesure de réflectivité (présence d'obstacles, de masques, dépôts sur l'antenne...), ou aux écarts entre la pluie au sol et la pluie en altitude.

  Malgré la multiplicité des sources d'erreurs, le radar permet néanmoins d'établir une cartographie "qualitative" de la pluie. Il tend à devenir de plus en plus opérationnel pour la mesure de la pluie.

- Mesures par satellite : Les satellites sont généralement des dispositifs passifs. Ils captent des radiations provenant de la terre essentiellement selon deux gammes d'onde :

  · L'infrarouge : L'analyse d'une image IR permet de distinguer les pixels froids des autres. Leur nombre peut être relié à une quantité d'eau précipitée en se basant sur le fait que les cellules convectives intenses sont plus froides que les

| | Résolution temporelle | Résolution spatiale | Phénomènes observés |
|---|---|---|---|
| SPOT P | entre 2.5 et 26 jours | 10 m | Statiques et localisés |
| IRS-1C | entre 5 et 24 jours | 5.8 m | Statiques et localisés |
| Landsat TM | 16 jours | 30 m | Statiques et localisés |
| NOAA | 4 fois par jour | 1 km | Innondations, sécheresses |

Tableau 2.2: Résolution temporelle et spatiale de quelques satellites.

sommets des nuages avoisinant (à haute altitude). Dans le cas de cellules convectives relativement isolées, ces relations sont bien explicitées. Par contre, dans le cas de nuages stratiformes ou de forçage orographique, ces relations sont quasi-inexistantes (Bourel 94 [14]). Des résultats intéressant sont obtenus par exemple au Sahel par le satellite Météosat[1]. La résolution spatiale s'étend de 5 à 10 km de coté et la résolution temporelle est de l'ordre de la demi-heure.

- Les micro-ondes : C'est un système actif puisqu'il consiste à embarquer un radar sur un satellite à orbite polaire. Ce système sert à sonder les nuages. La résolution spatiale est de l'ordre de 20 km et la résolution temporelle est de 12 heures.
- Citons aussi le visible qui sert à la classification des nuages.

Les résolutions en espace et en temps des satellites sont cependant trop grandes pour les besoins de l'hydrologie : la fréquence minimale d'acquisition d'images reste supérieure à 30 minutes tandis que la résolution spatiale avoisine au mieux les 5 km de coté (voir tableau 2.2). Mais cette technique de mesure est en constante amélioration, puisque les satellites de la nouvelle génération atteindront une taille de pixel de l'ordre de quelques mètres (le satellite SPOT 5 en 2002 prévoit une taille de pixel de 3 mètres). D'autre part, les passages des satellites sont peu nombreux et s'effectuent toujours à la même heure.

### 2.1.3 Echelles spatio-temporelles

Toutes les échelles spatio-temporelles interviennent dans le processus des précipitations. Par exemple, la constitution d'un modèle météorologique (pour la prévision d'un débit de pointe par exemple) associe plusieurs niveaux d'échelle : de $10^4$ à $10^2$ km$^2$pour la représentation de la cyclogénèse, et les mesures à des échelles inférieures pour la désagrégation à l'échelle du bassin versant. Mais les résultats obtenus à une certaine échelle pourraient être faussés par des mécanismes d'échelle inférieure, ou s'extrapoleraient mal pour expliquer des résultats à une échelle inférieure. Selon quel pas de temps et quel pas d'échelle spatiale est-il alors raisonnable de calculer des estimations de lames d'eau ? Ces gammes d'échelles concernent-elles les besoins pratiques de l'hydrologie ?

[1]Au Sahel, l'essentiel des précipitations provient des cumulo-nimbus dont le sommet est froid (environ -40°). Cependant, il existe des nuages à sommet froid non précipitant et tous les cumulo-nimbus ne précipitent pas. Il est donc nécessaire d'effectuer un callage par des mesures au sol.

| Bassin versant | Résolution spatiale | Echelle globale | Résolution en temps |
|---|---|---|---|
| Urbain | ha à 1 $km^2$ | $km^2$ | 5 à 30 min |
| Petie | $km^2$ | 10 à 100 $km^2$ | 30 min à 1 h |
| Moyen | du $km^2$ à la dizaine de $km^2$ | 1000 $km^2$ | 24 heures |

Tableau 2.3: Ordre de grandeur des tailles et de la résolution des estimations de lames d'eau de chaque type de bassin versant.

La réponse dépend du moyen de mesure (dans le cas de mesures issues d'un réseau de pluviomètres ou de pluviographes, elle dépend de la densité) mais surtout de la taille et de la nature du bassin versant.

Faure, 93 [33] considère trois types de bassins versants :

- Les bassins versants urbains ;
- Les petits bassins ruraux (par exemple montagneux) ;
- Les grands bassins versants.

On a synthétisé dans le tableau 2.3 l'échelle globale d'espace caractérisant chacun des bassins versants et les deux résolutions (en espace et en temps) nécessaires pour l'utilisation en pratique des sorties du modèle. La résolution temporelle correspond à la durée de la réponse du bassin versant.

### 2.1.4 Conclusion

Dans ce chapitre, nous avons dressé une synthèse de l'état actuel des connaissances sur la dynamique des champs de pluie, puis défini les contraintes liées à la mesure de la pluie et à son utilisation et réalisé une synthèse des sources d'erreurs de mesures ainsi que des ordres de grandeur d'échelles d'espace et de temps concernées.

L'état actuel des connaissances permet d'avoir une représentation globale de la dynamique des champs de pluie et des estimations de lame d'eau plus ou moins fiables selon les dispositifs de mesure. Toutefois, cette représentation météorologique et ces dispositifs de mesure fractionne le processus en un nombre fini d'échelles qui ne correspondent pas forcément aux besoins hydrologiques, et rendent difficile l'extrapolation des estimations à d'autres échelles. L'une des ambitions de ce travail est de les traiter simultanément en se fondant sur une continuité du processus de pluie entre échelles.

## 2.2 La modélisation stochastique de la pluie

Dans la partie précédente, nous avons rapidement présenté les difficultés soulevées par la modélisation physique de la pluie. Bien que les phénomènes entrant en jeu ne soient pas forcément régis par le hasard, une manière de décrire les processus hydrologiques consiste à construire un modèle de pluie stochastique. Selon E. Halphen : "Les probabilités mesurent l'ignorance humaine et la statistique est ce qui tient lieu de science aux ignorants que nous sommes."

### 2.2.1 Pourquoi une modélisation stochastique ?

L'approche stochastique de modélisation est susceptible a priori de présenter divers avantages en hydrologie :

- Tout d'abord, elle permettrait la réalisation de transferts d'information en des sites géographiques différents, ou bien à des échelles plus petites (spatiales, temporelles ou spatio-temporelles), atout non négligeable en raison du prix et de la difficulté d'acquisition des mesures.
- Face à l'extension des zones urbaines et à l'augmentation de la vulnérabilité, l'hydrologue a besoin d'estimations à une résolution spatiale et temporelle de plus en plus fine et de prévisions de plus en plus fiables. Les modèles stochastiques réalisent des estimations dont la précision dépend de la variabilité de la série, tout en quantifiant l'erreur, qui peut être retranscrite en terme de risque (Berger, 85 [11]).
- Un autre avantage de l'approche stochastique est le caractère opérationnel du modèle pour des applications hydrologiques. En effet, on modélise toujours par rapport à un objectif, et le modèle doit être bien distingué du phénomène à modéliser (de Marsily, 94 [26]). L'hydrologue s'intéresse à la prévision des crues et des étiages, à la gestion des réservoirs ou à l'estimation des extrêmes pour le dimensionnement d'ouvrages. Une prévision de la pluie pour l'annonce des crues en milieu urbain requiert une modélisation à petit pas de temps et d'espace, tandis que la détermination de la hauteur d'une digue sera réalisée à partir d'un modèle à l'échelle du bassin versant.

Mais modéliser des grandeurs hydrologiques en utilisant des modèles stochastiques n'est généralement pas une tâche facile. En fait, les variables hydrologiques possèdent une structure de dépendance et une variabilité des moins classiques sur le plan statistique. Les séries de débits ou de pluie sont fortement asymétriques (notamment à petit pas de temps, voir figure 2.2) et possèdent généralement une fonction d'autocorrélation convergeant lentement vers zéro.

Parmi les principaux modèles stochastiques proposés en hydrologie, on distingue les modèles de type gaussien et de type non-gaussien, au sein desquels apparaissent diverses structures de dépendance statistique (ARMA, Markov) qui possèdent chacune une interprétation physique simple.

Dans cette section, nous rappelons les principales méthodes d'estimation spatiale, avant de dresser une revue des différents modèles stochastiques de précipitation rencontrés dans la littérature : nous présenterons les modèles spatiaux, temporels, puis les modèles qui appréhendent la pluie en tant que champ spatio-temporel.

### 2.2.2 Méthodes statistiques d'estimation spatiale

Dans ce paragraphe, nous passerons en revue les méthodes de pondération, du kriegeage et du variogramme qui visent le plus souvent à obtenir des informations sur la variabilité spatiale de la pluie, mais peuvent aussi être employées pour la variabilité temporelle ou spatio-temporelle.

Figure 2.1: Asymétrie des cumuls en fonction du pas d'intégration (série de Athènes).

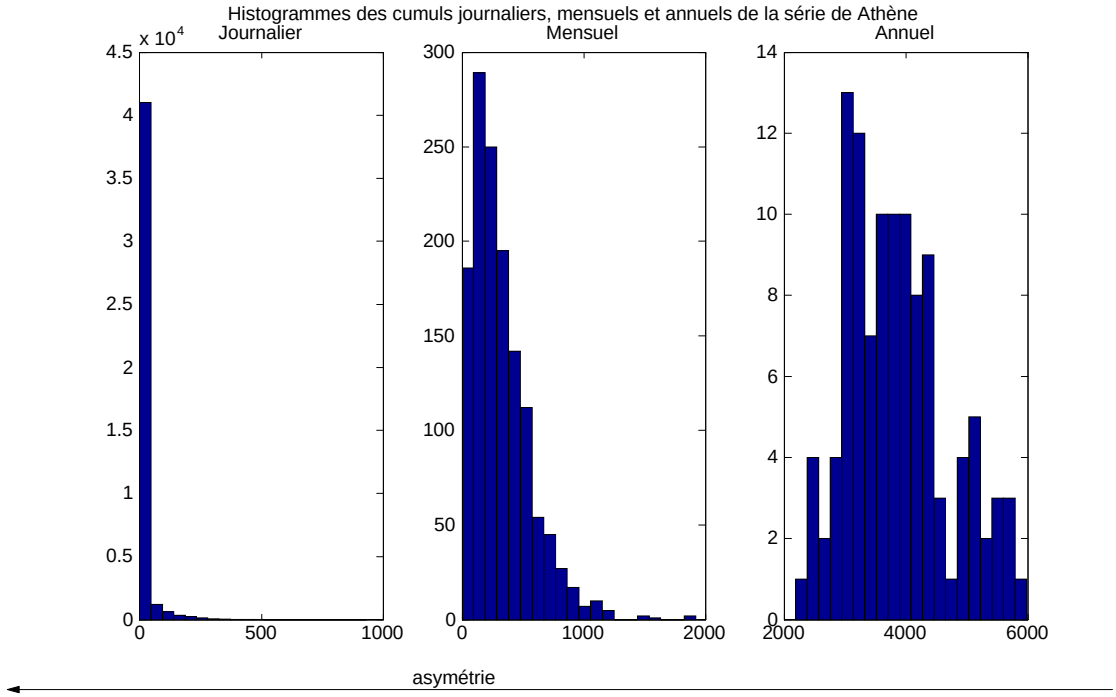



Figure 2.2: Asymétrie de la distribution des cumuls pluivométriques.

La démarche générale de ces méthode est la suivante : On aimerait déterminer, à partir d'un réseau dense de pluviomètres ou pluviographes, la hauteur de pluie tombée en tout point d'une région, ou bien en moyenne sur un domaine de cette région. On dispose de $n$ mesures $(z_i)_{i=1,\cdots,n}$ prises aux pluviomètres situés aux points $(x_i)_{i=1,\cdots,n}$ et on veut estimer :

- ponctuellement la mesure $z$ en un point $x$

$$z(x) = f(z_1, z_2, ..., z_n) = f(z_1(x_1), z_2(x_2), ..., z_n(x_n))$$

- ou en moyenne la lame d'eau $Z$ sur un domaine $X$

$$Z(X) = F(z_1, z_2, ..., z_n) = F(z_1(x_1), z_2(x_2), ..., z_n(x_n))$$

**2.2.2.a Méthodes de pondération :** Les méthodes de la moyenne pondérée, des polygones de Thiessen et des isohyètes, sont des méthodes déterministes de pondération. Elles sont fondées sur l'hypothèse de continuité du champ de pluie (en temps ou en espace) et peuvent servir à une estimation spatiale ou temporelle d'une hauteur, d'un cumul ou d'une intensité de pluie. Pour un pas de temps donné, un champ pluviométrique est *continu* si la mesure en un point est représentative de la lame d'eau précipitée sur une surface assez grande.

Remarquons que ces méthodes peuvent cependant aussi être considérées comme aléatoires, l'aléa portant sur les hauteurs $z_i$ ou sur la répartition des postes $x_i$.

**Méthode de la moyenne pondérée :** On estime la variable aléatoire au point $x$ à partir d'une moyenne pondérée sur les postes voisins $(x_i)_{i=1...n}$

$$\widehat{z} = f(z_1, ..., z_n) = \frac{\sum_{i=1}^{n} w(d_{xx_i}) \times z_i}{\sum_{i=1}^{n} w(d_{xx_i})}$$

avec

- $\widehat{z}$ la hauteur estimée au point $x$
- $d_{xx_i}$ la distance de $x$ à $x_i$.
- $w$ une fonction de poids (par exemple $w(d) = \frac{1}{d}$ ) est décroissante en fonction de la distance. Le poids affecté à chaque point $x_i$ s'écrit

$$\frac{w(d_{xx_i})}{\sum_{i=1}^{n} w(d_{xx_i})}$$

**Méthode des polygones de Thiessen :** Cette méthode, proposée par Thiessen en 1911 [82], est une généralisation de la précédente lorsque $w(d) = \frac{1}{d^k}$ et $k \to \infty$. Elle permet de réaliser une *estimation sur un domaine*. Les poids sont définis comme suit : on affecte un polygone d'influence à chaque observation $z_i$ d'un bassin versant ; le polygone d'influence est un domaine défini par toutes les médiatrices des segments joignant les couples de points autour de chaque observation. Le poids affecté à chaque observation est le rapport de la surface du polygone d'influence sur la surface du bassin $\frac{S_i}{S}$ :

$$\widehat{Z} = \sum_{i=1}^{n} \frac{S_i}{S} \times z_i$$

**Méthode des isohyètes :** Pour chaque couple d'observations, on découpe le segment les joignant en intervalles réguliers et on relie ensuite les valeurs identiques pour obtenir les courbes isohyètes $(H_i)_{i=1,\cdots,n}$ correspondant à des hauteurs $(z_i)_{i=1,\cdots,n}$ . En notant $S_i$ l'aire délimitée par les isohyètes $H_i$ et $H_{i+1}$, on estime la lame d'eau par

$$\widehat{Z} = \sum_{i=1}^{n} \frac{S_i}{S} \times \frac{z_i + z_{i+1}}{2}$$

**2.2.2.b Méthodes d'estimation optimales :** Les méthodes du kriegeage et du variogramme sont toutes deux des méthodes d'estimation optimales, et ont été introduites il y a 30 ans dans des domaines aussi variés que l'industrie minière par Matheron, 65 [64], ou la météorologie par Gandin, 65 [38]. On n'en fait ici qu'une brève présentation. Pour les détails des calculs, on se référera à Delhomme, 78 [27].

On considère qu'un champ de pluie correspond à la réalisation d'une fonction aléatoire (*FA*) notée $Z$, à valeurs $Z(x_i)$, et on détermine l'estimateur $\widehat{Z}$ de la hauteur $Z$ en $x$ qui minimise l'erreur quadratique d'estimation :

$$\min_i E\left\{\left(\widehat{Z(x)} - Z(x_i)\right)^2\right\}$$

**Méthode du kriegeage (simple) :** Elle repose sur l'hypothèse de stationnarité d'ordre un et deux de la *FA* $Z$ dans l'espace :

- L'espérance est constante

$$\forall x \quad E(Z(x)) = m(x) = m$$

- La covariance entre deux points $x_i$ et $x_j$ de ce champ ne dépend pas séparément des deux points mais seulement de leur distance $|x_i - x_j|$ :

$$cov\left[Z(x_i), Z(x_j)\right] = E\left[Z(x_i) \cdot Z(x_j)\right] - EZ(x_i) \cdot EZ(x_j) = K(|x_i - x_j|)$$

On détermine l'estimateur $\widehat{Z}$ de la hauteur $Z$ en $x$ qui minimise l'erreur quadratique d'estimation $\min E\left\{\left(\widehat{Z}-Z\right)^2\right\}$ avec une contrainte de non biais $E(\widehat{Z}-Z)=0$, et on recherche un estimateur de la forme

$$\widehat{Z}(x) = \sum_{i=1}^{n} \alpha_i Z(x_i)$$

On obtient un système sous contraintes permettant de calculer les coefficients $(\alpha_i)_{i=1,\cdots,n}$ qui est résolu par la méthode des multiplicateurs de Lagrange en se ramenant à une équation matricielle :

$$\begin{bmatrix} C & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \Lambda \\ \mu \end{bmatrix} = \begin{bmatrix} C_0 \\ 1 \end{bmatrix}$$

où $C$ et $C_0$ représentent respectivement la matrice de covariance des postes de mesures (dimension $n \times n$) et le vecteur des covariances entre le point à interpoler $x$ et les postes $x_i$ (dimension $n$). On estime $C$ et $C_0$ à partir des réalisations de $Z$ en ajustant un modèle $C(d)$ sur les $\dfrac{n(n-1)}{2}$ valeurs mesurées de la covariance.

L'inconvénient majeur de la méthode du kriegeage est qu'elle est fondée sur une hypothèse trop contraignante. Il existe cependant d'autres versions de cette méthode qui s'adaptent mieux aux variables hydrologiques (le krigeage lagrangien de Amani et Lebel, 96 [1] par exemple).

**Méthode du variogramme :** Le variogramme est un outil pour quantifier la variabilité (spatiale, temporelle ou spatio-temporelle) d'un champ.

Il remplace la fonction de covariance dans le cas de processus non stationnaires car on se place sous l'hypothèse intrinsèque (Matheron, 65 [64]). Cette dernière est plus faible, car c'est une hypothèse de stationnarité d'ordre un et deux qui ne porte que sur les accroissements $Z$ :

Pour tous points $x_i$ et $x_j$ distants de $|h|$ :

$$\begin{aligned} E\left(Z(x_i)-Z(x_j)\right) &= m(|h|) \\ Var\left(Z(x_i)-Z(x_j)\right) &= 2\gamma\left(|h|\right) \end{aligned}$$

La variance ne dépend que de la distance $|h|$.

**Définition 1** *Le variogramme $\gamma(|h|)$ de l'événement est défini par*

$$\gamma\left(|h|\right) = \frac{1}{2} Var\left(Z(x_i)-Z(x_j)\right)$$

Sous l'hypothèse $m(|h|)=0$, il s'écrit encore $\gamma\left(|h|\right) = \frac{1}{2}E\left[\left(Z(x_i)-Z(x_j)\right)^2\right]$. Le variogramme de $h$ est l'accroissement quadratique moyen entre deux points distants de $|h|$.

On minimise l'erreur quadratique sous contrainte de non biais et sous l'hypothèse intrinsèque. On obtient par la méthode des multiplicateurs de Lagrange un système linéaire

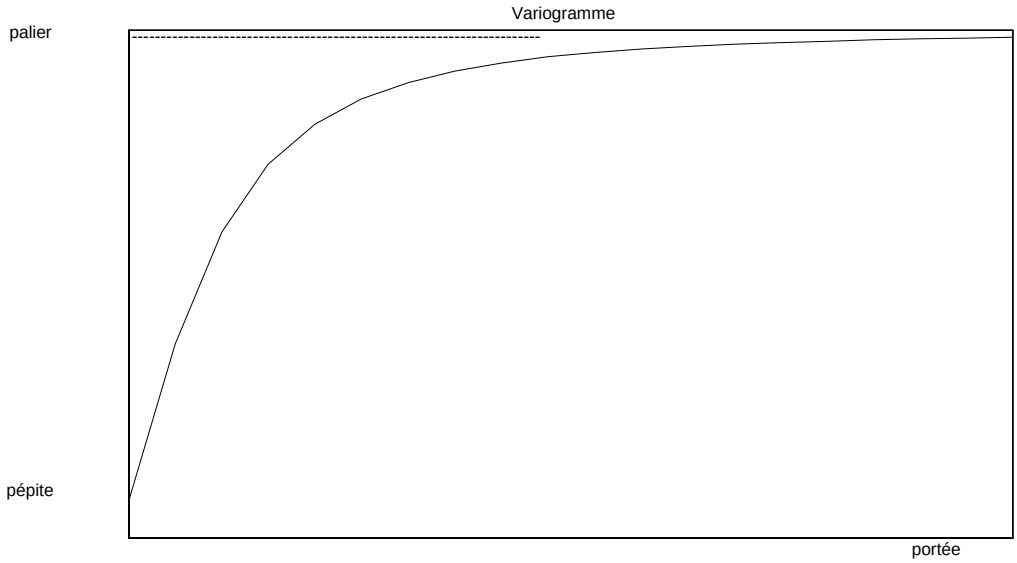



Figure 2.3: Pépite, portée et palier d'un variogramme.

permettant d'estimer les coefficients $(\alpha_i)_{i=1..n}$ :

$$\begin{bmatrix} \gamma_{11} & \cdots & \gamma_{1m} & 1 \\ . & . & . & . \\ \gamma_{m1} & \cdots & \gamma_{mm} & 1 \\ 1 & ... & 1 & 0 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ . \\ \alpha_n \\ \mu \end{bmatrix} = \begin{bmatrix} \gamma_{10} \\ \\ \gamma_{m0} \\ 1 \end{bmatrix}$$

avec $\mu$ coefficient de Lagrange et $\gamma_{ij} = \frac{1}{2}E\left[Z\left(x_i\right) - Z\left(x_j\right)\right]$

La résolution de cette équation nécessite le choix d'un variogramme.

D'un point de vue pratique, ce variogramme est estimé en :

- Regroupant les $\dfrac{n(n-1)}{2}$ couples de mesures en un nombre $N_C$ de classes d'interdistance.
- Affectant à chaque distance moyenne de classe $d_m$ la valeur estimée et on estime $\gamma(d_m)$ par la moyenne sur les classes des carrés des écarts.

  $\hat{\gamma}(d_m) = \dfrac{1}{2N_k}\sum\left(z(x) - z(x+h)\right)^2$

  avec $N_k$ le nombre de couples présents dans la classe $k$.

- On dispose de $N_C$ valeurs de $\hat{\gamma}(d_m)$ correspondant à l'une des $N_C$ valeurs $d_m$. On ajuste un variogramme théorique (sphérique, exponentiel, gaussien, ...) au variogramme expérimental (figure 2.3). Les modèles couramment utilisés sont définis par trois paramètres :

- le palier, qui est égal à la variance du champ. Il n'apparaît pas si le phénomène est très variable.
- la portée, qui représente le seuil à partir duquel le palier est atteint.
- la pépite, qui renseigne sur la continuité de $Z$. Une discontinuité en zéro, qui peut être due à la discontinuité du milieu, à un problème d'échelle ou à des erreurs de mesures.

En conclusion, ces différents outils, conçus pour appréhender la variabilité spatiale d'un champ, reposent tous sur des hypothèse d'homogénéité (plus ou moins affaiblie) de ce dernier. Or on a constaté que ces hypothèses sont mises à mal lorsqu'il s'agit d'un champ pluviométrique, surtout à petit pas de temps (on pourra se reporter à l'article de Hubert et Carbonnel, 88 [44] pour une étude sur les données EPSAT-Niger).

### 2.2.3 Modèles temporels

On distingue les modèles d'occurrence de pluie des modèles de hauteur non nulles. Ces deux processus peuvent être soit modélisés séparément, puis recombinés, soit modélisés simultanément. On ne présentera ici que les modèles en un seul site.

#### Modèles d'occurrence de pluie

Pour modéliser l'occurrence de pluie en un point fixé, on considère la variable aléatoire discrétisée à un certain pas de temps (généralement la journée) $X_t = 0$ ou 1 selon qu'il pleuve ou non sur ce pas de temps (ou bien selon que le cumul de pluie dépasse ou non un certain seuil). Deux approches ont été considérées ces dernières années : la modélisation des périodes sèches et des périodes humides et les séries à temps discret.

La première approche consiste à s'intéresser aux séquences de jours (ou heures) de pluie consécutifs. On ajuste une loi à la durée des périodes humides et des périodes sèches. Cette voie a été largement exploitée par des auteurs tels que Green en 64 [41] ou Eagleson en 78 [29] qui utilisent respectivement une loi exponentielle et une loi de Weibull pour les durées des périodes sèches et humides. Ces modèles sont en fait des processus de renouvellement dans le sens où états secs et pluvieux alternent, aucune transition dans le même état n'étant possible. Un autre modèle de renouvellement créé par Galloy, Martin et Lebreton en 82 [37] utilise la loi binomiale négative pour les durées des périodes sèches et humides.

Cependant ces méthodes reposent sur l'hypothèse d'indépendance des périodes pluvieuses entre elles, rarement vérifiée à petit pas de temps (à pas de temps une heure, deux séquences humides peuvent correspondre au même événement pluvieux). De plus, la durée variable des événements mériterait que l'on conditionne les cumuls à cette durée (Wilks, 89 [86]).

La seconde approche utilise les séries à temps discret, série binaire de 0 ou de 1. L'occurrence de pluie peut être modélisée par :

- Une suite de variables aléatoires de Bernoulli. C'est le modèle le plus simple mais il requiert une indépendance complète. Il est mieux adapté aux grands pas de temps (jour et plus).

- Les chaînes de Markov : Elles sont très utilisées du fait de leur simplicité. Ce sont des chaînes à deux états $X_t = 0$ ou 1. Le précurseur de cette modélisation semble être Quetelet en 1852 (d'après Woolhiser, 91 [87]). Van-Thanh-Van Nguyen, 83 [84] détermine la loi de probabilité des cumuls journaliers appartenant à une période pluvieuse de $n$ jours, en décrivant l'occurrence de pluie par une chaîne de Markov stationnaire d'ordre 1. Mais selon le degré de persistance désiré, les chaînes utilisées peuvent être d'ordre supérieur à un et même non homogènes. Stern et Coe [78] représentent l'alternance des jours secs et pluvieux par une chaîne de Markov non homogène en se plaçant à l'échelle d'un jour. Ces modèles servent à calculer la loi des cumuls sur $n$ jours, la loi du maximum annuel des précipitations. Les théorèmes limites fournissent des lois asymptotiques pour les hauteurs extrêmes.

  Ces modèles ne décrivent pas les persistances à long terme (par exemple les activités cycloniques persistantes durant certaines saisons) ni les effets de classement à petit pas de temps (car, par exemple, un événement pluvieux dû à un événement pluvieux précédent a une vraisemblance plus forte).

- Les processus $DARMA(p, q)$, $p$ étant l'ordre de l'autorégression et $q$ l'ordre de la moyenne mobile créés par Buishand en 1978 [16] tiennent mieux compte des persistances à long terme que les chaînes de Markov d'ordre élevé mais la structure du modèle est difficile à interpréter physiquement.

**Modèles de hauteur non nulles**

L'approche la plus courante est de décrire la distribution des hauteur $Y(t)$ indépendamment de celle de l'occurrence $X(t)$. Plusieurs types de dépendance caractérisent ces modèles :

- Les $Y(t)$ sont indépendantes : Divers lois sont utilisées telles que la loi Exponentielle, Gamma ou Weibull (Zucchini et al., 92 [89]).

- Les $Y(t)$ sont dépendantes : On utilise des modèles ARMA pour modéliser une dépendance entre les $Y(t)$ non nuls. Ils sont surtout utilisés pour modéliser les débits mais peu pour les hauteurs de pluie à cause de la quasi indépendance.

- Les $Y(t)$ sont indépendantes mais dépendent de l'occurrence du jour précédent $X(t-1)$ : Ce sont les processus dépendants en chaîne, introduits par Katz en 1977 ([48] et [49]) qui conditionnent $Y(t)$ par $X(t-1)$. Ces modèles nécessitent un grand nombre de paramètres.

  D'autres expressions de la dépendance peuvent être trouvées dans la littérature, telles que celle de Buishand, 78 [16] qui classe les événements pluvieux selon les types suivants :

  · I : jour pluvieux isolé.
  · II : jour pluvieux précédé ou suivi d'un jour sec.
  · III : jour pluvieux précédé et suivi d'un jour sec.

**Modèles Mixtes**

On peut modéliser l'occurrence $X(t)$ et les hauteurs $Y(t)$ séparément ou bien simultanément $Z(t) = (X(t), Y(t))$. La plupart des modèles mixtes s'appuient sur le cadre mathématique construit par Le Cam en 1961 [17] qui adapte à l'hydrologie un processus d'agrégation de type Neyman-Scott. Ce dernier, créé en astronomie pour modéliser la formation d'étoiles, présente l'avantage de prendre en compte la dynamique des champs de pluie. De nombreux auteurs se sont attachés à analyser formellement la qualité des estimateurs (Cowpertwait, 91 [20]) ou à simuler des données à pas de temps fin, notamment en hydrologie urbaine (Thauvin et al. 97 [81]). Dans un modèle de type Neyman-Scott (figure 2.4), durées de temps sec et de temps pluvieux alternent. Les origines de périodes pluvieuses (notées '0'), arrivent selon un processus de Poisson. A l'intérieur d'un intervalle pluvieux, la pluie se décompose en une succession d'averses (représentées sur le schéma par le symbole '$*$'). Chaque averse est caractérisée par sa durée et son intensité. L'intensité totale en chaque point est la somme des intensités de chaque averse active.

Le processus de Neyman-Scott présente des variantes définies par la loi des intensités attribuées à chaque cellule de pluie. Citons le Neyman-Scott White Noise (NSWN) de Rodriguez-Iturbe et al. en 84 [74] ou le Neyman-Scott Rectangular Pulse (NSRP) de Rodriguez-Iturbe et al. en 87 [73]. Il fait partie d'une vaste classe de processus ponctuels présentés sur un plan théorique dans l'ouvrage de Cox et Isham [21] parmi lesquels on peut citer le processus de Poisson double (ou processus de Cox). Ce dernier est un processus de Poisson à intensité aléatoire. Kavvas et Delleur [50] appliquent le processus de Cox pour construire un modèle stochastique à deux niveaux, le RCM (Renewal Cox process with Markovian Intensity) pour l'occurrence de pluie dans lequel la vitesse d'occurrence des orages est déterminée par un autre processus appelé processus climatologique. Smith et Karr, 83 [77] exploitent ce modèle pour calculer le nombre d'orages et les durées d'inter-arrivées sur des données issues du bassin versant de la rivière Potomac.

Le processus de Neyman-Scott se généralise naturellement en un processus spatial ou spatio-temporel comme nous pourrons le constater dans la section suivante.

### 2.2.4 Modèles spatio-temporels

La littérature propose peu de modèles spatio-temporels. Ils ont été créés généralement en vue de simulations afin de tester des stratégies d'estimation à partir d'appareils de mesure de types différents. Les modèles présentés ici sont construits autour de structure de variabilité observées aux différentes échelles choisies selon les observations (cellules pluvieuses, bandes de cellules en espace, événement pluvieux en temps...).

Le plus simple de ces modèles est celui de Cox et Isham [22]. Il généralise le processus de Neyman-Scott en espace et en temps ; il est fondé sur des hypothèses très restrictives. Les averses arrivent selon un processus de Poisson en espace et en temps de paramètre $\lambda$. Chaque averse est modélisée par une région circulaire de pluie de rayon $R$ qui se déplace à une vitesse aléatoire $v = (v_x, v_y)$ sur une durée aléatoire $D$ jusqu'à disparition de la cellule. L'intensité de pluie $X$ est supposée constante. C'est au prix de ces simplifications drastiques que le nombre de paramètres reste raisonnable (six au total : $\lambda$, $R$, $v_x$, $v_y$, $D$ et $X$).

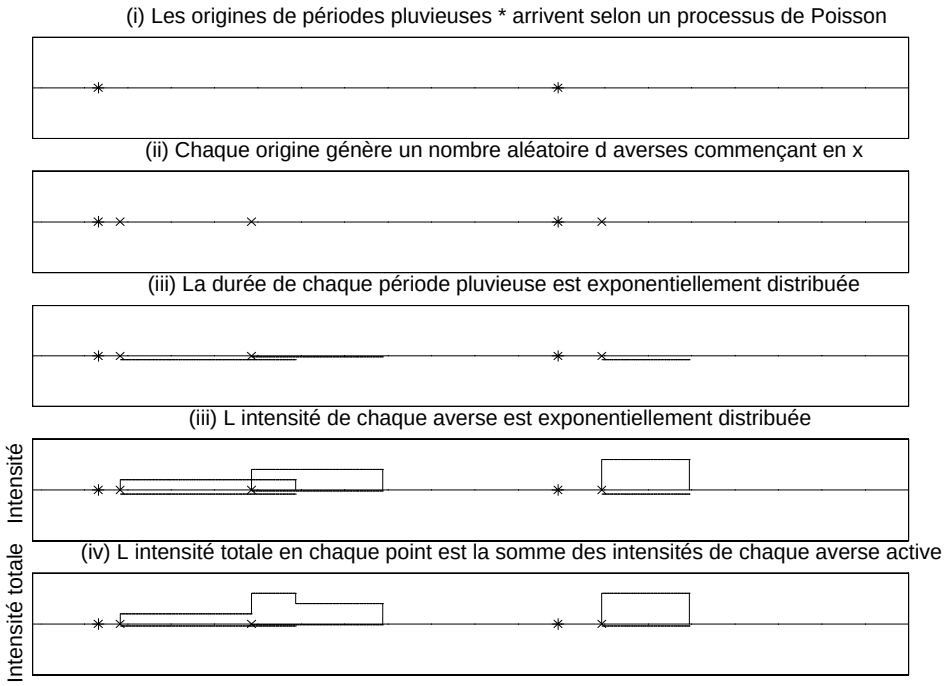



Figure 2.4: Processus de Neyman-Scott.

Dans le modèle de Waymire, Gupta, Rodriguez Iturbe, 84 [85], chaque cellule pluvieuse est caractérisée par son intensité, sa durée de vie ainsi que des paramètres de taille et de forme. Elles sont groupées en bandes pluvieuses de la petite méso-échelle (*section* 2.1.1). Leur nombre et leur emplacement suit un processus spatial de Poisson. Ces bandes pluvieuses de la petite méso-échelle sont elles-mêmes groupées dans des zones de pluie de la grande méso-échelle, leur nombre et leur emplacement suivant aussi un processus de Poisson de paramètres différents. L'intensité à l'intérieur de la cellule décroît exponentiellement depuis le centre jusqu'à atteindre zéro sur les bords. Au total, ce modèle comporte 12 paramètres. La moyenne et la variance théorique sont calculées. L'hypothèse de turbulence de Taylor (Taylor, 1938) stipule que la covariance en temps en un point fixe sur une durée $\Delta t$ est égale à la covariance à une date arbitraire entre deux points séparés par une distance $\Delta x$ telle que $\Delta x$ soit égal à $\Delta t$ à une constante près : $\Delta x = U \cdot \Delta t$. L'hypothèse de Taylor est vérifiée par les champs de pluie simulés par ce modèle.

Les chaînes de Markov peuvent elles aussi être généralisées en espace-temps. Pour des applications de ces modèles en hydrologie, on pourra consulter les travaux de Zucchini et Guttorp en 91 [88] ou la thèse de Bayomog, 94 [6] qui emploie des processus de Markov spatio-temporels pour modéliser l'occurrence de pluie.

### 2.2.5 Conclusion

Contrairement aux modèles temporels, on trouve peu de modèles spatiaux ou spatio-temporels dans la littérature. Les modèles temporels sont développés en vue de faire de l'estimation et de la prévision tandis que les modèles spatiaux ou spatio-temporels sont en général des modèles de simulation.

La prise en compte de la structure d'échelle du phénomène pluvieux est importante (pour la désagrégation par exemple), et permettrait d'établir un pont entre les connaissances des hydrologues et des météorologues. Les modèles stochastiques de pluie basés sur la classification de Austin et Houze font appel à un trop grand nombre de paramètres (au minimum 12, modèle de Waymire et al., 84 [85]), et ne donnent de bons résultats que sur les échelles sur lesquelles ils ont été calibrés. Les modèles hydrologiques mettent en oeuvre des outils statistiques standards : des distributions d'intensités ou de cumuls de type exponentiel (loi exponentielle, exponentielle mélangée, normale, gamma, kappa, etc...), ainsi qu'une structure de dépendance dite à court terme telle que ARMA ou Markov. Ces modèles donnent des résultats satisfaisants pour l'estimation ou la prévision des événements courants (faibles périodes de retour), mais sont défaillants pour les événements rares. En fait, les choix de ces lois et dépendances sont arbitraires et parfois ne s'accordent même pas avec les constatations empiriques. Des études de plus en plus nombreuses tendent à montrer que les corrélations entre observations éloignées (en espace ou en temps) décroissent à une vitesse inférieure à celle obtenue avec des données indépendantes ou un modèle de type ARMA ou Markov. Un modèle stochastique adapté à cette forme de dépendance est donc nécessaire.

Dans un souci d'adéquation à la réalité physique du phénomène et dans le but d'intégrer nos connaissances a priori, un modèle de pluie devrait représenter une série à différentes échelles, avec les mêmes valeurs pour certains paramètres. Mais comment élaborer des modèles fondés sur la probable existence d'invariant(s) d'échelle *?* Les fractals et mul-

tifractals, présentés dans la section suivante, constituent un outil susceptible de répondre à cette question.

## 2.3 Les modèles fractals

Dans cette section, on présente les principales définitions ainsi que les résultats de la théorie fractale. La géométrie fractale fournit un moyen de quantification du degré d'irrégularité (encore appelé rugosité) d'un ensemble. De plus elle permet d'extraire un paramètre indépendant de l'échelle : la dimension fractale. Des applications directes d'estimation de cette dernière sur des données sahéliennes d'occurrence de pluie (données Badinage) sont réalisées dans cette section, et leurs résultats comparés à ceux existant dans la littérature.

Nous présentons ensuite les cascades multifractales qui permettent de modéliser non seulement le support d'occurrence d'un phénomène mais aussi son intensité. Ces modèles stochastiques possèdent des caractéristiques fractales telles que des paramètres invariants d'échelle et une grande irrégularité. Enfin, une revue des principaux auteurs ayant travaillé avec les fractals et les multifractals en hydrologie (ainsi que leurs résultats) est ensuite dressée.

### 2.3.1 Les fractals

Les fractals, dont la racine latine *fractus* signifie irrégulier, ont été introduits par Mandelbrot en 1975 [59] pour caractériser des ensembles aux propriétés inhabituelles en géométrie classique. L'intérêt porté à cette théorie, initialement fondée sur des contre-exemples introduits par des mathématiciens entre les années 1875 et 1950, a été stimulé par ses multiples possibilités d'application à des objets naturels (tels que le contour des nuages, les côtes, les réseaux hydrographiques).

Dans cette section, on rappelle les définitions d'objet et de dimension fractals en se référant à l'ouvrage de Falconer, 90 [32]. Puis une étude des artefacts de la principale méthode d'estimation de la dimension fractale (la méthode du comptage de boîtes) est menée sur des simulations d'ensemble de Cantor.

#### Objet fractal

Plusieurs définitions d'objet fractal ont été proposées. Nous choisirons ici celle de Mandelbrot [59], selon laquelle un fractal est un ensemble dont la dimension fractale (qui reste à définir) est strictement supérieure à sa dimension euclidienne. Mais certains auteurs, tel que Falconer, 90 [32], pensent que cette définition est trop restrictive et préfèrent concevoir un fractal $F$ comme un ensemble possédant quelques unes des caractéristiques suivantes :

- $F$ a une structure fine, c'est-à-dire des détails à des échelles arbitrairement petites.
- $F$ est trop irrégulier pour être décrit en langage géométrique traditionnel, à la fois localement et globalement.

- Souvent $F$ présente une forme d'auto-similarité (exacte ou statistique, cf *Annexe E*). Par exemple, dans certains cas, $F$ est défini comme la limite d'une procédure récursive.

La notion de fractal est, quoi qu'il en soit, intimement liée à celle de dimension fractale. On trouve dans la littérature plusieurs définitions de cette dernière.

### Dimension fractale

Dans ce paragraphe, on se limitera à présenter la dimension de boîte et la dimension de Hausdorff mais le lecteur pourra se référer à Falconer [32] pour d'autres dimensions fractales.

**Définition 2** *Soit $F$ un ensemble borné non vide de $\mathbb{R}$ (resp. $\mathbb{R}^n$). Considérons la suite de segments (resp. de cubes) de maille $\delta$ :*

$$\begin{aligned} [m\delta, (m+1)\,\delta] \;\; avec\; m &\in \mathbb{N} \\ (resp.\;\; [m_1\delta, (m_1+1)\,\delta] \times \cdots \times [m_n\delta, (m_n+1)\,\delta] \, avec\; m_1, \cdots, m_n &\in \mathbb{N}) \end{aligned}$$

*et soit $N_\delta(F)$ le nombre de tels segments (resp. de cubes) intersectant $F$. Les dimensions de boîte supérieures et inférieures sont respectivement définies par :*

$$\begin{aligned} \underline{\dim}_B F &= \underline{\lim}_{\delta \to 0} \frac{\log N_\delta(F)}{-\log \delta} \\ \overline{\dim}_B F &= \overline{\lim}_{\delta \to 0} \frac{\log N_\delta(F)}{-\log \delta} \end{aligned}$$

*(voir[2]). Si elles sont égales, on définit la dimension de boîte par :*

$$\dim_B F = \lim_{\delta \to 0} \frac{\log N_\delta(F)}{-\log \delta}$$

La dimension de boîte est la notion la plus fréquemment utilisée pour définir la dimension fractale, car elle est très facile à calculer en pratique (contrairement à la dimension de Hausdorff comme on le verra par la suite). La méthode du comptage de boîtes, qui permet parfois de calculer la dimension de boîte d'un ensemble, consiste à utiliser une suite $\delta_n = \frac{1}{b^n}, \quad b > 1$ et à déterminer :

$$-\lim_{n \to \infty} \frac{\log N_{\delta_n}(F)}{n \log b}$$

[2] où $\underline{\lim}$ désigne la limite inférieure : soit $E_\delta = \{f(r), 0 < r < \delta\}$

$$\underline{\lim}_{\delta \to 0} f(\delta) = \lim_{\delta \to 0} (\inf E_\delta)$$

et $\overline{\lim}$ la limite supérieure :

$$\overline{\lim}_{\delta \to 0} f(\delta) = \lim_{\delta \to 0} (\sup E_\delta)$$

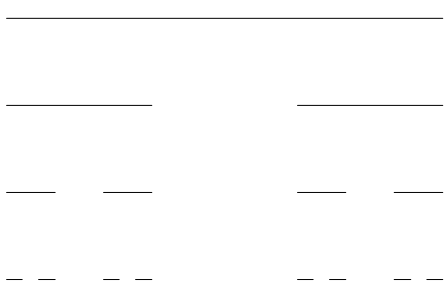


Figure 2.5: Construction de l'Ensemble de Cantor.

**Exemple 3** *L'Ensemble de Cantor est un ensemble qui se construit par itérations : En partant d'un segment de longueur a que l'on divise en trois partie égales et auquel on retire le segment central, on obtient un ensemble formé de deux segments disjoints de longueur $a/3$ et séparés par un vide de même longueur (figure 2.5). On réitère indéfiniment l'opération pour obtenir un ensemble appelé Ensemble de Cantor (ou parfois poussière de Cantor).*

*La dimension fractale de cet ensemble peut être déterminée par la méthode du comptage de boîtes en choisissant une suite de mailles de longueurs $\delta_n = \frac{a}{3^n}$, avec n entier, et en remarquant qu'une division par 3 de la longueur $\delta_n$ de la jauge conduit à une multiplication par 2 du nombre de boîtes $N_\delta(F)$ intersectant F, donc :*

$$2 = 3^{\dim_B F} \text{ et } \dim_B F = \frac{\log 2}{\log 3} = 0.63$$

**Exemple 4** *Courbe de Weierstrass : La fonction de Weierstrass est définie par :*

$$\begin{aligned} f &: [0,1] \longrightarrow \mathbb{R} \\ f(t) &= \sum_{k=1}^{\infty} \lambda^{(s-2)k} \sin\left(\lambda^k t\right) \end{aligned}$$

*avec $\lambda > 1$ et $1 < s < 2$. On montre (Falconer [32]) que son graphe (figure 2.6) est un ensemble fractal de dimension de boîte égale à s.*

D'une manière générale, la dimension fractale quantifie la puissance selon laquelle la mesure d'un ensemble diverge vers l'infini. Elle peut être interprétée comme le degré d'irrégularité de cet ensemble, conditionnellement cependant à la dimension classique (ou entière) de l'espace auquel elle appartient (tableau 2.4).

Pour un objet non fractal, la dimension de boîte est égale à sa dimension euclidienne (par exemple, un segment (resp. un carré) a pour dimension de boîte 1 (resp. 2). Pour un ensemble fractal, elle est non entière.

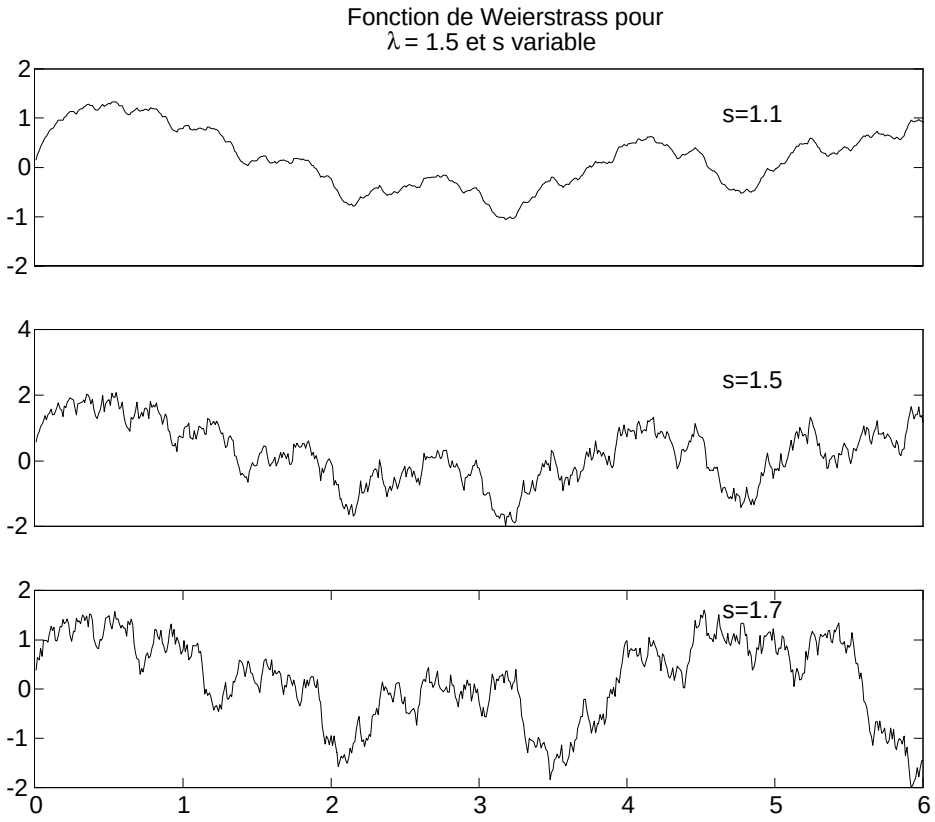



Figure 2.6: Graphes de la fonction de Weirstrass pour $s$ variable.

| Objet | Poussière sur une droite | Courbe plane très irrégulière | Surface très feuilletée |
|---|---|---|---|
| $\dim_B F$ | $0 \leq \dim_B F \leq 1$ | $1 \leq \dim_B F \leq 2$ | $2 \leq \dim_B F \leq 3$ |

Tableau 2.4: Gammes de variation des dimensions fractales.

Cependant, la dimension de boîte est difficile à manipuler mathématiquement. En effet, on n'a généralement pas :

$$\dim_B U_{i=1}^{\infty} F_i = \sup_{i \geq 1} \dim_B F_i$$

Elle rend ainsi parfois très mal compte du caractère fractal de certains ensembles.

**Exemple 5** *L'ensemble* $F = \left\{1, \frac{1}{2}, \cdots, \frac{1}{k}, \cdots\right\}$ *a pour dimension de boîte* $\frac{1}{2}$*, est donc considéré comme fractal*[3]*.*

La dimension de Hausdorff est une autre définition de la dimension fractale. Définie à partir de la mesure de Hausdorff, elle présente l'avantage d'être facile à manipuler et d'être définie pour tout ensemble. Elle est cependant difficile à calculer.

**Mesure de Hausdorff :** Rappelons que si $U$ est un ensemble non vide inclus dans $\mathbb{R}^n$, alors son $|U|$ **diamètre** est défini par :

$$|U| = \sup \{|x - y| \;\; x, y \in U\}$$

Si $\{U_i\}$ est une suite d'ensembles de diamètres inférieurs à $\delta$ recouvrant $F$ (i.e. $F \subset U_{i=1}^{\infty} U_i$) avec $0 < |U_i| \leq \delta$, on dit qu'on a un **recouvrement** de $F$.

**Définition 6** *Soit* $F$ *un ensemble inclus dans* $\mathbb{R}^n$ *et soient* $s$ *et* $\delta$ *deux nombre positifs. Soit :*

$$H_{\delta}^{s}(F) = \inf \left\{ \sum_{i=1}^{\infty} |U_i|^s , \{U_i\} \;\; \textit{recouvrement de } F \right\}$$

*La mesure de Hausdorff s-dimensionnelle est définie par :*

$$H^s(F) = \lim_{\delta \to 0} H_{\delta}^{s}(F)$$

**Dimension de Hausdorff :** On montre qu'il existe une valeur de coupure $\dim_H(F)$ entre deux comportement relatifs à la mesure de $F$ (figure 2.7) :

$$\begin{aligned} \text{Si } \alpha &< \dim_H(F), \text{ alors } H^s(F) = \infty \\ \text{Si } \alpha &> \dim_H(F), \text{ alors } H^s(F) = 0 \end{aligned}$$

[3]Preuve : On considère un recouvrement par les intervalles de taille $\delta > \frac{1}{2}$. Il existe un entier $k$ tel que $\frac{1}{k(k-1)} > \delta \geq \frac{1}{(k+1)k}$ . Ainsi, chaque intervalle ne peut recouvrir plus d'un seul élément de $F$, donc :

$$\frac{\log N_\delta(F)}{-\log \delta} \geq \frac{\log k}{\log[(k+1)k]} \text{ et } \underline{\dim}_B F \geq \frac{1}{2}$$

D'autre part, si $\frac{1}{2} > \delta > 0$ on choisit $k$ tel que $\frac{1}{k(k-1)} > \delta \geq \frac{1}{(k+1)k}$. Alors $k+1$ intervalles recouvrent $[0, \frac{1}{k}]$ laissant $(k-1)$ intervalle recouvrir les autres points de $F$. Ainsi :

$$\frac{\log N_\delta(F)}{-\log \delta} \leq \frac{\log(2k)}{\log[(k-1)k]} \text{ et } \overline{\dim}_B F \leq \frac{1}{2}$$

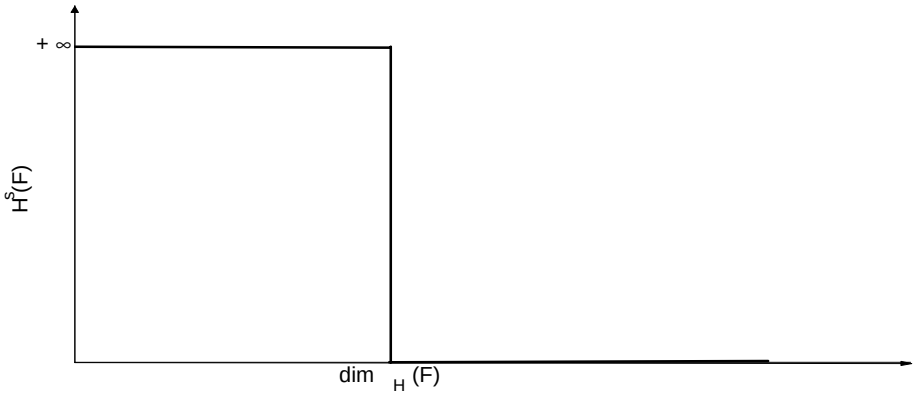



Figure 2.7: Mesure et dimension de Hausdorff.

**Définition 7** *La dimension de Hausdorff est définie par :*

$$\dim_H(F) = \inf\{s : H^s(F) = 0\} = \sup\{s : H^s(F) = \infty\}$$

**Exemple 8** *Pour l'Ensemble de Cantor, on montre que, si* $s = \frac{\log 2}{\log 3}$ *alors* $H^s(F) \in \left[\frac{1}{2}, 1\right]$. *Donc* $\dim_H(F) = s$.

### Etude de la méthode du comptage de boîtes sur simulations

On simule un ensemble de Cantor à $3^9$ éléments (soit un peu moins de 20 000 données) et on lui applique la méthode du comptage de boîtes afin de déterminer empiriquement sa dimension fractale et tester ainsi les défauts de la méthode. Plusieurs facteurs d'échelle $m$ de taille de boîtes sont choisis (tailles de boîte variant de 2 en 2 ou de 3 en 3, etc...). On obtient (figure 2.8) des courbes à peu près linéaires pour $m = 2, 5, 7$ et un alignement parfait dans le cas où la taille des boîtes se démultiplie par 3 (qui est le facteur d'échelle d'un Ensemble de Cantor).

Pour chaque facteur d'échelle $m$, on calcule la pente moyenne entre points consécutifs et le nombre moyen de boîtes correspondant. Sur le tableau 2.5, on constate une surestimation (resp. sous-estimation) systématique du nombre moyen de boîtes non vides lorsque les boîtes se démultiplient par un facteur d'échelle supérieur (resp. inférieur) à 3. En *Annexe A*, on démontre ce résultat : le nombre de boîtes non vides de taille $\frac{1}{m^k}$ avec $m > 3$ (resp. $m < 3$), est supérieur (resp. inférieur) au nombre de boîtes non vides de taille $\frac{1}{3^k}$.

La méthode du comptage de boîtes présente donc en pratique quelques artefacts relatifs au choix du facteur d'échelle dont il faudra tenir compte lors des applications du paragraphe suivant.

### Conclusion

La dimension fractale se révèle être parfois un moyen efficace pour extraire d'un ensemble une caractéristique invariante d'échelle. La méthode du comptage de boîtes est un

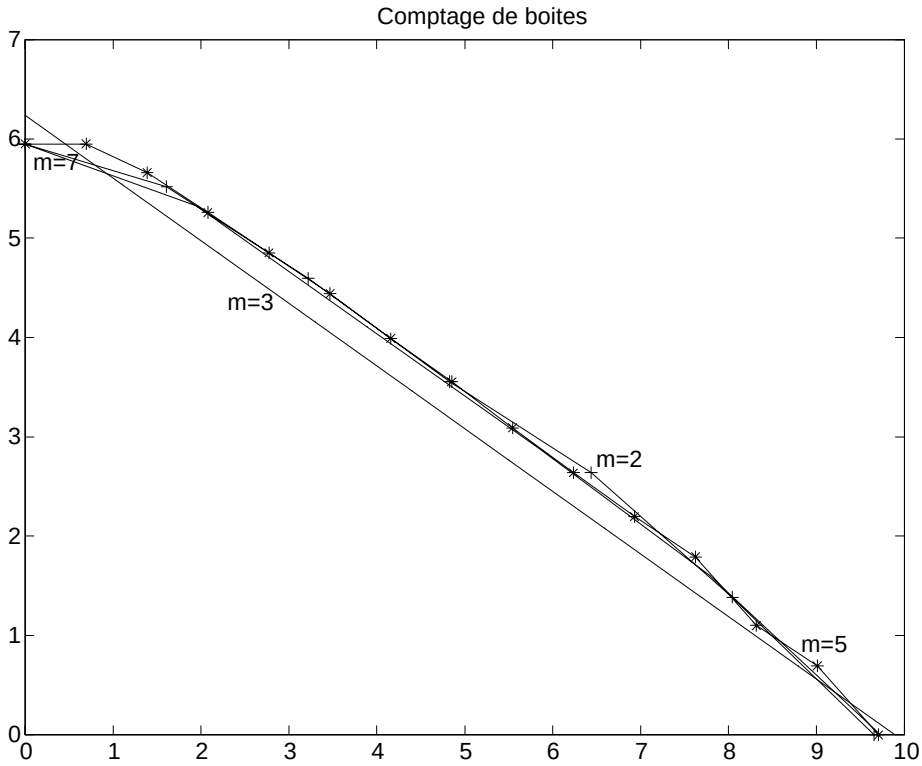



Figure 2.8: Comptage de boîtes sur un Cantor simulé à $3^9$ données, avec boîtes se dé-$m$-uplant.

| Nombre de données | $2^{14}$ | $3^9$ | $5^6$ | $7^5$ |
|---|---|---|---|---|
| $a$ | 2 | 3 | 5 | 7 |
| Pente moyenne | -0.61 | -0.63 | -0.62 | -0.61 |
| $N$ | 101.1. | 102.3 | 112.4 | 112.3 |

Tableau 2.5: Comptage de boîtes sur l'Ensemble de Cantor pour plusieurs facteurs d'échelle $a$ ($N$ = nombre moyen de boîtes non vides).

moyen simple d'estimation de la dimension fractale qui ne conduit cependant pas toujours à des résultats fiables. En effet, le critère de choix du facteur d'échelle n'est pas établi et l'estimateur de boîte manque de robustesse quand la série est tronquée. De plus, la forte intermittence des séries pluviométriques rend inutile l'exploration du comportement fractal en dessous d'un pas de temps minimal. Dans le but d'accéder à une quantification de la qualité d'estimation, cette méthode a aussi été étudiée dans un cadre probabiliste (Falconer, 90 [32] ou Harte, 97 [42]). Nous présentons dans le paragraphe suivant une application directe de la géométrie fractale à l'analyse de la répartition temporelle de séries pluviométriques (occurrence de pluie ou de basculement d'auget).

### 2.3.2 Application à quelques séries d'occurrence

On représente l'occurrence de pluie comme un nuage de segments disjoints supportés par l'axe représentatif du temps. La méthode du comptage de boîtes est appliquée à cet ensemble évoquant une poussière de Cantor, afin d'analyser la répartition temporelle d'occurrence de pluie et trouver d'éventuels invariants d'échelle.

**Application 1 : Données *Badinage*** Ce sont cinq mois de cumuls de pluie à pas de temps 5 minutes de l'année 1991, sur quatre mois de la saison pluvieuse (entre le 14.04.91 et le 02.09.91). On dispose des relevés en six stations au Niger. Le cumul minimum observé sur ces données est de 5 millimètres.

On applique la méthode du comptage de boîtes au nuage d'occurrence de cumuls non nuls $F$ en reportant sur un graphe log-log la taille $\delta$ des boîtes (ici des segments de longueur $\delta = 3^k$ ) *vs* $N_\delta(F)$ le nombre de boîtes contenant au moins un segment du nuage d'occurrence de pluie (figure 2.9).

Les six graphes de comptage de boîtes se présentent sous forme de lignes brisées. Cette linéarité du graphe confirme la présence d'auto-similarité dans les séries temporelles d'occurrence de pluie. D'autre part, les deux ruptures de pente sont constatées aux mêmes abscisses sur les 6 courbes : aux pas de temps 1 heure 40 minutes ($e^3 * 5 = 100,43\,\text{min}$) et 1 jour 10 heures ($e^6 * 5 = 2017\,\text{min}$). La pente $-1$ observée sur la dernière portion du graphe est due au fait que toutes les boîtes sont pleines lorsque l'on considère des pas de temps supérieurs à un mois (il pleut au moins une fois tous les 10 jours en cette saison). Quant à la première rupture, elle indique que l'occurrence de pluie apparaît par paquets denses sur des gammes de pas de temps allant de 1 heure à 1 jour. Une explication peut être trouvée dans l'analyse des événements pluvieux et de leur durée. En effet, cette rupture peut traduire un changement de régime pluvieux car elle correspond à la durée moyenne d'un événement pluvieux dans cette région (1 heure 40 minutes). On a représenté dans le tableau 2.6 les pentes sur les deux premières portions du graphique.

Sur la deuxième portion (pas de temps compris entre 1 h 40 min et 1 j 10 h), la pente moyenne est de $-0.27$. Ce résultat est proche de ceux de Hubert et Carbonnel en 1989 [45] : sur la série horaire de Ouagadougou (Burkina Faso), ils estiment la dimension de boîte sur ces pas gammes de temps à 0.22.

Sur la première portion du graphe (pas de temps de l'ordre de la minute), on constate une pente moyenne de $-0.55$, résultat qui se rapproche de la pente 0.47 déterminée

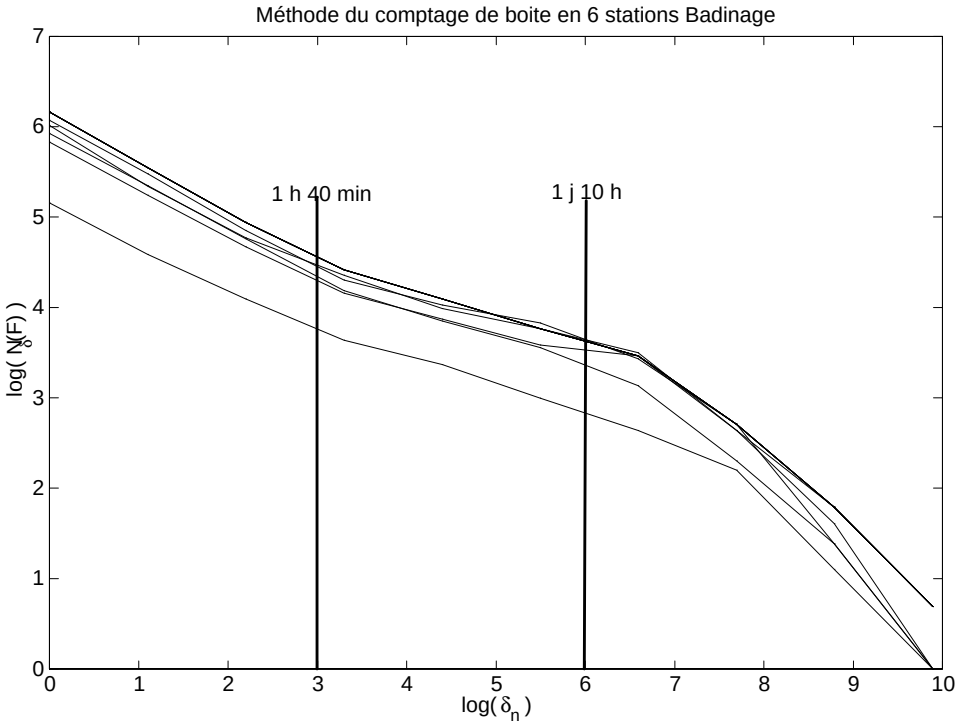



Figure 2.9: Comptage de boites sur les données Badinage.

| Station | Longueur | Longueur tronquée | $-P_1$ | $-P_2$ |
|---|---|---|---|---|
| 1 | 39 376 | 19 683 | 0.56 | 0.29 |
| 2 | 39 383 | 19 683 | 0.56 | 0.32 |
| 3 | 36 791 | 19 683 | 0.55 | 0.21 |
| 4 | 22 334 | 6 561 | 0.51 | 0.30 |
| 5 | 39 394 | 19 683 | 0.54 | 0.26 |
| 6 | 40 769 | 19 683 | 0.58 | 0.26 |
| Moyennes | 36 341 | 17 496 | 0.55 | 0.27 |

Tableau 2.6: Comptage de boites sur les données Badinage ($P_1$ = pente sur $[5min; 1h40min]$ et $P_2$= pente sur $[1h40min; 1j10h]$).

| $m$ | 2 | 3 | 5 | 7 |
|---|---|---|---|---|
| $-P$ | 0.30 | 0.29 | 0.28 | 0.31 |

Tableau 2.7: Comptage de boites sur la station 1 ($P$ = pente sur $[1h40, 1j10h]$ et $m$ facteur d'échelle).

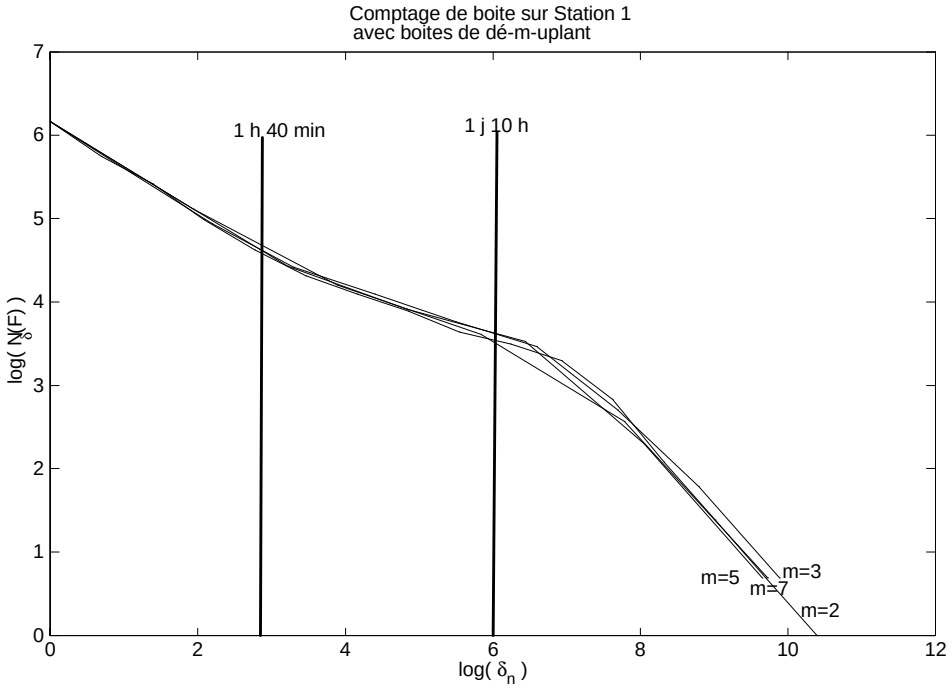



Figure 2.10: Comptage de boites sur la station 1.

sur les premiers points (pas de temps de l'ordre de l'heure) du graphe de comptage de boîte à Ouagadougou ([45]).

On vérifie que cette la faible pente sur la portion 2 n'est pas due à un artefact de la méthode du comptage de boîtes du même type que ceux rencontrés lors des simulations sur Cantor. En effet, on a vu dans la section précédente que le choix du facteur d'échelle (s'il existe) influe sur l'estimation de la dimension de boîte d'un ensemble de Cantor. On applique donc la méthode du comptage de boîtes sur la station 1 en faisant varier le facteur d'échelle $m$ dans l'ensemble $\{2, 3, 5, 7\}$ (figure 2.10). Les résultats (tableau 2.7) ne varient quasiment pas avec $m$. Il est possible que la présence d'aléas au sein de la série atténue l'artefact lié au choix du facteur d'échelle.

L'explication la plus plausible de cette sous-estimation reste celle d'un artefact lié à la troncature de la série, artefact d'autant plus présent que les pas de temps sont grands.

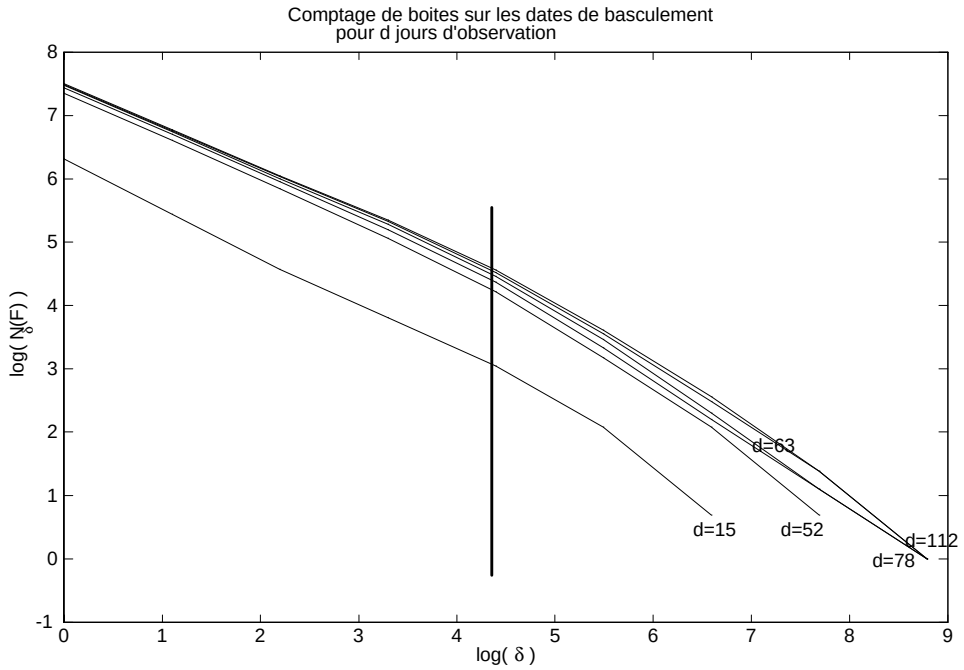



Figure 2.11: Comptage de boites sur la série de la Réunion.

**Application 2 : Données *Réunion***

L'exploration du comportement fractal de l'occurrence de pluie aux petits pas de temps requiert des données à petits pas de temps. Les données de La Réunion sont des données à basculement d'auget. L'auget bascule dès que le cumul de pluie atteint 0.5 mm et un appareil enregistre les dates de basculement. La résolution temporelle de l'appareil est de 1 seconde, donc l'intensité maximale décelable est de 0.5 mm/s. On applique la méthode du comptage de boîtes au nuage $F$ formé par les dates de basculement représentées sur l'axe du temps (un point pour chaque date exprimée en secondes). La densité de $F$ reflète l'intensité de pluie ou de l'occurrence ou non de hauteurs de pluie dépassant un certain seuil.

Les tailles de boîte varient en se démultipliant par 3, en partant de 15 minutes jusqu'à 70 jours, en faisant varier la durée $d$ d'observation entre 15 et 112 jours (figure 2.11).

La pente sur les trois derniers points (boîtes de taille supérieure à 7 jours) se stabilise à 1 puisqu'il pleut au moins une fois tous les 7 jours.

La pente de la portion de graphe correspondant à des cumuls de pluie sur des pas de temps compris entre 15 min et 1 jour (5 premiers points) décroît avec la longueur de la série mais on constate qu'elle se stabilise à 0.65 (tableau 2.8).

**Conclusion :** Les séries temporelles d'occurrence de pluie possèdent un structure fractale mais avec des dimensions fractales qui diffèrent avec la gamme de pas de temps considérée. Les périodes de ruptures semblent correspondre à la durée des événements pluvieux.

| $N$ | $D$ (en jours) | $-P$ |
|---|---|---|
| 10 000 | 112 | 0.65 |
| 9 000 | 78 | 0.65 |
| 8 000 | 63 | 0.67 |
| 7 000 | 52 | 0.68 |
| 6 000 | 40 | 0.69 |
| 2 000 | 15 | 0.76 |

Tableau 2.8: Méthode du comptage de boites sur la série de la Réunion ($N$ =Nombre de basculement, $D$ =Durée d'observation $P$ =Pente sur $[2.5min; 6h]$).

L'estimation de la dimension fractale est peu variable entre données sahéliennes (données Badinage et Ouagadougou). Il serait intéressant d'étudier un grand nombre de séries en ces lieux géographiques pour conclure quant à l'invariance spatiale de ce paramètre.

Mais l'interprétation de ces résultats d'estimation fractale du support de l'occurrence de pluie ne doit se faire que relativement à un seuil : celui de mesurabilité des cumuls ou des intensités des séries étudiées. En effet, la dimension fractale d'un ensemble d'occurrence de dépassement d'un seuil $s$ varie avec ce dernier. Il serait utile d'intégrer la prise en compte de l'intensité du phénomène dans l'approche fractale. C'est l'ambition de l'approche multifractale.

### 2.3.3 Modèle en cascade multifractale

Les modèles en cascades multifractales fournissent les plus simples structures de dépendance retranscrivant des propriétés d'invariance de la distribution des observations. Ce sont des modèles possédant des propriétés multifractales telles qu'elles sont rappelées en *Annexe D* (pour toute précision supplémentaire sur les modèles multifractals généraux, on pourra se référer à l'ouvrage de Schertzer et Lovejoy 93 [76]). Dans cette section, on commence par définir les cascades multiplicatives multifractales avant de constater leur propriété distinctive : l'invariance d'échelle. On dresse ensuite une revue de leurs principales utilisations en de l'hydrologie.

#### Définition

Une cascade multiplicative multifractale de générateur $\eta$ est une suite de variables aléatoires $(\mu_{j,k})$ définies par la relation de récurrence (figure 2.12) :

$$\begin{aligned} \mu_0 &= \mu_{0,0} > 0 \\ \text{et } \forall j &\geq 1 \quad \forall k = 0 \cdots 2^j - 1 \quad \mu_{j,k} = \mu_{j-1,\lfloor k/2 \rfloor} \cdot \eta_{j,k} \end{aligned}$$

L'interprétation d'une telle suite est immédiate en hydrologie :

- Supposons que l'on dispose d'un cumul $\mu_0$ sur $[0, T[$ (à l'échelle $T$ ou au niveau 0).
- Au niveau 1, il est désagrégé en $\mu_{1,0}$ sur $\left[0, \frac{T}{2}\right[$et de $\mu_{1,1}$ sur $\left[\frac{T}{2}, T\right[$.

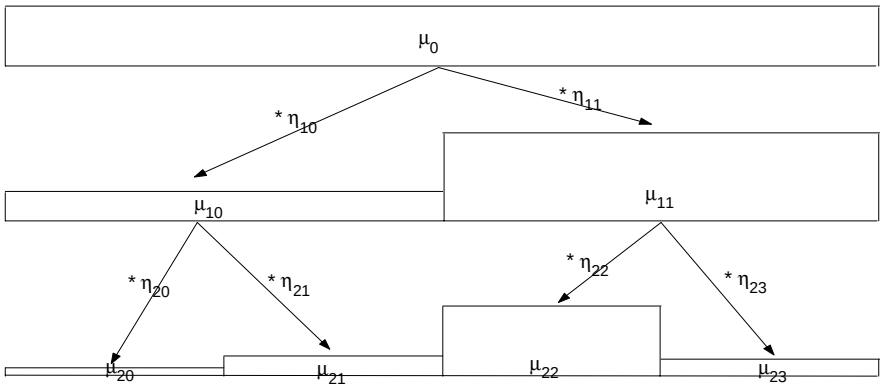



Figure 2.12: Cascade multifractale.

- etc ...
- En désagrégeant jusqu'au niveau d'homogénéité $n$, on obtient une suite de variables aléatoires $\left(\mu_{n,k}\right)_{k=0\cdots 2^n-1}$, les cumuls sont de $\mu_{n,0}$ sur $\left[0, \frac{T}{2^n}\right[$, ..., de $\mu_{N,2^n-1}$ sur $\left[(2^n-1)\frac{T}{2^n}, T\right[$.

Il est important de remarquer qu'en hydrologie, on ne s'intéresse pas à une désagrégation "infinie" ($n$ tend vers l'infini) mais plutôt à un niveau de désagrégation limite $N$ nommé niveau d'homogénéité de la cascade. Pour une étude de l'existence et de la définition d'un processus en cascade (désagrégation infinie), on pourra se reporter aux travaux de Kahane, 85 [47].

### Générateurs

Les générateurs $\left(\eta_{j,k}\right)_{j=1\cdots N,k=0\cdots 2^j-1}$ sont choisis indépendants deux à deux et identiquement distribués :

$$\left(\eta_{j,k}\right)_{j=1\cdots N,k=0\cdots 2^j-1} \ \ iid$$

Leur distribution $f$ est choisie log $-$infiniment divisible c'est-à-dire invariante par multiplication (par exemple, la loi log $-$normale puisque la loi normale est infiniment divisible). Ce choix de distribution, associé à la convention $\mu_0 \equiv 1$, permet d'aboutir à un résultat d'invariance d'échelle. En effet, les variables désagrégées $\mu_{N,i}$ à tout niveau $N$

suivront toutes la même loi $f$ puisque par itérations :

$$\forall N > 0 \quad \mu_{N,*} = \prod_{j=1}^{N} \eta_{j,*}$$

**Cascades universelles**

Ces cascades ont été construites dans le but de détecter et d'estimer d'éventuels paramètres communs à toutes les séries hydrologiques (à différents pas de temps et en différents site, sur des séries de pluie ou de débit) qui seraient donc des invariants directement liés à la phénoménologie des champs de pluie.

Dans ces cascades, la loi du générateur est une loi stable (parfois appelée loi de Lévy, Feller 71 [35]). Ces lois ne sont définies que par leur fonction caractéristique :

$$\Phi_{\alpha,\beta,c,\gamma}(x) = \exp\left\{ i\gamma x - c\,|x|^{\alpha}\left[1 + i\beta\frac{t}{|t|}\tan\left(\frac{\pi\alpha}{2}\right)\right]\right\}$$

avec $0 \leq \alpha \leq 2$, $-1 \leq \beta \leq 1$, $c \geq 0$ et $\gamma \in \mathbb{R}$.

Ces distributions (dont la loi normale fait partie puisque $\Phi_{norm}(x) = \exp\left\{-c\,|x|^2\right\}$) peuvent être vues comme des généralisations de la loi normale dans la mesure ou elles sont stables[4] et attractives[5].

**Définition 9** *La fonction d'échelle des moment est définie par :*

$$K(q) = \log_{2^N} E[\mu_N]$$

On montre que, par le théorème de Lévy Khinchine :

$$\begin{aligned} K(q) &= \frac{C_1}{\alpha - 1}\left(q^{\alpha} - q\right) \text{ si } \alpha \neq 1 \\ K(q) &= C_1 q \log q \text{ si } \alpha = 1 \end{aligned}$$

[4]Stabilité : La loi $f_\alpha$ est dite stable lorsque pour tout échantillon $(X_i)_{1\leq i\leq n} \overset{iid}{\sim} f_\alpha$, on a :

$$\frac{\sum_{i=1}^{n} X_i}{n^{1/\alpha}} \sim f_\alpha$$

[5]Attractivité : La loi $f_\alpha$ est dite attractive lorsque pour tout échantillon $(X_i)_{1\leq i\leq n}$ vérifiant :

$$\begin{aligned} E[X_1] &= 0 \text{ et } E\left[X_1^2\right] = \infty \\ \exists\alpha &< 2, \quad E[X_1^{\alpha}] = 1 \text{ et } \forall\beta > \alpha \quad E\left[X_1^{\beta}\right] = \infty, \end{aligned}$$

on a : $\dfrac{\sum_{i=1}^{n} X_i}{n^{1/\alpha}} \overset{n\to\infty}{\longrightarrow} f_\alpha$

Le cas $\alpha = 0$ correspond au modèle mono-fractal $\beta$ (voir[6]) et le cas $\alpha = 2$ correspond aux cascades log-normales (voir[7]).

Les paramètres $\alpha$ et $C_1$, qui caractérisent à eux seuls le processus, sont dits paramètres universels.

**Décroissance algébrique des queues de distribution**

Une distribution est dite de type algébrique si sa fonction de survie décroît algébriquement :

$$G(x) \overset{x\infty}{\propto} x^{-q} \text{ (ou } \lim_{x\to\infty} \frac{G(x)}{x^{-q}} = c^{te})$$

On a donc en particulier le résultat suivant : Si le générateur suit une loi de type algébrique invariante par multiplication, alors la cascade ainsi construite est une suite de variables aléatoires de type algébrique à tout niveau d'agrégation. Le paramètre de décroissance $q$ est en particulier invariant d'échelle[8].

### 2.3.4 Revue bibliographique de modèles hydrologiques multifractals

L'emploi des fractals et des multifractals pour la modélisation n'a commencé que récemment. Ils attirent à la fois les physiciens pour qui la notion d'invariance d'échelle est familière et les statisticiens qui y trouvent le moyen d'intégrer une très forte variabilité dans une structure de dépendance relativement simple.

Les estimations du paramètre de décroissance algébrique $q_D$ se recoupent entre les divers auteurs et les différents pas de temps (tableau 2.9) : Elles sont comprises entre 2.1 et 3.6 pour des cumuls de pluie à pas de temps supérieur à 15 minutes mais diminuent nettement pour les pas de temps très petits (dans les travaux de Rodriguez-Iturbe, 91 [72], $q_D = 1.5$ pour un pas de temps de 15 secondes).

[6] C'est le plus simple des modèles en cascade multifractale. Le générateur $\eta$ suit une loi de Bernoulli :

$$\begin{aligned} P(\eta = \lambda^c) &= \lambda^{-c} \\ P(\eta = 0) &= 1 - \lambda^{-c} \end{aligned}$$

[7] Dans ce modèle en cascade multifractale, le générateur $\eta$ suit une loi log-normale :

$$\begin{aligned} \eta &= e^{\theta} \\ \theta &\sim Normale \end{aligned}$$

[8] Par exemple, si le générateur est $\log-$Gamma,

$$\eta \sim \log\Gamma(n, q) \text{ (ou bien } \log\eta \sim \Gamma(n, q))$$

alors, à tout niveau $N$ de désagrégation :

$$\mu_{N,*} = \prod_{j=1}^{N} \eta_{j,*} \sim \log\Gamma(Nn, q)$$

Les cumuls désagrégés suivent tous une loi $\log-$Gamma de même paramètre de décroissance algébrique $q$.

Des valeurs du même ordre de grandeur sont trouvée sur les séries de débit. L'existence d'un paramètre invariant d'échelle et spatial, lié à la nature même du phénomène pluvieux n'est donc pas inacceptable.

| | $\widehat{q_D}$ | débit/pluie | Pas de temps | Site |
|---|---|---|---|---|
| Bolgov et al. [13],1999 | | débit | Max annuels | Russie |
| Tessier et al. [80],1996 | 2.7 | débit | 1 jour | France |
| Pandey et al. [69],1998 | 3.1 | débit | 8 jours | EU |
| Hubert et Carbonnel [44],1988 | 3.4 | pluie | 1 jour | Burkina Faso |
| Van Monfort and Witter [65],1986 | 2.9 | pluie | 1 heure | Pays-Bas |
| Ladoy et al.[53],1993 | 3.0 | pluie | 12 heures | Nîmes |
| Rodriguez-Iturbe [72],1991 | 1.5 | pluie | 15 s | Boston, EU |
| Georgakakos et al.[39],1994 | [1.7 ; 2.9] | pluie | haute résolution | Iowa City, EU |
| de Lima [25],1998 | 3.1 | pluie | 15 min | V. Formoso |
| | 3.6 | pluie | 1 jour | V. Formoso |
| | 3.6 | pluie | 1 mois | V. Formoso |
| | 2.6 | pluie | 1 heure | Assink |
| | 2.1 | pluie | 15 min | Nancy |

Tableau 2.9: Revue bibliographique des modèles multifractals

Mais l'approche de modélisation multifractale n'a pas encore été totalement explorée, notamment au niveau de la critique des résultats d'estimation. C'est l'une des ambitions de ce travail où, dans le chapitre suivant, on se penche sur l'étude de la qualité des estimateurs standards du paramètre de décroissance algébrique $q_D$, avant de proposer un outil original de détection et d'estimation de comportement algébrique. On verra dans la section suivante que les distributions à décroissance algébrique constituent un cadre théorique bien exploré en probabilités : elles appartiennent au domaine d'attraction de la loi de Fréchet. Bien que leurs propriétés s'apparentent à celles des lois de la famille exponentielle, elles attribuent un poids radicalement différent aux grandes observations, ce qui n'est pas sans conséquences pratiques en hydrologie.

# Chapitre 3

# Modélisation des extrêmes hydrologiques

L'objectif de ce chapitre est de modéliser les extrêmes hydrologiques. De nombreux résultats et outils portant sur la modélisation des extrêmes existent dans la littérature probabiliste et statistique. Cependant, les séries hydrologiques sont courtes et les estimateurs proposés dans la littérature statistique sont, pour la plupart, entachés d'erreurs d'estimation rédhibitoires compte tenu des exigences pratiques hydrologiques. Ce chapitre propose donc une nouvelle méthode de modélisation des extrêmes, comparable en terme d'erreur à celle du maximum de vraisemblance, mais présentant l'avantage d'être semi-paramétrique (dans le sens où l'on ne fait d'hypothèses que sur l'appartenance à une classe de modèles et non pas sur un modèle particulier). Ces deux méthodes, appliquées à de nombreuses séries pluviométriques, permettent de conclure que les queues de distribution des séries de pluie ont un comportement algébrique et que le paramètre statistique caractérisant la décroissance algébrique des queues de distribution est un invariant spatial et temporel.

## 3.1 Introduction

L'abondance croissante d'informations sur la magnitude de nombreux évènements naturels extrêmes récents et sur leurs conséquences dramatiques porte à croire que la fréquence de ceux-ci a augmenté brusquement. En fait, il se peut que seules leur relation et leur description soient devenues plus complètes. L'accroissement de la vulnérabilité des sites, du fait de leur utilisation de plus en plus intense, conduit aussi le plus grand nombre à penser que l'intensité des phénomènes naturels est en augmentation. La prévention contre les phénomènes extrêmes, au moins dans le domaine de l'hydrologie, est encore fondée sur des modèles de prédiction indiscutés. La non conformité d'événements extrêmes aux prédictions de ces modèles standard pousse des personnes même averties à croire que les conditions naturelles de génération de ces phénomènes exceptionnels ont subi des modifications. Ainsi l'accroissement de l'information sur les catastrophes naturelles, une plus grande vulnérabilité à ces catastrophes, les écarts entre les observations et les prédictions des modèles utilisés font croire à un dérèglement de la nature.

Pour l'hydrologie opérationnelle il s'agit de proposer un outil de description et de

prédiction, dégagé d'hypothèses trop arbitraires (souvent adoptées pour le confort qu'elles procurent), et rendant plausibles des intensités de phénomènes, intensités qui, dans le cadre des modèles actuels, auraient des probabilités infinitésimales.

La théorie des valeurs extrêmes fournit un cadre probabiliste visant à construire un modèle asymptotique pour les variables aléatoires extrêmes. Plus précisément, elle établit les conditions sous lesquelles les extrêmes d'un échantillon aléatoire $(X_1, \cdots, X_n)$ convergent vers une distribution limite non dégénérée quand la taille de l'échantillon $n$ tend vers l'infini. Les lois rencontrées dans la littérature hydrologique sont le plus souvent de type exponentiel (loi exponentielle, normale, Weibull ...). La loi normale a connu un développement considérable dans les applications hydrologiques, car elle a la propriété d'être invariante par addition. Elle assure ainsi à l'hydrologue une invariance dans la nature de la répartition des erreurs; mieux encore, elle garantit, grâce au Théorème de la Limite Centrale, que la distribution de la somme d'erreurs de distributions quelconques converge vers la loi Normale :

$$\frac{X_1 + \cdots + X_n - n\mu}{\sqrt{n}\sigma} \overset{n\to\infty}{\longrightarrow} N(0,1)$$

ce qui n'est vérifié que si la queue de la distribution $F$ des variables aléatoires $(X_1, \cdots, X_n)$ n'est pas trop lourde. Cette condition peut ne pas être vérifiée lorsque l'analyse statistique porte sur les $k^{ième}$ extrêmes de l'échantillon $(X_1, \cdots, X_n)$ (au sens des $k$ plus grandes observations), et l'on peut se demander plus généralement s'il existe des constantes $a_n$ et $b_n$ telles que le rapport :

$$\frac{X_1 + \cdots + X_n - b_n}{a_n}$$

tende en loi vers une distribution non dégénérée.

Dans ce chapitre, nous présentons dans ses grandes lignes la théorie des valeurs extrêmes en *section* 3.3, en détaillant le domaine d'attraction de la loi de Fréchet (l'ouvrage de Galambos [36] est la référence en ce domaine). Puis nous présentons en *section* 3.4 l'application de cette théorie probabiliste en statistique. Nous passons en revue les principaux outils statistiques (en nous référant à l'ouvrage de Embrechts [30]) avant de développer un outil graphique original de caractérisation et d'estimation de l'indice de décroissance algébrique. Nous comparons, sur des données simulées, la méthode proposée avec des méthodes paramétriques (maximum de vraisemblance ou moments) et des méthodes semi-paramétriques. En *section* 3.5, on estime le paramètre de décroissance algébrique de quelques longues séries pluviométriques à différents pas de temps (à pas journalier en une station de Dédougou (Burkina-Faso), à pas horaire sur une série d'Alabama, et à pas égal à la minute pour une station de l'Ile de la Réunion), ainsi qu'en divers sites (232 longues séries annuelles à travers le monde). Les résultats obtenus sont ensuite confrontés aux estimations découlant du traitement brut de séries de basculements d'auget en *section* 3.6, en adaptant au préalable les estimateurs à ces séries.

## 3.2 Notations

Dans tout ce chapitre, on adoptera les notations et définitions suivantes :

**Notation 10** *Considérant une suite de N vap (variables aléatoires positives)* $(X_i)_{i=1\cdots N}$ *indépendantes et identiquement distribuées (iid) de* ***densité*** $f$*, on notera :*

$$(X_i)_{i=1\cdots N} \overset{iid}{\sim} f$$

*et* $M_N$ *le maximum de ce* $N-$*échantillon*

$$M_N = \max_{i=1\cdots N} (X_i)$$

**Hypothèse :** Cette densité $f$, à support dans $\mathbb{R}_+$, sera supposée ici intégrable contre $t^2$ et de primitive intégrable : $\int_1^{+\infty} t^2 f(t)dt < \infty$ et $\int_1^{+\infty}\int_s^{+\infty} f(t).ds.dt < \infty$

**Définition 11** *Sous cette hypothèse, on définit sur* $\mathbb{R}_+^*$ *les fonctions suivantes :*

***fonction de survie*** *:* $G(s) = \int_0^{+\infty} f(t).1_{t>s}dt = \int_s^{+\infty} f(t)dt$

***fonction de répartition*** *:* $F(s) = \int_0^{+\infty} f(t).1_{t\leq s}dt = \int_0^{s} f(t)dt$

***cumul de dépassement*** $H$ *:* $H(s) = \int_0^{+\infty} t.f(t).1_{t>s}dt = sG(s) + \int_s^{+\infty} G(t)dt$

**Notation 12** *La notation* $\lfloor x \rfloor$ *désignera la partie entière de* $x$*.*

**Notation 13** *La notation* $1_E$ *désignera la fonction indicatrice de* $E$ *:*

$$1_E(x) = \begin{cases} 1 & si\ x \in E \\ 0 & si\ x \notin E \end{cases}$$

## 3.3 Rappels sur la théorie des valeurs extrêmes

La modélisation des extrêmes en hydrologie constitue un problème complexe car plus les données sont irrégulières, plus l'estimation de leurs maxima est entachée d'erreurs. On sait que la qualité des outils statistiques classiques d'estimation et de prévision se dégrade quand la variabilité et/ou la dépendance au sein de la série étudiée augmente. Il est donc nécessaire de renforcer le socle théorique pour réduire et quantifier cette erreur. A partir d'une série d'observations $(X_1, ..., X_N)$, l'hydrologue cherche à quantifier, de façon relative, le "caractère extrême" de l'une d'entre elles. Par exemple, la cote atteinte par la crue de 1910 à Paris dépasse de 2 mètres celle de 1876 ; dans quelle mesure peut-on dire que cette crue est exceptionnelle ? On est donc amené à s'intéresser à la distribution des données, ou tout au moins à quelques caractéristiques d'intérêt de cette distribution.

Selon Galambos [36], la théorie des valeurs extrêmes semble de prime abord contradictoire : en effet elle vise à "comprendre la régularité du comportement extrême". Mais sous certaines hypothèses, elle fournit un cadre de modélisation simple que nous présentons dans cette section.

### 3.3.1 Loi des extrêmes

On rappelle dans cette section les principaux résultats de la théorie des extrêmes (Galambos [36]). On cherche à savoir s'il est possible de caractériser, à une normalisation près, les comportements des extrêmes d'un échantillon d'une loi quelconque. On s'intéresse à la distribution asymptotique du maximum $M_N$ de $N$ *vap* $(X_i)_{i=1\cdots N}$ indépendantes et identiquement distribuées, que l'on notera de façon générique $X$. Soit $F$ la fonction de répartition de $X$. On sait que la fonction de répartition $F_N$ de $M_N$ s'écrit, pour tout $x > 0$ :

$$F_N(x) = F^N(x)$$

ce qui indique que, lorsque $N$ tend vers l'infini, le maximum $M_N$ tend vers la borne à droite $x_F$ du support de la fonction de répartition $F$ ($x_F$ est éventuellement infini)[1] :

$$x_F = \sup\{x \in \mathbb{R}, F(x) < 1\}$$

Dans la théorie des valeurs extrêmes, on cherche à trouver, pour le maximum $M_N$ convenablement normalisé, quelles sont les lois limites possibles quand $N$ tend vers l'infini.

**Définition 14** *Les lois de Fréchet, Weibull et Gumbel sont respectivement définies par les fonctions de répartition (voir figure 3.1) :*

$$\begin{aligned}
\phi_\alpha(x) &= \begin{cases} 0 \ si\ x \le 0 \\ \exp\left[-x^{-\alpha}\right] \ si\ x > 0 \end{cases} \ avec\ \alpha > 0 \\
\psi_\alpha(x) &= \begin{cases} \exp\left[-(-x)^{-\alpha}\right] \ si\ x \le 0 \\ 1 \ si\ x > 0 \end{cases} \ avec\ \alpha > 0 \\
\Lambda(x) &= \exp\left[-e^{-x}\right] \ pour\ x \in \mathbb{R}
\end{aligned}$$

**Définition 15** *On dit qu'une distribution $f$ appartient au domaine d'attraction maximal d'une loi extrême $H$ s'il existe des suites de réels $c_N$ et $d_N$ telles que*

$$\begin{aligned}
\frac{M_N - d_N}{c_N} &\xrightarrow[N\to\infty]{L} H \\
avec\ M_N &= \max_{1\le i\le N} X_i \quad et\ X_i \overset{iid}{\sim} f
\end{aligned}$$

**Théorème 16** *(de Fisher-Tippet). Lois limites pour les maxima : Soit $(X_N)_N$ un échantillon de vap de loi $f$. S'il existe des suites de réels $c_N$ et $d_N$ et une loi extrême $H$ telle que $f$ appartienne au domaine d'attraction maximal de $H$, alors $H$ est l'une des trois distribution Fréchet, Weibull ou Gumbel.*

---

[1] En effet,

$$\forall x < x_F, \quad F_N(x) \overset{N\to\infty}{\longrightarrow} 0$$

et dans le cas où $x_F < \infty$,

$$\forall x \ge x_F, \quad F_N(x) = 1$$

donc $M_N$ tends vers $x_F$ presque sûrement, avec $x_F \le \infty$.

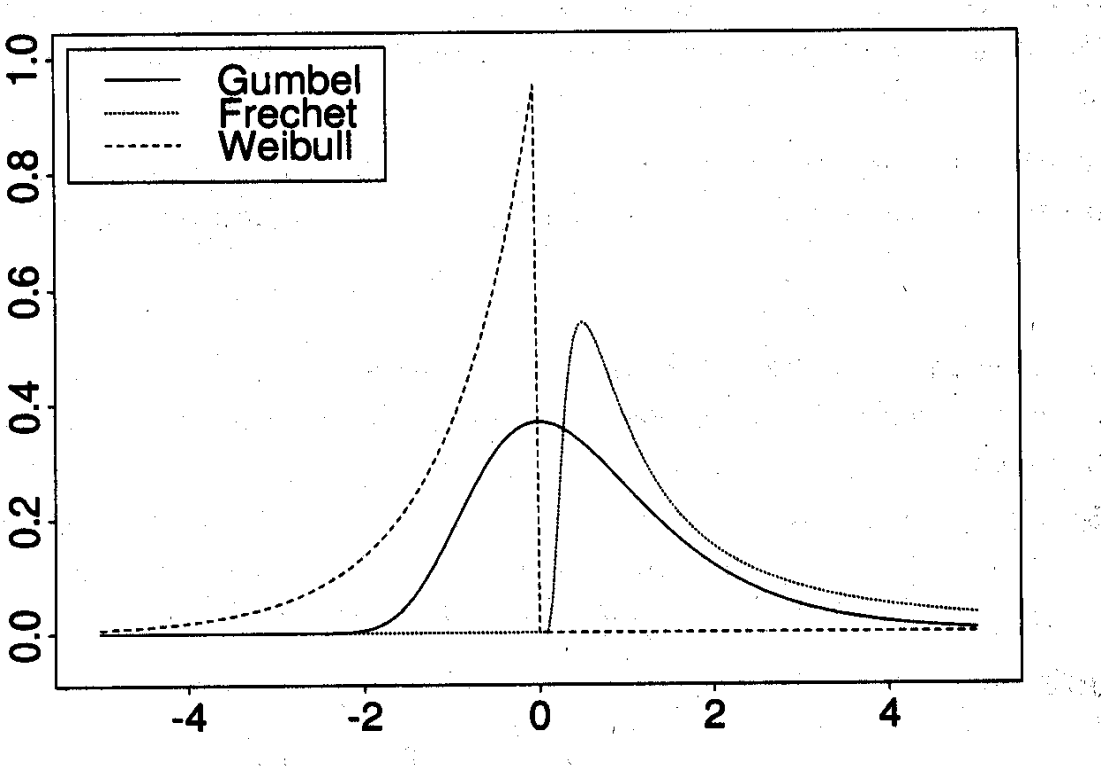



Figure 3.1: Densité de probabilité des lois de Gumbel, Fréchet et Weibull (Embrechts et al. [38]).

Mais il est possible de synthétiser ces trois lois extrêmes en une seule famille :

**Définition 17** *(Représentation de Jenkinson et Von Mises). On définit la loi généralisée des valeurs extrêmes (loi GVE) par :*

$$H_\xi(x) = \begin{cases} \exp\left[-(1+\xi x)^{-1/\xi}\right] & si\ \xi > 0 \\ \exp\left[-\exp(-x)\right] & si\ \xi = 0 \end{cases} \quad avec \quad 1+\xi x > 0$$

On montre que les cas $\xi > 0$, $\xi = 0$ et $\xi < 0$ correspondent respectivement aux distributions Fréchet $\phi_{\frac{1}{\xi}}$, Gumbel $\Lambda$ ou Weibull $\psi_{\frac{-1}{\xi}}$.

La loi *GVE* fournit donc un moyen efficace d'unification des trois lois extrêmes. Son introduction a été motivée par des considérations pratiques, car on peut estimer le paramètre $\xi$ (par le maximum de vraisemblance par exemple), et faire ainsi de l'inférence paramétrique dans un cadre non paramétrique.

### 3.3.2 Le domaine d'attraction de la loi de Fréchet

Réciproquement, étant donnée une loi $H$ appartenant à l'un des trois types Fréchet, Gumbel ou Weibull, quelles conditions imposer à la loi $F$ pour que le maximum $M_N$ converge vers $H$ ? Différentes caractérisations des trois domaines d'attraction (Fréchet, Gumbel, Weibull) sont proposées (voir [36]) mais dans ce travail, on ne détaillera que le domaine d'attraction de la loi de Fréchet (en raison de son potentiel d'applications dans le domaine de l'hydrologie stochastique). Dans cette section, on commence par définir quelques familles de lois avant de présenter quelques conditions suffisantes d'appartenance au domaine d'attraction de Fréchet.

#### 1- Fonction moyenne des dépassements et lois algébriques

**Définition 18** *Soit $X$ une vap telle que $EX < \infty$. La fonction moyenne des dépassements (FMD) est définie par*

$$\begin{array}{rccl} e: & [0,\infty[ & \longrightarrow & \mathbb{R} \\ & s & \longmapsto & e(s) = E(X-s|X>s) \end{array}$$

Cette fonction quantifie en quelque sorte la vitesse de décroissance de la queue de la distribution et donc l'irrégularité de la variable aléatoire.

On peut utiliser aussi la fonction moyenne des dépassements relatifs $h$ (*FMDR*) définie pour tout $s > 0$ par :

$$h(s) = \frac{e(s)}{s}$$

Rappelons deux propositions qui établissent une relation bi-univoque entre la *FMD* $e$ et la fonction de survie $G$, garantissant ainsi une équivalence entre l'étude de ces deux fonctions.

**Proposition 19** *Pour tout $s \geq 0$, $e(s) = \dfrac{H(s)}{G(s)} - s$ ([2])*

[2]En effet, pour toute vap $Y$ telle que $E(Y) < \infty$, $E(Y) = \int_0^\infty G_Y(y)dy$

**Proposition 20** $\forall s \geq 0 \quad G(s) = e(0).\exp\left\{-\int_0^s \frac{du}{e(u)}\right\}$ [3]

Le tableau suivant fournit les *FMD* de quelques distributions standards.

Pour une loi exponentielle $Exp(\lambda)$, la *FMD* est une fonction constante (égale à $\frac{1}{\lambda}$). Dès qu'une distribution attribue plus de poids que la loi exponentielle aux observations au delà d'un certain seuil $s_0$, sa *FMD* est croissante sur le domaine $[s_0, +\infty[$. Elle est d'autant plus croissante que ce poids est lourd. Les lois de type Pareto ont typiquement une *FMD* qui converge vers 0.

**Définition 21** *On dit qu'une loi est de type algébrique de coefficient $q > 1$ quand sa fonction de survie au delà de 1 s'écrit :*

$$\forall s \geq 1 \quad G(s) = r(s).s^{-q}$$

*avec $r$ vérifiant :*

$$\begin{cases} (C1)\ r(s) \overset{s\to\infty}{\longrightarrow} K \text{ et } r \text{ croissante sur } [1, +\infty[ \\ (C2)\ \forall s \geq 1 \qquad r'(s) \leq As^{-\theta} \text{ avec } A > 0 \text{ et } \theta > 1 \end{cases}$$

La *FMDR* d'une loi de type algébrique (voir *section 3.3.2*) est décroissante sur $[1, +\infty[$[4].

De plus, pour une loi de type algébrique de paramètre $q$, la *FMDR* $h$ tend vers $\frac{1}{q-1}$ à l'infini. On a même équivalence sous certaines conditions (pour la démonstration, on pourra se reporter à l'*Annexe B*) :

**Proposition 22** *Une distribution est de type algébrique si sa FMDR vérifie les deux conditions suivantes :*

$$\begin{cases} (C3)\ \forall s \geq 1 \qquad 0 \leq q - 1 - \frac{1}{h(s)} \leq Cs^{-\nu} \text{ avec } C > 0 \text{ et } \nu > 0 \\ (C4)\ \forall s \geq 1 \qquad 0 \leq -h'(s) \leq Ds^{-\theta'} \text{ avec } D > 0 \text{ et } \theta' > 1 \end{cases}$$

---

Comme : $\forall x > 0 \quad G_{X-a|X>a}(x) = \frac{P(X > x+a)}{P(X > a)} = \frac{G(x+a)}{G(a)}$

$\forall a \geq 0 \quad e(a) = \int_0^\infty G_{X-a|X>a}(x)dx$ ce qui nous conduit au résultat.

[3]En effet : $\forall a \geq 0 \quad \int_0^a \frac{du}{e(u)} = \int_0^a \frac{G(u)}{\int_u^\infty G(v)dv} du = \left[-\log\left(\int_u^\infty G(v)dv\right)\right]_0^a$

Donc $\forall a \geq 0 \quad \int_0^a \frac{du}{e(u)} = \log e(0) - \log\left[e(a)G(a)\right]$

[4]En effet : $\forall s' \geq s \geq 1 \quad h(s) - h\left(s'\right) = \int_s^{s'} \frac{G(t)}{s^{1-q}.r(s)}dt \geq \int_s^{s'} \frac{G(t)}{K}dt \geq 0$

| Nom | Fonction de survie $G(x)$ | FMD $e(s)$ | $\lim_{s\to\infty} h(s)$ |
|---|---|---|---|
| Uniforme | $(1-x).1_{0<x<1}+1_{x<0}$ | $\frac{1}{2}(1-s)$ | 0 |
| Benktander II $\begin{cases}\alpha>0\\0<\beta<1\end{cases}$ | $x^{-(1-\beta)}.\exp\left(-\frac{\alpha}{\beta}x^{\beta}\right)$ | $\frac{1}{\alpha}s^{1-\beta}$ | 0 |
| Weibull $\begin{cases}\lambda>0\\0<\tau<1\end{cases}$ | $\exp\left(-\lambda x^{\tau}\right)$ | $\frac{s^{1-\tau}}{\lambda\tau}\left[1+o\left(1\right)\right]$ | 0 |
| Exponentielle $\lambda>0$ | $\exp\left(-\lambda x\right)$ | $\frac{1}{\lambda}$ | 0 |
| Gamma $m,\lambda>0$ | $\frac{\lambda^m}{\Gamma(m)}\int_x^{\infty}u^{m-1}\exp\left(-\lambda u\right)du$ | $\frac{1}{\lambda}+\frac{m-1}{a\lambda^2}\left[1+o\left(1\right)\right]$ | 0 |
| Logistique | $\frac{2}{1+\exp(x)}$ | $\left(1+e^{s}\right)\log\left(1+e^{-s}\right)$ | $\infty$ |
| lognomale $\begin{cases}\mu\in R\\\sigma>0\end{cases}$ | $\int_x^{\infty}\frac{1}{\sqrt{2\pi\sigma u}}\exp\left(\frac{1}{2\sigma^2[\log u-\mu]^2}\right)du$ | $\frac{s}{\log s}\left[1+o\left(1\right)\right]$ | 0 |
| Benktander I $\alpha,\beta>0$ | $\gamma x^{-(1+\alpha+\beta\log x)}.(\alpha+2\beta\log x)$ | $\frac{s}{\alpha+2\beta\log s}$ | 0 |
| Pareto $\alpha>1$ | $x^{-\alpha}$ | $\frac{s}{\alpha-1}$ | $\frac{1}{\alpha-1}$ |
| Burr $\beta,\lambda,\tau>0$ | $\left(\frac{\beta}{\beta+x^{\tau}}\right)^{\lambda}$ | $\frac{(s^{\tau}+\beta)^{\frac{1}{\tau}}}{\lambda\tau-1}\left[1+o\left(1\right)\right]$ | $\frac{1}{\lambda\tau-1}$ |
| Pareto gén. $\alpha,\beta>0$ | $\left(1+\frac{\xi x}{\beta}\right)^{-\frac{1}{\xi}}$ | $\frac{\xi s}{1-\xi}\left[1+o\left(1\right)\right]$ | $\frac{\xi}{1-\xi}$ |
| Loggamma $m,\lambda>0$ | $\frac{\lambda^m}{\Gamma(m)}\int_x^{\infty}(\log u)^{m-1}.u^{-\lambda-1}du$ | $\frac{s}{\lambda-1}\left[1+o\left(1\right)\right]$ | $\frac{1}{\lambda-1}$ |
| Log-logistique $\alpha>1$ | $1/\left(1+\beta x^{\alpha}\right)$ | $\frac{s}{\alpha-1}\left[1+o\left(1\right)\right]$ | $\frac{1}{\alpha-1}$ |
| Fréchet $\alpha>1$ | $1-\exp\left(-x^{-\alpha}\right)$ | $\frac{s}{\alpha-1}\left[1+o\left(1\right)\right]$ | $\frac{1}{\alpha-1}$ |

Tableau 3.1: *FMD* pour quelques lois.

**2- Propriétés**

On montre (Galambos, 78 [36] et David, 81 [24] *p.260*) que le domaine d'attraction de la loi de Fréchet est constitué entre autres :

- des lois $f$ satisfaisant à la condition dite de Von Mises :

$$\lim_{x\to\infty} \frac{xf(x)}{G(x)} = \alpha > 0$$

(où $G$ désigne la fonction de survie).

- des lois dites de type Pareto, c'est-à-dire dont la fonction de survie décroît algébriquement au voisinage de l'infini[5] :

$$G(x) = l(x).x^{-\alpha} \text{ avec } l \text{ à variations lentes ie : } \frac{l(\lambda y)}{l(y)} \overset{y\to\infty}{\longrightarrow} 1 \quad \forall\lambda > 0$$

Le schéma 3.2 détaille quelques conditions suffisantes d'appartenance au domaine d'attraction de la loi de Fréchet.

**3- Exemples de lois appartenant au domaine de Fréchet**

Les exemples les plus simples de lois appartenant au domaine de Fréchet sont les lois de Pareto stricte et généralisée.

**Définition 23** *On dit que $X$ suit une loi de Pareto stricte de paramètre $\alpha$ $Pa(\alpha)$ si et seulement si sa fonction de survie s'écrit :*

$$\forall x \geq 1 \quad G(x) = x^{-\alpha}$$

On vérifie facilement que cette loi satisfait la condition de Von Mises. Elle est simple d'utilisation car elle se ramène à la loi exponentielle par transformation logarithmique[6].

Le support de la loi de Pareto étant trop particulier ($[1, +\infty[$), on doit introduire des paramètres supplémentaires pour modifier la forme de cette distribution. La prise en compte d'un paramètre de localisation $\nu$ peut se faire en considérant les données translatées $(X_i - \nu)_{1\leq i\leq N}$.

**Définition 24** *On introduit un paramètre d'aplatissement $\beta$ dans la loi définie précédemment pour définir la Pareto généralisée à deux paramètres $PG(\xi, \beta)$ :*

$$G_{\xi,\beta}(x) = \begin{cases} \left(1 + \xi\frac{x}{\beta}\right)^{-1/\xi} & si\ \xi \neq 0 \\ 1 - \exp\left(-\frac{x}{\beta}\right) & si\ \xi = 0 \end{cases} \quad avec \begin{cases} x \geq 0\ si\ \xi \geq 0 \\ 0 \leq x \leq -\beta/\xi\ si\ \xi < 0 \end{cases}$$

[5]Un développement de Taylor au voisinage de l'infini de la fonction de survie de Fréchet permet d'avoir une visualisation rapide de cette décroissance algébrique :

$$1 - \Phi_\alpha(x) = 1 - \exp\left(-x^{-\alpha}\right) \overset{x\infty}{\sim} x^{-\alpha}$$

[6]Si $X \sim Pa(\alpha)$ alors $Y = \log X \sim Exp(\alpha)$. En effet : $\forall y \geq 0 \quad P(Y > y) = P(X > e^y) = e^{-\alpha y}$ car $e^y \geq 1$

Cette loi possède un paramètre $\xi$ invariant par homothétie[7], résultat utile lorsque l'on change l'unité de mesure des observations. Mais l'emploi de la loi de Pareto généralisée pour la modélisation des dépassements est surtout justifié par la proposition suivante ([30], p.165).

**Proposition 25** *Soit $F_u$ la fonction de répartition des dépassements de $u$ :*

$$F_u(s) = P(X - u \leq s | X > u)$$

*$F_u$ appartient au domaine d'attraction de $H_\xi$ si et seulement si :*

$$\lim_{u \to \infty} \sup_{s \geq 0} \left| F_u(s) - G_{\xi,\beta(u)}(s) \right| = 0$$

*pour une fonction positive $\beta$.*

En pratique, on translate les données de $\nu$, que l'on prend égal au seuil de dépassement $u$. On peut mentionner aussi d'autres lois telles que la loi log-gamma ou la loi de Fréchet elle-même, qui appartiennent au domaine d'attraction de Fréchet, et sont utilisées en hydrologie. La loi log-logistique et la loi Burr (Embrechts et al., [30]) sont utilisées en assurances ou en finances (voir tableau 3.1).

Les lois du domaine d'attraction de Fréchet possèdent quelques propriétés souvent constatées en pratique sur les séries hydrologiques, telles qu'une divergence des moments à partir d'un certain ordre (cet ordre est le coefficient de décroissance algébrique), ou bien une faible vitesse de convergence du Théorème de la limite centrale (propriété qui est liée à la variance de la série, puisque la vitesse de convergence du théorème est inversement proportionnelle à cette dernière). Elles sont d'interprétation physique simple, puisqu'elles découlent des modèles multifractals (voir section 2.3.3).

## 3.4 Estimation du comportement algébrique

Disposant d'un ensemble d'observations $(X_i)_{i=1\cdots N}$, l'analyste cherche à appréhender le type de comportement des données extrêmes. Pour cela, il doit répondre, dans l'ordre, aux deux questions suivantes :

- dans quel domaine d'attraction est-il raisonnable de se placer *?*
- quelle loi utiliser ?

[7] En effet : si $X \sim PG(\beta, \xi)$ et si $Y = \Delta X$, alors $Y \sim PG\left(\beta', \xi\right)$ avec $\beta' = \Delta\beta$. En effet, sa fonction de survie s'écrit : $G_Y(s) = G_X\left(\frac{s}{\Delta}\right) = \begin{cases} \left(1 + \xi\frac{s}{\beta'}\right)^{-1/\xi} & \text{si } \xi \neq 0 \\ 1 - \exp\left(-\frac{s}{\beta'}\right) & \text{si } \xi = 0 \end{cases}$ avec $\begin{cases} s \geq 0 \text{ si } \xi \geq 0 \\ 0 \leq s \leq -\beta'/\xi \text{ si } \xi < 0 \end{cases}$

Il aura recours à des outils statistiques d'estimation, mais aussi à des outils d'exploration afin d'avoir une visualisation du comportement asymptotique de la loi. Dans cette section, on présente les principaux outils statistiques adaptés à l'exploration, puis à l'estimation, des valeurs extrêmes, avant de proposer une procédure originale de détection et d'estimation du comportement algébrique.

Tout au long de cette section, nous supposerons que les observations $(X_i)_{i=1\cdots N}$ sont des *vap* indépendantes et identiquement distribuées de densité $f$.

### 3.4.1 Revue d'outils statistiques d'exploration

Avant de s'engager dans une étude statistique détaillée des données, il est important d'en avoir une visualisation. Les outils d'analyse exploratoire des extrêmes sont nombreux et constituent un centre d'intérêt commun à plusieurs disciplines : séries financières, ou en assurances (Chambers et al. 83 [31], Cleveland 93 [19] ou Tufts 83 [83]). Nous présentons dans ce paragraphe les méthodes les plus classiques que sont le graphe de la fonction de survie, les graphes de quantiles (ou de probabilités), le graphe de la fonction moyenne des dépassements, et l'estimateur de Hill.

#### Graphe de la fonction de survie

L'histogramme est fréquemment utilisé pour estimer directement la distribution $f$, mais la méthode la plus simple pour visualiser la queue d'une distribution consiste à reporter sur un graphe l'estimateur de sa fonction de survie $G$. Les estimateurs utilisés dans la plupart des études à application hydrologique sont :

- La fréquence empirique des dépassements :

$$\widehat{G_0}(s) = \frac{1}{N}\sum_{i=1}^{N} 1_{X_i>s}$$

  qui est un estimateur non biaisé de la fonction de survie $G(s)$ au delà de $s$.

- L'estimateur de Weibull : si $i$ est le rang de l'observation $X_{(i)}$, il s'écrit :

$$\widehat{G_W}(X_{(i)}) = \frac{N+1-i}{N+1}$$

Dans un modèle de Pareto strict $Pa(\alpha)$[8], on peut proposer deux estimateurs du paramètre $\alpha$. Le plus simple s'écrit :

$$\widehat{\alpha} = -\frac{\log \widehat{G_0}\,(s)}{\log\,(s)}$$

[8]Ces estimateurs peuvent aussi être utilisés dans un modèle de type algébrique (voir *section 3.2*). En effet, si $G\,(s) = r\,(s)\,s^{-\alpha}$ alors $-\dfrac{\log G\,(s)}{\log\,(s)} \overset{s\to\infty}{\longrightarrow} \alpha$.

Cet estimateur est consistant[9], car les $(X_i)_{i=1\cdots N}$ sont supposées *iid,* et par la loi forte des grands nombres on a :

$$\frac{1}{N}\sum_{i=1}^{N} 1_{X_i>s} \overset{N\to\infty}{\longrightarrow} G(s) \text{ et } \widehat{\alpha} \overset{N\to\infty}{\longrightarrow} -\frac{\log G(s)}{\log(s)} = \alpha \quad \forall s \geq 0$$

Cependant, il induit un biais dans l'estimation de $\alpha$, ce dernier étant systématiquement négatif[10]. Pour l'estimateur de Weibull, les résultats concernant le biais et la consistance dans un modèle de Pareto stricte sont identiques.

Le paramètre $\alpha$ d'une loi de Pareto peut aussi être estimé par la pente de la droite de régression ajustée au graphe log-log de l'un des estimateurs de $G$. Mais lorsque la série de variables aléatoires est très irrégulière, cette méthode est mal adaptée, car les graphes présentent des paliers horizontaux (ou dans le cas de la formule de Weibull, trop peu de points sur la zone d'ajustement de la droite de régression). En effet, soient $s_1$ et $s_2$ deux seuils entre lesquels on ne dispose d'aucune observation. Le graphe log-log de $\widehat{G_0}$ présentera un palier horizontal sur $[s_1, s_2]$ tandis que celui de $\widehat{G_W}$ ne représentera que la borne droite de ce palier (figure 3.3). Ces paliers ne traduisent donc que notre manque d'information sur le phénomène et la difficulté d'interprétation des mesures (intensité calculée selon un pas de temps trop grand, série trop courte, mal tronquée, ou relevés de hauteurs à instants fixes) (voir [11]).

En conclusion, l'utilisation de la fonction de survie seule (pour l'estimation du paramètre de Pareto) est insuffisante, soit à cause de l'erreur de régression, soit à cause du biais.

Présentons à présent trois outils d'exploration graphiques classiques : les graphes de quantile et probabilité, le graphe de la fonction moyenne des dépassements et le graphe de l'estimateur de Hill. On verra que ces méthodes sont reliées les unes aux autres par une simple transformation.

### Les graphes de quantiles et de probabilités[12]

Disposant d'observations $(X_i)_{1\leq i\leq N}$, on aimerait voir graphiquement si une distribution donnée (dont on note $F$ la fonction de répartition) s'ajuste bien à cette série. Pour cela, on se place sous l'hypothèse $H$ qu'elles sont *iid* de fonction de répartition $F$

[9] Un estimateur $\widehat{E}_N$ de $E$ est dit consistant si $\widehat{E}_N \overset{N\to\infty}{\longrightarrow} E$.

[10] En effet, par l'inégalité de Jensen :

$$E\left[\log \widehat{G_0}(s)\right] \geq \log\left[E\widehat{G_0}(s)\right] \text{ et } E(\widehat{\alpha}) \leq \alpha$$

[11] En effet, même si l'on dispose d'une longue série d'observations, l'ajustement de la droite de régression sur la queue de distribution ne sera fait que sur un petit nombre de points si cette série est très irrégulière.

[12] En anglais $qq-plot$ et $pp-plot$.

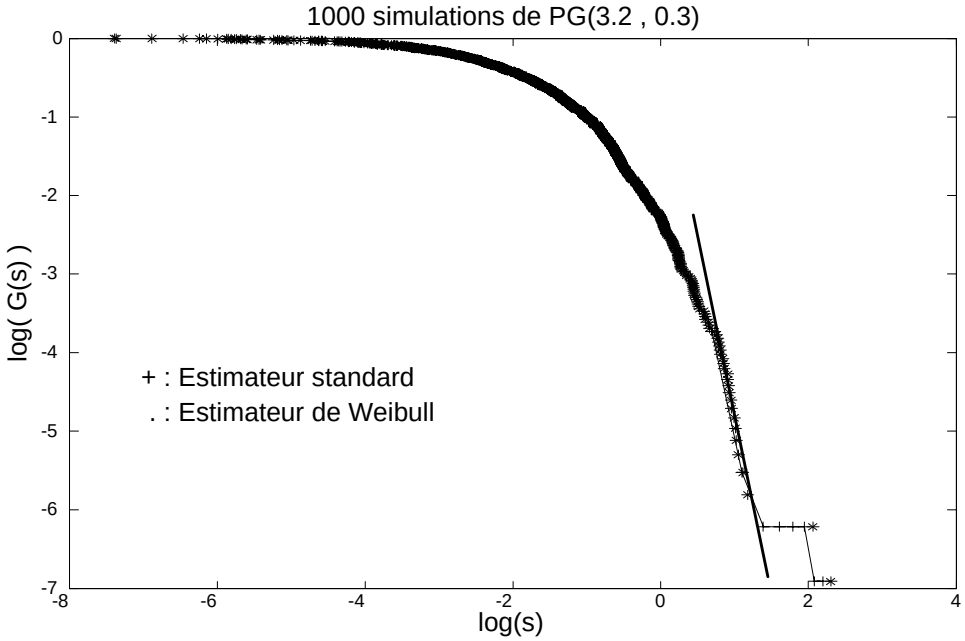



Figure 3.3: Comparaison des estimateurs de Weibull et standard de la fonction de survie.

(inversible). Un calcul[13] montre que :

$$E\left[F(X_i^*)\right] = \frac{i}{N+1}$$

avec $(X_i^*)_{1\leq i\leq N}$ le $N$-échantillon classé par ordre croissant. L'idée des graphes de probabilités (pour la loi $F$) est de reporter sur un graphe les points de coordonnées :

$$\left(F(X_i^*), \frac{i}{N+1}\right)$$

Pour les graphes de quantiles, le principe est identique : on représente les variations des $X_i^*$ en fonction des $F^{-1}\left(\frac{i}{N+1}\right)$.

La courbure du graphe renseigne sur la validité de l'hypothèse $H$ : si la courbe est à peu près linéaire, l'hypothèse $H$ est plausible, et si le graphe devient concave (resp. convexe) vers la droite, la distribution $F$ attribue trop (resp. pas assez) de poids aux grandes observations (figure 3.4).

[13]Notons $U_i = F(X_i)$. On a : $(U_i)_{i=1\cdots n} \overset{iid}{\sim} U(0,1)$. Soit $(U_i^*)_{1\leq i\leq n}$ le $n$-échantillon classé par ordre croissant. On a :

$$\forall i \in \{1,\cdots,n\}, \quad F(X_i^*) \overset{L}{=} U_i^*$$

et le résultat découle du fait que : $E\left[U_i^*\right] = \frac{i}{n+1}$.

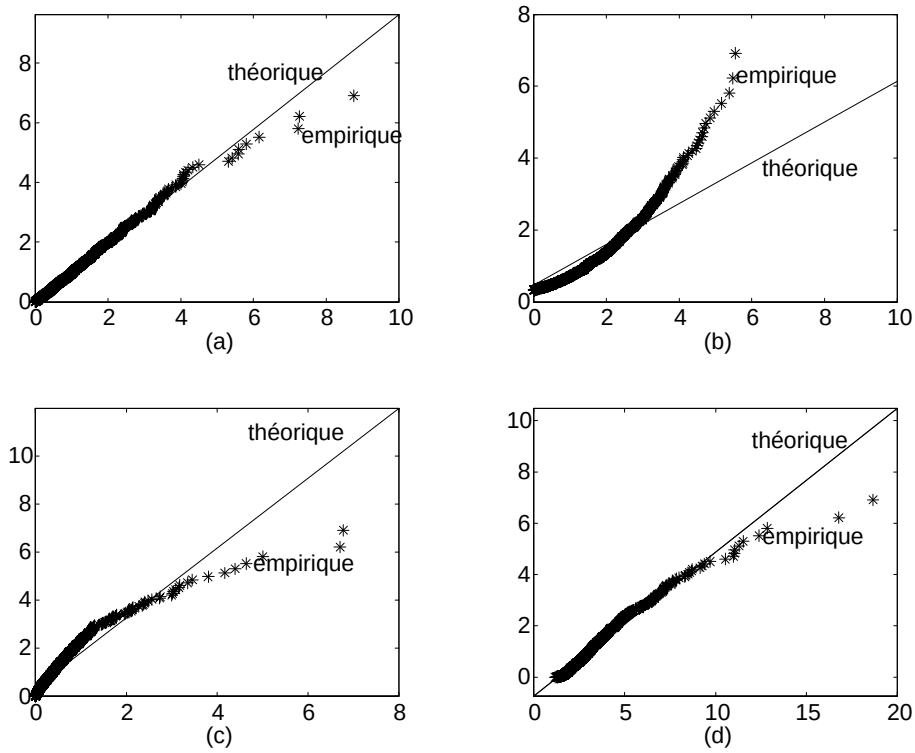



Figure 3.4: Graphe de quantiles d'un 1000-échantillon de loi exponentielle comparé à une loi exponentielle (a), normale (b), pareto (c) et log-gamma (d).

Il est à noter qu'un changement des paramètres de forme et de localisation n'affecte pas la linéarité du graphe, mais seulement la valeur de la pente et du coefficient à l'origine.

Graphes de quantiles et de probabilités pour quelques lois :

- Loi exponentielle : La fonction de répartition d'une *vap* de loi exponentielle de paramètre 1 s'écrit : $F(x) = 1 - e^{-x} \quad \forall x \geq 0$. On représentera donc les points :

$$\left(X_i^*, -\log(1 - \frac{i}{N+1})\right)_{1 \leq i \leq N}$$

  Si l'on veut réaliser un graphe de quantiles de comparaison à une loi exponentielle de paramètre $\alpha$, alors on peut se ramener au cas précédent en considérant les va $\alpha X_i$ qui suivent une loi exponentielle de paramètre 1

$$\left(\alpha X_i^*, -\log(1 - \frac{i}{N+1})\right)_{1 \leq i \leq N}$$

- Loi de Pareto : Vérifier que $(X_1, ..., X_N)$ suit une loi de Pareto de paramètre $\alpha$ revient à vérifier que $(\log X_1, ..., \log X_N)$ suit une loi exponentielle de paramètre $\alpha$. On représentera donc les points :

$$\left(\alpha \log(X_i^*), -\log(1 - \frac{i}{N+1})\right)_{1 \leq i \leq N}$$

**La fonction moyenne des dépassements**

La fonction moyenne des dépassements (*FMD*), que nous avons définie en *section 3.3.2*, fournit un moyen efficace d'exploration des données. Le principe de la méthode de la *FMD* consiste à comparer graphiquement la *FMD* estimée aux *FMD* théoriques (voir tableau de la *section 3.3.2* et figure 3.6).

Disposant d'un $N$-échantillon $(X_i)_{1 \leq i \leq N}$, la *FMD* en un seuil $s \geq 0$ est naturellement estimée par :

$$\widehat{e_N}(s) = \frac{\sum_{i=1}^{N} (X_i - s) \, .1_{X_i > s}}{\sum_{i=1}^{N} 1_{X_i > s}}$$

Souvent, la *FMD* est estimée aux valeurs $X^*_{N-k}$, $k = 1, \cdots N-1$. Le numérateur s'écrit : $\sum_{j=1}^{k} \left(X_{N-j+1} - X^*_{N-k}\right)$ tandis que le dénominateur vaut $k$. On représente alors les variations de :

$$E_{N,k} = \widehat{e_N}(X^*_{N-k}) = \frac{1}{k} \sum_{i=1}^{N} X_{N-j+1} - X^*_{N-k}$$

Figure 3.5:

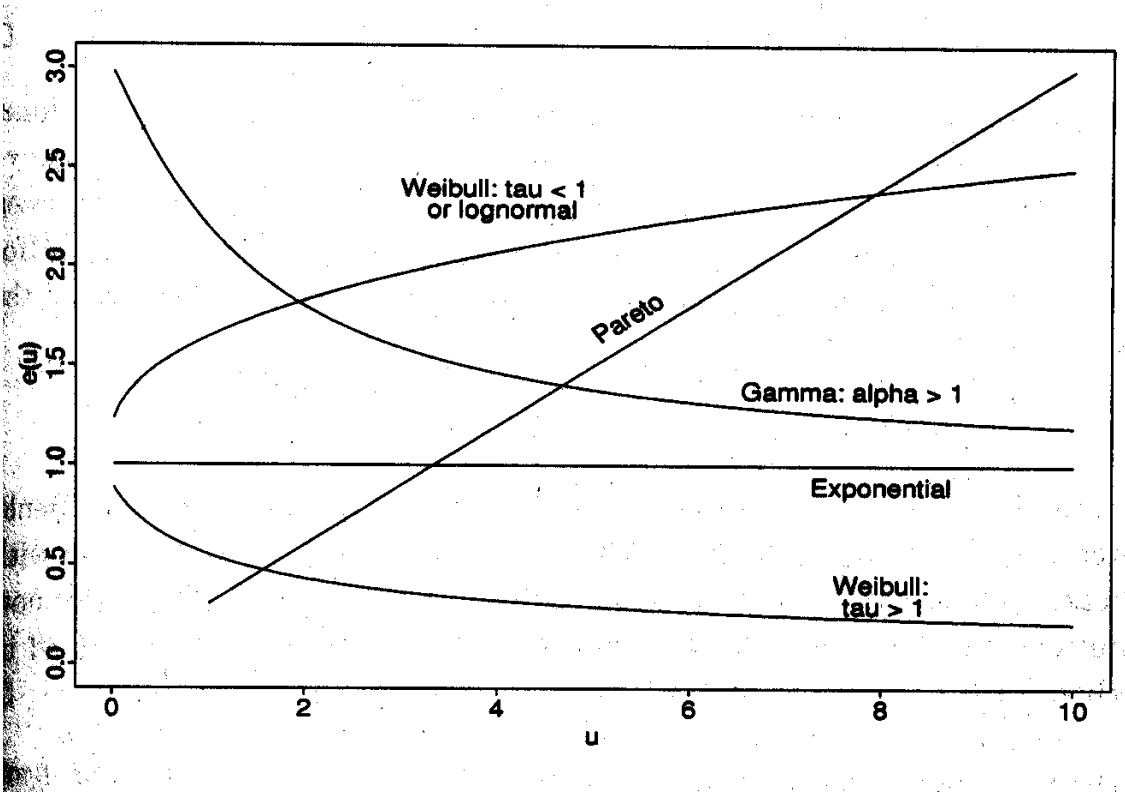



Figure 3.6: Graphes des *FMD* de quelques lois (Embrechts et al. [38]).

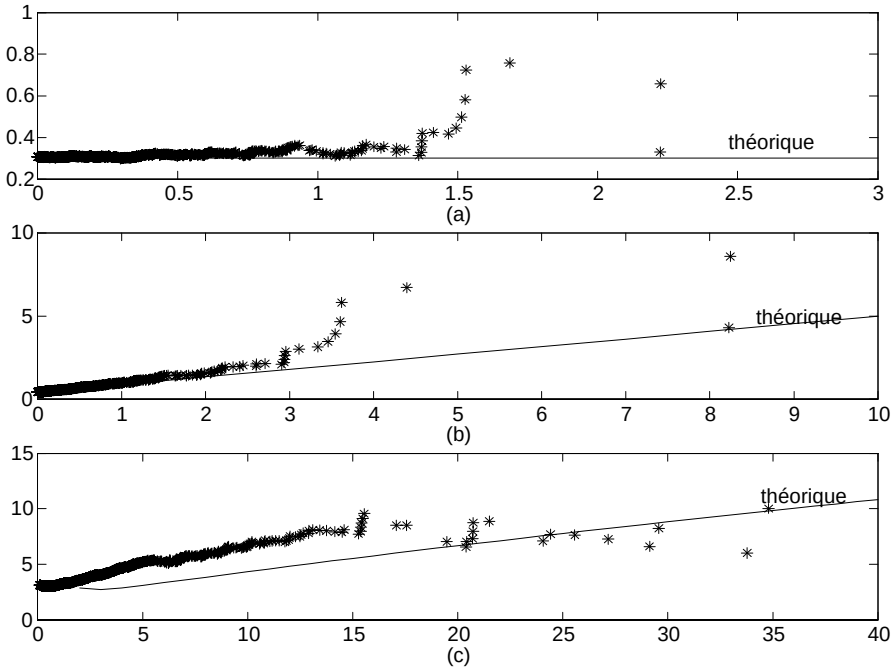



Figure 3.7: *FMD* empirique sur des 5000-échantillons de loi (a) exponentielle, (b) Pareto et (c) log-gamma.

quantité qui n'est autre que la pente à droite du point $\left(X^*_{N-k}, -\log\left(\frac{k+1}{N}\right)\right)$ du graphe de quantiles pour la loi exponentielle. Les graphes de la *FMD* sont utilisés comme un moyen graphique permettant de distinguer les modèles attribuant un grand poids aux extrêmes, des modèles leur attribuant un faible poids (voir 3.7). Les résultats sont à interpréter avec précaution en raison du faible nombre de données extrêmes et, comme on le verra en *section 3.4.3*, du biais croissant de $\widehat{e_N}(s)$.

**L'estimateur de Hill**

Cet estimateur est défini par (Embrechts et al., [30]) :

$$H_{k,N} = \frac{1}{k}\sum_{j=1}^{k} \log\left(X^*_{N-j+1}\right) - \log\left(X^*_{N-k}\right)$$

L'allure asymptotique du graphe $(k, H_{k,N})$ renseigne sur la forme de la distribution. En effet, dans un modèle de type Pareto de paramètre $\alpha$, $H_{k,N}$ tend presque sûrement vers $\frac{1}{\alpha}$. L'estimation du paramètre $\frac{1}{\alpha}$ par la limite de l'estimateur de Hill est satisfaisante dans ce modèle. Mais le graphe de l'estimateur de Hill est un mauvais outil d'exploration car $H_{k,N}$ diverge dès que l'on s'écarte du modèle Pareto (voir figure 3.8).

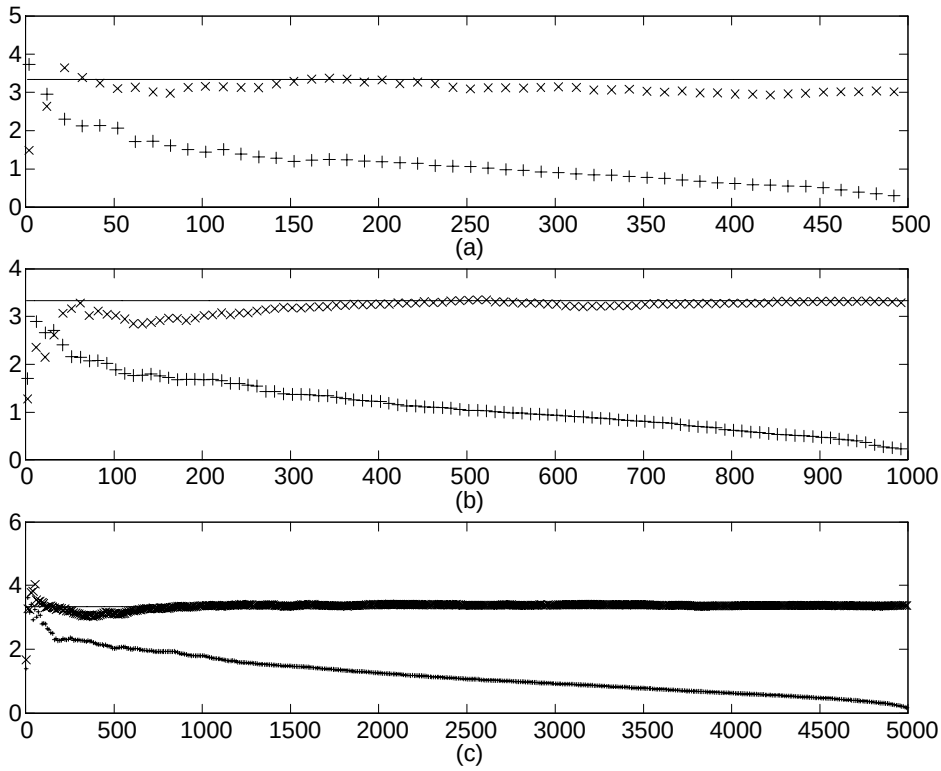



Figure 3.8: Estimateur de Hill pour $N$ simulations de lois de fonction de survie $G(s) = s^{-3.33}$ (ligne 'x') et $G(s) = (s+1)^{-3.33}$ (ligne '+') pour : (a) $N = 500$, (b) $N = 1000$ et (c) $N = 5000$.

Remarquons que l'estimateur de Hill est aussi la pente du graphe de quantiles Pareto à droite du point $\left(-\log\left(\frac{k+1}{N+1}\right), \log\left(X^*_{N-k}\right)\right)$. C'est aussi l'estimateur de la fonction moyenne des dépassements de la variable transformée $\log(X)$ en $\log\left(X^*_{N-k}\right)$.

### 3.4.2 Revue d'outils d'estimation

Dans cette section, on passe en revue les principales méthodes d'estimation paramétrique et semi-paramétrique du paramètre de décroissance de la queue de distribution d'un ensemble d'observations.

#### Revue d'outils semi-paramétriques

Les différentes méthodes d'exploration citées dans la section précedente sont capables de constituer des outils d'estimation semi-paramétrique performants lorsqu'elles sont associées à des techniques visant à améliorer la qualité de l'estimation. L'article de Beirlant

et al. [8] faire figure de référence pour une revue en détail de toutes ces méthodes. Nous nous attacherons dans ce paragraphe à décrire brièvement quelques unes de ces méthodes.

On reprend les notations de la section précédente : disposant d'un $N$-échantillon $(X_i)_{1\leq i\leq N}$, $(X_i^*)_{1\leq i\leq N}$ désignera ce $N$-échantillon classé par ordre croissant. Les graphes de quantiles, de la fonction moyenne des dépassements et de l'estimateur de Hill constituent des méthodes d'estimation de l'indice de Pareto. Les estimateurs découlant de ces méthodes appartiennnent à la classe dite des estimateurs du noyau, car ils peuvent tous s'écrire sous la forme :

$$(\widehat{\alpha}_{k,n})^{-1} = \frac{\sum_{j=1}^{k} \frac{j}{k} K\left(\frac{j}{k}\right) \cdot \left[\log\left(X^*_{n-j+1}\right) - \log\left(X^*_{n-j}\right)\right]}{\int_0^1 K(t)\, dt}$$

On sait ainsi que ces estimateurs possèdent des propriétés asymptotiques de normalité sous certaines conditions (voir l'article de Beirlant et al. [8] p. 1660 pour plus de détails et l'article de Csorgo et Viharos [23]).

Cependant, le problème du choix du seuil de troncature (ou du nombre de grandes observations à retenir) pour optimiser la qualiter de l'estimation est difficile. On sait cependant optimiser ce choix dans le cas de la méthode du graphe des quantiles Pareto, où $\alpha$ est estimé comme la pente de régression de ce graphe (par l'utilisation d'algorithmes des moindres carrés pondérés, cf [8]). Il en est de même pour les méthodes de la fonction moyenne des dépassements (voir Beirlant et al. [7]) et pour l'estimateur de Hill, puisque ces estimateurs s'obtiennent par simple transformation de l'estimateur du graphe de quantiles Pareto (on a vu dans la section précédente que l'estimateur de Hill est la pente du graphe de quantiles Pareto à droite du point $\left(-\log\left(\frac{k+1}{N+1}\right), \log\left(X^*_{N-k}\right)\right)$ ainsi que l'estimateur de la fonction moyenne des dépassements de la variable transformée $\log(X)$ en $\log\left(X^*_{N-k}\right)$).

### Méthodes paramétriques

La méthode d'estimation la plus simple consiste à se placer dans le cadre paramétrique. Supposons que sur la série des cumuls journaliers de pluie de Dédougou, on cherche à modéliser les plus hautes observations (constituant par exemple 5% de celles-ci). En tenant compte de la linéarité du graphe log-log de la fonction de survie (fig. 3.30) sur ces plus fortes valeurs, on est conduit à proposer un modèle de Pareto, ou de Pareto généralisée (voir *section* 3.3.2). Si par contre, on ne s'intéresse qu'aux maxima (par exemple, à une série de précipitations maximales annuelles), le modèle proposé sera un modèle de loi généralisée des extrêmes, puisque la représentation de Jenkinson et Von Mises (*section* 3.3.1) permet d'avoir une représentation unifiée des lois extrêmes.

**Dans le modèle loi généralisée des extrêmes :** Si l'on veut créer un modèle pour les maxima, la loi généralisée des extrêmes est un modèle pertinent. De plus, l'estimation du coefficient $\xi$ nous renseignera sur le domaine d'attraction des variables (voir *section 3.3.1*).

On se placera dans un modèle prenant en compte des paramètres de forme et de localisation :

$$H_{\xi,\mu,\psi}(x) = \begin{cases} \exp\left[-\left(1+\xi\dfrac{x-\mu}{\psi}\right)^{-1/\xi}\right] & si\ \xi > 0 \\ \exp\left[-\exp\left(-\dfrac{x-\mu}{\psi}\right)\right] & si\ \xi = 0 \end{cases} \quad \text{où } 1+\xi\frac{x-\mu}{\psi} > 0$$

1. Méthode des moments pondérés : Le principe de la méthode des moments pondérés est proche de celui de la méthode des moments. L'estimation est réalisée à partir des moments de la *vap* $X$ pondérée par $H^r_{\xi,\mu,\psi}(X)$ avec $r$ un réel positif :

$$w_r(\xi,\mu,\psi) = E\left[XH^r_{\xi,\mu,\psi}(X)\right]$$

En effet, ces derniers se calculent facilement en remarquant que :

$$\int_{-\infty}^{+\infty} xH^r_{\xi,\mu,\psi}(x)\, dH^r_{\xi,\mu,\psi}(x) = \int_{-\infty}^{+\infty} H^{-1}_{\xi,\mu,\psi}(y)\, .y^r.dy$$

Or

$$H^{-1}_{\xi,\mu,\psi}(y) = \begin{cases} \mu + \dfrac{\psi}{\xi}\left[(-\log y)^{-\xi} - 1\right] & si\ \xi > 0 \\ \mu - \psi\log(-\log y) & si\ \xi = 0 \end{cases}$$

L'intégrale converge lorsque $\xi < 1$ et $\xi \neq 0$ et :

$$w_r(\xi,\mu,\psi) = \frac{1}{r+1}\left[\mu - \frac{\psi}{\xi}\left(1-\Gamma(1-\xi)(1+r)^{\xi}\right)\right]$$

En calculant les trois premiers moments pondérés ($r = 0, 1, 2$):

$$\begin{aligned} w_0(\xi,\mu,\psi) &= \mu - \frac{\psi}{\xi}\left[1-\Gamma(1-\xi)\right] \\ w_1(\xi,\mu,\psi) &= \frac{1}{2}\left[\mu - \frac{\psi}{\xi}\left(1-\Gamma(1-\xi).2^{\xi}\right)\right] \\ w_2(\xi,\mu,\psi) &= \frac{1}{3}\left[\mu - \frac{\psi}{\xi}\left(1-\Gamma(1-\xi).3^{\xi}\right)\right] \end{aligned}$$

et en inversant le système, l'estimateur de $\xi$ est déterminé par :

$$\frac{3^{\widehat{\xi}}-1}{2^{\widehat{\xi}}-1} = \frac{3\widehat{w}_2(\xi,\mu,\psi) - \widehat{w}_0(\xi,\mu,\psi)}{2\widehat{w}_1(\xi,\mu,\psi) - \widehat{w}_0(\xi,\mu,\psi)}$$

Les estimateurs par la méthode des moments pondérés de $(\mu,\psi)$ s'écrivent :

$$\begin{cases} \widehat{\psi} = \dfrac{(2\widehat{w}_1 - \widehat{w}_0)\,\widehat{\xi}}{\Gamma\left(1-\widehat{\xi}\right).\left(2^{\widehat{\xi}}-1\right)} \\ \widehat{\mu} = \widehat{w}_0 + \dfrac{\widehat{\psi}}{\widehat{\xi}}\left[1-\Gamma\left(1-\widehat{\xi}\right)\right] \end{cases}$$

Les moments pondérés sont estimés par :

$$\widehat{w}_r\left(\xi,\mu,\psi\right)=\frac{1}{N}\sum_{i=1}^{N}X_{i,N}U_{i,N}^r$$

compte tenu du fait que$\left(H_{\xi,\mu,\psi}\left(X_{1,N}\right),\cdots,H_{\xi,\mu,\psi}\left(X_{N,N}\right)\right)\overset{l}{=}\left(U_{1,N},\cdots,U_{N,N}\right)$.

2. Méthode du maximum de vraisemblance : Contrairement aux estimateurs précédents qui ont pu être déterminés dans le modèle $GVE$, l'estimateur du maximum de vraisemblance ne peut être calculé dans le cas général. Dans le modèle de Gumbel ($\xi=0$), la densité de probabilité s'écrit :

$$h_{0,\mu,\psi}(x)=\frac{1}{\psi}\exp\left(-\frac{x-\mu}{\psi}\right)\exp\left[-\exp\left(-\frac{x-\mu}{\psi}\right)\right]$$

Il en découle la log-vraisemblance :

$$\log L_{\mu,\psi}\left(X_1,\cdots,X_N\right)=-N\log\psi-\sum_{i=1}^{N}\frac{X_i-\mu}{\psi}-\sum_{i=1}^{N}\exp\left(-\frac{X_i-\mu}{\psi}\right)$$

Pour déterminer les paramètres $(\mu,\psi)$ maximisant cette dernière, il suffit de résoudre (numériquement) le système :$\begin{cases}\displaystyle\sum_{i=1}^{N}\frac{X_i-\mu}{\psi}\left[1-\exp\left(-\frac{X_i-\mu}{\psi}\right)\right]=N\\ \displaystyle\sum_{i=1}^{N}\exp\left(-\frac{X_i-\mu}{\psi}\right)=N\end{cases}$

3. Simulations : Dans le but de comparer les différents estimateurs, on simule 50 20-échantillons de loi de Gumbel de paramètres $(\mu,\psi)$ avec $\psi=0.45$ et $\mu=\alpha\log 100=2.07$[14]. La taille de l'échantillon a été choisie à 20 dans le but d'utiliser les résultats de cette étude dans des applications hydrologiques où les séries recouvrent pour la plupart une vingtaine d'années (et possèdent donc 20 maxima annuels). L'estimation de $\xi$ par la méthode des moments pondérés (dans le modèle $GVE$) est mauvaise : La moyenne des 50 estimations est de -0.13 tandis que la variance est de 0.29.

   Les estimations obtenues par les méthodes du maximum de vraisemblance ($MV$) et des moments pondérés ($MP$) de $\mu$ et $\psi$ sont données par le tableau 3.2. Les estimations du maximum de vraisemblance sont meilleures. La méthode des moments pondérés n'est donc pas utilisable pour la modélisation de séries de maxima annuels couvrant 20 années (voir figure 3.9).

Ce modèle, simple d'utilisation, fournit donc de bons résultats par le maximum de vraisemblance. Cependant, la loi des extrêmes généralisée ne sert qu'à modéliser les maxima (maxima annuel par exemple), ce qui restreint les applications en hydrologie. Pour la modélisation des périodes de retour, la loi de Pareto généralisée est plus adaptée.

[14]En simulant un échantillon de loi exponentielle moyenne $\frac{1}{\psi}$ et de taille $N$ et en prenant le maximum de cet échantillon, on obtient une simulation d'une loi de Gumbel de paramètres $(\mu,\psi)$ avec $\mu=\psi\log N$ et ici $N=100$.

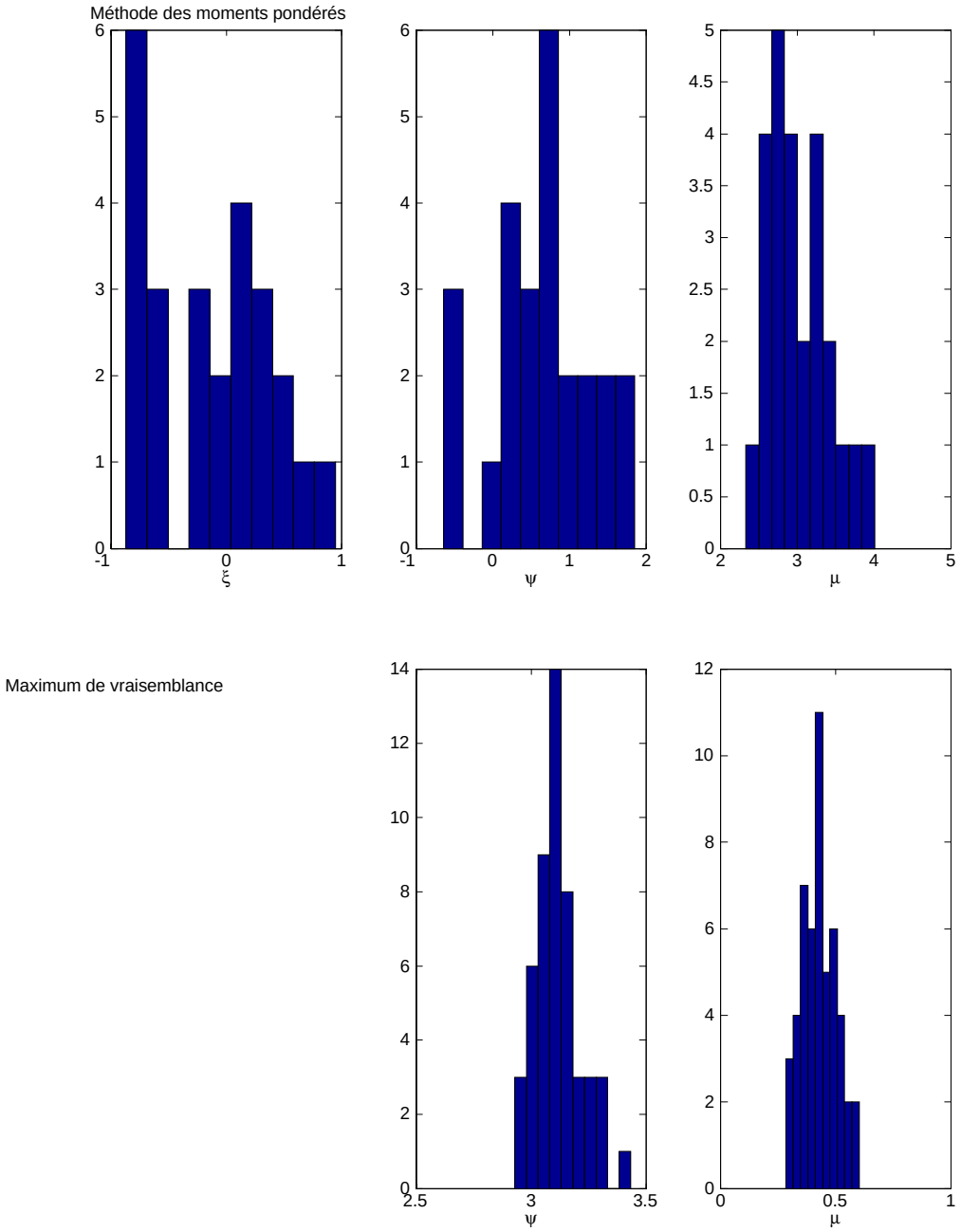



Figure 3.9: Histogrammes des estimations de $\xi, \psi$ et $\mu$ par les méthodes des moments pondérés et du maximum de vraisemblance dans un modèle $GVE$ $(\xi = 0, \mu = 2.07, \psi = 0.45)$.

| Méthode | Moyennes des $\widehat{\mu}_{20}$ | Variances des $\widehat{\mu}_{20}$ | Moyennes des $\widehat{\psi}_{20}$ | Variances des $\widehat{\psi}_{20}$ |
|---|---|---|---|---|
| MV | 3.12 | 0.0096 | 0.43 | 0.0057 |
| MP | 3.02 | 0.01601 | 0.81 | 0.2524 |

Tableau 3.2: Moyennes et variances des estimations dans un modèle $GVE\,(\xi = 0, \mu = 2.07, \psi = 0.45)$ ($MV$ : Maximum de vraisemblance et $MP$ : moments pondérés).

**Dans le modèle de Pareto généralisée :** On a vu en *section* 3.3 que la loi de Pareto généralisée (de fonction de survie $G(x) = \left(1 + \frac{\xi}{\beta}x\right)^{-\frac{1}{\xi}}$) apparaissait comme la distribution limite des dépassements d'un seuil. Elle est donc très utile en pratique. L'estimation de ses paramètres $(\xi, \beta)$ par la méthode des moments et celle du maximum de vraisemblance est simple à mettre en oeuvre.

1. Méthode des moments : Le calcul des moments d'ordre $k$ d'une loi de Pareto généralisée à deux paramètres $PG(\xi, \beta)$ conduit aux résultats :

$$\begin{cases} e_1 = EX = \dfrac{\beta}{1-\xi} \text{ pour } \xi < 1 \\ e_2 = EX^2 = \dfrac{2\beta^2}{(1-\xi)(1-2\xi)} \text{ pour } \xi < \frac{1}{2} \end{cases}$$

   On en déduit les estimateurs des moments de $(\xi, \beta)$ :

$$\begin{cases} \widehat{\beta} = \dfrac{\widehat{e}_1\widehat{e}_2}{2\left(\widehat{e}_2 - \widehat{e}_1^2\right)} \\ \widehat{\xi} = \dfrac{\widehat{e}_2 - 2\widehat{e}_1^2}{2\left(\widehat{e}_2 - \widehat{e}_1^2\right)} \end{cases} \text{ avec } \begin{cases} \widehat{e}_1 = \frac{1}{N}\displaystyle\sum_{i=1}^{N} X_i \\ \widehat{e}_2 = \frac{1}{N}\displaystyle\sum_{i=1}^{N} X_i^2 \end{cases}$$

   Cette méthode aboutit souvent à des résultats aberrants (estimation de $\xi$ supérieure à $\frac{1}{2}$).

2. Méthode du maximum de vraisemblance : La log-vraisemblance d'un $N-$échantillon $(X_i)_{1\leq i\leq N}$ de loi $PG(\xi, \beta)$ s'écrit dans le cas $\xi \neq 0$ :

$$\log L_{\xi,\beta,\nu}(X_1, \cdots, X_N) = -N\log\beta - (1/\xi + 1)\sum_{i=1}^{N}\log\left(1 + \frac{\xi}{\beta}X_i\right)$$

   Pour trouver son maximum, il faut résoudre (numériquement) le système :

$$\begin{cases} (\xi+1)\displaystyle\sum_{i=1}^{N}\frac{X_i}{\beta + \xi X_i} = N \\ (\xi+1)\displaystyle\sum_{i=1}^{N}\frac{X_i}{\beta + \xi X_i} = \frac{1}{\xi}\displaystyle\sum_{i=1}^{N}\log\left(1 + \frac{\xi}{\beta}X_i\right) \end{cases}$$

| Taille de l'échantillon | Moments | Max de Vrais |
|---|---|---|
| $\widehat{\beta}_{1000}$ | $(2.13, 0.026)$ | $(2.01, 0.021)$ |
| $\widehat{\xi}_{1000}$ | $(0.27, 0.003)$ | $(0.31, 0.003)$ |
| $\widehat{\beta}_{500}$ | $(2.06, 0.016)$ | $(2.00, 0.008)$ |
| $\widehat{\xi}_{500}$ | $(0.29, 0.003)$ | $(0.31, 0.001)$ |

Tableau 3.3: Estimations moyennes de $\beta$ et $\xi$ par les méthodes des moments et du maximum de vraisemblance dans un modèle $PG(\beta, \xi)$ sur 50 échantillons simulés de tailles 500 et 1000 de $PG(2, 0.31)$.

| Taille de l'échantillon | Moments | Max de Vrais |
|---|---|---|
| $\widehat{\beta}_{1000}$ | $(6.74, 13.4662)$ | $(1.99, 0.0119)$ |
| $\widehat{\xi}_{1000}$ | $(0.48, 0.0002)$ | $(0.90, 0.0021)$ |

Tableau 3.4: Estimations moyennes de $\beta$ et $\xi$ sur 50 1000−échantillons simulés de $PG(2, 0.90)$ (par les méthodes des moments et du maximum de vraisemblance dans un modèle de Pareto généralisée).

3. Comparaison sur des simulations : On simule 50 1000-échantillon de loi de Pareto généralisée[15] de paramètres $(\beta, \xi) = (2, 0.31)$. Contrairement aux résultats obtenus par la méthode du maximum de vraisemblance, ceux obtenus par la méthode des moments sont médiocres pour 500 observations, mais s'améliorent considérablement quand la taille de l'échantillon augmente (voir tableau 3.3 et figure 3.10). Pour des échantillons de taille 1000, il y a peu de différences entre les deux estimateurs (figure 3.11). Par contre, la différence entre l'estimateur du maximum de vraisemblance et l'estimateur des moments s'accentue lorsque $\xi$ augmente. Le tableau 3.4 fournit les résultats obtenus sur 50 1000-simulations de loi de Pareto généralisée de paramètres $\xi = 0.90$ et $\beta = 2$. L'estimateur du maximum de vraisemblance de $\xi$ converge (avec la taille de l'échantillon) tandis que l'estimateur des moments tend vers $\frac{1}{2}$. Quant à l'estimateur des moments de $\beta$, il semble diverger (figure 3.12).

### 3.4.3 Nouvel outil statistique

Dans cette section, un outil statistique original d'estimation du paramètre de décroissance algébrique est proposé. Il est fondé sur le résultat de la *section 3.3.2,* reliant le comportement asymptotique de la *FMDR* $h$ à l'appartenance ou non de cette loi au domaine d'attraction de Fréchet. Ce résultat permet de plus de déterminer l'indice

[15] On simule une loi de Pareto généralisée à partir d'une loi exponentielle en remarquant que si $X \sim Exp\left(\frac{1}{\xi}\right)$ alors $Y = \frac{\beta}{\xi}\left(e^X - 1\right) \sim PG(\beta, \xi)$.

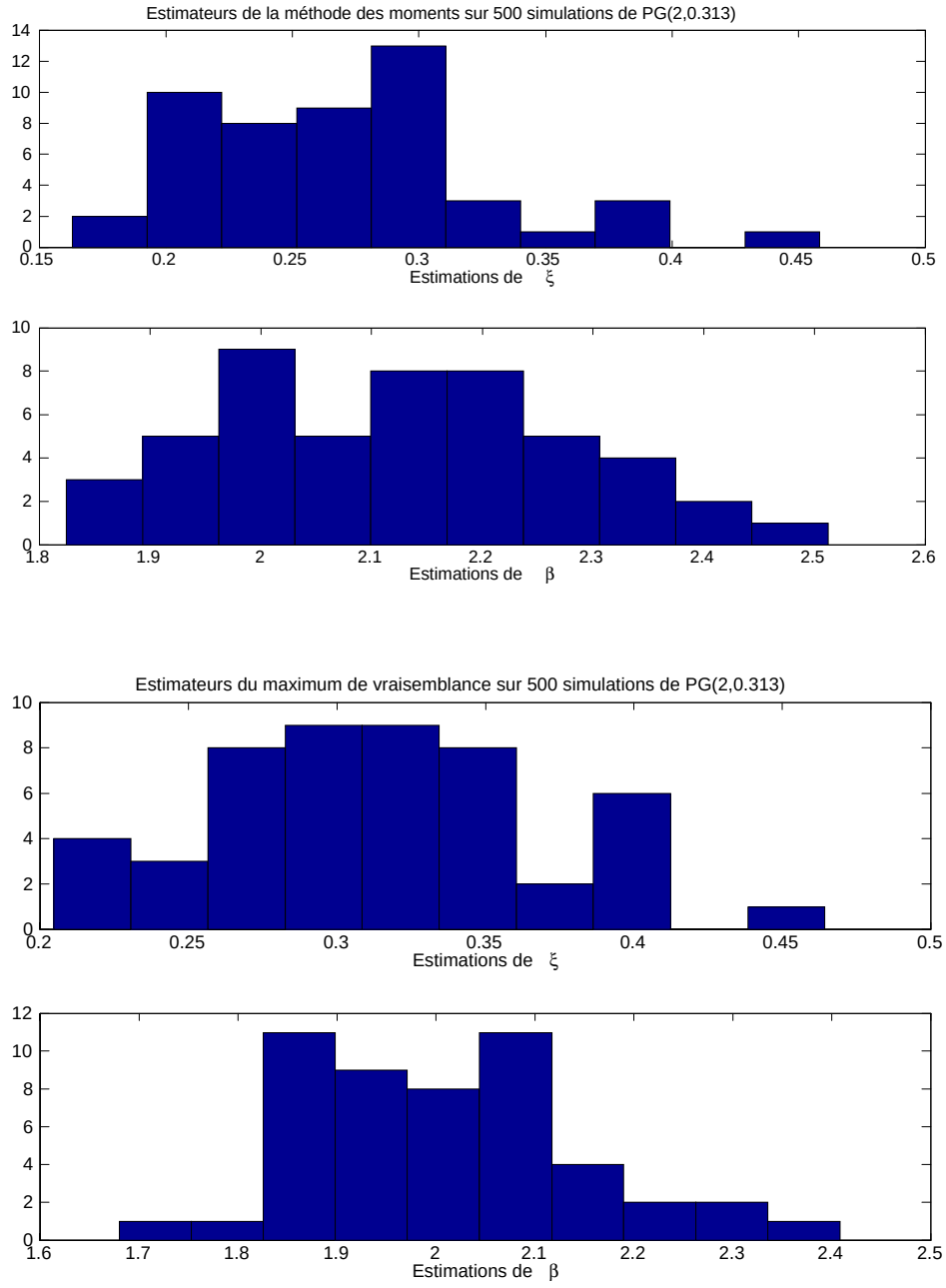



Figure 3.10: Histogrammes des estimations de $\beta$ et $\xi$ par les méthodes des moments et du maximum de vraisemblance dans un modèle $PG(\beta, \xi)$ sur 50 échantillons simulés de $PG(2, 0.31)$ de taille 500.

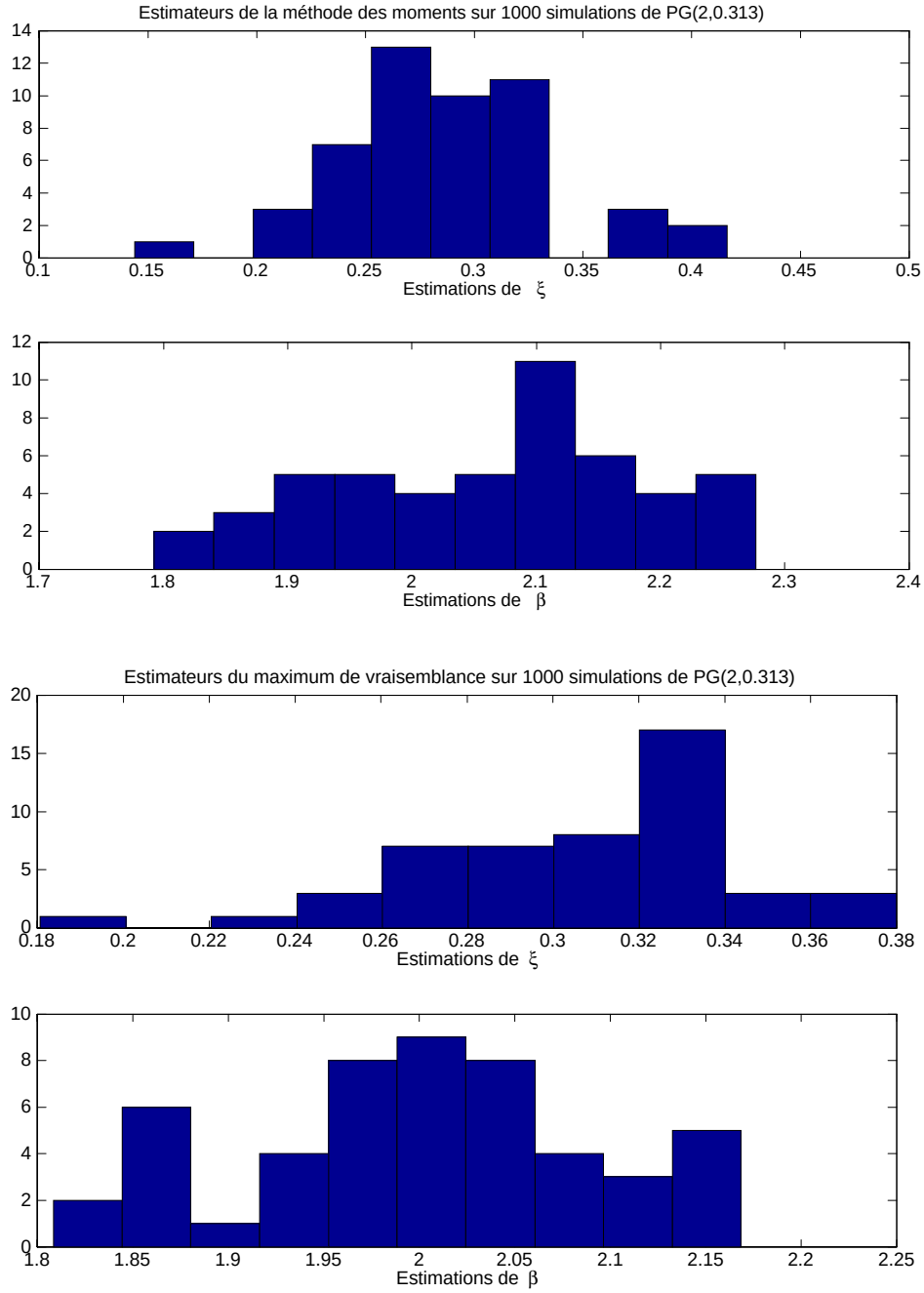



Figure 3.11: Histogrammes des estimations de $\beta$ et $\xi$ par les méthodes des moments et du maximum de vraisemblance dans un modèle $PG(\beta, \xi)$ sur 50 échantillons simulés de $PG(2, 0.31)$ de taille 1000.

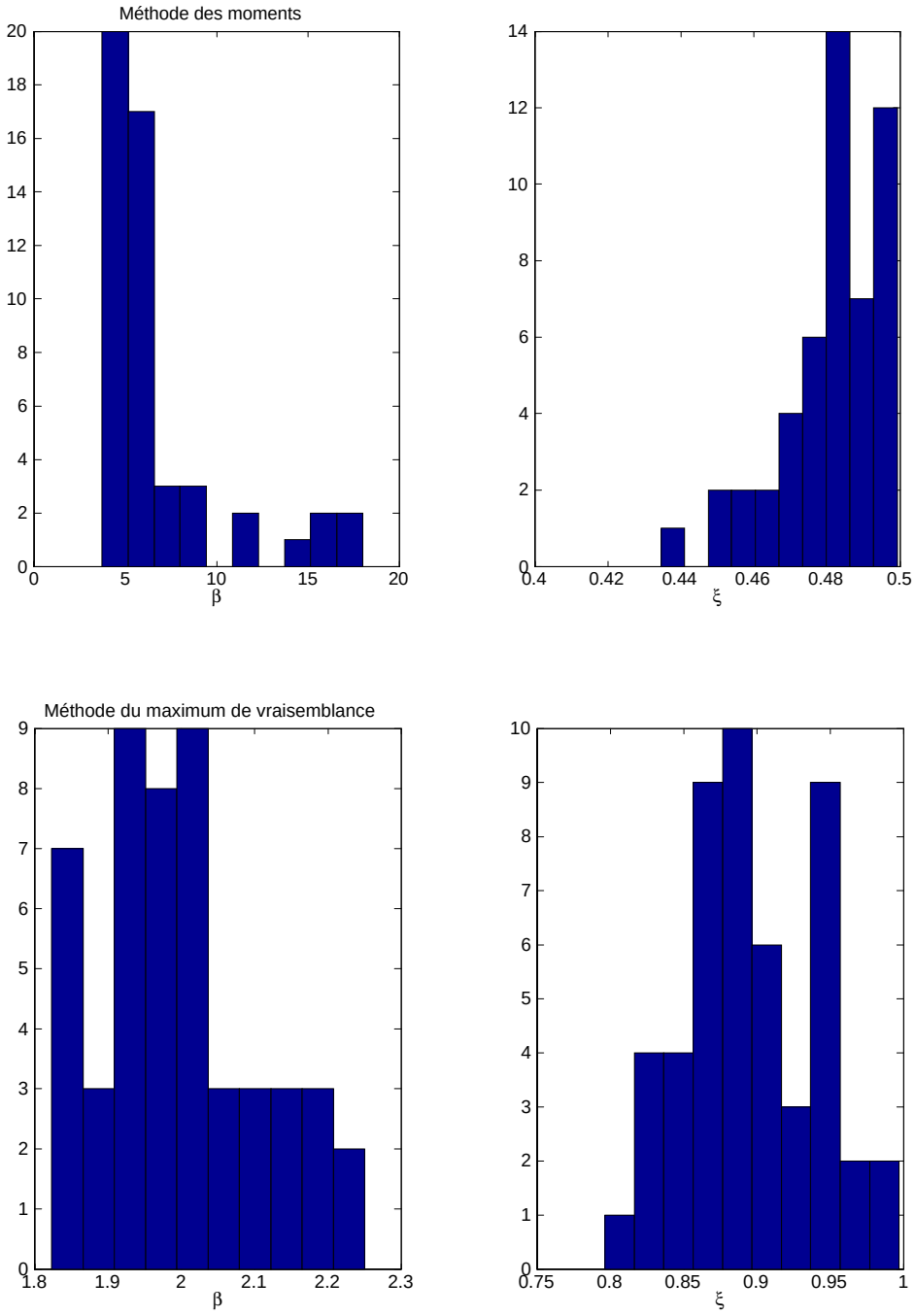



Figure 3.12: Histogrammes des estimations de $\beta$ et $\xi$ par les méthodes des moments et du maximum de vraisemblance dans un modèle $PG(\beta, \xi)$ sur 50 échantillons simulés de $PG(2, 0.90)$ de taille 1000.

de décroissance algébrique $q$ car, dans le cas où la distribution appartient au domaine d'attraction de Fréchet, $h$ tend vers $\frac{1}{q-1}$ à l'infini.

La méthode présentée dans cette section va donc consister à estimer, pour différents seuils de troncature, la *FMDR* $h$. Après avoir calculé le biais et la variance de cet estimateur, une méthode de débiaisement numérique ainsi qu'un algorithme de détermination du seuil de troncature adapté à l'estimation de $q$ sont appliqués.

**Estimation**

Disposant d'une série de $N$ observations $(X_i)_{i=1\cdots N}$, pour tout seuil $s > 0$,

$$\begin{cases} \widehat{G}(s) = \frac{1}{N}\sum_{i=1}^{N} 1_{X_i>s} \\ \widehat{H}(s) = \frac{1}{N}\sum_{i=1}^{N} X_i.1_{X_i>s} \end{cases}$$

sont des estimateurs sans biais de la fonction de survie et de la primitive de dépassement en $s$. On pourra estimer la *FMDR* par :

$$\widehat{h}(s) = \frac{\widehat{H}(s)}{s\widehat{G}(s)} - 1$$

**Remarque 26** *L'interprétation de cet estimateur peut être la suivante : $\widehat{h}$ est la vitesse de décroissance des durées de dépassement par rapport aux cumuls de dépassement (somme des longueurs des bases et des aires des rectangles colorés de la figure 3.13).*

*En effet, ces variables aléatoires s'écrivent pour tout $s > 0$ :*

$$d_N(s) = \sum_{i=1}^{N} 1_{X_i>s} \text{ et } p_N(s) = \sum_{i=1}^{N} X_i 1_{X_i>s}$$

*Soient $\widehat{x}_N(s)$ et $\widehat{y}_N(s)$ les logarithmes de ces dernières normalisées respectivement par la durée totale d'observation $D = N$ et le cumul total $P = \sum_{i=1}^{N} X_i$ :*

$$\begin{cases} \widehat{x}_N(s) = \log\left[\widehat{G}(s)\right] \\ \widehat{y}_N(s) = \log\left[\dfrac{\widehat{H}(s)}{\overline{X}^N}\right] \end{cases}$$

*Ces estimateurs convergent respectivement vers*[16] *:*

$$\begin{cases} x(s) = \log\left[G(s)\right] \\ y(s) = \log\left[\dfrac{H(s)}{H(0)}\right] \end{cases}$$

[16] Car les transformées $(U_i = 1_{X_i>s})_{i=1\cdots N}$ et $(V_i = X_i 1_{X_i>s})_{i=1\cdots N}$ sont elles aussi *iid.* En vertu de la loi (forte) des grands nombres, pour tout seuil $s > 0$ :

$$\widehat{x}_N(s) \overset{N\to\infty}{\longrightarrow} \log EU_1(s) = x(s) \quad p.s.$$

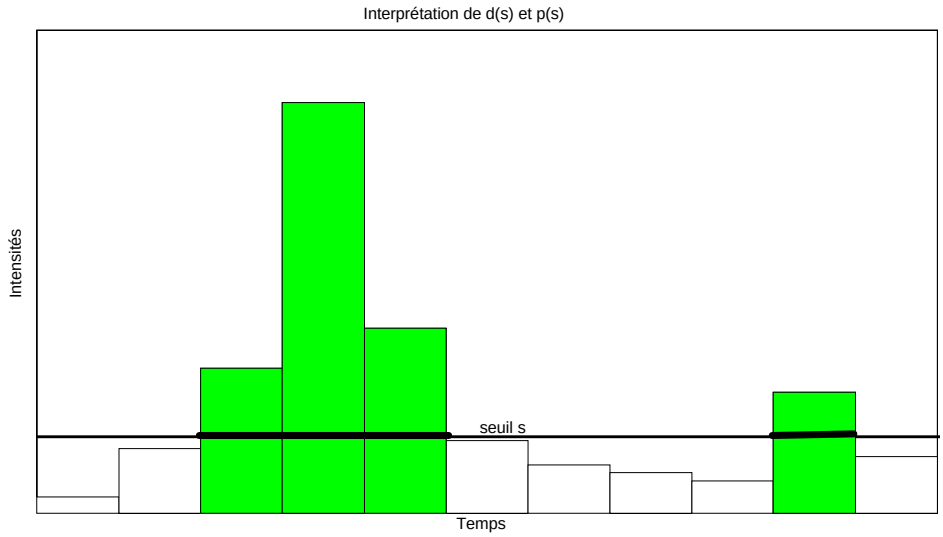



Figure 3.13: Durée et cumul de dépassement.

*et la pente* $a(s)$ *du graphe* $(x(s), y(s))$ *est simplement reliée à la FMDR* $h$ *:*

$$a(s) = \frac{1}{1 + h(s)}$$

**Biais et variance de** $\widehat{h}$

Le biais et la variance de $\widehat{h}$ s'écrivent (voir *Annexe C)* :

$$\forall s > 0 \qquad E\left[\widehat{h}(s) - h(s)\right] = -F^N(s).h(s)$$

$$\forall s > 0 \qquad Var\left[\widehat{h}(s)\right] = \frac{2S_N}{s^2 G} \int_s^\infty H - h^2.\left[S_N - F^N\left(1 - F^N\right)\right]$$

$$\text{avec } S_N(s) = \sum_{k=1}^{N} \frac{1}{k} C_N^k G^k(s) F^{N-k}(s)$$

Ce biais est négatif, donc on sous-estime systématiquement $h$. Les variations du biais et de la variance avec le seuil $s$ diffèrent selon la loi $F$. Dans le cas algébrique strict,

et par une intégration par parties :

$$\widehat{y}_N(s) \overset{N\to\infty}{\longrightarrow} \log EV_1(s) - \log EX_1 = y(s) \quad p.s.$$

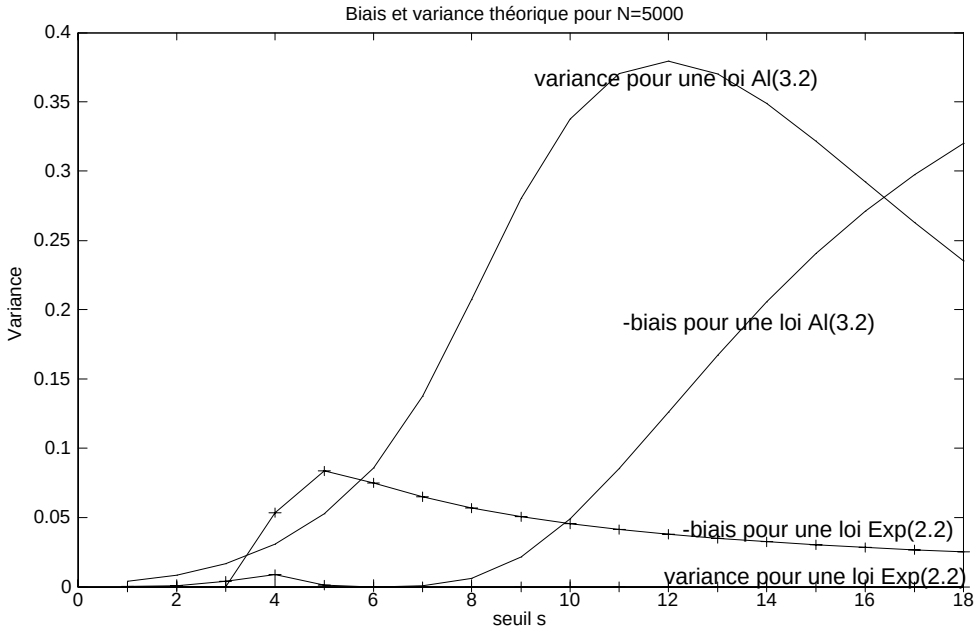



Figure 3.14: Variations du biais de la variance avec le seuil.

le biais diverge avec le seuil tandis que la variance de $\widehat{h}$ s'écrit :

$$Var\left[\widehat{h}(s)\right] = \frac{2\left(s+1\right)^{2} S_N}{s^2\left(q-1\right)\left(q-2\right)} - \frac{\left(s+1\right)^{2}}{s^2\left(q-1\right)^{2}} \cdot \left[S_N - F^N\left(1-F^N\right)\right]$$

et tend vers l'infini quand $s$ et $N$ tendent vers l'infini. (voir figure 3.14).

**Débiaisement**

On cherche à débiaiser $\widehat{h}$ en tout seuil $s$, mais ceci est impossible dans le cas général car on ne connaît pas la relation à inverser. Il est nécessaire de se placer dans un cadre paramétrique. On choisit ici les modèles exponentiel et algébrique strict[17] où les $FMDR$ s'écrivent respectivement

$$\begin{aligned} h(s) &= \frac{1}{as} \\ h\left(s\right) &= \frac{s+1}{\left(q-1\right)s} \end{aligned}$$

[17]La loi algébrique stricte de coefficient $q$ est définie par sa densité : $f(s) = \begin{cases} q\left(s+1\right)^{-q-1} & \text{si } s \geq 0 \\ 0 \text{ sinon} \end{cases}$

C'est une loi de type algébrique :
$\forall s \geq 1 \quad r(s) = (1+\frac{1}{s})^{-q} \overset{s\to\infty}{\longrightarrow} 1$ et $r$ croissante sur $[1, +\infty[$.
$\forall s \geq 1 \quad s^2.r^{'}(s) = q(1+\frac{1}{s})^{-q-1} \leq q$

- Pour débiaiser numériquement $\widehat{h}$ dans les cas exponentiel et algébrique strict, il faudra respectivement inverser les relations :

$$\begin{aligned}\widehat{h}(s) &= h(s).\left[1-\left[1-e^{-as}\right]^N\right] = h(s).\left[1-\left[1-e^{-\frac{1}{h(s)}}\right]^N\right] \\ \widehat{h}(s) &= h(s).\left[1-\left(1-(s+1)^{-\frac{s+1}{s.h(s)}-1}\right)^N\right]\end{aligned}$$

  Une limite non-nulle pour $\widehat{h}_{deb.}$ permettra de conclure (à l'aide d'un test) que les données ont un comportement algébrique.

- L'estimation des paramètres du modèle à partir de l'estimateur débiaisé $\widehat{h}_{deb.}$ est réalisée en inversant les relations précédentes en un seuil raisonnablement grand[18] (pas trop grands car le biais diverge avec le seuil). Il faudra donc, dans un premier temps, déterminer une limite supérieure formelle pour ce seuil en fonction du nombre de données $N$. En utilisant les résultats de la *section 3.3.1* sur la loi du maximum (ici Gumbel et Fréchet), on choisira respectivement dans les cas exponentiel et algébrique, strict un seuil du type :

$$\begin{aligned}s_0 &= quantile(Gumbel) + \log(N) \\ s_0 &= quantile(Fréchet) \times N^{1/q}\end{aligned}$$

Le débiaisement de $\widehat{h}$ aura donc pour double objectif de conclure quant à la pertinence d'un modèle de type algébrique (à l'aide d'un test), et d'estimer dans le cas d'un comportement algébrique l'indice de décroissance algébrique $q$.

### Algorithme de détermination de $q$

Le choix du seuil de débiaisement repose sur un algorithme de type adaptatif (Lepski et Spokoiny [56]) : On commence par se placer au seuil $s_0 = Ent(N^\gamma - 1)$, $\gamma$ étant un réel appartenant à $]0,1[$. On calcule l'estimateur $\widehat{q_1}$ du coefficient $q$ en ce seuil $s_0$, puis on itère en estimant $q$ en $s_1 = \left\lfloor N^{1/\widehat{q_1}} - 1 \right\rfloor$, ceci jusqu'à stabilisation de la valeur de l'estimateur de $q$.

### Simulations

**Simulations exponentielles :** On simule un 5000-échantillon de loi exponentielle de paramètre $a = 0.45$. Sur la figure 3.15, on a représenté les résultats du débiaisement par inversion numérique. La correction de l'estimateur est bonne.

[18] Le choix d'un grand seuil permet de s'écater du modèle algébrique strict pour estimer $q$ dans un modèle de type algébrique. En effet, pour toutes ces lois, $h(s) \overset{s\to\infty}{\longrightarrow} \frac{1}{q-1}$

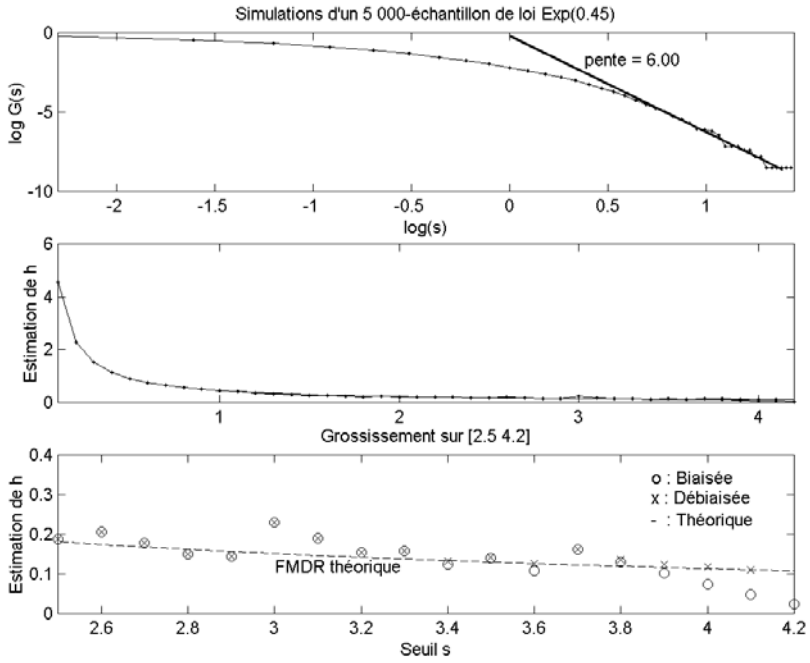



Figure 3.15: Estimation de $h$ sur un 5000-échantillon de loi $Exp(0.45)$.

On simule 50 8000-échantillons de loi $Exp(0.45)$, et on estime la limite de la *FMDR* par le nouvel outil. On obtient une estimation biaisée et une estimation débiaisée numériquement. Elles ont respectivement pour moyennes 0.10 et 0.13 (ce qui confirme bien la sous estimation systématique de la *FMDR*), et pour variances 0.0026 et 0.0009 (figure 3.16). Le débiaisement numérique est efficace, car il permet de diviser par dix l'erreur (voir figure 3.16). Contrairement à la méthode consistant à estimer $\xi$ par les moments pondérés (présentée en *section* 3.4.2), cette méthode d'exploration des données permet donc de conclure que les données n'appartiennent pas au domaine d'attraction de Fréchet, puisque la *FMDR* est quasi-nulle en des seuils relativement faibles (ils ne dépassent pas 5). Cependant, il serait nécessaire de construire un test pour déterminer si la *FMDR* est significativement non-nulle.

**Simulations de lois algébriques strictes :** La figure 3.17 a été obtenue à partir d'un 5000-échantillon. On a débiaisé numériquement l'estimateur de la fonction moyenne de dépassement relative $\widehat{h}$ avec une résolution de $10^{-4}$ obtenu à partir de 5000 simulations de loi algébrique stricte de coefficient $q = 3.20$ en cherchant à partir de $s$, $\widehat{p}$ et $N$ (pour une résolution fixée à $10^{-4}$), le paramètre $p$ vérifiant la relation ci-dessus. On constate que l'estimation de $q$ par la pente de régression sur le graphe de la fonction de survie est mauvaise ($\widehat{q}_G = 2.39$). Le débiaisement numérique fournit de bons résultats. Toutefois l'estimation débiaisée s'éloigne de la valeur théorique quand le seuil augmente (en $s = 18$), ce qui s'explique par l'augmentation du biais de $\widehat{h}$ avec le seuil (voir figure 3.14).

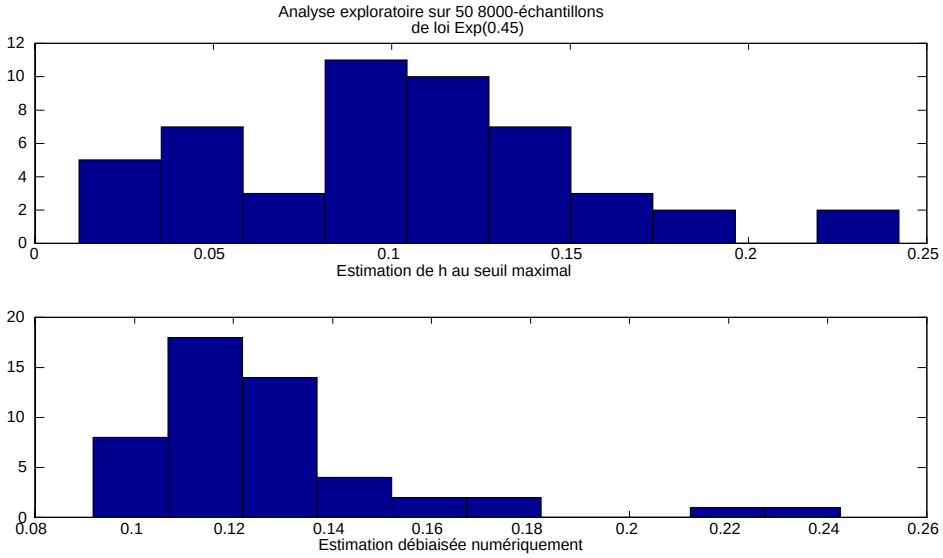



Figure 3.16: Répartition des estimations de la *FMDR* $h$ au seuil maximal sur 50 8000-échantillons simulés de loi $Exp(0.45)$.

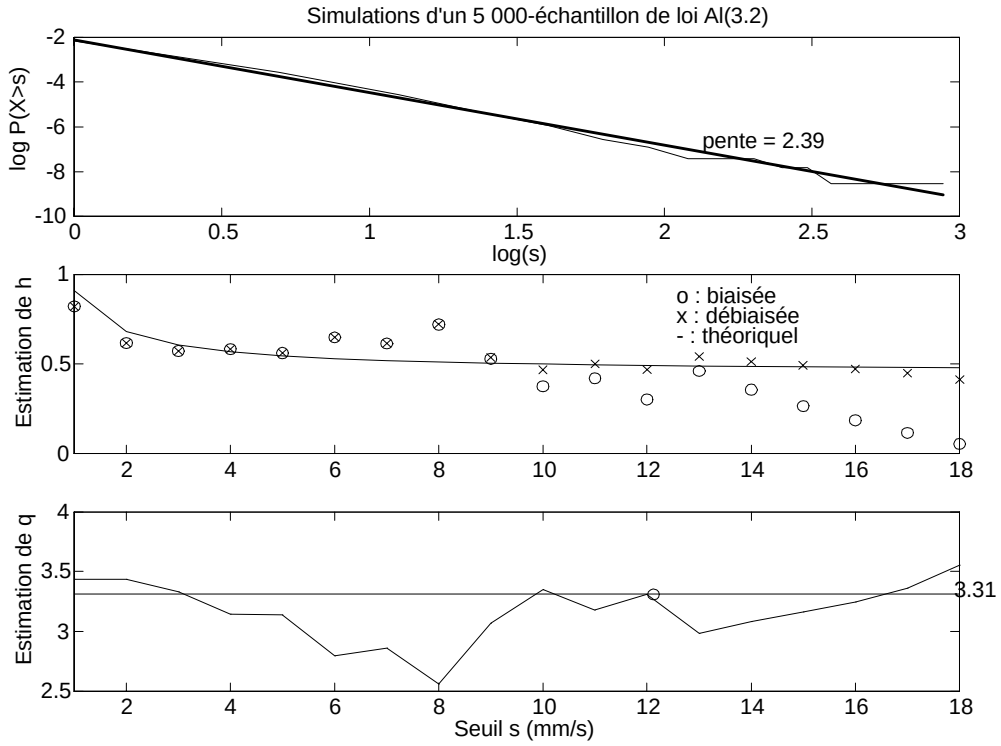



Figure 3.17: Estimation de la *FMDR* $h$ et de $q$ sur un 5000-échantillon simulé de loi $Al(3.2)$.

| Taille de l'échantillon | Moyenne des $\xi_i$ | Variance des $\xi_i$ |
|---|---|---|
| 500 | 0.31 | 0.0010 |
| 1000 | 0.31 | 0.0005 |

Tableau 3.5: Moyennes et variances des estimations de $\xi$ par la *FMDR* sur 50 échantillons simulés de loi $Al(3.2)$ ($\xi = 0.31$) de tailles 500 et 1000.

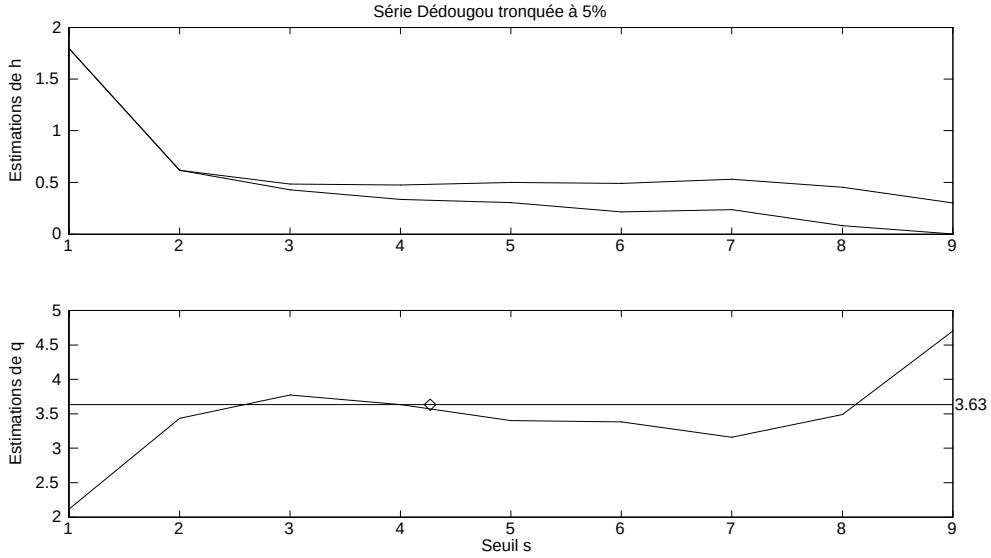



Figure 3.18: Répartition des estimations de $\xi$ par la *FMDR* sur 50 échantillons simulés de loi $Al(3.2)$ ($\xi = 0.31$) de taille 1000.

Pour comparer les performances de cette méthode d'estimation avec celle du maximum de vraisemblance dans un contexte d'application à des séries hydrologiques, on simule 50 1000-échantillons de loi $Al(3.2)$ (c'est-à-dire de $PG(0.31, 1)$). Les moyennes et variances des estimations de $\xi$ sont fournies par le tableau 3.5. Les estimations du paramètre $\xi = \frac{1}{q}$ ne sont pas biaisées à $10^{-2}$ près. La variance est faible (figure 3.18).

**Conclusion**

La méthode statistique présentée dans cette section comporte deux volets : l'exploration et l'estimation dans le cadre semi-paramétrique des lois de type algébrique.

- La méthode d'exploration ne fournit pas de résultats graphiques satisfaisants : les résultats des estimations (biaisée et débiaisée) de la *FMDR* $h$ doivent être interprétés en tenant compte de la grandeur du seuil. Il serait nécessaire de construire un test.
- La méthode d'estimation du paramètre de décroissance algébrique fournit par contre, sur les simulations, des biais et des variances d'estimation de $\xi$ du même ordre que

celle du maximum de vraisemblance (sur des échantillons de 500 ou 1000 données). Mais ce nouvel outil permet en plus de s'écarter du cadre paramétrique de la loi de Pareto généralisée pour se placer dans le cadre semi-paramétrique des lois de type algébrique.

### 3.4.4 Conclusion

Dans cette section, nous avons étudié des outils statistique permettant de :

- déterminer dans quel domaine d'attraction il est raisonnable de se placer, ainsi que la loi à utiliser pour modéliser un ensemble de données. Les outils statistiques d'exploration très classiques que sont les graphes de la fonction de survie, des quantiles ou le graphe de la *FMD,* permettent une visualisation du comportement asymptotique de la loi. Cependant, ils ne fournissent pas de résultats satisfaisants sur les données à grande variabilité. Quant à l'estimateur de Hill (présenté ici sous sa forme classique), il ne fournit de bons résultats que dans le cas d'un comportement Pareto, et est très sensible aux écarts à ce modèle. La nouvelle méthode d'exploration, basée sur la visualisation du comportement asymptotique de la *FMDR,* ne peut constituer un outil graphique et doit être couplée à un test. La méthode des moments pondérés dans le modèle *GVE* ne fournit pas de bons résultats sur des échantillons de maxima de la taille de ceux exploités en hydrologie.

- estimer le paramètre de décroissance algébrique $\xi$ des lois de type algébrique. La méthode des moments fournit de mauvais résultats sur les paramètres d'une Pareto dès que la variance de cette dernière augmente ; celle du maximum de vraisemblance fournit de bons estimateurs, mais impose de se placer dans un cadre paramétrique. Or, on peut se demander s'il est nécessaire de se placer dans le cadre paramétrique alors que c'est le comportement asymptotique de la loi qui est étudié. La *FMDR* fournit, asymptotiquement, un estimateur de $\xi$. Cette méthode est semi-paramétrique dans la mesure où son domaine de validité s'étend au cadre des lois de type algébrique. Elle est, de plus, aussi performante que la méthode du maximum de vraisemblance. Il serait nécessaire cependant d'étudier à quel point les résultats pré-asymptotiques observés dans la pratique sont révélateurs des comportements asymptotiques.

Ajoutons enfin que, pour mettre en évidence les propriétés asymptotiques de la *FMDR* des lois de type algébrique, il est nécessaire de disposer de longues séries afin de faire apparaître de grands seuils car, plus le seuil maximal est grand, plus l'estimation du paramètre $q$ est bonne. La nécessité de longues séries provient aussi du fait de l'erreur non négligeable due à l'application de la loi des grands nombres. D'autre part, cette méthode n'est pas adaptée aux séries ne présentant pas une quasi-indépendance.

## 3.5 Application à des séries hydrologiques

Dans cette section, on examine le comportement des extrêmes de quelques longues séries hydrologiques à pas de temps différents : 1 minute, 1 heure, 1 jour et 1 an. Deux

méthodes sont employées pour l'exploration des données : la visualisation du graphe de la fonction de survie et la méthode de la *FMDR* présentée précédemment. Cette méthode est aussi utilisée pour l'estimation du paramètre de décroissance algébrique. Les résultats obtenus sont ensuite comparés aux estimations du maximum de vraisemblance.

### 3.5.1 Hauteurs journalières de pluie

Dans cette section, on étudie de longues séries hydrologiques provenant de régimes climatiques différents : Dédougou (Burkina-Faso), Athènes (Grèce), Bra (Italie) et Larnaca (Chypre).

#### Dédougou

Les données de Dédougou (Burkina-Faso) sont constituées de 23 années (de 1968 à 1990) de cumuls journaliers de pluie. Les relevés sont effectués à 6 heures et sont exprimés en $10^{ème}$ de millimètres.

On constitue l'échantillon des extrêmes à 5% en tronquant ces données au seuil de 180 mm (qui est le seuil correspondant à 5% des plus hautes observations), et en les translatant par ce seuil de troncature.

L'analyse exploratoire de l'échantillon des extrêmes par le graphe de l'estimateur standard de la fonction de survie (figure 3.19) conduit à privilégier un modèle algébrique pour la queue de distribution. Cette constatation est corroborée par l'estimation non nulle de $\widehat{h}_{débiaisé}$ en de grands seuils (figure 3.19) :

$$\widehat{h}_{débiaisé}\,(1700) = 0.4$$

Etant donnés les résultats des simulations de la section 3.4.2 et le faible nombre de données dont on dispose, on s'abstient d'appliquer la méthode des moments pondérés pour l'exploration des données.

On applique la méthode du maximum de vraisemblance dans un modèle de Pareto généralisée $PG\,(q,\beta)$ dont la densité s'écrit :

$$f_{q,\beta}\,(s) = \frac{1}{\beta}\left(1+\frac{s}{\beta q}\right)^{-q-1}$$

Les résultats des estimations du maximum de vraisemblance de $(q,\beta)$ sont de :

$$\begin{aligned}\widehat{q}_{MV} &= 2*10^{7}\\ \widehat{\beta}_{MV} &= 1.30\end{aligned}$$

La méthode durée-cumul est appliquée à ces dépassements. On observe sur la figure 3.19 une convergence de l'algorithme vers la valeur :

$$\widehat{q}_{FMDR} = 3.63$$

Afin de comparer la qualité des ajustements, on reporte sur un même graphe $\log - \log$ trois courbes : celle de la fonction de survie empirique, ainsi que celles des fonctions

de survie Pareto généralisée, dont les coefficients $\beta$ et $\xi$ sont d'une part estimés par le maximum de vraisemblance et d'autre part par la méthode de la $FMDR$ (couplée à une estimation de $\beta$ par la méthode du maximum de vraisemblance dans un modèle de Pareto généralisée). On constate que l'estimation par la $FMDR$ a tendance à sur-évaluer les extrêmes, tandis que la méthode du maximum de vraisemblance conduit à de bons résultats (légère sous-estimation toutefois).

**Athènes**

La série de Athènes est constituée de 44 640 cumuls journaliers exprimés en $10^{ème}$ de mm (soit environ 122 ans). L'analyse exploratoire des 5% des plus hautes observations conduit à constater un comportement algébrique de la queue de distribution : le graphe de l'estimateur standard de la fonction de survie (figure 3.21) est asymptotiquement linéaire et la limite de $\widehat{h}_{débiaisé}$ est non nulle (figure 3.20).

L'estimation de $q$ par la méthode de la $FMDR$ est de 4.22 (figure 3.20), résultat inférieur à l'estimation du maximum de vraisemblance qui est de :

$$\begin{aligned} \widehat{q}_{MV} &= 8.83 \\ \widehat{\beta}_{MV} &= 0.85 \end{aligned}$$

En comparant ces deux estimateurs par l'examen de la qualité d'ajustement de la fonction de survie estimée à la fonction de survie empirique, on constate que la $FMDR$ surestime la fonction de survie tandis que le maximum de vraisemblance conduit à un meilleur ajustement. Ces résultats sont toutefois à intégrer dans un modèle de gestion du risque car, en termes de coût, il peut se révéler plus avantageux d'adopter l'estimation de la $FMDR$.

**Bra**

La série de Bra (Italie) est constituée de 47 480 cumuls journaliers exprimés en $10^{ème}$ de mm (soit environ 130 ans). L'analyse exploratoire des 5% des plus hautes observations conduit là aussi à constater un comportement algébrique de la queue de distribution (allure asymptotiquement linéaire de l'estimateur standard de la fonction de survie sur la figure 3.21 et limite de $\widehat{h}_{débiaisé}$ non nulle sur la figure 3.20).

L'estimation de $q$ par la méthode de la $FMDR$ est de 3.34 (figure 3.20), résultat inférieur à l'estimation du maximum de vraisemblance $\widehat{q}_{MV}$ :

$$\begin{aligned} \widehat{q}_{MV} &= 6.31 \\ \widehat{\beta}_{MV} &= 0.87 \end{aligned}$$

La comparaison entre les deux estimateurs $\widehat{q}_{MV}$ et $\widehat{q}_{FMDR}$ sur la base de la fonction de survie fait apparaître, cette fois, une nette sous-estimation de la fonction de survie par la méthode du maximum de vraisemblance et un meilleur ajustement par la méthode de la $FMDR$.

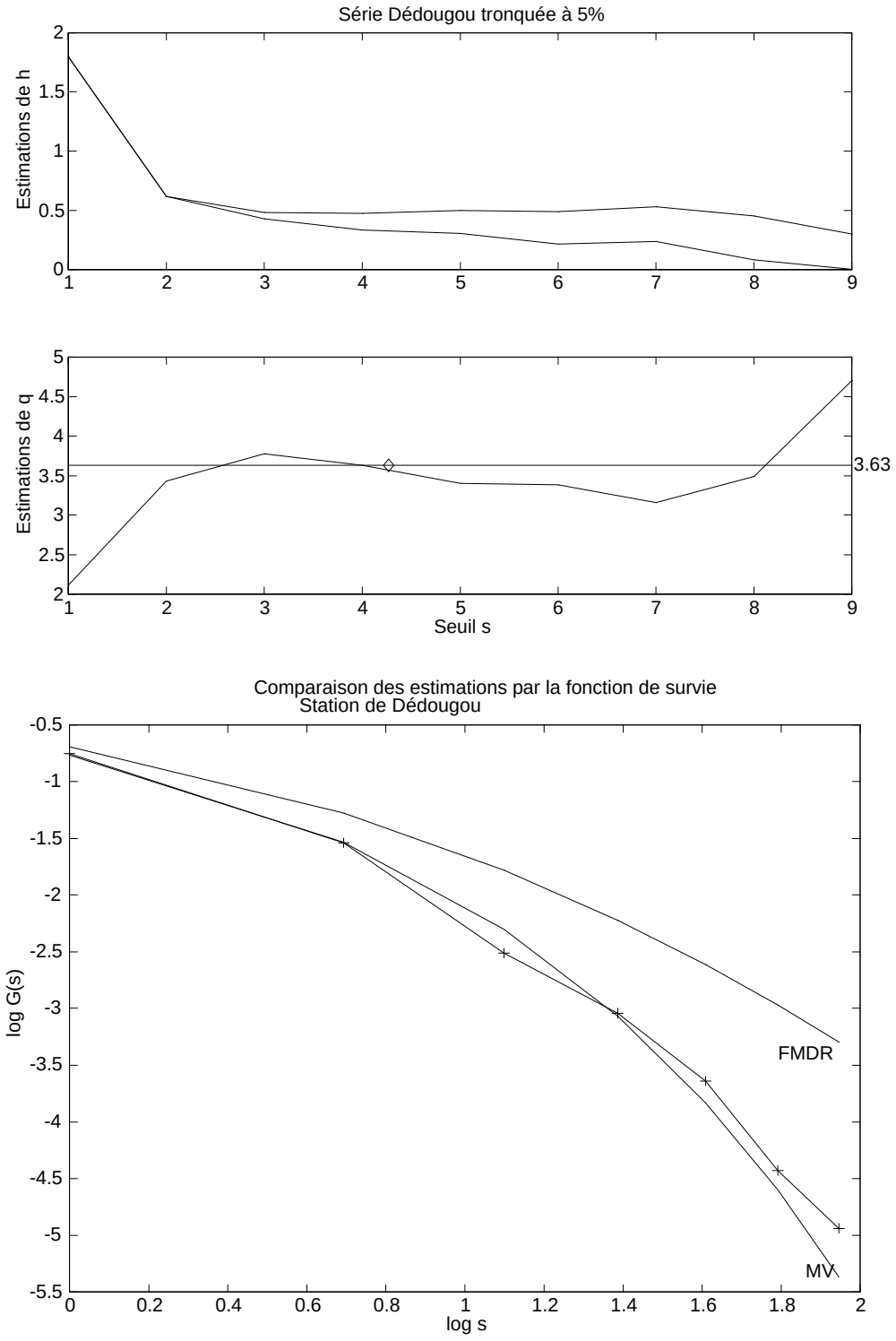



Figure 3.19: Série de Dédougou (5% des plus hautes observations) : Estimation de $q$ par la $FMDR$ et comparaison des estimations du $MV$ et de la $FMDR$ par la fonction de survie ('+' : estimateur standard de la fonction de survie).

**Larnaca**

Cette série de Larnaca (Chypre) comporte 40 680 cumuls journaliers exprimés en $10^{ème}$ de mm (soit 111 années de mesure). La distribution des 5% des plus hautes observations a un comportement algébrique : sur la figure 3.22, l'estimateur standard de la fonction de survie est asymptotiquement linéaire et sur la figure 3.22, la limite de $\widehat{h}_{débiaisé}$ est non nulle.

L'estimation de $q$ par la méthode de la *FMDR* est de 3.29 (figure 3.22), résultat qui se rapproche de l'estimation du maximum de vraisemblance $\widehat{q}_{MV}$ :

$$\begin{aligned} \widehat{q}_{MV} &= 4.56 \\ \widehat{\beta}_{MV} &= 0.82 \end{aligned}$$

La comparaison entre ces deux estimateurs sur la base de la fonction de survie fait apparaître cette fois une nette surestimation de la fonction de survie par la méthode de la *FMDR* méthode et un meilleur ajustement par la méthode du maximum de vraisemblance.

### 3.5.2 Données horaires

**Série d'Alabama**

Les données d'Alabama constituent une série de cumuls à pas de temps horaire (19 368 données). En n'en retenant que les 5 % supérieurs (seuil de troncature à 0.1), il en reste 968. La queue de distribution a un comportement algébrique (figure 3.23). On applique la méthode du maximum de vraisemblance et de la *FMDR* (figure 3.23). On obtient les estimations du maximum de vraisemblance suivantes :

$$\begin{aligned} \widehat{q}_{MV} &= 1.99 \\ \widehat{\beta}_{MV} &= 0.74 \end{aligned}$$

$\widehat{q}_{MV}$ est nettement plus faible que $\widehat{q}_{FMDR}$ :

$$\widehat{q}_{FMDR} = 3.12$$

mais l'estimation par la *FMDR* est meilleure (figure 3.23).

**Série de Bordeaux**

Cette longue série horaire couvre une durée d'enregistrement d'environ 5 ans (42349 données). Elle présente un comportement algébrique (figure 3.24). On obtient, par les méthodes du maximum de vraisemblance et de la *FMDR,* les estimations suivantes (figure 3.24) :

$$\begin{aligned} \widehat{q}_{MV} &= 2.47 \\ \widehat{\beta}_{MV} &= 0.23 \\ \widehat{q}_{FMDR} &= 3.16 \end{aligned}$$

L'estimation du maximum de vraisemblance est trop faible, tandis que celle de la *FMDR* s'ajuste bien aux extrêmes.

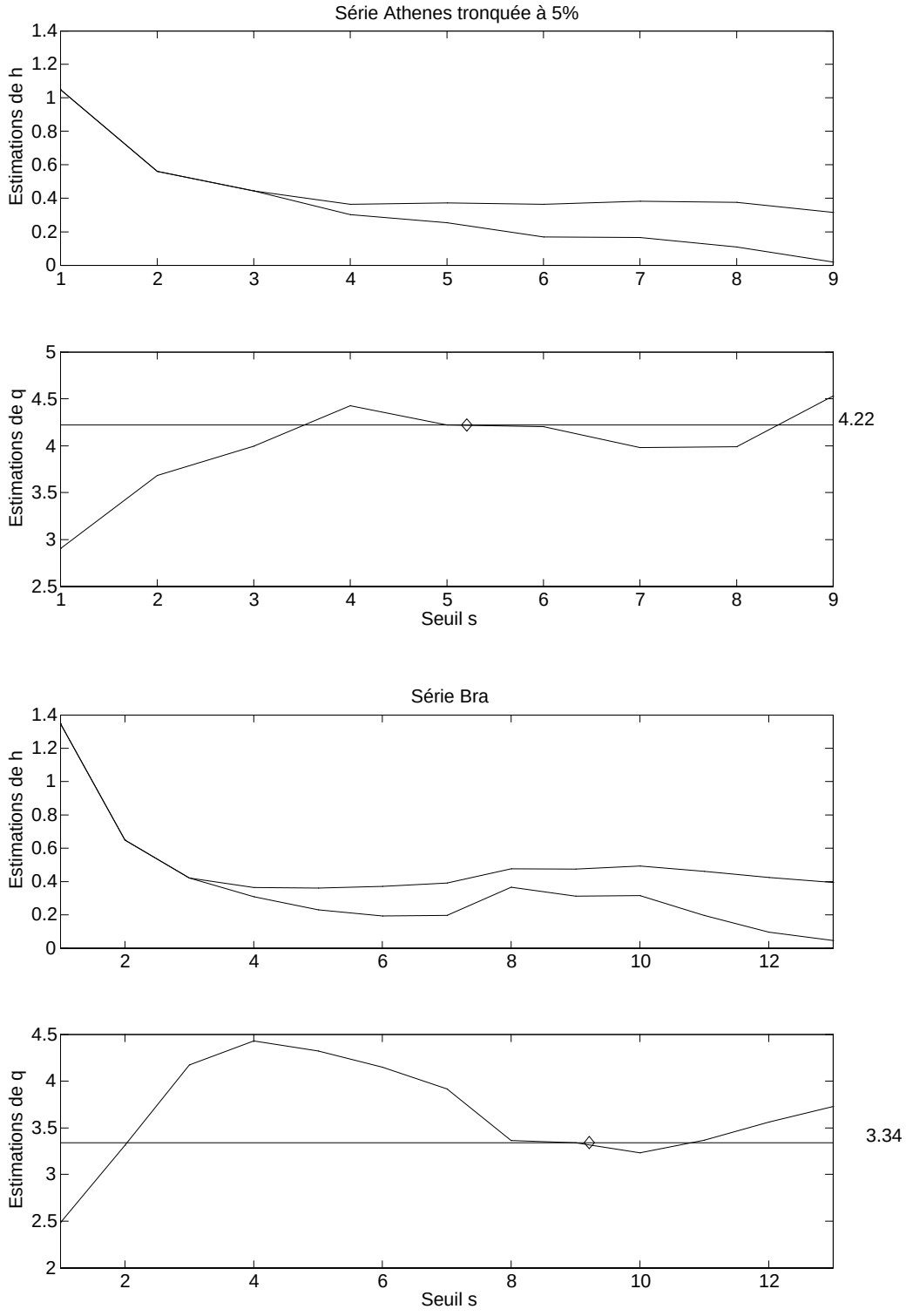
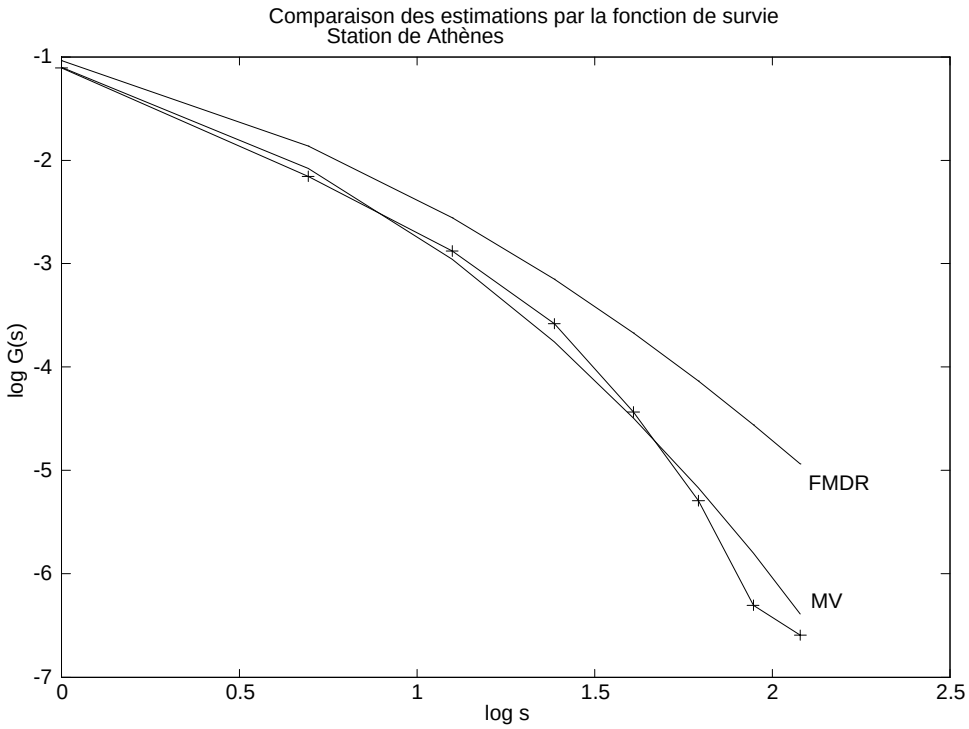



Figure 3.20: Estimation de $q$ par la *FMDR* sur les séries de Athènes et Bra.



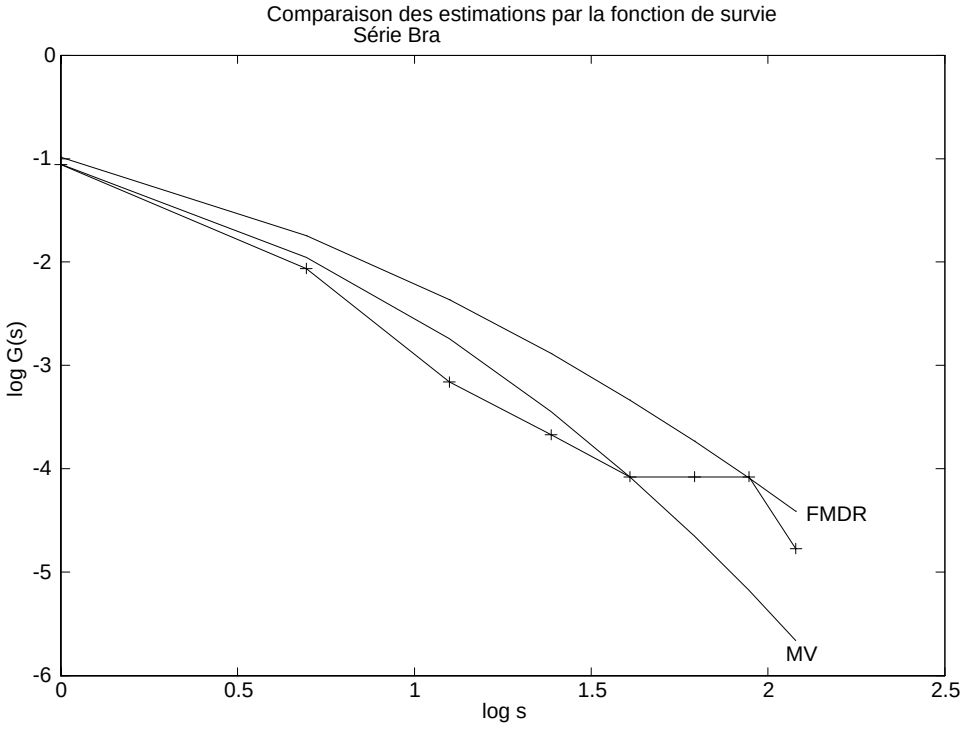




Figure 3.21: Athènes et Bra sur 5% des plus hautes observations : Comparaison des estimations du *MV* et de la *FMDR* de la fonction de survie comparées à la fonction de survie empirique.

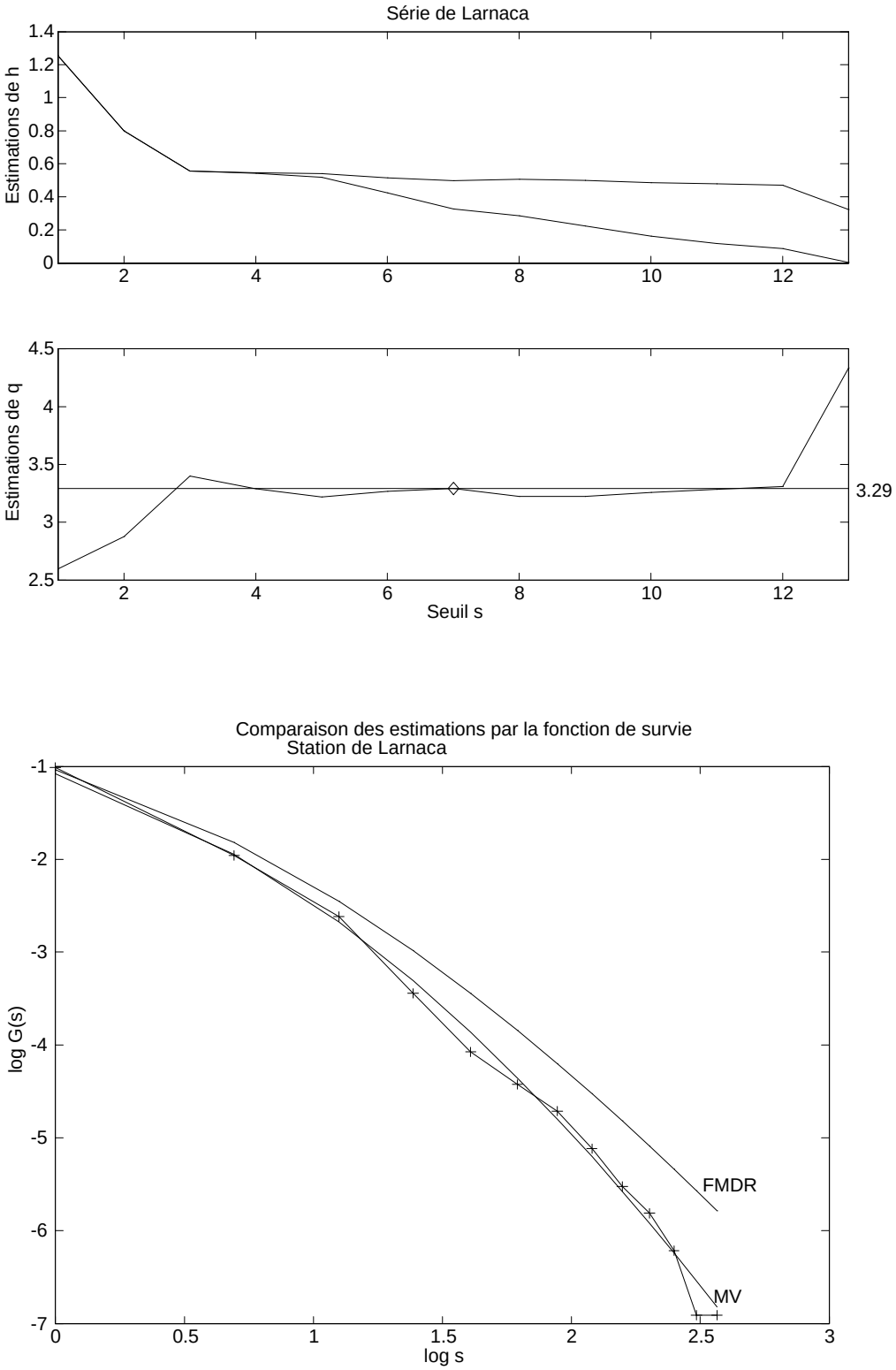



Figure 3.22: Série de Larnaca (5% des plus hautes observations) : Estimation de $q$ par la $FMDR$ et comparaison des estimations du $MV$ et de la $FMDR$ par la fonction de survie ('+' : estimateur standard de la fonction de survie).

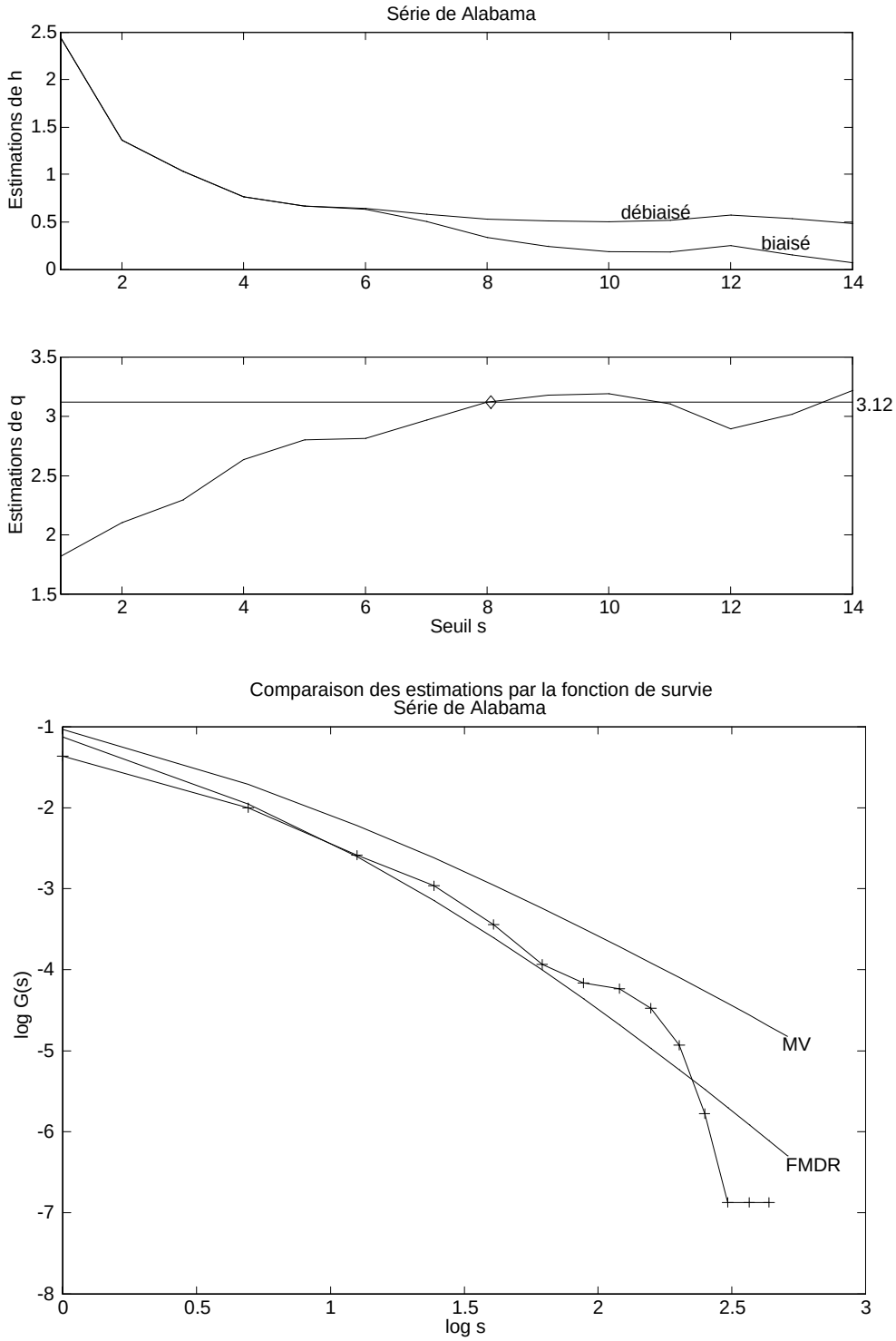



Figure 3.23: Série de Alabama (5% des plus hautes observations) : Estimation de $q$ par la $FMDR$ et comparaison des estimations du $MV$ et de la $FMDR$ par la fonction de survie ('+' : estimateur standard de la fonction de survie).

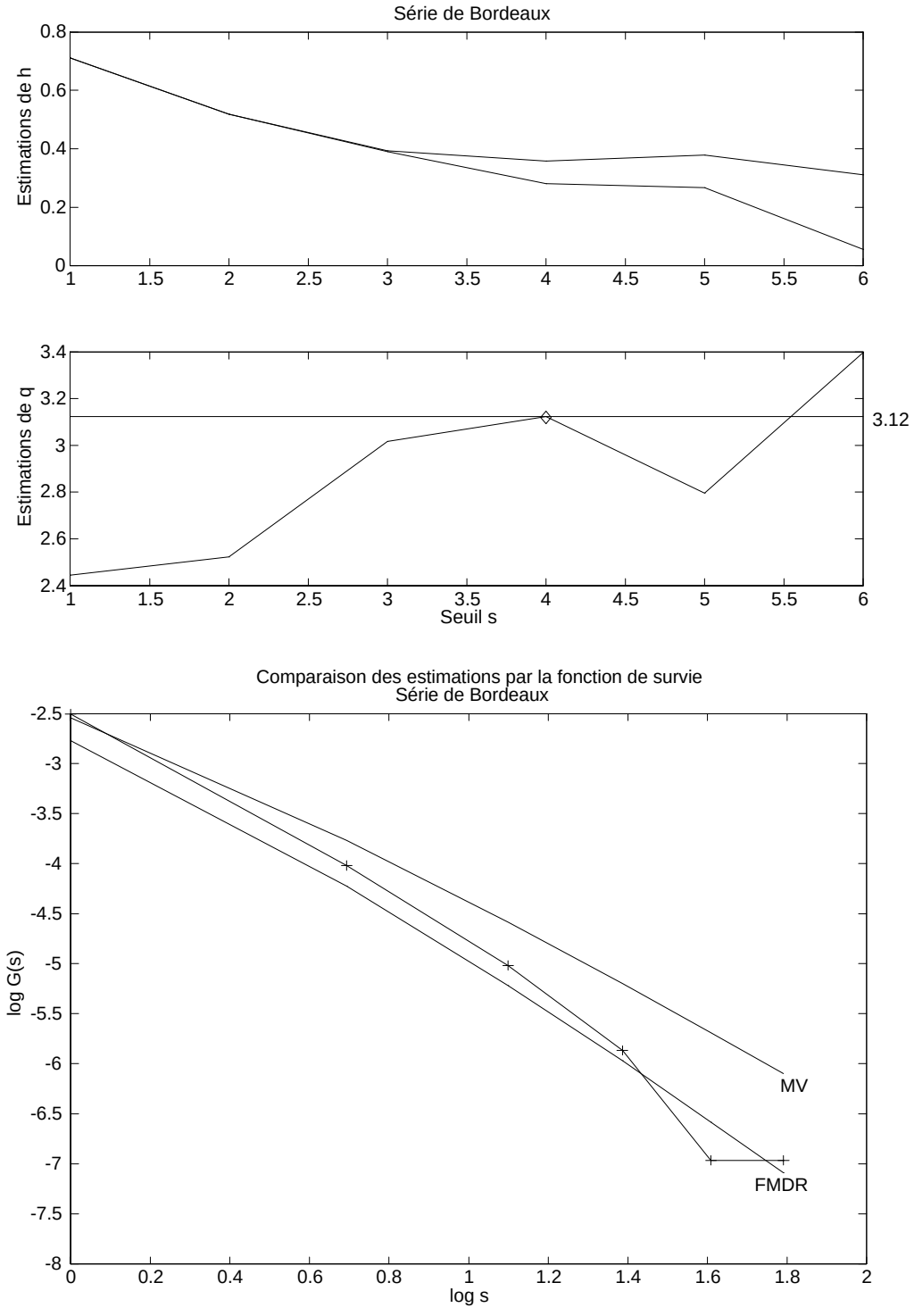



Figure 3.24: Série de Bordeaux (5% des plus hautes observations) : Estimation de $q$ par la $FMDR$ et comparaison des estimations du $MV$ et de la $FMDR$ par la fonction de survie ('+' : estimateur standard de la fonction de survie).

### 3.5.3 Données de l'Ile de la Réunion

On dispose d'une série de 2.5 ans de précipitations en une station de la partie Est de La Réunion[19]. Ce sont des données à basculement d'auget dont les dates de basculement sont exprimées en secondes, puis converties en intensités pluvieuses intégrées sur un pas de temps de 76 secondes (pour des raisons liées à la taille du fichier de données). La contenance maximale de l'auget est de 0.5 mm. Ce dispositif de mesure ne pourra donc pas déceler d'intensités supérieures à 0.5 mm/s[20].

L'estimation de $q$ sur la série où le pourcentage retenu est de 5% varie entre 1.20 et 2.80 (figure 3.25) et l'algorithme converge vers :

$$\widehat{q}_{FMDR} = 1.85$$

L'estimation du maximum de vraisemblance est de :

$$\begin{aligned} \widehat{q}_{MV} &= 3.88 \\ \widehat{\beta}_{MV} &= 0.50 \end{aligned}$$

La comparaison sur la base de la fonction de survie fait apparaître l'ampleur de la différence entre les deux estimations $\widehat{q}_{FMDR}$ et $\widehat{q}_{MV}$ (figure 3.25). En effet, la fonction de survie empirique présente un palier horizontal sur les dernières observations, qui est nettement sous-estimé par la méthode du maximum de vraisemblance. L'ajustement par la méthode de la *FMDR* sur les grands seuils est bon, malgré une légère sous-estimation sur les derniers seuils, probablement due à la troncature supérieure des données imposée par le dispositif de mesure.

### 3.5.4 Application à 232 longues séries annuelles

La base de données dont nous disposons est composée de séries de cumuls pluviométriques annuels relevés en 368 stations à travers le monde. Elle a été constituée par l'intermédiaire de plusieurs organismes, principalement l'UNESCO, l'IRD, le CIEH et Météo-France (voir *Annexe E* pour plus de détail sur les séries).

Les longueurs de ces séries se répartissent entre 39 et 299 ans et sont en moyenne de 106 ans. En raison des limites de validité des méthodes statistiques employées, nous ne retenons pour cette étude que les séries de plus de 100 ans (232 séries au total). Pour ces séries, la durée moyenne d'observation est de 106 ans, avec toutefois une forte concentration autour de 100 ans (plus de 60 % d'entre elles sont de longueur inférieure à 110 ans). Les séries de plus de 200 ans sont au nombre de 6 (voir la figure 3.26). Géographiquement, les stations considérées recouvrent tous les continents, bien qu'une forte proportion d'entre elles se situent en Europe (figure 3.26).

La méthode d'estimation du paramètre de décroissance algébrique par la FMDR est appliquée à ces longues séries annuelles. Il apparaît que les estimations sont en moyenne

[19]Données fournies par A. Barcelo (auquel j'adresse mes remerciements).

[20]Cette série comporte 15 intensités de 0.5 mm/s. On peut suspecter que ces dernières sont supérieures à cette valeur.

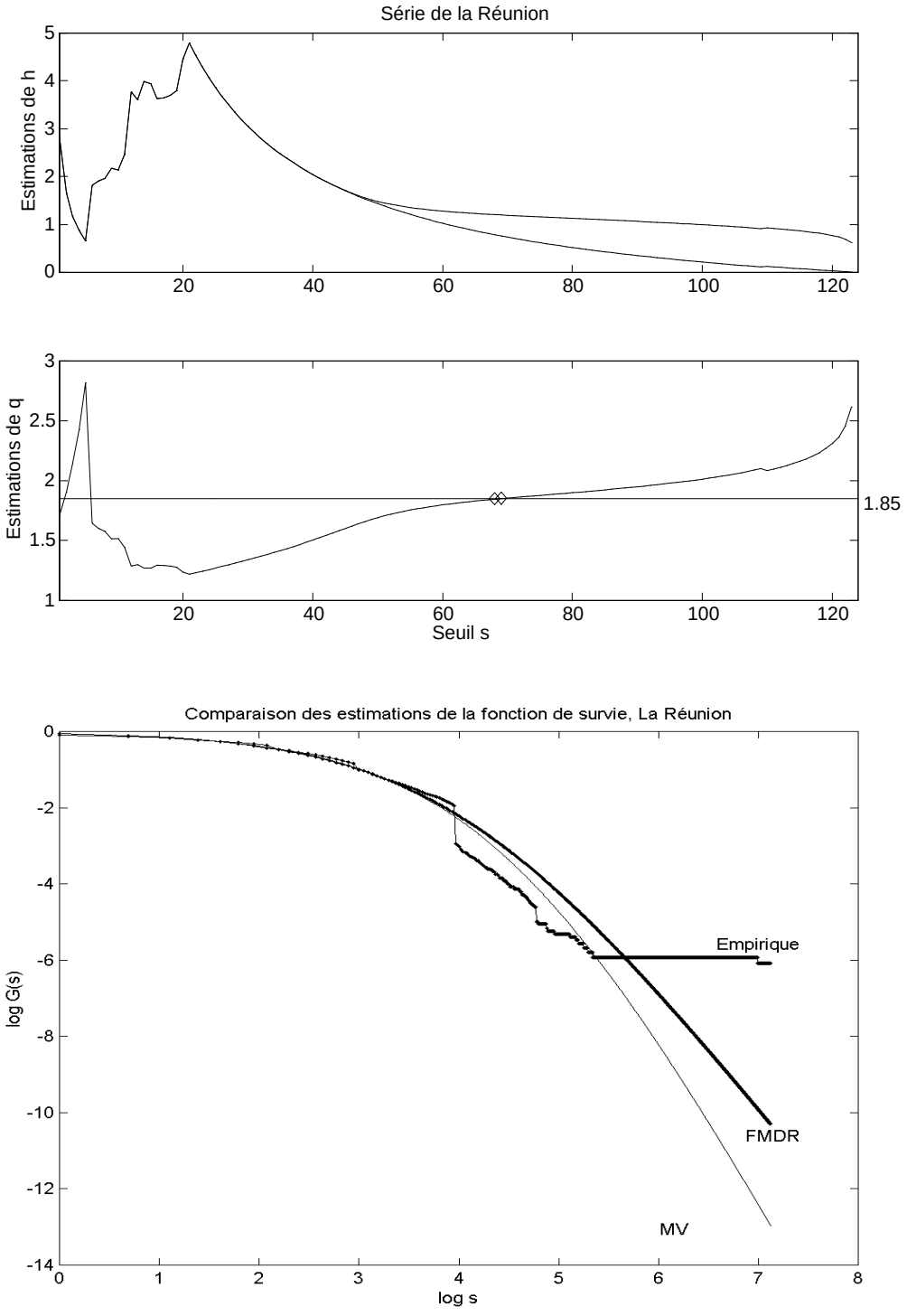



Figure 3.25: Série de la Réunion (5% des plus hautes observations) : Estimation de $q$ par la $FMDR$ et comparaison des estimations du $MV$ et de la $FMDR$ par la fonction de survie ('+' : estimateur standard de la fonction de survie).

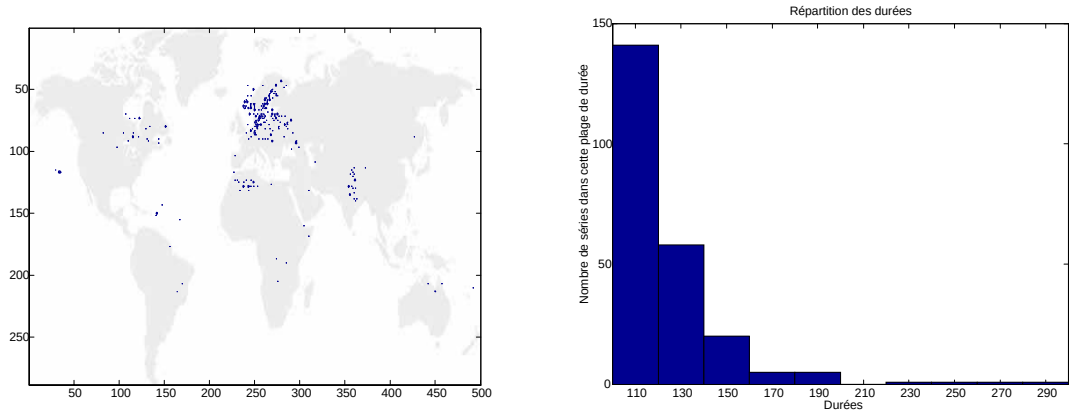



Figure 3.26: Les 230 stations étudiées ainsi que la répartition de la durée des séries.

égales à 3.26 avec une faible variance (égale à 0.12, voir figure 3.27). L'histogramme est unimodal et excentré vers la droite, ce qui s'explique par le fait que la majorité des longueurs des séries sont inférieures à 150 ans, et que l'estimation est d'autant plus faible que la série est courte (la justification théorique à été donnée au paragraphe 3.4.3, et la figure 3.28 permet de s'en rendre compte en pratique).

Afin de mieux visualiser l'effet de la longueur sur la valeur de l'estimateur, on réalise un histogramme des estimations en trois dimensions. Sur la figure 3.29, on a regroupé les 8 histogrammes des estimations de $q$ en partitionnant les séries annuelles en 8 classes de longueur : la première classe comprend les séries de longueur de 100 à 110 ans, la seconde de 111 à 120 ans, jusqu'à la $7^{ème}$ qui comprend les séries de 160 à 170 ans. La $8^{ème}$ est constituée des séries de 170 ans à 299 ans. Il apparaît bien que ces estimations sont en moyenne d'autant plus fortes que leur longueur est plus grande.

### 3.5.5 Application à des séries de débits

Dans cette section, on examine la forme de dépendance de séries de débits moyens journaliers de cours d'eau[21]. On retient de la base de données les 19 séries les plus longues (voir tableau 3.6).

L'exploration du caractère algébrique des queues de distributions de ces données ainsi que l'estimation du paramètre de décroissance algébrique est réalisée par le nouvel outil présenté en section 3.4.3. Là encore, le seuillage est fixé à 5% des plus grandes observations. Les résultats obtenus sont donnés par le tableau 3.7.

Là encore, les estimations sont très proches, elles se situent en moyenne autour de 2.6 malgré les différences de régimes de ces rivières.

[21]Séries extraites de la base de données hydrologique HYDRO du Ministère de l'Agriculture.

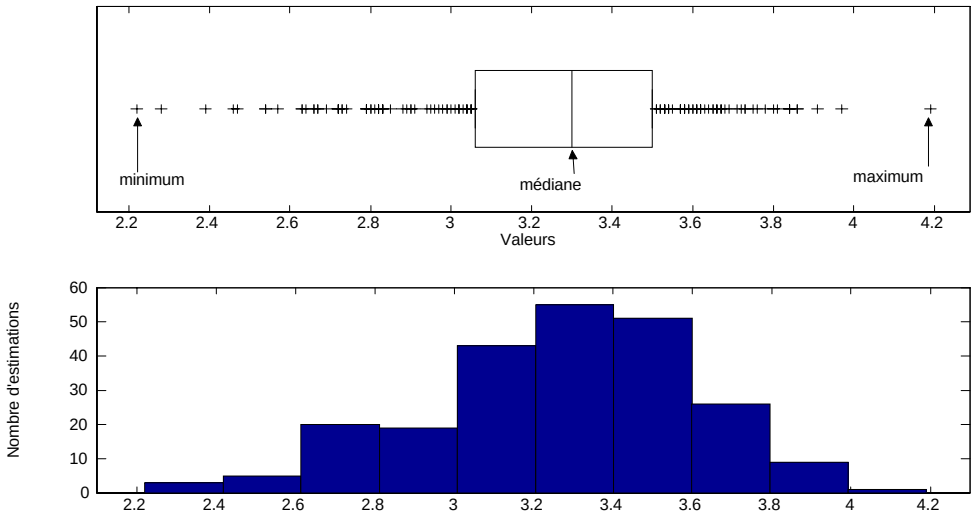



Figure 3.27: Boite à moustaches et histogramme des estimations de $q$ sur les 232 séries annuelles.

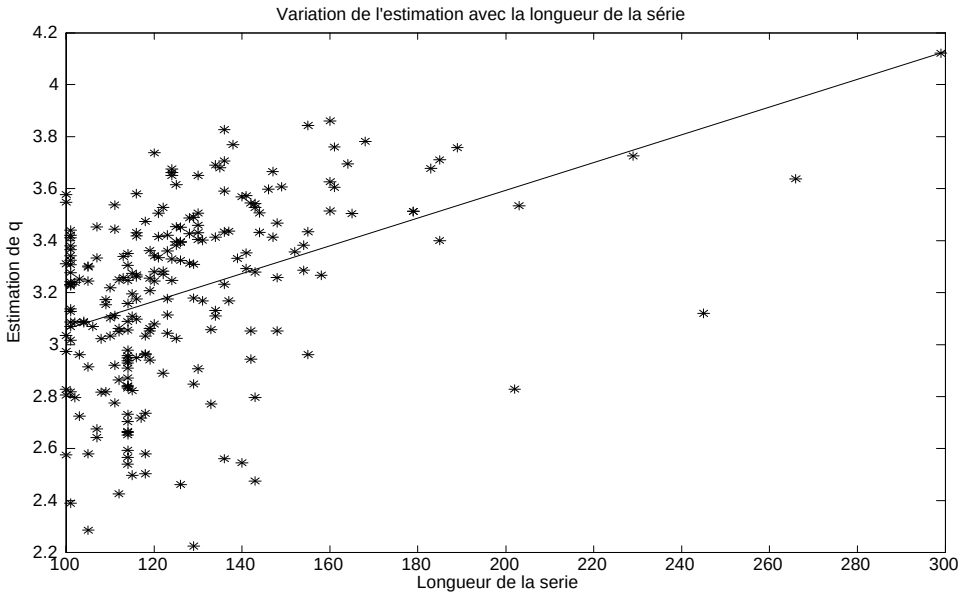



Figure 3.28: Effet de la longueur de la série sur l'estimation de $q$.

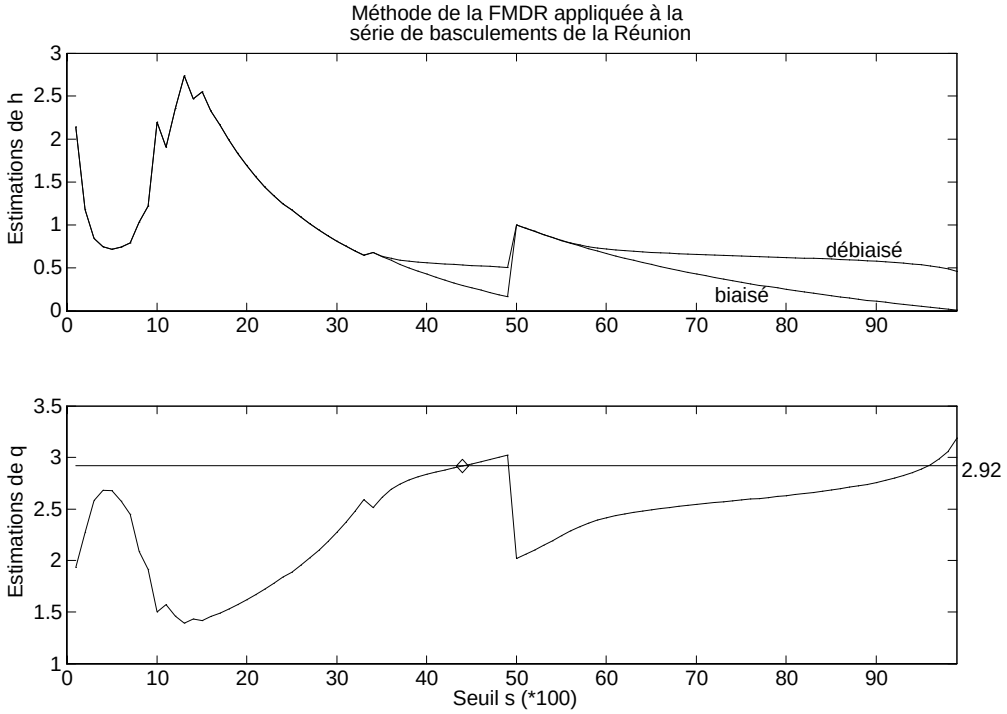



Figure 3.29: Histogramme des estimations de $q$.

| Rivière | Commune | Département | Bassin versant ($km^2$) |
|---|---|---|---|
| Authie | Dompierre | Somme | 726 |
| Averole | Bessans | Savoie | 45 |
| Corrèze | Corrèze | Corrèze | 167 |
| Corrèze | Brive | Corrèze | 947 |
| Dordogne | St Sauve | Puy de Dôme | 87 |
| Dordogne | Argentat | Corrèze | 4420 |
| Drôme | Luc en Diois | Drôme | 194 |
| Garonne | Le Mas d'Agenais | Lot et Garonne | 52000 |
| Herbissonne | Allibaudière | Aube | 87 |
| Isère | Val d'Isère | Savoie | 46 |
| Seine | Pont d'Austerlitz | Paris | 43800 |
| Seine | Poses | Eure | 65000 |
| Seine | Bar sur Seine | Aube | 2320 |
| Soude | Soudron | Marne | 105 |
| Suippe | Orainville | Aisne | 802 |
| Superbe | St Saturnin | Marne | 320 |
| Ubaye | Barcelonnette | Alpes Hte Prov. | 549 |
| Vienne | Le Palais sur Vienne | Haute-Vienne | 2300 |
| Zorn | Waltenheim | Haut-Rhin | 688 |

Tableau 3.6: Caractéristiques des stations.

### 3.5.6 Conclusion

Les comportements des queues de distribution des séries journalières de Dédougou, Athènes, Bra ou Larnaca, des séries horaires de Alabama et de Bordeaux, ainsi que de la série à pas de temps fin de l'Ile de la Réunion sont donc bien de type algébrique. De plus, les estimations du coefficient de décroissance algébrique sont proches : sur les séries étudiées et sur des gammes d'échelle temporelle allant de l'année à l'heure, c'est un quasi-invariant d'échelle égal à 3.46 en moyenne, l'écart type des estimations étant de 0.17. L'étude des séries annuelles, issues de régimes climatiques différents, permet de constater une invariance spatiale de ce paramètre, l'estimation se situant autour de 3.5 et l'écart type étant de 0.12 (voir tableau 3.8).

Pour une gamme d'échelle de l'ordre de la minute, cette invariance n'est pas conservée mais l'ajustement d'une loi de Pareto généralisée sur les extrêmes est bon, et la méthode de la *FMDR* plus performante que celle du maximum de vraisemblance sur ces séries irrégulières.

La plupart des séries étudiées précédemment sont cependant courtes sur le plan statistique. Les pluviomètres à basculement fournissent de nombreuses longues séries pluviométriques à pas de temps très fin, et peuvent permettre de constater si l'invariance d'échelle du paramètre de décroissance algébrique est conservée sur de fines gammes d'échelle.

| Rivière | $\widehat{q}$ | Altitude | Longueur (ans) | VM/an(*) |
|---|---|---|---|---|
| Authie | 4.13 | 12 | 27 | 234 |
| Averole | 2.69 | 1950 | 16 | 64 |
| Corrèze | 2.95 | 465 | 47 | 177 |
| Corrèze | 2.98 | 101 | 72 | 671 |
| Dordogne | 3.70 | 780 | 61 | 117 |
| Dordogne | 2.20 | 173 | 76 | 3370 |
| Drôme | 2.81 | 237 | 76 | 90 |
| Garonne | 2.43 | 17 | 74 | 19872 |
| Herbissonne | 2.54 | 91 | 21 | 11 |
| Isère | 2.25 | 1831 | 40 | 59 |
| Seine | 3.93 | 26 | 60 | 8986 |
| Seine | 3.09 | 18 | 49 | 12874 |
| Seine | 2.80 | 148 | 31 | 2583 |
| Soude | 2.22 | 110 | 19 | 19 |
| Suippe | 2.63 | 60 | 23 | 23 |
| Superbe | 3.66 | 79 | 17 | 17 |
| Ubaye | 3.84 | 1133 | 78 | 78 |
| Vienne | 2.80 | 230 | 70 | 70 |
| Zorn | 3.45 | 147 | 67 | 67 |

Tableau 3.7: Estimations du paramètre de décroissance algébrique $\widehat{q}$ sur les séries de débits (* : volume moyen écoulé par an ($10^6$ $m^3$).

## 3.6 Données à basculement d'auget

On dispose de peu de séries pluviométriques à pas de temps fin (inférieur à 15 minutes). Les séries à basculement d'auget peuvent fournir quelques cumuls sur un pas de temps égal à la précision de mesure du temps (en général 1 seconde). Cette section vise à permettre l'exploration et l'estimation du comportement algébrique de séries pluviométriques à petit pas de temps, sur la base de la série brute de dates de basculements. En effet, la conversion de toute la série des dates de basculement en une série d'intensités moyennes ou de cumul à un pas de temps fixé implique forcément une perte d'information. Plus précisément, on a un objectif double :

- explorer les longues séries disponibles pour estimer le coefficient de décroissance algébrique avec plus de précision
- savoir si le comportement multifractal concerne les fines gammes d'échelle.

### 3.6.1 Distribution des fréquences de basculement

Avant de déduire la loi des fréquences de basculement de celle des intensités pluvieuses, on s'intéresse à la distribution du nombre de basculements sur un intervalle de temps donné.

| Série | Nb de données | Durée (en ans) | $\widehat{q}_{FMDR}$ | $\widehat{q}_{MV}$ |
|---|---|---|---|---|
| Athènes | 44640 | 122 | 4.22 | 8.83 |
| Larnaca | 40680 | 111 | 3.29 | 4.56 |
| Bra | 47480 | 130 | 3.34 | 6.31 |
| Dédougou | 8395 | 23 | 3.63 | $2.10^7$ |
| Alabama | 19368 | 2 | 3.12 | 1.99 |
| Bordeaux | 42349 | 5 | 3.16 | 2.47 |
| la Réunion | 52645 | 2.5 | 1.85 | 3.88 |
| Séries annuelles | $> 100$ | $> 100$ | 3.26 | – |

Tableau 3.8: Récapitulatif des estimations obtenues.

**Loi du nombre de basculements**

On dispose d'une série d'intensités pluvieuses $(X_i)_{i=1\cdots N}$ discrétisées à pas de temps $P$.

Pour tout $t$ strictement positif, on note $K_0^t$ le nombre de basculements sur l'intervalle $[0, t[$. On convient que le premier basculement a lieu à la date $t = 0$ ($K_0^0 \equiv 1$). Ainsi $K_0^t$ est une variable aléatoire strictement positive.

On choisit $P = 1$, car on peut se ramener au cas $P$ quelconque en considérant les intensités $\left(\frac{X_i}{P}\right)_{i=1\cdots N}$.

Pour tout $t$ strictement positif et pour tout $0 \leq i < N - t$, on note $K_i^t$ le nombre de basculements sur l'intervalle $I_i(t) = [i, i + t[$. On suppose que les variables aléatoires $\left(K_i^t\right)_{i=0\ldots N-t-1}$ sont indépendantes et identiquement distribuées :

$$\left(K_i^t\right)_{i=0\ldots N-t-1} \overset{iid}{\sim} K^t$$

Notons que l'hypothèse de stationnarité semble raisonnable : elle consiste à supposer que le nombre de basculements sur un intervalle de temps $I_i(t)$ ne dépend pas de son origine mais de sa longueur. L'hypothèse d'indépendance est, quant à elle, risquée pour des valeurs de $t$ inférieures à la journée.

On note $Y_j^t$ les intensités moyennes sur les intervalles $I_i(t)$ :

$$\left(Y_j^t = \frac{1}{t}\sum_{i=j}^{t+j-1} X_i\right)_{j=1\cdots\left\lfloor\frac{N}{t}\right\rfloor}.$$

Ces intensités s'expriment comme le cumul $\delta K_i^t$ sur l'intervalle $I_i(t)$ divisé par la longueur $t$ de ce dernier :

$$Y_i^t = \delta K_i^t . \frac{1}{t} \qquad i = 1 \cdots \left\lfloor \frac{N}{t} \right\rfloor$$

Donc, sous les hypothèses précédentes, les $\left(Y_j^t\right)_{j=1\cdots\left\lfloor\frac{N}{t}\right\rfloor}$ sont elles aussi indépendantes et identiquement distribuées, et pour tout $t$ positif, la fonction de survie $G_{K^t}$ du

nombre de basculements $K^t$ sur une durée $t$ se déduit de celle des intensités $Y^t$ à pas de temps $t$ par une simple homothétie :

$$\forall k \geq 1 \quad G_{K_t}(k) = G_{Y^t}\left(\frac{k\delta}{t}\right)$$

En effet :

$$\forall k \geq 1 \quad P(K_t > k) = P\left(Y^t > \frac{k\delta}{t}\right)$$

Il en résulte que, si on suppose que pour tout $t$ positif, les intensités à pas de temps $t$ $(Y_j)_j$ suivent une loi asymptotiquement algébrique de paramètre $q$ :

$$G_{Y^t}(u) \overset{u\infty}{\sim} Au^{-q}$$

alors le nombre de basculements $K_t$ survenus avant la date $t$ suit aussi une loi algébrique de même paramètre :

$$G_{K^t}(k) \overset{u\infty}{\sim} \left(A\delta^{-q}t^q\right).k^{-q}$$

Mais la réciproque est tout autant applicable : On peut ainsi, à $t$ fixé, déterminer la distribution des intensités de pluie à pas de temps $t$, à partir d'une série de date de basculement d'auget, en analysant la série des nombre de basculements sur tout intervalle de longueur $t$. La valeur de $t$ choisie ne devra pas être cependant trop faible pour que l'hypothèse d'indépendance soit vérifiée.

**Loi des fréquences de basculements**

On note $D_k$ la date du $k^{ième}$ basculement. On a une seule réalisation de chacune des variables aléatoires $D_k$. Si on se place sous l'hypothèse précédente de stationnarité, on peut prendre pour origines des temps successivement les $k^{ièmes}$ basculements. On aura ainsi plusieurs réalisations de la variable $D_1$, date du deuxième basculement, qui est aussi la durée $T_1$ écoulée entre le premier et le deuxième basculement. Sous l'hypothèse de stationnarité, les durées inter-basculement $(T_i)_i$ sont identiquement distribuées, et leur loi découle de celle du nombre de basculements $K_t$ par la relation :

$$\forall t \geq 1 \quad G_T(t) = P(T_1 > t) = P\left(K_0^t = 1\right) = 1 - G_{Y^t}\left(\frac{\delta}{t}\right)$$

La distribution des durées inter-basculement est la loi inverse de $Y^t$, car l'inverse des durées inter-basculements (que l'on appellera les fréquences de basculements) suivent la loi de $Y^t$ :

$$\forall f \leq 1 \quad P(\frac{1}{T} > f) = 1 - G_T\left(\frac{1}{f}\right) = G_{Y_t}(\delta f) \qquad (1)$$

On a donc obtenu le résultat suivant : Sous hypothèse de stationnarité, les fréquences de basculements $\frac{1}{T}$ suivent la même loi que les intensités à pas de temps $t$ normalisées par le contenu de l'auget $\left(\frac{1}{\delta}Y_j^t\right)_{j=1\cdots\left\lfloor\frac{N}{t}\right\rfloor}$.

**Exemples** :

- Dans le cas où les intensités à pas de temps $t$ $\left(Y_j^t\right)_j$ suivent une loi asymptotiquement algébrique de paramètre $q$, la fréquence de basculements $\frac{1}{T}$ est asymptotiquement algébrique de même paramètre. En effet, par (1) :

$$\forall f \leq 1 \quad P(\frac{1}{T} > f) = G_{Al}(\delta f)$$

- Dans le cas où les intensités $(X_i)_i$ sont indépendantes et suivent une loi exponentielle $Exp(a)$, les cumuls à pas de temps $t > 0$ $\left(tY_j^t\right)_j$ suivent une loi Gamma $\Gamma(a,t)$ et par (1) les fréquences de basculements suivent une loi $\Gamma(at\delta, t)$ :

$$\forall f \leq 1 \quad P(\frac{1}{T} > f) = G_{\Gamma(a,t)}(t\delta f) = G_{\Gamma(at\delta,t)}(f)$$

- Cas Normal : $(X_i)_i \overset{iid}{\sim} N\left(\mu, \sigma^2\right)$ alors pour tout $t > 0$, $\left(Y_j^t\right)_j \overset{iid}{\sim} N\left(\mu, \frac{\sigma^2}{t}\right)$ et les fréquences de basculements suivent une loi $N\left(\mu, \frac{\sigma^2}{t\delta^2}\right)$ car :

$$\forall f \leq 1 \quad P(\frac{1}{T} > f) = G_{N\left(\mu, \frac{\sigma^2}{t}\right)}(\delta f) = G_{N\left(\mu, \frac{\sigma^2}{t\delta^2}\right)}(f)$$

**Traduction graphique :** En reportant dans un graphe log – log la fonction de survie des fréquences de dépassements, on doit observer un comportement linéaire dans le cas algébrique, et une décroissance plus rapide dans le cas exponentiel.

### 3.6.2 Application aux données

Les mesures de dates de basculement, bien que plus précises pour les faibles intensités, imposent cependant une borne supérieure aux intensités (égale à la contenance de l'auget sur la plus petite durée mesurable) dont il faudra tenir compte. De plus, en raison de la saturation éventuelle de l'auget, en pratique il faut s'attendre à un problème de troncature supérieure des intensités et des fréquences de basculement. On négligera donc les seuils trop élevés.

#### Série de la Réunion

La série de fréquence de basculement de la Réunion comporte 54 459 données. On constate que le comportement de la queue de distribution des fréquences de basculement de l'auget est de type algébrique (figure 3.30). Ce résultat d'analyse exploratoire est confirmé par la limite de $\widehat{h}_{débiaisé}$ qui est de 0.5. L'estimation du coefficient de décroissance algébrique $q$ par les méthodes du maximum de vraisemblance et de la *FMDR* sont voisines :

$$\begin{aligned} \widehat{q}_{MV} &= 3.96 \\ \widehat{\beta}_{MV} &= 1.30 \\ \widehat{q}_{FMDR} &= 2.92 \end{aligned}$$

mais la qualité de l'ajustement à la fonction de survie est différente (figure 3.30).

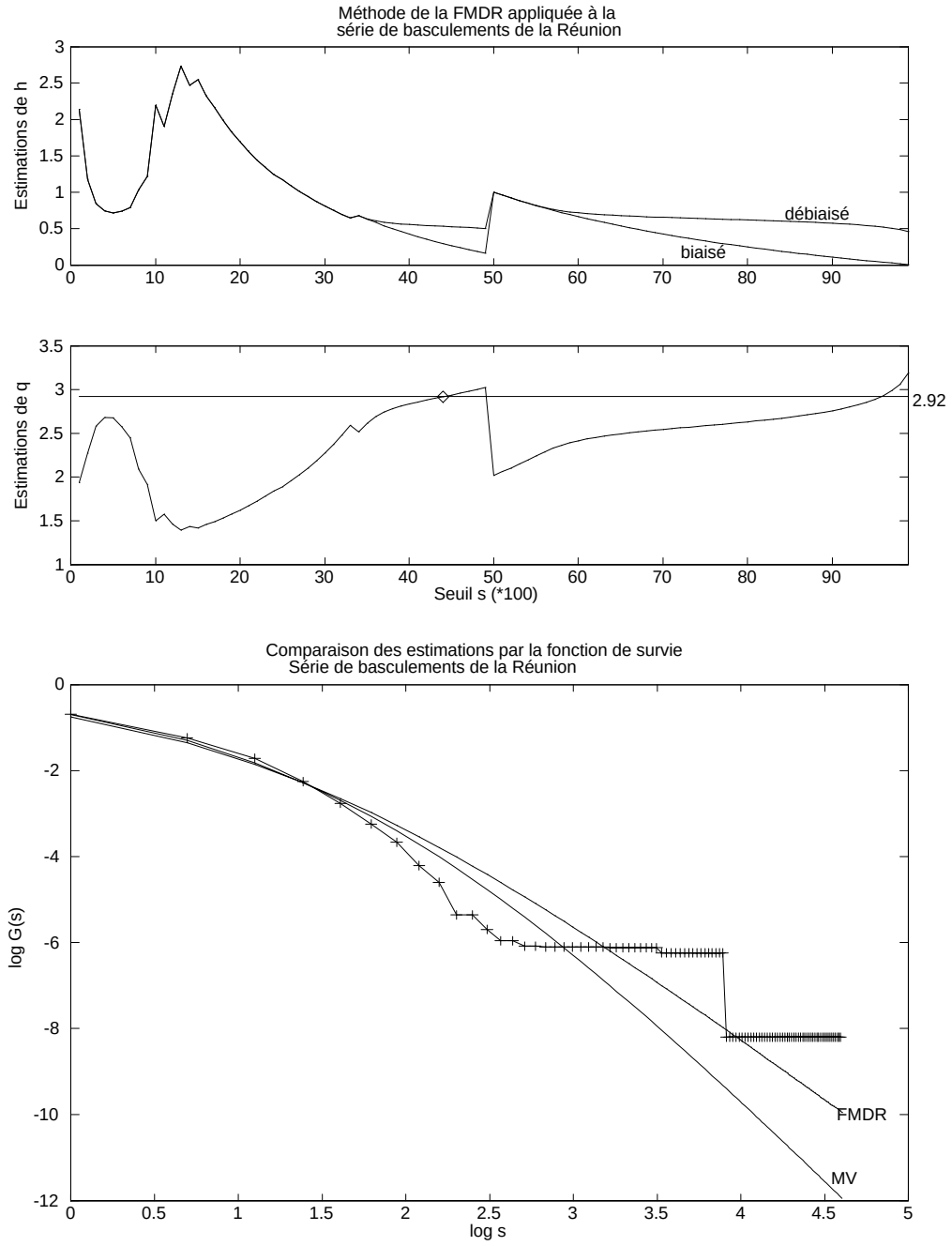



Figure 3.30: Estimation de la fonction de survie de la série des grandes fréquences de basculement de la Réunion (troncature à 5%) : Méthodes du maximum de vraisemblance (MV) et de la FMDR.

**Données EPSAT-Niger**

Les données EPSAT-Niger (Estimation des Précipitations par SATellite-expérience NIGER) résultent conjointement d'un réseau de pluviomètres à basculement d'auget (93 postes sur 16 000 km$^2$) et d'un radar météorologique bande C (pour plus de précisions, on pourra se reporter à l'article de Lebel et al. [55]). L'enregistrement des dates de basculement d'auget est effectué sur trois ans et constitue une vaste base de données pour notre étude. Sur la figure 3.31, on constate que le comportement de la queue de distribution des fréquences de basculements de l'auget est de type algébrique ($\widehat{h}_{débiaisé} = 0.6$). L'estimation du coefficient de décroissance algébrique $q$ par la méthode du maximum de vraisemblance est inférieure à celle de la *FMDR* :

$$\begin{aligned} \widehat{q}_{MV} &= 2.15 \\ \widehat{\beta}_{MV} &= 1.40 \\ \widehat{q}_{FMDR} &= 2.50 \end{aligned}$$

mais la qualité de l'ajustement de la fonction de survie estimée par la *FMDR* à la fonction de survie empirique est meilleure que par le maximum de vraisemblance.

## 3.7 Conclusions

Les résultats de l'étude sur un grand nombre de séries accréditent les hypothèses :

- d'un comportement algébrique de la distribution des séries d'extrêmes pluviométriques
- d'invariance d'échelle, sur une gamme d'échelle s'étalant de 1 heure à 1 an, de l'exposant de décroissance algébrique $q$
- d'invariance spatiale de ce dernier.

Sur le plan pratique, l'écart entre les modèles algébriques et les modèles *traditionnels* de pluie est important, surtout si l'on s'intéresse à la modélisation des événements rares. Pour comparaison, une période de retour estimée millennale par un modèle exponentiel ne sera que centenale par un modèle algébrique. L'ampleur de la différence d'estimation des périodes de retour des événements rares nécessite d'intégrer de telles distributions dans les modèles d'analyse de risque hydrologique ou de coût des ouvrages de génie civil.

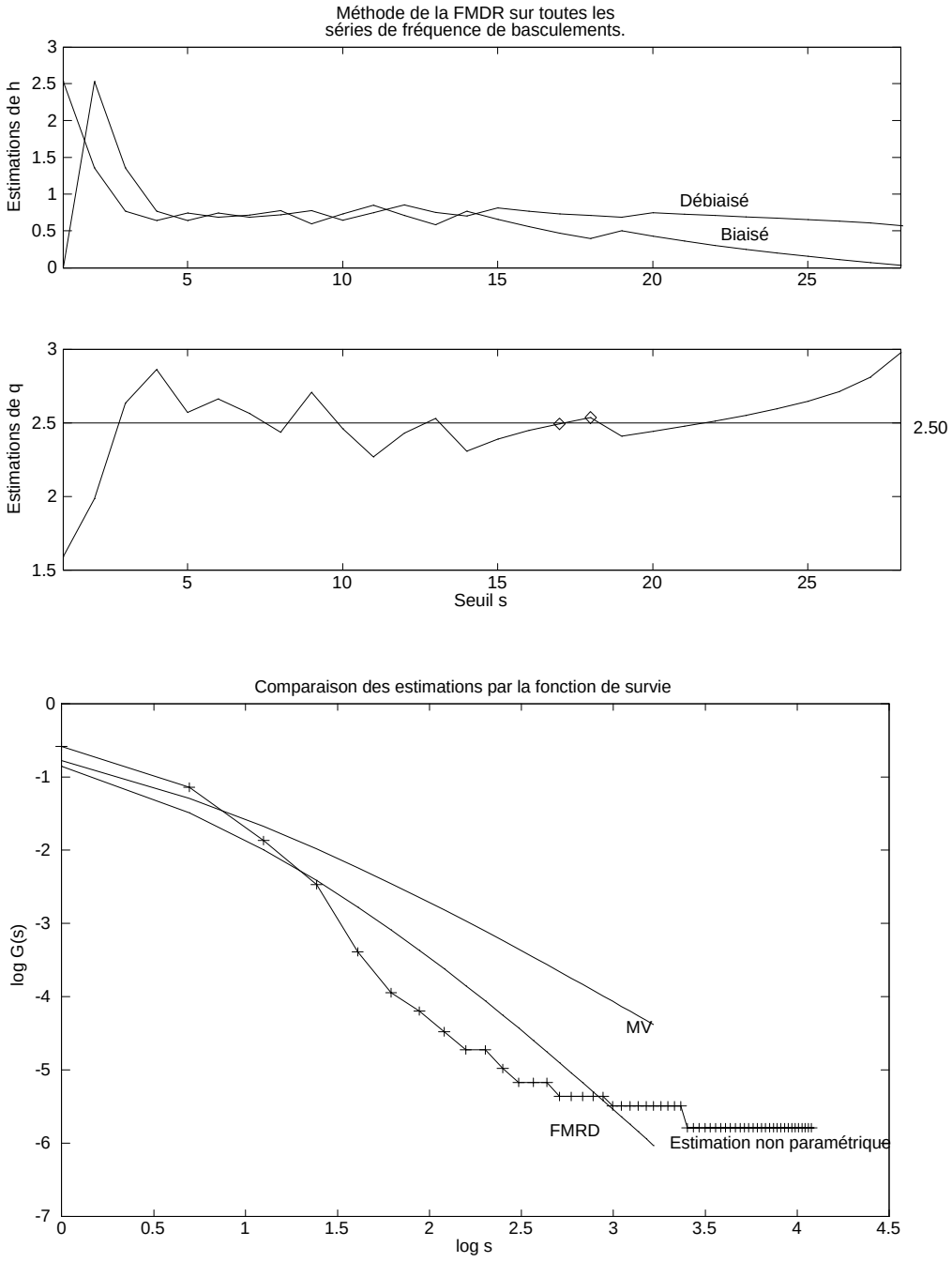



Figure 3.31: Estimation de la fonction de survie de la série des grandes fréquences de basculement EPSAT (troncature à 5%) : Méthodes du maximum de vraisemblance (MV) et de la FMDR.

# Chapitre 4

# Les cascades multifractales et la longue dépendance

En hydrologie, tout comme en géophysique, en climatologie ou en économie, la planification expérimentale est impraticable en situation réelle, et l'on doit considérer les observations au fur et à mesure qu'elles se présentent. Le statisticien a tendance à vouloir modéliser les séries à l'aide de variables aléatoires indépendantes, car elles permettent une analyse statistique simple. Or les données hydrologiques sont influencées par plusieurs facteurs, dont certains échappent au contrôle et même à la connaissance de l'hydrologue. Travailler avec l'hypothèse que les données sont indépendantes est donc irréaliste ; au contraire, l'étude de la structure de dépendance de ces dernières pourrait restituer ou refléter la structure interne du phénomène pluvieux.

Mais la principale motivation de la modélisation en cascades multifractales est liée à l'invariance d'échelle du paramètre de décroissance algébrique, constatée dans la partie précédente. Les cascades multifractales sont les modèles les plus simples retranscrivant cette propriété. Mais rendent-ils compte d'autres propriétés statistiques constatées en pratique ? La longue dépendance est une propriété statistique non classique susceptible d'être présentée par tout système constitué d'un grand nombre de facteurs en interaction. Elle se caractérise par une décroissance de la fonction de corrélation avec le temps $t$ si faible qu'elle n'en est plus sommable.

De telles propriétés ont été mises en évidence dans des séries financières ou en télécommunication sur des séries de trafic (Lang, 94 [54] ou Resnick, 97 [71]). Plusieurs hydrologues et statisticiens s'intéressant à des domaines de la géophysique ont constaté un comportement de longue dépendance au sein des séries étudiées.

Ces propriétés de longue dépendance ne sont pas décrites par les modèles usuels (ARMA, Markov), et l'élaboration de modèles possédant une telle propriété n'a commencé que dans les années 50, avec la constatation du phénomène de Hurst (Hurst, 51 [46]). Parmi ces modèles, on peut citer le mouvement Brownien fractionnaire ou encore les agrégations de processus AR. Les modèles en cascades multifractales ont fait l'objet d'une telle étude : il a été décelé une décroissance algébrique de la fonction de corrélation (Marsan, 96 [62]). Mais peut-on pour autant en conclure à la présence de longue dépendance ?

Le présent chapitre rappelle la définition mathématique de la longue dépendance,

puis passe en revue les principales études des manifestations ce phénomène en hydrologie ces dernières décennies. Nous présenterons par la suite quelques modèles stochastiques rendant compte de la propriété de longue dépendance. Nous analyserons dans quelle mesure les cascades multifractales possèdent la propriété de longue dépendance. Enfin, la longue dépendance de quelques longues séries de débits sera estimée à l'aide d'une méthode statistique performante (le log-périodogramme global), et les résultats seront confrontés aux estimations issues d'outils statistiques plus classiques.

## 4.1 Notations

Avant de définir la longue dépendance, nous introduisons dans cette section les notations ainsi que les hypothèses choisies. Tout au long de cette partie, on adoptera les notations suivantes : la série d'observations hydrologiques aux instants $i$ (cumul de pluie ou débit de rivière) sera notée $(X_i)_{i=1\cdots n}$ . On considérera qu'elle constitue une suite de variables aléatoires identiquement distribuées de fonction de répartition $F$, et l'on ne s'intéresse qu'au cas où les moments d'ordres 1 et 2 existent :

$$\mu = E[X_i] < \infty \text{ et } \sigma^2 = Var[X_i] < \infty$$

Les covariances (resp. corrélations) aux instants $i$ et $j$ entre les observations $X_i$ et $X_j$ seront notées :

$$\gamma(i,j) = E[(X_i - \mu)(X_j - \mu)] \text{ (resp. } \rho(i,j) = \frac{1}{\sigma^2}\gamma(i,j)\text{ )}$$

On se place dans le cadre de séries stationnaires. Sous cette hypothèse, les covariances (resp. corrélations) s'écrivent :

$$\gamma(i,j) = \gamma(|i-j|) \text{ (resp. } \rho(i,j) = \rho(|i-j|)\text{ )}$$

## 4.2 Définitions de la longue dépendance

On trouve dans la littérature diverses définitions de la longue dépendance. On présente ici deux définitions asymptotiques (Taqqu, 00 [79]) qui caractérisent le comportement limite de la fonction de corrélation $\gamma$ quand la distance tend vers l'infini. Comme nous l'avons énoncé précédemment, la longue dépendance se traduit par une décroissance algébrique de la fonction de corrélation $\gamma$ vers zéro, et par une vitesse de convergence vers zéro si faible que $\sum_{-\infty}^{+\infty} \gamma(k)$ diverge :

**Définition 27** *Soit $(X_i)_i$ une série stationnaire. On dit que $(X_i)_i$ est à longue dépendance si et seulement si :*

$$\sum_{-n}^{+n} \gamma(k) \sim n^{\alpha} L_1(n), \quad n \to \infty, \quad 0 < \alpha < 1$$

*avec $L_1$ une fonction à variations lentes c'est-à-dire bornée sur tout intervalle et telle que :*

$$\forall a > 0 \quad \frac{L(ax)}{L(x)} \overset{x\to\infty}{\longrightarrow} 1$$

Mais on trouve dans la littérature une autre définition qui ne lui est pas exactement équivalente :

**Définition 28** *On dit que $(X_i)_i$ stationnaire est à longue dépendance si et seulement si :*
$\gamma(k) \sim k^{-\beta} L_2(k), \quad k \to \infty, \quad 0 < \beta < 1$
*avec $L_2$ une fonction à variations lentes.*

$(X_i)_i$ est encore dite à longue mémoire, à dépendance à long terme, ou à forte dépendance.

Pour citer des contre-exemples, les modèles couramment utilisés en hydrologie tels que les ARMA (Box et Jenkins, 70 [15]) ou les chaînes de Markov ne sont pas à longue dépendance. Ils sont à courte dépendance car leur fonction de corrélation décroît exponentiellement (en $e^{-\alpha t}$, $\alpha > 0$).

Ces deux définitions sont équivalentes dans le cas d'une fonction de corrélation $\gamma$ monotone par la proposition suivante (Taqqu, 00 [79]) :

**Proposition 29** *Si $\gamma(k)$ est asymptotiquement monotone quand $k \to \infty$, alors les deux définitions sont équivalentes et on a les relations suivantes :*
$\alpha = 1 - \beta, \quad L_1(x) = 2(1-\beta)^{-1} L_2(x)$

Remarquons que lorsqu'on s'intéresse à un comportement de longue dépendance, on ne cherche pas à quantifier les corrélations dans le but de constater si elles sont plus ou moins proches de zéro (par exemple en regardant les $\pm 2/\sqrt{n}-$intervalles de confiance du graphe de $\widehat{\rho}(k)$ vs $k$), mais plutôt à quantifier la vitesse de convergence de $\widehat{\rho}$ vers zéro quand la distance $k$ tend vers l'infini. En effet, chacune des corrélations peut être arbitrairement petite et appartenir à l'intervalle de confiance. On ne peut donc pas visualiser la longue dépendance sur le corrélogramme. Par contre, comme nous le verrons en section 5.5.1, le graphe de la variance constitue un moyen d'accéder à l'estimation du paramètre de longue dépendance $H$. Cet outil est fondé sur le théorème suivant :

**Proposition 30** *Soit $(X_i)_i$ une série à longue dépendance. On a alors :*

$$\lim_{n\to\infty} \left[ \frac{Var(\overline{X})}{c_\gamma . n^{-2H-2}} \right] = \frac{1}{H(2H-1)}$$

La variance de la moyenne empirique décroît vers zéro à une vitesse inférieure à $n^{-1}$. La vitesse de décroissance est proportionnelle à $n^{-\gamma}$, $\gamma = 2H + 2$.

## 4.3 La longue dépendance des séries hydrologiques

L'étude de la longue dépendance des séries hydrologiques a commencé avec les travaux de H.E. Hurst [46], qui s'intéressa à la validité de l'hypothèse d'indépendance adoptée en pratique pour la modélisation des débits maximum annuels.

S'intéressant à la capacité idéale d'un réservoir, c'est-à-dire à la taille optimale d'un réservoir conçu pour absorber les débits $X_1, X_2, \cdots, X_k$ arrivant entre les instants 1 et $k$, tout en ayant une vidange constante, un contenu identique aux instants 1 et $k$, et ne débordant jamais, il définit l'étendue $R$ :

$$\begin{aligned} R(t,k) &= \max_{0\leq i\leq k}\left[Y_{t+i} - Y_t - \frac{i}{k}\left(Y_{t+k} - Y_t\right)\right] - \min_{0\leq i\leq k}\left[Y_{t+i} - Y_t - \frac{i}{k}\left(Y_{t+k} - Y_t\right)\right] \\ \text{avec } Y_t &= \sum_{i=1}^{t} X_i \end{aligned}$$

et $S$ l'écart type des débits observés :

$$S(t,k) = \left[\frac{1}{k}\sum_{i=t+1}^{t+k}\left(X_i - \overline{X}_{t,k}\right)^2\right] \text{ où } \overline{X}_{t,k} = \frac{1}{k}\sum_{i=t+1}^{t+k} X_i$$

Sur la longue série des débits du Nil (constituée conjointement d'observations systématiques et paléohistoriques et couvrant les années 622 jusqu'à 1281), il constata un comportement linéaire de la courbe $\log - \log$ de l'étendue normalisée $R/S$ par rapport au nombre d'observations $k$, et définit le coefficient de Hurst $H$ comme la pente de cette dernière :

$$Q_t(k) = \frac{R(t,k)}{S(t,k)} \propto k^H$$

L'estimation du coefficient de Hurst $H$ sur la série du Nil ainsi que d'autres séries de débits conduisit à des valeurs de l'ordre de 0.7. Sur d'autres séries d'observations géophysiques, des estimations de $H$ du même ordre furent trouvées.

Les modèles statistiques classiques ne rendent pas compte de cette caractéristique. En effet, en 1951 Feller [34] montra que dans le cas de séries totalement indépendantes, le coefficient $H$ était de $\frac{1}{2}$. Divers auteurs cherchèrent à trouver des modèles retranscrivant cet effet, dans le cadre des modèles stationnaires (étant donné que cet effet peut aussi être expliqué par une non-stationnarité ou un comportement pré-asymptotique de modèle autorégressif d'ordre 1 (Matalas, 67 [63])). Feller [34] s'orienta vers une structure de dépendance markovienne, mais Barnard en1956 [5] montra qu'une telle dépendance conduisait aussi à une estimation de $\frac{1}{2}$. Mandelbrot ([60] et [61]) fût le premier à proposer un modèle de longue dépendance stationnaire, le mouvement brownien fractionnaire, que nous présentons dans la section qui suit.

## 4.4 Le mouvement brownien fractionnaire

Le mouvement brownien fractionnaire, proposé par Mandelbrot [61], est un modèle à accroissements stationnaires et à longue dépendance. Nous rappelons succinctement

la définition de ce modèle avant d'énoncer deux théorèmes permettant de distinguer les processus à longue dépendance des autres.

**Définition 31** *$(Y_t)_t$ est un mouvement brownien $B(t)$ si et seulement si :*
*1/ $Y_t$ est une variable aléatoire gaussienne*
*2/ $Y_0 = 0$ p.s*
*3/ $(Y_t)_t$ est à accroissements stationnaires*
*4/ $E(Y_t - Y_s) = 0$*
*5/ $Var(Y_t - Y_s) = \sigma^2 |t - s|$*

**Définition 32** *$(Y_t)_t$ est un mouvement brownien fractionnaire $B_H(t)$ si et seulement si*
*1/ $Y_t$ est une variable aléatoire gaussienne*
*2/ $Y_0 = 0$ p.s*
*3/ $(Y_t)_t$ est à accroissements stationnaires*
*4/ $E(Y_t - Y_s) = 0$*
*5/ $Var(Y_t - Y_s) = \sigma^2 |t - s|^{2H}$*

**Proposition 33** *Un mouvement brownien fractionnaire $B_H(t)$ est à longue dépendance et son paramètre de longue dépendance $\beta$ s'écrit (Taqqu, 00 [79]) :*

$$\beta = 2 - 2H \ (ou\ \alpha = 2H - 1)$$

Les deux théorèmes limites suivants (Beran, 94 [10]) permettent de distinguer les processus à longue dépendance (exposant de Hurst supérieur à $\frac{1}{2}$) des processus à courte dépendance ($H = \frac{1}{2}$).

**Théorème 34** *Soit $(X_t)_t$ tel que ${X_t}^2$ soit ergodique et $t^{-\frac{1}{2}} \sum_{s=1}^{t} X_s \overset{t\to\infty}{\longrightarrow} B(t)$. Alors,*

*$k^{-\frac{1}{2}} Q_t(k) \overset{k\to\infty}{\longrightarrow} \xi$ avec $\xi$ v.a. non dégénérée.*

Les conditions de ce théorème sont vérifiées par la plupart des processus stationnaires, en particulier dans tous les cas où le Théorème de la Limite Centrale est vérifié, puisqu'il existe une variable aléatoire $Z$ telle que :

$$k^{-\frac{1}{2}} Q_t(k) \overset{k\to\infty}{\longrightarrow} Z$$

**Théorème 35** *Soit $(X_t)_t$ tel que ${X_t}^2$ soit ergodique et $t^{-H} \sum_{s=1}^{t} X_s \overset{t\to\infty}{\longrightarrow} B_H(t)$. Alors,*

*$t^{-H} Q_t(k) \overset{k\to\infty}{\longrightarrow} \xi$ avec $\xi$ v.a. non dégénérée.*

Donc le graphe $(\log k, \log Q_t(k))$ a un comportement asymptotique rectiligne de pente $H$, exposant de Hurst, qui est égal à $\frac{1}{2}$ pour tout processus stationnaires à courte dépendance et supérieur à $\frac{1}{2}$ pour tout processus stationnaires à longue dépendance.

D'autres caractéristiques statistiques propres aux séries de débits ou d'extrêmes pluviométriques sont aussi retranscrites par le mouvement brownien fractionnaire. En effet,

la plupart des séries de débits présentent de longues périodes de basses eaux suivies de longues périodes de hautes eaux. De plus, bien que ces séries semblent stationnaires et ne présentent ni cycle ni tendance, sur de petits intervalles de temps, on distingue de petits cycles et des tendances locales. Le mouvement brownien fractionnaire reproduit ces caractéristiques appelées "Effet Joseph" par Mandelbrot [61].

Plus généralement, vérifient la propriété de longue dépendance (voir *Annexe* **??**).

Parmi les modèles stationnaires de longue dépendance, citons les accroissements stationnaires de processus auto-similaires (d'exposant $H > \frac{1}{2}$), ou encore le modèle d'agrégation de processus auto-régressifs d'ordre 1 (Granger, 80 [40]). Ce dernier modèle possède l'avantage d'être simple à interpréter physiquement et pourrait répondre à l'interrogation de Klemes [52] sur la forme de mécanisme physique à la source de la retransmission de l'influence d'une température moyenne d'une année sur des dizaines ou des centaines d'années.

## 4.5 Les cascades sont-elles à longue dépendance ?

La pluie, comme tout autre phénomène géophysique, possède une grande variabilité, sur des gammes d'échelles (spatiales et temporelles) très étendues. Comme on l'a vu en *section 2.3.3*, les modèles les plus simples rendant compte de cette caractéristique sont les modèles en cascades multifractales.

L'objectif de cette section est de déterminer la structure de dépendance entre les cumuls désagrégés à une échelle donnée, et de comparer cette dernière à une situation où l'on a indépendance (de type processus ponctuel de Poisson). Les cascades multifractales possèdent-elles toutes la propriété de longue dépendance ? Quelles sont les conditions nécessaires sur les générateurs pour la retranscription de cette propriété ?

### 4.5.1 Définition de la cascade

On rappelle la définition d'une cascade multifractale : Une cascade multifractale de générateur $\eta$ est une suite de variables aléatoires $\left(\mu_{j,k}\right)$ définies par la relation de récurrence (figure 2.12 de la *section 2.3.3*) :

$$\begin{array}{rcl} \mu_0 & = & \mu_{0,0} > 0 \\ \text{et } \forall j & \geq & 1 \quad \forall k = 0 \cdots 2^j - 1, \quad \mu_{j,k} = \mu_{j-1,\lfloor k/2 \rfloor} . \eta_{j,k} \end{array}$$

En désagrégeant jusqu'au niveau d'homogénéité $N$, on obtient une suite de variables aléatoires $\left(\mu_{N,k}\right)_{k=0\cdots 2^N-1}$, les cumuls sont de $\mu_{N,0}$ sur $\left[0, \frac{T}{2^N}\right[$, ..., de $\mu_{N,2^N-1}$ sur $\left[(2^n-1)\frac{T}{2^N}, T\right[$.

Il est important de remarquer qu'en hydrologie, on ne s'intéresse pas à une désagrégation "infinie" ($j$ tend vers l'infini) mais plutôt à un niveau de désagrégation limite $N$ nommé niveau d'homogénéité de la cascade.

**Hypothèse** : Les générateurs $\left(\eta_{j,k}\right)_{j=1\cdots N, k=0\cdots 2^j-1}$ sont indépendants deux à deux et identiquement distribués :

$$\left(\eta_{j,k}\right)_{j=1\cdots N, k=0\cdots 2^j-1} \; iid$$

On note $e_1$ et $e_2$ les moments d'ordre 1 et 2 du générateur $\eta$ :

$$e_1 = E\eta \text{ et } e_2 = E\left(\eta^2\right)$$

### 4.5.2 Type de dépendance des cumuls

#### Notions de conservativité des cumuls

Lorsqu'on travaille sur des cumuls de pluie, une contrainte vient s'imposer dans la construction de la cascade : la condition de conservativité. En effet, une cascade de cumuls doit être conservative dans le sens où à chaque niveau $j$ de son développement $j = 1 \cdots N$, la somme des cumuls de ce niveau vaut $\mu_0$ :

$$\forall j = 1 \cdots N \quad \sum_{k=0}^{2^j-1} \mu_{j,k} = \mu_0$$

Sous cette condition, on ne peut supposer l'indépendance des générateurs par ligne. On a recours à une autre notion de conservativité, la conservativité en moyenne, qui s'écrit :

$$\forall j = 1 \cdots N \quad \sum_{k=0}^{2^j-1} E\mu_{j,k} = \mu_0$$

Ce choix est accrédité par le fait que la notion physique de conservativité ne prend son sens que pour des échelles macroscopiques. Ici, c'est la conservativité en moyenne qui nous intéresse car les intensités sont désagrégées à des échelles de plus en plus ténues.

Comme les générateurs d'une même ligne sont identiquement distribués, cette condition revient à :

$$\forall j = 1 \cdots N \quad 2^j . E\mu_{j,*} = \mu_0$$

Or, par itérations, le premier cumul désagrégé $j$ fois $\mu_{j,0}$ s'écrit : $\mu_{j,0} = \mu_0 . \prod_{l=1}^{j} \eta_{l,0}$.
La condition de conservativité devient, par indépendance entre les niveaux :

$$\forall j = 1 \cdots N \quad \prod_{l=1}^{j} e_1 = 2^{-j}$$

et revient à une condition sur le moment d'ordre 1 des générateurs :

$$e_1 = \frac{1}{2}$$

**Remarque 36** *Si l'on impose aussi une conservativité sur les moments d'ordre 2 des cumuls :*

$$\forall j = 1 \cdots N \quad \sum_{k=0}^{2^j-1} E\mu_{j,k}^2 = \mu_0^2$$

*on doit alors avoir :* $\forall j = 1 \cdots N \quad \prod_{l=1}^{j} e_2 = 2^{-j}$
*ce qui s'écrit :* $\forall j = 1 \cdots N \quad e_2 = \frac{1}{2}$

On se contentera d'imposer ici des générateurs de moment d'ordre 2 inférieur à 1.

**Variance des cumuls**

La variance du premier cumul $\mu_{N,0}$ de la cascade développée à un niveau $N$ s'écrit :

$$Var\left(\mu_{N,0}\right) = \mu_0^2 \left[\prod_{j=1}^{N} e_2 - e_1^{2N}\right] \text{ car } \mu_{N,0} = \mu_0 . \prod_{j=1}^{N} \eta_{j,0}$$

et comme les $\left(\mu_{N,k}\right)_k$ sont identiquement distribués :

$$Var\left(\mu_{N,*}\right) = Var\left(\mu_{N,0}\right) = \mu_0^2 \left[e_2^N - 2^{-2N}\right]$$

La variance des cumuls tend vers zéro lorsque $N$ tend vers l'infini.

**Covariance des cumuls**

Soient deux cumuls $\mu_{N,k}$ et $\mu_{N,0}$ d'un même niveau de développement de la cascade $N$. Le nombre d'ancêtres différents $q\left(k\right)$ entre $\mu_{N,k}$ et $\mu_{N,0}$ s'écrit :

$$q\left(k\right) = \left\lfloor \frac{\log k}{\log 2} \right\rfloor$$

et le nombre d'ancêtres en commun entre $\mu_{N,k}$ et $\mu_{N,0}$ est $N - q\left(k\right) - 1$. On a :

$$\begin{aligned}
\mu_{N,0} &= \mu_{N-1,0}.\eta_{N,0} = \cdots = \mu_0 . \prod_{j=1}^{N} \eta_{j,0} \\
\mu_{N,k} &= \mu_{N-1,k}.\eta_{N,k} = \cdots = \mu_0 . \left(\prod_{j=1}^{N-q(k)-1} \eta_{j,0}\right) \left(\prod_{j=N-q}^{N} \eta_{j,*}\right) \text{ avec } * \neq 0
\end{aligned}$$

La covariance entre ces deux observations espacées de $k$ s'écrit donc :

$$\gamma_{N0}\left(k\right) = Cov\left(\mu_{N,0}, \mu_{N,k}\right) = \mu_0^2 . Cov\left(XY_1, XY_2\right)$$

avec :

$$\begin{aligned}
X &= \prod_{j=1}^{N-q(k)-1} \eta_{j,0} \\
Y_1 &= \prod_{j=N-q(k)}^{N} \eta_{j,0} \\
Y_2 &= \prod_{j=N-q(k)}^{N} \eta_{j,*} \text{ avec } * \neq 0
\end{aligned}$$

Comme on est dans le cas d'indépendance des générateurs entre les niveaux $j$,

$$Cov\left(XY_1, XY_2\right) = VarX.EY_1.EY_2$$

et comme :

$$\begin{aligned} EY_1 &= EY_2 = 2^{-q(k)-1} \text{ (car } E\eta_{j,*} = e_1 = \frac{1}{2}) \\ VarX &= \prod_{j=1}^{N-q-1} e_{j,2} - \left(\prod_{j=1}^{N-q-1} e_1\right)^2 = e_2^{N-q(k)-1} - 2^{-2N+2q(k)+2} \end{aligned}$$

on obtient la covariance entre les deux observations $\mu_{N,0}$ et $\mu_{N,k}$ :

$$\gamma_{N0}\left(k\right) = \mu_0^2 \left[2^{-2q(k)-2}.e_2^{N-q(k)-1} - 2^{-2N}\right]$$

Par stationnarité des cumuls du niveau $N$ $\left(\mu_{N,j}\right)_{j=1\cdots 2^N-1}$, cette relation est vérifiée pour tout couple d'observations espacées de $k$ observations ($\mu_{N,l}$ et $\mu_{N,l+k}$ avec $l = 0 \cdots 2^N - 1 - k$). A $N$ fixé, la décroissance de $\gamma$ avec $k$ est donc en $\left(2^2 e_2\right)^{-q(k)}$ soit en $k^{-2-\log_2(e_2)}$. La décroissance de la covariance entre les intensités est donc algébrique de paramètre :

$$\log_2\left(e_2\right) + 2$$

Cette fonction de covariance étant monotone, les deux définitions de la longue dépendance citées en *section 4.2* sont équivalentes. On obtient donc le résultat suivant : une cascade multifractale conservative en moyenne est à longue dépendance si le moment d'ordre 2 du générateur vérifie :

$$0 < \log_2\left(e_2\right) + 2 < 1$$

c'est-à-dire si :

$$\frac{1}{4} < e_2 < \frac{1}{2}$$

son paramètre de longue dépendance s'écrit alors :

$$\beta = \log_2\left(e_2\right) + 2 \text{ (ou bien } \alpha = -1 - \log_2\left(e_2\right) \quad H = -\frac{\log_2\left(e_2\right)}{2})$$

### 4.5.3 Cascades particulières

#### Cascades log-Gamma

Nous présentons dans cette section une cascade particulière : celle dont les générateurs $\eta$ suivent une loi log-Gamma $\log\Gamma\left(m, \alpha\right)$ ($\alpha > 1$ et $m \geq 1$) dont la densité est définie par :

$$f_{m,\alpha}(x) = \frac{\alpha^m}{\Gamma\left(m\right)} \left(\log x\right)^{m-1} x^{-\alpha-1} \qquad x \geq 1$$

Remarquons que :

- C'est une loi de type algébrique de paramètre de décroissance $\alpha$.
- Elle est reliée à la loi Gamma par une simple transformation log :
  Si $X \sim \Gamma(m, \alpha)$ alors $\eta = e^X \sim \log\Gamma(m, \alpha)$

Le calcul de ses moments est simple :

$$\eta \sim \Gamma(m, \alpha) \qquad E(\eta^q) = \left(\frac{\alpha}{\alpha - q}\right)^m \quad \forall 0 < q < \alpha - 1$$

On choisit les générateurs de la cascade comme suit : Les $\left(\eta_{j,k}\right)_{j,k}$ sont des variables aléatoires *iid* $\Gamma(m, \alpha)$ :

$$\left(\eta_{j,k}\right)_{j,k} \overset{iid}{\sim} \log\Gamma(m, \alpha)$$

On veut une cascade de cumuls conservative en moyenne, donc le moment d'odre 1 de la loi des générateurs doit être égal à $\frac{1}{2}$ :

$$\left(\frac{\alpha}{\alpha - 1}\right)^m = \frac{1}{2}$$

condition qui revient à une relation entre les paramètres $\alpha$ et $m$ de la loi des générateurs :

$$\frac{1}{m} = -\log_2\left(\frac{\alpha}{\alpha - 1}\right)$$

Le moment d'ordre 2 se calcule aisément. Pour que la cascade soit conservative en moyenne, il doit être inférieur à $\frac{1}{2}$. Or, sous la condition précédente :

$$e_2 = \left(\frac{\alpha}{\alpha - 2}\right)^m = \frac{1}{2}.\left(\frac{\alpha - 1}{\alpha - 2}\right)^m > \frac{1}{2}$$

On n'a donc pas de longue dépendance dans une cascade conservative log-Gamma.

**Cascade log-Normale**

La cascade log-Normale a été introduite par Kolmogorov, 62 [51] et Obhoukov, 62 [67] en turbulence. La loi des générateurs est définie par :

$$\eta = e^X \text{ avec } X \sim N\left(\gamma, \sigma^2\right)$$

La fonction génératrice des moments se calcule facilement :

$$E(\eta^q) = e^{\gamma q + \frac{\sigma^2}{2}\log 2q^2}$$

En particulier,

$$\begin{aligned} e_1 &= e^{\gamma + \frac{\sigma^2}{2}\log 2} \\ \text{et } e_2 &= e^{2\gamma + \frac{3\sigma^2}{2}\log 2} \end{aligned}$$

Comme la cascade est conservative en moyenne ($e_1 = \frac{1}{2}$) :

$$e_2 = \left(\frac{1}{2}\right)^2 .e^{\frac{\sigma^2}{2}\log 2} = 2^{\frac{\sigma^2}{2}-2}$$

et $e_2 < \frac{1}{2}$ si et seulement si $\sigma^2 < 2$. Une cascade conservative en moyenne log-Normale est donc à longue dépendance si et seulement si son paramètre $\sigma^2$ est inférieur à 2. Dans ce cas, son paramètre de longue dépendance s'écrit :

$$\beta = \frac{\sigma^2}{2}$$

**Cascade log-Poisson**

Cette cascade a été introduite en turbulence par Dubrulle, 94 [28]. La loi des générateurs est définie par :

$$\eta = e^{\gamma + aX_\lambda} \text{ avec } X_\lambda \sim P(\lambda)$$

La fonction génératrice des moments se calcule :

$$E(\eta^q) = e^{\gamma q + \frac{\lambda}{\log 2}(2^{aq}-1)}$$

En particulier,

$$\begin{aligned} e_1 &= e^{\gamma + \frac{\lambda}{\log 2}(2^a-1)} = \frac{1}{2} \\ \text{et } e_2 &= e^{2\gamma + \frac{\lambda}{\log 2}(2^{2a}-1)} = \left(\frac{1}{2}\right)^2 .e^{\frac{\lambda}{\log 2}(2^{2a}-2^{a+1}+1)} \end{aligned}$$

En choisissant $a$ et $\lambda$ tels que $0 < \lambda < \frac{(\log 2)^2}{2^{2a} - 2^{a+1} + 1}$, la cascade log-Poisson est à longue dépendance. Dans ce cas, son paramètre de longue dépendance s'écrit :

$$\beta = \frac{\lambda}{(\log 2)^2}\left(2^{2a} - 2^{a+1} + 1\right)$$

## 4.6 Application aux données

Dans cette section, on présente deux outils d'exploration et d'estimation de la longue dépendance couramment utilisés en statistique, avant de les appliquer à l'estimation de la longue dépendance de cascades multifractales simulées et de séries de débits.

### 4.6.1 Outils statistiques

Il existe dans la littérature de nombreux estimateurs de la longue dépendance. En tenant compte des résultats de l'étude comparée de leurs performances sur des séries simulées de Bardet et al. [4], on choisit dans cette section d'employer l'estimateur du log-périodogramme global. Les résultats seront ensuite confrontés aux estimations par l'écart ajusté réduit.

**L'écart ajusté réduit : la statistique $R/S$**

On a vu que cet estimateur, introduit par Hurst [46], est à l'origine de la découverte de la notion de longue dépendance (*section 4.3*). Il peut être utilisé pour estimer le paramètre de longue dépendance $H$ ou bien, à cause de ses faibles performances, comme outil exploratoire pour détecter une longue dépendance. On rappelle son expression :

$$Q_t(k) = \frac{R(t,k)}{S(t,k)}$$

avec :

$$R(t,k) = \max_{0\le i\le k}\left[Y_{t+i} - Y_t - \frac{i}{k}(Y_{t+k} - Y_t)\right] - \min_{0\le i\le k}\left[Y_{t+i} - Y_t - \frac{i}{k}(Y_{t+k} - Y_t)\right]$$

et

$$S(t,k) = \left[\frac{1}{k}\sum_{i=t+1}^{t+k}\left(X_i - \overline{X}_{t,k}\right)^2\right] \text{ où } \overline{X}_{t,k} = \frac{1}{k}\sum_{i=t+1}^{t+k} X_i$$

La pente asymptotique du graphe $\log - \log$ de $(k, Q_t(k))$ constitue une estimation de l'exposant $H$ de Hurst, directement relié, comme on l'a vu en *section 4.3*, à la détection de longue dépendance.

Mais lors de la mise en oeuvre de la méthode, deux problèmes essentiels apparaissent : quand le comportement asymptotique apparaît-il, et quelle est la qualité de l'estimateur $\widehat{H}$ ? Concernant cette dernière question, l'étude sur simulations de Lang [54] conduit à considérer cet estimateur comme peu fiable, car la variance d'estimation décroît peu avec la taille de l'échantillon.

**Le log-périodogramme global**

Cette méthode d'estimation a été proposée par Moulines et Soulier 98 [66]. Elle est basée sur le périodogramme de la série d'observations $(X_1, \cdots, X_N)$ :

$$I_N(\lambda) = \frac{1}{2\pi N}\sum_{i=1}^{N}\left|X_j e^{-ij\lambda}\right|^2$$

où $\lambda$ est la fréquence. Une série à longue dépendance possède une densité spectrale proportionnelle à $|\lambda|^{1-2H}$ au voisinage de l'origine. Comme $I(\lambda)$ est un estimateur de la densité spectrale $g$, que l'on écrit dans le cas stationnaire sous la forme semi-paramétrique :

$$g(\lambda) = g_*(\lambda).\left|1 - e^{-i\lambda}\right|^{1-2H}$$

un estimateur de $H$ découle d'une régression sur le graphe $\log - \log$ du périodogramme. Dans la méthode du $\log -$périodogramme, la régression est effectuée sur les fréquences de Fourier $\lambda_j$, où le périodogramme est calculable à partir des données observées :

$$\log I_N(\lambda_j) = \log g_*(0) + (1-2H)\log\left|1 - e^{-i\lambda_j}\right| + \log\left(\frac{g_*(\lambda_j)}{g_*(0)}\right) + \log\left(\frac{I_N(\lambda_j)}{g(\lambda_j)}\right)$$

La méthode du log −périodogramme global consiste à prendre en compte toutes les fréquences (et pas uniquement celles proches de zéro, comme dans la méthode du log −périodogramme local, où l'on veut négliger la part mémoire courte). Il est alors nécessaire d'estimer la part "mémoire courte" $\log(g_*)$. Cette dernière fonction est projetée sur la base des $h_j(\lambda_j) = \frac{\cos(j\lambda_j)}{\sqrt{\pi}}$ dont on ne retient que $m$ premières composantes :

$$\log(g_*) = \sum_{j=0}^{m} \theta_j h_j$$

Dans la méthode FEXP, on choisit la base des $h_j(\lambda_j) = \cos\left(\frac{j\lambda_j}{\sqrt{\pi}}\right)$. Pour estimer $H$, on réalise ensuite une régression linéaire ordinaire :

$$Y_{N,k} = \left(1 - e^{-ix_k}\right)\left(1 - 2\widehat{H}\right) + \sum_{j=0}^{m} \widehat{\theta}_j h_j(x_k) + \varepsilon_{N,k}$$

avec $x_k = (2k+1)\pi/N$.

Contrairement à l'estimateur du rang ajusté, l'estimateur du log-périodogramme global conduit à des estimations fiables, le paramètre $m$ étant calculé par un algorithme adaptatif. Pour plus de précisions sur la qualité de cet estimateur (et d'autres estimateurs de la longue dépendance), on pourra se reporter à l'ouvrage de Bardet et al. [4].

### 4.6.2 Séries simulées

Dans ce paragraphe, on estime la longue dépendance de séries de cascades multifractales simulées par le log-périodogramme global[1]. On simule 10 8192-échantillons de cascades log-Normale, log-Poisson et log-Gamma. Les résultats de l'estimation de la longue dépendance sont d'autant meilleurs que la longue dépendance est faible (cascade log-Gamma). Mais, sur les 10 cascades log-Normales simulées, l'écart à la valeur théorique $H - 0.5$ atteint en moyenne 0.06 (tableau 4.1). La taille de l'échantillon simulé (8192) est peut être trop faible pour interpréter les résultats du log périodogramme global. Dans le paragraphe suivant, les tailles de séries de débits varieront entre 5 840 et 28 470.

| Série | $H-0.5$ | $\overline{GLP}$ | nb (*) | var | longueur |
|---|---|---|---|---|---|
| log-Normale($\mu=1,\sigma=1$) | 0.25 | 0.19 | 10 | 0.025 | 8192 |
| log-Poisson$\left(\gamma=0,a=2,\lambda=\frac{(\log 2)^3}{9}\right)$ | 0.15 | 0.17 | 10 | 0.025 | 8192 |
| log-Gamma($n=5,\alpha=1$) | 0 | 0.01 | 10 | $1.32*10^{-4}$ | 8192 |

Tableau 4.1: Estimation par le log-périodogramme global sur des cascades simulées (* : nb=nombre d'échantillons simulés).

[1]Je remercie E. Moulines et J.M. Bardet qui ont fournit un programme Matlab de calcul de l'estimateur du log-périodogramme global (programme qui sera prochainement mis à disposition sur internet).

### 4.6.3 Séries de débits

L'estimation de la longue dépendance sur des séries pluviométriques est difficile (Lang, 94 [54]), car elles s'écartent du cadre de définition des estimateurs (non-normalité mais surtout intermittence des données). Cependant, si un comportement de longue dépendance de séries de débits peut provenir d'un comportement de longue dépendance des séries de précipitations, il peut aussi résulter d'un échange avec un gros réservoir régulateur. Il est donc intéressant d'étudier si les séries de débit possèdent la propriété de longue dépendance et de tenter de relier cette dernière aux caractéristiques physiques des bassins versants.

Dans cette section, on examine la forme de dépendance des longues séries de débits moyens journaliers de cours d'eau[2] déjà étudiées en section 3.5.5 (voir tableau 3.6). Parmi ces séries, quelques unes présentent des caractéristiques hydrogéologiques capables d'expliquer une éventuelle mémoire à long terme, telles que les rivières alimentées par une nappe dans la craie. Ces séries ont été étudiées par Lang [54] qui a estimé la longue dépendance par diverses méthodes statistiques, notamment celle de l'estimateur de l'écart ajusté réduit ; ces résultats seront donc comparés aux estimations de la présente étude.

Avant d'estimer la longue dépendance, il est nécessaire de désaisonnaliser ces séries (car la saisonnalité a un effet sur l'estimation). Cette opération est effectuée de la façon la plus simple : en soustrayant à chaque observation à la date $t$ la moyenne de toutes les observations, à cette même date, sur toutes les années.

Les résultats du tableau 4.2 font clairement apparaître une longue dépendance des débits de rivières Authie, Herbissonne, Superbe, Suippe et Soude (puisque les estimations par le log-périodogramme global sont situées autour de 0.7). Lang [54] a constaté que ces rivières étaient des rivières sur craie, ce qui constitue une explication physique de la longue dépendance. De même, un groupe de grandes rivières apparaît : la Dordogne, la Drôme, la Garonne, la Seine, la Vienne et la Zorn présentent aussi une longue dépendance (estimations autour de 0.3). Ceci peut s'expliquer physiquement par une sorte d'effet réservoir. Les rivières de haute montagne, quant à elles, ne sont pas à longue dépendance (Averole, Isère et Ubaye). Contrairement aux estimations de l'écart ajusté réduit (Lang [54]), la proximité géographique des rivières est décrite par l'estimation du log-périodogramme global. En effet, les estimations des séries de l'Isère et de l'Averole sont proches, et il en est de même pour les deux séries de Corrèze et de Dordogne.

## 4.7 Conclusions

L'étude des longues séries de débits journaliers fait apparaître un résultat intéressant : Les rivières à gros volume et les rivières sur craie possèdent la propriété de longue dépendance. Les rivières de haute montagne, quant à elles, ne présentent pas la propriété de longue dépendance. On aurait pu suspecter cependant que les glaciers en amont soient à l'origine d'un effet réservoir en été, mais la forte saisonnalité de ces séries est peut être à l'origine de la non constatation de cet effet. L'estimateur du log-périodogramme fournit donc des résultats satisfaisant pour l'estimation de la longue dépendance, et ce malgré sa nature semi-paramétrique.

[2]Séries extraites de la base de données hydrologique HYDRO du Ministère de l'Agriculture.

| Rivière | GLP ($\widehat{H}$) | Altitude | Longueur (ans) | VM/an(*) |
|---|---|---|---|---|
| Authie | 0.75 | 12 | 27 | 234 |
| Averole | -0.11 | 1950 | 16 | 64 |
| Corrèze | 0.15 | 465 | 47 | 177 |
| Corrèze | 0.15 | 101 | 72 | 671 |
| Dordogne | 0.26 | 780 | 61 | 117 |
| Dordogne | 0.25 | 173 | 76 | 3370 |
| Drôme | 0.25 | 237 | 76 | 90 |
| Garonne | 0.25 | 17 | 74 | 19872 |
| Herbissonne | 0.70 | 91 | 21 | 11 |
| Isère | 0.15 | 1831 | 40 | 59 |
| Seine | 0.25 | 26 | 60 | 8986 |
| Seine | 0.29 | 18 | 49 | 12874 |
| Seine | 0.32 | 148 | 31 | 2583 |
| Soude | 0.54 | 110 | 19 | 19 |
| Suippe | 0.70 | 60 | 23 | 23 |
| Superbe | 0.50 | 79 | 17 | 17 |
| Ubaye | 0.22 | 1133 | 78 | 78 |
| Vienne | 0.27 | 230 | 70 | 70 |
| Zorn | 0.32 | 147 | 67 | 67 |

Tableau 4.2: Estimations de la longue dépendance par le log-périodogramme global (* : volume moyen écoulé par an en $hm^3$).

Les cascades multifractales conservatives en moyenne ne sont pas toutes à longue dépendance bien que leur fonction de covariance décroisse lentement (décroissance algébrique). C'est la valeur du moment d'ordre deux du générateur sous condition de conservativité en moyenne qui permet de déterminer si la cascade est à longue dépendance ou non. La modélisation par cascade multifractale est donc, dans certains cas, adaptée à la modélisation des séries à longue dépendance.

# Chapitre 5

# Conclusion

Les résultats de ce travail sur les possibilités d'application des modèles en cascades multifractales en hydrologie ne sont pas entièrement concordants :

La **non-normalité des séries pluviométriques** apparaît de façon claire sur la base des données étudiées. La théorie probabiliste des valeurs extrêmes permet de disposer d'outils statistiques exploratoires qui révèlent, sur la base des séries pluviométriques étudiées, une **bonne adaptation des modèles de type algébrique**, notamment pour l'estimation des grandes périodes de retour. Il est à noter que cette qualité d'ajustement est conservée aussi sur les grands pas de temps (séries annuelles).

L'application de l'outil semi-paramétrique développé dans ce travail à des données pluviométriques aboutit à la conclusion de l'**invariance d'échelle paramètre de décroissance algébrique**. Ce résultat n'est toutefois vérifié que sur une gamme d'échelle supérieure à l'heure, l'invariance n'apparaissant pas aux fines gammes d'échelle. Les estimations issues d'autres méthodes statistiques développées dans la théorie probabiliste des valeurs extrêmes (telles que le maximum de vraisemblance) ou d'une régression sur l'estimation de la fonction de survie ne permettent d'appuyer quant à eux qu'une quasi-invariance de ce paramètre.

Ces deux résultats pratiques sont en accord avec les résultats théoriques. En effet, comme on a pu le constater sur le plan stochastique, une structure de dépendance de type cascade multiplicative couplée à une loi de type Pareto implique la conservation du type de loi (lois de type Pareto à tous les stades de développement de la cascade), et mieux encore, l'invariance du paramètre de Pareto. **Cette étude incite donc à l'emploi de modèles en cascades multifractales des séries pluviométriques.**

En plus des grandeurs continues telles que les cumuls de pluie, les lois de type algébriques décrivent aussi des **caractéristiques discontinues** des champs de pluie telles que l'occurrence de pluie ou le nombre de basculements d'auget.

L'analyse des 232 séries annuelles nous a montré que le paramètre de décroissance algébrique est à **invariance spatiale**, ce qui peut signifier, à première vue, une homogénéité spatiale des séries de dépassements relatifs. Il conviendrait cependant d'être prudent et d'étayer cette hypothèse par une interprétation physique de cette homogénéité qui, de prime abord, ne cadre pas avec la diversité des régimes climatiques.

Par contre, la propriété de longue dépendance, qui est censée démarquer les pro-

cessus en cascades multifractales de la plupart des modèles classiques, n'est au contraire pas présente dans toutes les cascades. La longue dépendance ne constitue donc pas un critère d'appréciation de la qualité d'ajustement d'un modèle en cascade multifractal. De plus, l'étude des séries de débits hydrologiques fait apparaître que cette propriété n'est pas présente sur la totalité d'entre elles (il est cependant possible que ce soit dû au fait que les séries étudiées sont trop courtes ou possèdent une trop forte saisonnalité). Ajoutons que la présence d'une longue dépendance dans une série de débit ne garantit pas la présence d'une longue dépendance au sein de la série de pluie correspondante, car cette propriété ne se conserve pas par transformation (dans notre cas la transformation pluie→débit).

Sur un plan pratique, ces conclusions suggèrent des approches nouvelles pour l'estimation des grandes périodes de retour, pour l'agrégation et la désagrégation des séries pluviométriques (jusqu'à toutefois un niveau limite) et éventuellement pour l'extrapolation spatiale de l'information hydrologique ponctuelle.

Les répercussions de ce résultat concernent aussi la non-validité des modèles et des outils statistiques classiques. La justification théorique de la modélisation par la loi normale (ou par toute loi de la famille exponentielle) repose sur les propriétés de stabilité et d'attraction de cette dernière (entre autres le théorème de la Limite Centrale). La présence d'une longue dépendance couplée à la grande variabilité des séries pluviométriques invalide le théorème de la Limite Centrale aux faibles pas de temps et diminue considérablement sa vitesse de convergence aux pas de temps supérieurs.

Pour que les conclusions de ce travail concernant les possibilités d'application des modèles en cascades multifractales pour les champs de pluie soient plus tranchées, il est donc nécessaire d'analyser avec plus de précision la forme de dépendance des séries de pluie, et pour cela de développer des outils statistiques de détection et d'estimation de la longue dépendance, adaptés à ces courtes séries à forte saisonnalité et intermittence.

# Bibliographie


[1] A. AMANI and T. LEBEL. Lagrangian kriging for the estimation of sahelian rainfall at small time steps. *Journal of Hydrology*, 192:125–157, 1996.

[2] H. ANDRIEUX. *Interprétation des mesures du radar Rodin de Trappe pour la connaissance en temps réel des précipitations en Seine-Saint-Denis et Val-de-Marne*. PhD thesis, Ecole Nationale des Ponts et Chaussées, 1986.

[3] P. AUSTIN and R. HOUZE. Analysis of the structure of precipitation pattern in new england. *J. Appl. Meteorology*, 11(1):926–935, 1972.

[4] J.M. BARDET, G. LANG, G. OPPENHEIM, A. PHILLIPE and M. TAQQU. Semi-parametric estimation of long range dependence : A survey. *a paraitre*, 2000.

[5] G.A. BARNARD. Discussion of Hurst. *Proc. Inst. Civ. Eng.*, 5:552–553, 1956.

[6] S. BAYOMOG. *Modélisation et analyse des données spatio-temporelles*. PhD thesis, Université Paris-Sud, 1994.

[7] J. BEIRLANT, P. VYNCKIER and L. TEUGELS. Excess functions and estimation of the extreme-value index. *Bernoulli*, 2(4):293–318, 1996.

[8] J. BEIRLANT, P. VYNCKIER and L. TEUGELS. Tail index estimation, pareto quantile plots and regression diagnostics. *JASA*, 91(436):1659–1667, 1996.

[9] H. BENDJOUDI et P. HUBERT. A propos de la distribution statistique des cumuls pluviométriques annuels, faut-il en finir avec la normalité? *Revue des sciences de l'eau*, 1998.

[10] J. BERAN. *Statistics for long-memory processes*. Chapman Hall, New York, first edition, 1994.

[11] J.O. BERGER. *Statistical Decision Theory and Bayesian Analysis*. Springer Verlag, New York, second edition, 1985.

[12] F. BLANCHET. *Elaboration d'une mesure de réflectance de la lame d'eau en hydrologie urbaine*. PhD thesis, Université Paris-XIII, 1993.

[13] M.V. BOLGOV, V.F. PISARENKO and M.I. FORTUS. On the distribution function of extreme flood discharge. In Lars Gottschalk JC Olivry, editor, *IAHS Publication*, pages 181–190. Ducan Reed and D. Rosbjerg, 1999.

[14] L. BOUREL. *Recherches méthodologiques sur l'estimation des précipitations par radar sur un bassin versant en région montagneuse en vue de la prévision des crues.* PhD thesis, Institut National Polytechnique de Toulouse, 1994.

[15] G. BOX and G. JENKINS. *Time series analysis ; Forecasting and control.* Holden-Day, San franscico, first edition, 1970.

[16] T.A. BUISHAND. Some remarks on the use of daily rainfall data models. *Journal of Hydrology*, 36:295–308, 1978.

[17] L. Le CAM. A stochastic description of precipitation. In *IVth symposium on probability and statistics*, Berkeley, 1961. University of California.

[18] H.R. CHO. Stochastic dynamics of precipitation : An example. *Water Resources Research*, 21(8):1225–1232, 1985.

[19] W.S CLEVELAND. *Visualizing data.* Hobart Press, Summit N.J, first edition, 1993.

[20] P. COWPERTWAIT. Further developments of the Neyman-scott clustered point process for modeling rainfall. *Water Resources Research*, 27(7):1431–1438, 1991.

[21] D.R. COX and V. ISHAM. *Point Processes.* Chapman and Hall, London, first edition, 1980.

[22] D.R. COX and V. ISHAM. A simple spatial-temporal model of rainfall. *Proc. R. Soc. London*, A(415):317–328, 1988.

[23] S. CSORGO and L. VIHAROS. Asymptotic normality of least-squares estimators of tail indices. *Bernoulli*, 3(3):351–370, 1997.

[24] H.A. DAVID. *Order Statistics.* John Wiley, New York, first edition, 1981.

[25] M.I.P. de LIMA. *Multifractals and the temporal structure of rainfall.* PhD thesis, Wageningen Agricultural University, 1998.

[26] G. de MARSILY. Quelques réfléxions sur l'utilisation des modèles en hydrologie. *Revue des sciences de l'eau*, 7(1):219–234, 1994.

[27] J.P. DELHOMME. Applications de la théorie des variables régionalisées dans les sciences de l'eau. *Bull. du BRGM*, 3(4):341–375, 1978.

[28] B. DUBRULLE. Intermittency in fully developed turbulence : log-poisson statistics and generalized scale covariance. *Phys. Rev. Lett.*, 73:959–962, 1994.

[29] P.S. EAGLESON. Climate, soil and vegetation. 2-the distribution of annual precipitation derived from observed storm sequences. *Water Resources Research*, 14(5):713–719, 1978.

[30] P. EMBRECHTS, C. KLUPPELBERG and T. MIKOSCH. *Modelling Extremal Events for Insurance and Finance.* Springer, Berlin, first edition, 1997.

[31] J.M. CHAMBERS et al. *Graphical methods for data analysis*. Duxbury Press, Boston, first edition, 1983.

[32] K.J. FALCONER. *Fractal geometry: Mathematical foundations and applications*. John Wiley and Sons, England, first edition, 1990.

[33] D. FAURE. *Application à l'hydrologie du radar météorologique. Comparaison d'estimations radar et pluviométriques pour des lames d'eau horaires sur de petits bassins versant cévenol*. PhD thesis, Université de Genoble I, 1993.

[34] W. FELLER. The asymptotic distribution of the range of sums of independent random variables. *Ann. Math. Statist.*, 22:427–432, 1951.

[35] W. FELLER. *An introduction to probability theory and its applications*. Wiley, N.Y., first edition, 1971.

[36] J. GALAMBOS. *The Asymptotic Theory of Extreme Order Statistics*. Wiley and Sons, N.Y., first edition, 1978.

[37] E. GALLOY, S. MARTIN et A. LEBRETON. Analyse des séquences de jours secs consécutifs; application à 31 postes du réseau météorologique français. *La Météorologie*, 28(6):5–24, 1982.

[38] L.S. GANDIN. *Objective analysis of meteorological fields*. Israel prog. for sc. translation, Jerusalem, first edition, 1965.

[39] K.P. GEORGAKAKOS, A.A CARSTEANU, P.L. STURDEVANT and J.A. CRAMER. Observation and analysis of midwestern rain rates. *J. of Applied Meteorology*, 33:1433–1444, 1994.

[40] GRANGER. Long-memory relationships and the aggragation of dynamic models. *J. Econometrics*, 14:237–238, 1980.

[41] J.R. GREEN. A model for rainfall occurrence. *J. R. Statist. Soc.*, B(26):345–353, 1964.

[42] D.S. HARTE. *Multifractals, theory and application*. PhD thesis, Victoria University of Wellington, 1997.

[43] P. HUBERT. Des crues et des échelles. *La Houille Blanche*, 8(7):83–87, 1999.

[44] P. HUBERT et J.P. CARBONNEL. Caractérisation spatiale de la variabilité et de l'anisotropie des précipitations intertropicales. *C.R. Acad. Sci. Paris*, 2(307):909–914, 1988.

[45] P. HUBERT et J.P. CARBONNEL. Dimensions fractales de l'occurrence de pluie en climat soudano-sahélien. *Hydrologie Continentale*, 4(1):3–10, 1989.

[46] H.E. HURST. Long term storage capacity of reservoirs (with discussion). *Trans. Am. Soc. Civ. Eng.*, 116:770–808, 1951.

[47] J.P. KAHANE. Sur le chaos multiplicatif. *Ann. Sci. Math. Québec*, 9:435, 1985.

[48] R.W. KATZ. An application of chain-dependant processes to meteorology. *J. Appl. Prob.*, 14(14):598–603, 1977.

[49] R.W. KATZ. Precipitation as a chain dependent process. *Journal of Applied Meteorology*, 16(7):671–676, 1977.

[50] M.L. KAVVAS and J.W. DELLEUR. A stochastic cluster model of daily rainfall sequences. *Water Resources Research*, 17(4):1151–1160, 1981.

[51] A.N. KOLMOGOROV. A refinement of previous hypotheses concerning the local structure of turbulence in a viscous incompressible fluid at high reynolds number. *J. of fluid mechanics*, 13:82–85, 1962.

[52] W.F. KRAJEWSKI et J.D. CREUTIN. The hurst phenomenon - a puzzle ? *WRR*, 10:675–688, 1974.

[53] P. LADOY, F. SCHMITT, D. SCHERTZER and S. LOVEJOY. Analyse multifractale de la variabilité pluviométrique à Nîme. *CRAS, Paris*, II(317):775–782, 1993.

[54] G. LANG. *Estimation de la régularité et de la longue dépendance de processus gaussiens; application aux débits hydrologiques*. PhD thesis, Université Toulouse III, 1994.

[55] T. LEBEL, H. SAUVAGEOT, M. HOEPFFENER, M. DEBOIS, B.GUILLOT et P. HUBERT. Rainfall estimation in the sahel: the epsat-niger experiment. *Journal des sciences hydrologiques*, 37(3):201–215, 1992.

[56] O.V LEPSKI and V.G. SPOKOINY. Optimal pointwise adaptative methods in nonparametric estimation. *The Annals of Statistics*, 25(6):2512–2546, 19957.

[57] S. LOVEJOY. La géométrie fractale des nuages et des régions de pluies et les simulations aléatoires. *La Houille Blanche*, 5(6):431–436, 1983.

[58] S. LOVEJOY and D. SCHERTZER. Multifractal analysis techniques and the rain and cloud fields from 10^-3 to 10^6 m. In S. LOVEJOY and D. SCHERTZER eds., editors, *Non linear variability in Geophysics : scaling and fractals*, pages 111–144, The Netherlands, 1991. Kluwer Academic Publishers.

[59] B. MANDELBROT. *Les objets fractals : forme, hasard et dimension.* Flammarion, Paris, first edition, 1975.

[60] B.B. MANDELBROT. Une classe de processus stochastiques homothétiques à soi ; application à la loi climatologique de h.e. hurst. *Comptes Rendus de l'Académie des Sciences de Paris*, 260:3274–3277, 1965.

[61] B.B. MANDELBROT and J.W. Van NESS. Fractional brownian motions, fractional noises and application. *Soc. Ind. Appl. Math. Rev.*, 10:422–437, 1968.

[62] D. MARSAN, D. SCHERTZER and S. LOVEJOY. Causal space-time multifractal processes: predictability and forecasting of rain fields. *J. Geophysical research*, 101(D21):196–209, 1996.

[63] N.C. MATALAS and C.S. HUZEN. A property of the range of partial sums. *Proc. Int. Hydrol. Symp.*, 1:252–257, 1967.

[64] G. MATHERON. *Les variables régionalisées et leur estimation.* Masson, Paris, first edition, 1965.

[65] M.A.J VAN MONFORT and J.V. WITTER. The generalized pareto distribution applied to rainfall dephts. *Hydrological Sciences*, 31(2):151–162, 1986.

[66] E. MOULINES and P. SOULIER. Broadband log-periodogram regression of time series with long-range dependance. *J. Time Ser. Anal.*, 27(4):1415–1439, 1998.

[67] A.M. OBUKHOV. Some specific features of atmospheric turbulence. *J. of fluid mechanics*, 13:77–81, 1962.

[68] J. OLSSON, J. NIEMCZYNOWICZ, R. BERNDTSSON and M. LARSON. An analysis of the rainfall time structure by box counting : some practical implementations. *Journal of Hydrology*, 137(1-4):261–277, 1992.

[69] G. PANDEY, S. LOVEJOY and D. SCHERTZER. Multifractal analysis of daily river flows including extremes for basin of five to two million squre kilometres, one day to 75 years. *Journal of Hydrology*, 208(1-2):62–81, 1998.

[70] G. REMENIERAS. *L'Hydrologie de l'Ingénieur.* Eyrolles, 1965.

[71] S.I. RESNICK. Heavy tail modeling and teletraffic data. *The Annals of Statistics*, 25(5):1805–1869, 1997.

[72] I. RODRIGUEZ-ITURBE. Exploring the comlexity in the structure of rainfall. *Adv. Water Resour.*, 14:162–167, 1991.

[73] I. RODRIGUEZ-ITURBE, D.R COX and V. ISHAM. Some models for rainfall based on stochastic point processes. *Proc. R. Soc. Lond.*, A(410):269–288, 1987.

[74] I. RODRIGUEZ-ITURBE, V. GUPTA and E. WAYMIRE. Scale considerations in the modeling of temporal rainfall. *Water Resources Research*, 20(11):1611–1619, 1984.

[75] D. SCHERTZER and S. LOVEJOY. Physical modeling and analysis of rainand clouds by anisotropic scaling multiplicative processes. *Journal of Geophysical Research*, 92(D8):9693–9714, 1987.

[76] D. SCHERTZER and S. LOVEJOY. *Non linear variability in Geophysics : scaling and multifractal processes.* Institut scinentifique de Cargèse, Cargèse, first edition, 1993.

[77] J.A. SMITH and A.F. KARR. A point process model of summer season rainfall occurrences. *Water Resources Research*, 19(1):95–103, 1983.

[78] R.D. STERN and R.COE. A model fitting analysis of rainfall data. *J. Roy. Stat. Soc. serie A*, 147:1–34, 1984.

[79] M.S. TAQQU. Fractional brownian motion and long range dependence. A(403):27, 2000.

[80] Y. TESSIER, S. LOVEJOY, P. HUBERT, D. SCHERTZER and S. PECKNOLD. Multifractal analysis and modeling of rainfall and river flows and scaling, causual transfert functions. *Journal of Geophysical Research*, 101(D21):26427–26440, 1996.

[81] V. THAUVIN, E. GAUME and C. ROUX. A short time-step point rainfall stochastic model. *Publication interne du CERGRENE*, 1997.

[82] A.H. THIESSEN. Precipitation averages for large areas. *Monthly Weather Review*, 39:1082–1084, 1911.

[83] E.R. TUFTS. *The visual display of quantitative information.* Graphics Press, Cheshire, first edition, 1983.

[84] VAN-THANH-VAN NGUYEN. A stochastic description of temporal daily rainfall patterns. *Can. J. Civ. Eng.*, 11:234–238, 1984.

[85] E. WAYMIRE, V. GUPTA and I. RODRIGUEZ-ITURBE. A spectral theory of rainfall intensity at the meso-beta-scale. *Water Resources Research*, 20(10):1453–1465, 1984.

[86] D.S. WILKS. Conditioning stochastic daily precipitation models on total monthly precipitation. *Water Resources Research*, 25(6):1429–1439, 1989.

[87] D.A. WOOLHISER. Modeling daily precipitations-progress and problems. In P. Guttorp and A. Walden, editors, *Statistics in the environmental and earth sciences*, pages 71–89. Griffin, London, 1991.

[88] W. ZUCCHINI. A hidden Markov model for space-time precipitation. *Water Resources Research*, 27(8):1917–1923, 1991.

[89] W. ZUCCHINI, P. ADAMSON and L. McNEIL. A model of southern african rainfall. *Suid-Afrikaanse Tydskrif vir Wetenskap*, 88:103–109, 1992.

# Annexe A

# Comptage de boîtes sur l'Ensemble de Cantor

Dans cette annexe, on étudie le comportement de la méthode du comptage de boîtes lorsqu'elle est appliquée à l'Ensemble de Cantor avec des boîtes de dé-$a$-uplant (c'est-à-dire des boîtes de tailles $\left(\frac{1}{a^k}\right)_{k\geq 0}$) pour $a$ différent de 3. Plus précisément, on montre que le nombre de boîtes non vides de taille $\frac{1}{a^k}$ avec $a > 3$ (resp. $a < 3$), est supérieur (resp. inférieur) au nombre de boîtes non vides de taille $\frac{1}{3^k}$.

On appelle Ensemble de Cantor $C_N$ l'ensemble construit par itération jusqu'au niveau limite $N$ à partir d'ensembles $C_i$ de segments de longueur maximale $\frac{1}{3^i}$ :

$$\begin{aligned} C_0 &= [0,1] \\ C_1 &= \left[0,\frac{1}{3}\right] \cup \left[\frac{2}{3},1\right] \\ &\cdots \\ C_N &= \lim_{i\to N} C_i \end{aligned}$$

On définit les ensembles $D_i$ homothétiques de rapport $(3/a)^i$ des ensembles $C_i$ et $D_N$ la poussière limite des $D_i$ :

$$\begin{aligned} \forall i \quad &\geq \quad 0 \quad D_i = Hom_{(3/a)^i}(C_i) \\ D_N \quad &= \quad \lim_{i\to N} D_i \end{aligned}$$

Soit $N(k)$ le nombre de boîtes non vides de $C_N$ de taille $\frac{1}{3^k}$

$N^{"}(k)$ le nombre de boîtes non vides de $C_N$ de taille $\frac{1}{a^k}$

et $N^{'}(k)$ le nombre de boîtes non vides de $D_N$ de taille $\frac{1}{a^k}$

On a : $\forall k \geq 0 \quad N^{'}(k) = N(k)$.

De plus, comme $D_N$ est inclus dans $C_N$ (en effet les $D_i$ sont inclus dans les $C_i$ car les rapports $(3/a)^i$ des homothéties sont toujours inférieurs à 1), $\forall k \geq 0 \quad N^{'}(k) \leq N^{"}(k)$. Ainsi :

$$\forall k \geq 0 \quad N(k) \leq N^{"}(k)$$

On montre de la même manière que le nombre de boîtes non vides de taille $\frac{1}{2^k}$ est inférieur au nombre de boîtes non vides de taille $\frac{1}{3^k}$.

# Annexe B

# Caractérisation des lois de type algébrique

Dans cette annexe, on démontre la caractérisation des lois de type algébrique par le comportement limite de leur *FMDR* (proposition énoncée en section 3.3.2).

Rappelons qu'une loi est de type algébrique de coefficient $q > 1$ si sa fonction de survie au delà de 1 s'écrit :

$$\forall s \geq 1 \quad G(s) = r(s).s^{-q}$$

avec $r$ vérifiant :

$$\begin{cases} \text{(C1) } r(s) \overset{s\to\infty}{\longrightarrow} K \text{ et } r \text{ croissante sur } [1, +\infty[ \\ \text{(C2) } \forall s \geq 1 \qquad r'(s) \leq As^{-\theta} \text{ avec } A > 0 \text{ et } \theta > 1 \end{cases}$$

Montrons que les conditions $(C1)$ et $(C2)$ sont équivalentes aux conditions :

$$\begin{cases} \text{(C3) } \forall s \geq 1 \qquad 0 \leq q - 1 - \frac{1}{h(s)} \leq Cs^{-\nu} \text{ avec } C > 0 \text{ et } \nu > 0 \\ \text{(C4) } \forall s \geq 1 \qquad 0 \leq -h'(s) \leq Ds^{-\theta'} \text{ avec } D > 0 \text{ et } \theta' > 1 \end{cases}$$

Supposons que les conditions $(C1)$ et $(C2)$ sont vérifiées. Alors :

$$\int_x^\infty t^{-q} r(t)dt \geq r(1) . \int_x^\infty t^{-q} dt = \frac{r(1)}{q-1} x^{1-q}$$

et

$$\int_x^\infty t^{1-q} r'(t)dt \leq A. \int_x^\infty t^{-q} dt = \frac{A}{q+\theta-2} x^{-q-\theta+2}$$

En intégrant par parties,

$$\int_x^\infty t^{-q} r(t)dt = \frac{1}{q-1} \left\{ r(x) .x^{1-q} + \int_x^\infty t^{1-q} r'(t)dt \right\}$$

et il s'ensuit que :

$$q-1-\frac{1}{h\left(x\right)}=\frac{\int_{x}^{\infty}t^{1-q}r^{'}(t)dt}{\int_{x}^{\infty}t^{-q}r(t)dt}\leq C.x^{1-\theta}$$

avec $C=\dfrac{A.\left(q-1\right)}{r\left(1\right).\left(\theta+q-2\right)}$ ce qui montre que $(C3)$ est vérifiée.

Pour montrer que $(C4)$ est vérifiée, on calcule :

$$-h^{'}\left(x\right)=\frac{1}{q-1}\left\{\frac{r^{'}(x)}{r(x)}+\int_{x}^{\infty}t^{1-q}r^{'}(t)dt.\left[\frac{r^{'}(x)}{r^{2}(x).x^{1-q}}-\frac{q-1}{r(x).x^{2-q}}\right]\right\}$$

et il s'ensuit que $-h^{'}\left(x\right)\leq\dfrac{1}{q-1}\left\{\dfrac{A}{r(1)}+\dfrac{A.x^{1-2\theta}}{\left(q-1\right)r(1).x^{1-q}}\right\}$ ce qui montre que $(C4)$ est vérifiée en remarquant que le premier terme est dominant et que $r$ est croissante.

Supposons que les conditions $(C1)$ et $(C2)$ sont vérifiées. Posons $H$ la fonction définie par :

$$H\left(x\right)=\int_{x}^{\infty}t^{-q}r(t)dt$$

On a :

$$h\left(x\right)=\frac{-H(x)}{x.H^{'}(x)}$$

et :

$$\frac{1}{x}\left(q-1-\frac{1}{h(x)}\right)=\frac{H^{'}(x)}{H(x)}+\frac{\left(q-1\right).x^{q-2}}{x^{q-1}}$$

En intégrant cette relation, on obtient :

$$H(x)=H(1).x^{1-q}.\exp\left\{\int_{1}^{x}\left(q-1-\frac{1}{h(x)}\right)\frac{dt}{t}\right\}$$

et ainsi :

$$r\left(x\right)=\frac{H(1)}{h(x)}\exp\left\{\int_{1}^{x}\left(q-1-\frac{1}{h(x)}\right)\frac{dt}{t}\right\}$$

Donc $r$ est continue, croissante car $\frac{1}{h}$ est croissante et tend vers $q-1$. La condition $(C3)$ implique que l'intégrale converge quand $x$ tend vers l'infini, donc $r$ tend vers une limite finie $K$ et la condition $(C1)$ est satisfaite.

Pour montrer que la condition $(C2)$ est satisfaite, constatons que :

$$\frac{r^{'}(x)}{r(x)} = \frac{1}{x}\left(q - 1 - \frac{1}{h\,(x)}\right) - \frac{h^{'}(x)}{h(x)}$$

La condition $(C3)$ implique que le premier terme est majoré par $C.x^{-\theta}$. De plus, elle implique que $\frac{1}{h}$ et majorée par $q-1$ et la condition $(C4)$ entraîne que le second terme est majoré par $D.\,(q-1)\,.x^{-q}$, ce qui démontre que $(C2)$ est vérifiée.

# Annexe C

# Calculs du biais et de la variance de $\widehat{h}$

Dans cette annexe, on calcule formellement, et dans le cas général, le biais et la variance de l'estimateur $\widehat{h}$ présenté en section 3.4.3.

## C.1 Biais

Soit $s$ un réel supérieur à 1. On a :

$$\widehat{h}(s) = \frac{1}{s}\sum_{i=1}^{N}(X_i - s).1_{X_i>s} \times \frac{1}{1+\sum_{j\neq i}1_{X_j>s}}$$

et comme les $(X_i)_{i=1\cdots N}$ sont indépendantes :

$$E\left[\widehat{h}(s)\right] = H(s).E\left[\frac{1}{1+\sum_{j\neq 1}1_{X_j>s}}\right]$$

Or $\sum_{j\neq 1}1_{X_j>s} \sim Bin(N-1, G(s))$, donc :

$$E\left[\widehat{h}(s)\right] = H(s).\sum_{k=1}^{N}\frac{1}{1+k}C_{N-1}^{k}G^{k}(s)F^{N-1-k}(s) = \frac{H(s)}{sG(s)}\left[(F(s)+G(s))^{N} - F^{N}(s)\right]$$

## C.2 Variance

Soit $s$ un réel supérieur à 1. On a :

$$\widehat{h}^2(s) = \frac{1}{s^2}\sum_{i=1}^{N}\frac{(X_i - s).1_{X_i>s}}{\sum_j 1_{X_j>s}} = \frac{1}{s^2}(S_1 + S_2)$$

avec

$$S_1 = \sum_{i=1}^{N}\frac{(X_i - s)^2.1_{X_i>s}}{\left(1 + \sum_{j\neq i} 1_{X_j>s}\right)^2} \text{ et } S_2 = \sum_{i=1}^{N}\sum_{j\neq i}\frac{(X_i - s).(X_j - s).1_{X_i>s}.1_{X_j>s}}{\left(2 + \sum_{k\neq i,j} 1_{X_k>s}\right)^2}$$

et comme les $(X_i)_{i=1\cdots N}$ sont indépendantes et comme $\sum_{k\neq 1,2} 1_{X_k>s} \sim Bin(N-2, G(s))$, on a :

$$E[S_2] = N(N-1)H^2(s).\sum_{k=0}^{N-2}\frac{1}{(2+k)^2}C_{N-2}^k G^k(s)F^{N-2-k}(s)$$

Or :

$$\sum_{k=0}^{N-2}\frac{1}{(2+k)^2}C_{N-2}^k G^k(s)F^{N-2-k}(s) = \frac{1}{N(N-1)G^2(s)}\left[1 - G^N(s) - S_N(s)\right]$$

avec $S_N(s) = \sum_{k=1}^{N}\frac{1}{k}C_N^k G^k(s)F^{N-k}(s)$. Ainsi :

$$E[S_2] = \frac{H^2(s)}{G^2(s)}.\left[1 - F^N(s) - S_N(s)\right]$$

En intégrant deux fois par parties :

$$\forall i \geq 1 \quad E\left[(X_i - s)^2.1_{X_i>s}\right] = 2\int_s^\infty H$$

et $E[S_1] = 2N.\left(\int_s^\infty H\right).\sum_{k=1}^{N}\frac{1}{(1+k)^2}C_{N-1}^k G^k(s)F^{N-1-k}(s)$. Comme :

$$\sum_{k=1}^{N}\frac{1}{(1+k)^2}C_{N-1}^k G^k(s)F^{N-1-k}(s) = \frac{S_N(s)}{NG(s)}$$

le résultat découle directement puisque :

$$E\left[\widehat{h}^2(s)\right] = \frac{2S_N(s)}{s^2G(s)}\int_s^\infty H + \frac{H^2(s)}{s^2G^2(s)}.\left[1 - F^N(s) - S_N(s)\right]$$

# Annexe D

# Auto-similarité et Multifractalité

On rencontre un grand nombre de définitions non-équivalentes de l'auto-similarité dans la littérature. Parmi elles, la plus courante est celle selon laquelle un processus continu $\{Y(t), t \in T\}$ est auto-similaire (de paramètre $H$) si et seulement si :

$$Y(t) \stackrel{L}{=} a^{-H} Y(at) \quad \forall t \in T, \quad \forall a > 0, \quad 0 \leq H < 1 \tag{1}$$

**Remarque 37** *Ce processus ne peut être stationnaire. On suppose néanmoins que ses incréments le sont.*

Une deuxième définition de l'auto-similarité, plus appropriée dans ce contexte de séries chronologiques classiques, concerne un processus stationnaire $\{X(i), i \geq 1\}$. On définit le processus agrégé à pas de temps $m$ par :

$$X^{(m)}(k) = \frac{1}{m} \sum_{i=(k-1)m+1}^{km} X(i) \quad \forall k \geq 1 \tag{2}$$

Si $X$ est le processus d'incrémentation du processus $Y$ définit précédemment (ie $X(i) = Y(i+1) - Y(i)$), alors pour tout $m \geq 1$ :

$$X(t) \stackrel{L}{=} m^{1-H} X^{(m)} \tag{3}$$

**Définition 38** *Une série stationnaire $\{X(i), i \geq 1\}$ est dite **exactement auto-similaire** si et seulement si elle vérifie (3).*

*Elle est dite **asymptotiquement auto-similaire** si et seulement si (3) est vrai quand $m \to \infty$.*

*Elle est dite **auto-similaire du second ordre** si et seulement si $X$ et $m^{1-H} X^{(m)}$ ont même variance et auto-corrélation et **asymptotiquement auto-similaire du second ordre** si et seulement si la variance et l'auto-corrélation de $m^{1-H} X^{(m)}$ tend vers celle de $X$ quand $m \to \infty$.*

***Remarque 39*** *Si le processus $(X)$ est gaussien, les définitions sont equivalentes.*

**Définition 40** *Pour une série stationnaire $\{X(i), i \geq 1\}$, on définit les moments absolus d'ordre $q$ par :*

$$\mu^{(m)}(q) = E\left|X^{(m)}\right|^q = E\left|\frac{1}{m}\sum_{i=1}^{m} X(i)\right|^q$$

**Définition 41** *Troisième définition de l'auto-similarité (impliquée mais pas équivalente à la définition* (3)*) :*

$$\log\ \mu^{(m)}(q) = \beta(q)\log m + C(q)\ et \beta(q) = q(H-1)$$

**Définition 42** *Processus multifractal : Un processus positif est dit multifractal si et seulement si la relation* log log *entre ses moments d'ordre* $q$ *et son niveau d'agrégation* $m$ *est linéaire :*

$$\log\ \mu^{(m)}(q) = \beta(q)\log m + C(q)$$

**Définition 43** *Processus multifractal signé : Un processus est dit multifractal signé si et seulement si la relation* log log *entre ses moments d'ordre* $q$ *et son niveau d'agrégation* $m$ *est linéaire :*

$$\log\ \mu^{(m)}(q) = \beta(q)\log m + C(q)$$

***Remarque 44*** *Un processus multifractal signé est une généralisation d'un processus auto-similaire.*

***Remarque 45*** *Un processus stationnaire ne peut être auto-similaire ou asymptotiquement auto-similaire s'il n'est pas centré. En effet, par* (3)*,* $EX = m^{1-H}EX^{(m)} = m^{1-H}EX$ *(ou* $m^{1-H}EX \overset{m\to\infty}{\longrightarrow} EX$*). Cependant, il peut être multifractal.*

# Annexe E

# Liste des stations

Dans cette annexe, on fournit le détail des séries de plus de 100 ans utilisée dans la *section 3.5.4*.

Figure E.1: Liste des stations des séries annuelles.

Tableau 1 : Présentation de la base de données

| n° | stations | code | début | fin | durée | grd | pas | latitude | | longitude | | n° | stations | code | début | fin | durée | grd | pas | latitude | | longitude | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | deg | mn | deg | mn | | | | | | | | | deg | mn | deg | mn |
| 1 | KEW_GARDENS | GB | 1697 | 1995 | 299 | P | A | 51 | 30 | 0 | -18 | 101 | ROANNE | FR | 1851 | 1973 | 123 | P | A | 46 | 0 | 4 | 5 |
| 2 | Padova | IT | 1725 | 1990 | 266 | P | A | 45 | 25 | 11 | 53 | 102 | WILMINGTON | US | 1871 | 1993 | 123 | P | A | 34 | 17 | -77 | 54 |
| 3 | Marseille | FR | 1749 | 1993 | 245 | P | A | 43 | 18 | 5 | 23 | 103 | Adelaide | AU | 1839 | 1960 | 122 | P | A | -34 | 56 | 138 | 35 |
| 4 | OXFORD | GB | 1767 | 1995 | 229 | P | A | 51 | 42 | -1 | 12 | 104 | LUNEBERG | DE | 1854 | 1975 | 122 | P | A | 53 | 17 | 10 | 23 |
| 5 | Roma | IT | 1782 | 1984 | 203 | P | A | 41 | 54 | 12 | 29 | 105 | SALT_LAKE_CITY | US | 1875 | 1996 | 122 | P | A | 40 | 47 | -111 | 54 |
| 6 | Gibraltar | GI | 1791 | 1992 | 202 | P | A | 36 | 9 | -5 | 21 | 106 | BISMARCK | US | 1875 | 1996 | 122 | P | A | 46 | 47 | -100 | 48 |
| 7 | Strasbourg | FR | 1803 | 1991 | 189 | P | A | 48 | 25 | 7 | 46 | 107 | KLAUSTHAL | DE | 1855 | 1975 | 121 | P | A | 51 | 47 | 10 | 18 |
| 8 | Toulouse | FR | 1809 | 1993 | 185 | P | A | 43 | 36 | 1 | 26 | 108 | REVEL | FR | 1859 | 1973 | 121 | P | A | 43 | 24 | 2 | 0 |
| 9 | Palermo | IT | 1807 | 1991 | 185 | P | A | 38 | 7 | 13 | 21 | 109 | Paris | FR | 1873 | 1993 | 121 | P | A | 48 | 52 | 2 | 20 |
| 10 | MANSFIELD | GB | 1807 | 1989 | 183 | P | A | 53 | 5 | -1 | 6 | 110 | NEUCHATEL | CH | 1856 | 1975 | 120 | P | A | 47 | 0 | 6 | 54 |
| 11 | Lille | FR | 1801 | 1979 | 179 | P | A | 50 | 38 | 3 | 4 | 111 | Shanghai | CN | 1871 | 1990 | 120 | P | A | 31 | 14 | 121 | 28 |
| 12 | BOSTON | US | 1818 | 1996 | 179 | P | A | 42 | 24 | -71 | 5 | 112 | Praha | CZ | 1876 | 1995 | 120 | P | A | 50 | 5 | 14 | 26 |
| 13 | KENDAL_SCHOOL | GB | 1820 | 1987 | 168 | P | A | 54 | 17 | -2 | 42 | 113 | Athinai | GR | 1871 | 1990 | 120 | P | A | 37 | 58 | 23 | 43 |
| 14 | Castelnaudary | FR | 1829 | 1993 | 165 | P | A | 43 | 19 | 1 | 59 | 114 | Cape_Town | ZA | 1841 | 1960 | 120 | P | A | -33 | 56 | 18 | 29 |
| 15 | LEEDS_HOLLIES_PARK | GB | 1824 | 1987 | 164 | P | A | 53 | 47 | -1 | 36 | 115 | LINZ | AT | 1856 | 1974 | 119 | P | A | 48 | 17 | 14 | 18 |
| 16 | CHILGROVE_HOUSE | GB | 1834 | 1994 | 161 | P | A | 50 | 54 | 0 | -48 | 116 | LONINGEN | DE | 1857 | 1975 | 119 | P | A | 52 | 42 | 7 | 48 |
| 17 | CORK_AIRPORT | IE | 1836 | 1996 | 161 | P | A | 51 | 47 | -8 | 30 | 117 | GOTTINGEN | DE | 1857 | 1975 | 119 | P | A | 51 | 30 | 9 | 53 |
| 18 | FALMOUTH | GB | 1835 | 1994 | 160 | P | A | 50 | 5 | -5 | 5 | 118 | MAGYAROVAR | HU | 1859 | 1977 | 119 | P | A | 47 | 54 | 17 | 17 |
| 19 | DUBLIN_AIRPORT | IE | 1837 | 1996 | 160 | P | A | 53 | 24 | -6 | 11 | 119 | Bistrita | RO | 1875 | 1993 | 119 | P | A | 47 | 8 | 24 | 30 |
| 20 | FIRENZE | IT | 1821 | 1980 | 160 | P | A | 43 | 47 | 11 | 18 | 120 | TOPEKA | US | 1878 | 1996 | 119 | P | A | 39 | 5 | -95 | 35 |
| 21 | Napoli | IT | 1833 | 1990 | 158 | P | A | 40 | 51 | 14 | 17 | 121 | BADISCHEL | AT | 1858 | 1975 | 118 | P | A | 47 | 42 | 13 | 36 |
| 22 | LUEBECK | DE | 1840 | 1994 | 155 | P | A | 53 | 47 | 10 | 41 | 122 | MEREDITH_DARRA | AU | 1875 | 1992 | 118 | P | A | -37 | 49 | 144 | 8 |
| 23 | ARMAGH | GB | 1840 | 1994 | 155 | P | A | 54 | 17 | -6 | 35 | 123 | Constantine | DZ | 1876 | 1993 | 118 | P | A | 36 | 25 | 6 | 43 |
| 24 | Genova | IT | 1833 | 1987 | 155 | P | A | 44 | 25 | 8 | 57 | 124 | Seoul | KR | 1776 | 1893 | 118 | P | A | 37 | 30 | 127 | 0 |
| 25 | ALTHORP_PARK | GB | 1841 | 1994 | 154 | P | A | 52 | 17 | 0 | -48 | 125 | Sulina | RO | 1876 | 1993 | 118 | P | A | 45 | 9 | 29 | 40 |
| 26 | Budapest | HU | 1841 | 1994 | 154 | P | A | 47 | 30 | 19 | 5 | 126 | KARESUANDO | SE | 1879 | 1996 | 118 | P | A | 68 | 25 | 22 | 28 |
| 27 | Bordeaux | FR | 1842 | 1993 | 152 | P | A | 44 | 50 | 0 | -34 | 127 | SPRINGFIELD_MO | US | 1877 | 1994 | 118 | P | A | 37 | 12 | -93 | 24 |
| 28 | HULL_PEARSON_PARK | GB | 1847 | 1995 | 149 | P | A | 53 | 42 | 0 | -18 | 128 | Oran | DZ | 1877 | 1993 | 117 | P | A | 35 | 43 | 0 | -43 |
| 29 | Alger | DZ | 1846 | 1993 | 148 | P | A | 36 | 42 | 3 | 8 | 129 | LOVELY_BANKS | AU | 1877 | 1992 | 116 | P | A | -38 | 4 | 144 | 19 |
| 30 | San-Fernando | ES | 1838 | 1985 | 148 | P | A | 36 | 30 | -6 | 16 | 130 | ORNSLUND | DK | 1861 | 1976 | 116 | P | A | 55 | 42 | 10 | 36 |
| 31 | DE_BILT | NL | 1849 | 1996 | 148 | P | A | 52 | 5 | 5 | 10 | 131 | NIORT | FR | 1858 | 1973 | 116 | P | A | 46 | 17 | 0 | -18 |
| 32 | WAHNSDORF | DE | 1828 | 1974 | 147 | P | A | 51 | 5 | 13 | 41 | 132 | LONDONDERRY | GB | 1861 | 1976 | 116 | P | A | 55 | 0 | -7 | 35 |
| 33 | PEMBROKE_DALE_FORT | GB | 1849 | 1995 | 147 | P | A | 51 | 42 | -5 | 5 | 133 | ASKERSUND | SE | 1861 | 1976 | 116 | P | A | 59 | 0 | 14 | 48 |
| 34 | BREMEN | DE | 1830 | 1975 | 146 | P | A | 53 | 5 | 8 | 48 | 134 | NYKOPING_LANDSORT | SE | 1861 | 1976 | 116 | P | A | 58 | 47 | 17 | 0 |
| 35 | GRENOBLE_ST-GEOIRS | FR | 1845 | 1988 | 144 | P | A | 45 | 24 | 5 | 18 | 135 | LINKOPING_MALMSLATT | SE | 1861 | 1976 | 116 | P | A | 58 | 24 | 15 | 36 |
| 36 | Ljubljana | SI | 1851 | 1994 | 144 | P | A | 46 | 4 | 14 | 32 | 136 | CHATTANOOGA | US | 1879 | 1994 | 116 | P | A | 35 | 0 | -85 | 11 |
| 37 | MONTBARD | FR | 1831 | 1973 | 143 | P | A | 47 | 35 | 4 | 18 | 137 | BASEL | CH | 1861 | 1975 | 115 | P | A | 47 | 35 | 7 | 35 |
| 38 | Perpignan | FR | 1851 | 1993 | 143 | P | A | 42 | 42 | 2 | 53 | 138 | Tortosa | ES | 1880 | 1994 | 115 | P | A | 40 | 49 | 0 | 29 |
| 39 | Cahors | FR | 1851 | 1993 | 143 | P | A | 44 | 27 | 1 | 26 | 139 | VICHY_CHARMEIL | FR | 1859 | 1973 | 115 | P | A | 46 | 5 | 3 | 24 |
| 40 | Montpellier | FR | 1851 | 1993 | 143 | P | A | 43 | 37 | 3 | 53 | 140 | ISSOUDUN | FR | 1859 | 1973 | 115 | P | A | 46 | 54 | 2 | 0 |
| 41 | POUILLY | FR | 1831 | 1973 | 143 | P | A | 47 | 17 | 4 | 30 | 141 | Fiumicino | IT | 1871 | 1985 | 115 | P | A | 41 | 46 | 12 | 14 |
| 42 | Barcelona | ES | 1850 | 1991 | 142 | P | A | 41 | 23 | 2 | 11 | 142 | BAMBERG | DE | 1861 | 1974 | 114 | P | A | 49 | 54 | 10 | 53 |
| 43 | Toulon | FR | 1852 | 1993 | 142 | P | A | 43 | 7 | 5 | 56 | 143 | HUSUM | DE | 1861 | 1974 | 114 | P | A | 54 | 30 | 9 | 6 |
| 44 | Sibiu | RO | 1852 | 1993 | 142 | P | A | 45 | 48 | 24 | 9 | 144 | Saint-Sever | FR | 1880 | 1993 | 114 | P | A | 43 | 47 | 0 | -34 |
| 45 | HANNOVER | DE | 1856 | 1996 | 141 | P | A | 52 | 30 | 9 | 41 | 145 | LIMOGES | FR | 1860 | 1973 | 114 | P | A | 45 | 47 | 1 | 17 |
| 46 | Tokaj | HU | 1854 | 1994 | 141 | P | A | 48 | 6 | 7 | 25 | 146 | FEINS | FR | 1860 | 1973 | 114 | P | A | 48 | 17 | -1 | 36 |
| 47 | Debrecen | HU | 1854 | 1994 | 141 | P | A | 47 | 32 | 21 | 38 | 147 | NIZAMABAD | IN | 1871 | 1984 | 114 | P | A | 18 | 42 | 78 | 5 |
| 48 | Alicante | ES | 1856 | 1994 | 140 | P | A | 38 | 21 | 0 | -29 | 148 | CHIKMAGALUR | IN | 1871 | 1984 | 114 | P | A | 13 | 18 | 75 | 48 |
| 49 | Besancon | FR | 1845 | 1984 | 140 | P | A | 47 | 13 | 6 | 1 | 149 | SEONI | IN | 1871 | 1984 | 114 | P | A | 22 | 6 | 79 | 35 |
| 50 | GUTERSLOH | DE | 1837 | 1975 | 139 | P | A | 51 | 54 | 8 | 23 | 150 | DHARAMPURI | IN | 1871 | 1984 | 114 | P | A | 12 | 6 | 78 | 11 |
| 51 | GRONINGEN | NL | 1840 | 1977 | 138 | P | A | 53 | 12 | 6 | 35 | 151 | CHHINDWARA | IN | 1871 | 1984 | 114 | P | A | 22 | 6 | 79 | 0 |
| 52 | RHAYADER_TYNANT | GB | 1858 | 1994 | 137 | P | A | 52 | 17 | -3 | 35 | 152 | TUMKUR | IN | 1871 | 1984 | 114 | P | A | 13 | 23 | 77 | 5 |
| 53 | Auckland | NZ | 1853 | 1989 | 137 | P | A | -36 | 53 | 174 | 45 | 153 | TANJAVORE | IN | 1871 | 1984 | 114 | P | A | 10 | 48 | 79 | 5 |
| 54 | Fortaleza-CE | BR | 1849 | 1984 | 136 | P | A | -3 | 72 | -38 | 50 | 154 | HARIPAD | IN | 1871 | 1984 | 114 | P | A | 9 | 18 | 76 | 30 |
| 55 | CARDIFF_BUTE | GB | 1859 | 1994 | 136 | P | A | 51 | 30 | -3 | 12 | 155 | NALGONDA | IN | 1871 | 1984 | 114 | P | A | 17 | 6 | 79 | 18 |
| 56 | SHANNON_AIRPORT | IE | 1861 | 1996 | 136 | P | A | 52 | 42 | -8 | 53 | 156 | KHAMMAM | IN | 1871 | 1984 | 114 | P | A | 17 | 17 | 80 | 11 |
| 57 | VALENTIA_OBSERVATORY | IE | 1861 | 1996 | 136 | P | A | 51 | 54 | -10 | 11 | 157 | CHINGELPUT | IN | 1871 | 1984 | 114 | P | A | 12 | 41 | 80 | 0 |
| 58 | KARLSTAD | SE | 1861 | 1996 | 136 | P | A | 59 | 20 | 13 | 28 | 158 | WARDHA | IN | 1871 | 1984 | 114 | P | A | 20 | 42 | 78 | 35 |
| 59 | VISBY_AIRPORT | SE | 1861 | 1996 | 136 | P | A | 57 | 40 | 18 | 19 | 159 | ASIFABAD | IN | 1871 | 1984 | 114 | P | A | 19 | 23 | 79 | 18 |
| 60 | GORDON_C_KIRKHILL | GB | 1860 | 1994 | 135 | P | A | 57 | 35 | -3 | 5 | 160 | PARBHANI | IN | 1871 | 1984 | 114 | P | A | 19 | 17 | 76 | 48 |
| 61 | FT_WILLIAM | GB | 1861 | 1994 | 134 | P | A | 56 | 47 | -5 | 5 | 161 | BETUL | IN | 1871 | 1984 | 114 | P | A | 21 | 53 | 77 | 54 |
| 62 | Zagreb | HR | 1862 | 1995 | 134 | P | A | 45 | 48 | 15 | 58 | 162 | SHIMOGA | IN | 1871 | 1984 | 114 | P | A | 13 | 53 | 75 | 35 |
| 63 | Keszthely | HU | 1861 | 1994 | 134 | P | A | 46 | 46 | 17 | 15 | 163 | GUNTUR | IN | 1871 | 1984 | 114 | P | A | 16 | 17 | 80 | 30 |
| 64 | KILLARNEY | IE | 1861 | 1994 | 134 | P | A | 52 | 0 | -9 | 30 | 164 | MANDYA | IN | 1871 | 1984 | 114 | P | A | 12 | 30 | 76 | 54 |
| 65 | Huesca | ES | 1862 | 1994 | 133 | P | A | 42 | 5 | 0 | -19 | 165 | FARGO | US | 1881 | 1994 | 114 | P | A | 46 | 54 | -96 | 48 |
| 66 | Gap | FR | 1861 | 1993 | 133 | P | A | 44 | 34 | 6 | 5 | 166 | Salzburg | AT | 1881 | 1993 | 113 | P | A | 47 | 48 | 13 | 2 |
| 67 | CLERMONT_FERRAND | FR | 1858 | 1988 | 131 | P | A | 45 | 47 | 3 | 5 | 167 | QUIMPER | FR | 1861 | 1973 | 113 | P | A | 48 | 0 | -4 | 11 |
| 68 | PITEAA | SE | 1860 | 1990 | 131 | P | A | 65 | 19 | 21 | 28 | 168 | Wien | AT | 1882 | 1993 | 112 | P | A | 48 | 13 | 16 | 20 |
| 69 | Bra | IT | 1862 | 1991 | 130 | P | A | 44 | 42 | 7 | 51 | 169 | Larnaca | CY | 1882 | 1993 | 112 | P | A | 34 | 42 | 33 | 38 |
| 70 | HALMSTAD | SE | 1861 | 1990 | 130 | P | A | 56 | 40 | 12 | 55 | 170 | Nicosia | CY | 1882 | 1993 | 112 | P | A | 35 | 10 | 33 | 21 |
| 71 | VAEXJOE | SE | 1861 | 1990 | 130 | P | A | 56 | 52 | 14 | 48 | 171 | GALLE | LK | 1869 | 1980 | 112 | P | A | 6 | 0 | 80 | 11 |
| 72 | KARLSHAMN | SE | 1861 | 1990 | 130 | P | A | 56 | 12 | 14 | 53 | 172 | LINCOLN | US | 1878 | 1989 | 112 | P | A | 40 | 47 | -96 | 41 |
| 73 | VANERSBORG | SE | 1861 | 1990 | 130 | P | A | 58 | 24 | 12 | 18 | 173 | LUZERN | CH | 1861 | 1971 | 111 | P | A | 47 | 5 | 8 | 18 |
| 74 | LANSING | US | 1864 | 1993 | 130 | P | A | 42 | 47 | -84 | 35 | 174 | Limassol | CY | 1883 | 1993 | 111 | P | A | 34 | 40 | 33 | 1 |
| 75 | TORSHAVN | FO | 1868 | 1996 | 129 | P | A | 62 | 1 | -6 | 46 | 175 | VARDE | DK | 1861 | 1971 | 111 | P | A | 55 | 35 | 8 | 30 |
| 76 | Nice | FR | 1865 | 1993 | 129 | P | A | 43 | 42 | 7 | 15 | 176 | BODHOLT | DK | 1861 | 1971 | 111 | P | A | 56 | 5 | 9 | 18 |
| 77 | Catania | IT | 1866 | 1994 | 129 | P | A | 37 | 30 | 15 | 3 | 177 | Portela | PT | 1864 | 1974 | 111 | P | A | 38 | 46 | -9 | 8 |
| 78 | Filaret | RO | 1865 | 1993 | 129 | P | A | 44 | 43 | 26 | 10 | 178 | Bogota | CO | 1866 | 1975 | 110 | P | A | 4 | 36 | -74 | 5 |
| 79 | KRISTIANSTAD | SE | 1860 | 1988 | 129 | P | A | 56 | 0 | 14 | 11 | 179 | OCNA-SUGATAG | RO | 1881 | 1990 | 110 | P | A | 47 | 47 | 23 | 53 |
| 80 | KOLN | DE | 1848 | 1975 | 128 | P | A | 51 | 0 | 7 | 0 | 180 | Ain_Draham | TN | 1885 | 1994 | 110 | P | A | 36 | 47 | 8 | 42 |
| 81 | KLEVE | DE | 1849 | 1976 | 128 | P | A | 51 | 47 | 6 | 5 | 181 | Ajaccio | FR | 1885 | 1993 | 109 | P | A | 41 | 55 | 8 | 45 |
| 82 | BIDSTON_OBSERVATORY | GB | 1867 | 1994 | 128 | P | A | 53 | 24 | -3 | 5 | 182 | Gondrexange | FR | 1856 | 1965 | 109 | P | A | 48 | 42 | 6 | 54 |
| 83 | HALLE | DE | 1851 | 1976 | 126 | P | A | 51 | 30 | 11 | 53 | 183 | CONCORDIA | US | 1886 | 1994 | 109 | P | A | 39 | 35 | -97 | 41 |
| 84 | Murcia | ES | 1862 | 1987 | 126 | P | A | 37 | 59 | -1 | 8 | 184 | Tunis | TN | 1887 | 1994 | 108 | P | A | 36 | 48 | 10 | 11 |
| 85 | DULUTH | US | 1871 | 1996 | 126 | P | A | 46 | 47 | -92 | 5 | 185 | WICHITA | US | 1889 | 1996 | 108 | P | A | 37 | 42 | -97 | 24 |
| 86 | CHICAGO_O'HARE | US | 1871 | 1996 | 126 | P | A | 41 | 54 | -87 | 35 | 186 | TORNQUIST | AR | 1889 | 1995 | 107 | P | A | -38 | 4 | -62 | 13 |
| 87 | NASHVILLE | US | 1871 | 1996 | 126 | P | A | 36 | 12 | -86 | 48 | 187 | LERWICK | GB | 1890 | 1996 | 107 | P | A | 60 | 7 | -1 | 10 |
| 88 | CAPPOQUIN | IE | 1870 | 1994 | 125 | P | A | 52 | 12 | -7 | 48 | 188 | GOETEBORG_GAVE | SE | 1890 | 1996 | 107 | P | A | 57 | 46 | 11 | 52 |
| 89 | ENNISCORTHY | IE | 1870 | 1994 | 125 | P | A | 52 | 30 | -6 | 35 | 189 | Worcester | ZA | 1880 | 1986 | 107 | P | A | -33 | 39 | 19 | 26 |
| 90 | ARDARA | IE | 1870 | 1994 | 125 | P | A | 54 | 47 | -8 | 18 | 190 | OKLAHOMA_CITY | US | 1891 | 1996 | 106 | P | A | 35 | 24 | -97 | 35 |
| 91 | DENVER | US | 1872 | 1996 | 125 | P | A | 39 | 47 | -105 | 0 | 191 | Neufchateau | FR | 1859 | 1963 | 105 | P | A | 48 | 21 | 5 | 42 |
| 92 | GRAND_RAPIDS | US | 1870 | 1994 | 125 | P | A | 42 | 54 | -85 | 30 | 192 | Windhoek | NA | 1892 | 1996 | 105 | P | A | -22 | 34 | 17 | 6 |
| 93 | KIEL | DE | 1851 | 1974 | 124 | P | A | 54 | 17 | 10 | 6 | 193 | KVIKKJOKK | SE | 1886 | 1990 | 105 | P | A | 66 | 56 | 17 | 43 |
| 94 | PUTBUS | DE | 1854 | 1977 | 124 | P | A | 54 | 42 | 13 | 29 | 194 | LANDER | US | 1892 | 1996 | 105 | P | A | 42 | 47 | -108 | 41 |
| 95 | LAVAL | FR | 1850 | 1973 | 124 | P | A | 48 | 0 | 0 | -12 | 195 | AMARILLO | US | 1892 | 1996 | 105 | P | A | 35 | 12 | -101 | 41 |
| 96 | PORTREE | GB | 1861 | 1984 | 124 | P | A | 57 | 24 | -6 | 11 | 196 | Beograd | YU | 1888 | 1992 | 105 | P | A | 44 | 48 | 20 | 28 |
| 97 | Calcutta-Alipore | IN | 1837 | 1960 | 124 | P | A | 22 | 32 | 88 | 24 | 197 | Sadovetz | BG | 1892 | 1995 | 104 | P | A | 43 | 18 | 24 | 23 |
| 98 | PEYREHORADE | FR | 1851 | 1973 | 123 | P | A | 43 | 35 | -1 | 6 | 198 | Chiflik | BG | 1892 | 1995 | 104 | P | A | 43 | 48 | 26 | 2 |
| 99 | Avignon | FR | 1871 | 1993 | 123 | P | A | 43 | 57 | 4 | 48 | 199 | GOULBURN_KIPPILAW | AU | 1886 | 1988 | 103 | P | A | -34 | 45 | 149 | 36 |
| 100 | VENDOME | FR | 1851 | 1973 | 123 | P | A | 47 | 47 | 1 | 6 | 200 | Chipka | BG | 1893 | 1995 | 103 | P | A | 42 | 43 | 25 | 20 |

# Résultats de la théorie des valeurs extrêmes.