Potentiel de réserves d'un bassin pétrolier : modélisation et estimation

Vincent Lepez

Université Paris XI - ORSAY École du Pétrole et des Moteurs

N° D'ORDRE : 7059

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

THÈSE

présentée pour obtenir

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

SPÉCIALITÉ : MATHÉMATIQUES

par

Vincent LEPEZ

Sujet : **POTENTIEL DE RÉSERVES D'UN BASSIN PÉTROLIER : MODÉLISATION ET ESTIMATION**

Soutenue le **10 Décembre 2002** devant le jury composé de :

M.	Denis BABUSIAUX	Président
M.	Philippe BESSE	Rapporteur
Mme.	Gwénaëlle CASTELLAN - GUÉRIN	Examinateur
Mme.	Fabienne COMTE	Rapporteur
M.	Pierre DELFINER	Invité au jury
M.	Jean-Pierre INDJEHAGOPIAN	Examinateur
M.	Pascal MASSART	Directeur de Thèse

À mon Père et Professeur,

À mon Tonton du Havre et Connaisseur...

Merci...

Je tiens en tout premier lieu à remercier Pascal Massart, mon directeur de thèse, à bien des titres. D'abord pour ses qualités exceptionnelles tant d'enseignant que de scientifique, évidemment. Je veux aussi le remercier pour toute la confiance qu'il m'a témoigné. Je crois qu'il sait faire en sorte que ses étudiants donnent le meilleur d'eux-mêmes, du moins, cela a été le cas pour moi et je lui en suis profondément reconnaissant.

Un immense merci aussi à Denis Babusiaux, Pierre-René Bauquis et Pierre Delfiner, qui ont été mes interlocuteurs très privilégiés pendant ces années de travail. Grâce à eux, et plus généralement à l'accueil enthousiaste du Centre Économie et Gestion de l'ENSPM et celui de la DGEP de Total-FinaElf, de nombreux et très riches échanges avec chercheurs, enseignants et professionnels de domaines connexes ont été possibles.

Je veux aussi témoigner ma reconnaissance et mon amitié à l'égard de Gwénaëlle Castellan-Guérin, dont les travaux ont été au coeur (avec certainement aussi pas mal d'organes...) des développements théoriques de cette thèse, sur lesquels j'ai parfois bien souffert... Et sur lesquels elle m'a efficacement épaulé.

Je remercie très sincèrement Fabienne Comte et Philippe Besse qui m'ont fait l'honneur de bien vouloir rapporter cette thèse. Merci aussi à Jean-Pierre Indjehagopian pour les discussions et conseils chaleureux, pour sa confiance, et le fait qu'il ait accepté d'être membre de mon jury.

Merci enfin à Jean-Luc Karnik et Jean-Pierre Favennec, respectivement directeurs de l'ENSPM et du CEG pour m'avoir permis de terminer ce travail dans d'excellentes conditions.

J'en aurais bassiné, des gens que j'aime, avec cette thèse... Tous m'ont porté, apporté ou supporté. Je ne peux pas citer tout le monde ici, mais voici un petit florilège, forcément non exhaustif, ne m'en veuillez pas...

- *les IFP's dont Valérie la breizhou, Florane la hô'môrnâise, Maribé du pays de la moutarde et du bon vin et enfin Fabrice avec qui on se serrait les coudes contre ces deux dernières lô ;*
- *les Orsay's dont Estelle et ses rumeurs, Antoine et son appétit qui dépasse l'entendement (au fait, tu es sûr que tu seras pris en prépa Agreg ?...) et puis évidemment Catherine et ses syndromes ADB dont les symptômes ne m'ont jamais affecté, pour longtemps encore j'espère ;*
- *les SdeC's, ils se reconnaîtront ;*
- *les Family's, mes très proches ;*
- *enfin, spéciale dédicace bien-sûr à Madame Oune.*

Y'a forcément de vous tous dans les pages qui suivent.

Table des matières

Table des figures

Introduction

L'objectif de cette thèse est de construire un modèle statistique de la distribution des tailles des gisements d'hydrocarbures qui existent dans le sous-sol ainsi que de celle des découvertes déjà effectuées. L'estimation des paramètres de ce modèle via l'estimation de la densité des observations par sélection de modèles de polynômes par morceaux par maximum de vraisemblance pénalisé nous permet de déduire des estimations du nombre de gisements restant à découvrir, par classe de taille.

Le contexte

Les sources d'énergies fossiles que sont le pétrole et le gaz naturel font partie de la classe des ressources épuisables. Cette caractéristique a des implications majeures dans la formation de l'équilibre économique entre offre et demande. De façon très schématique, la pénurie (ou la peur de la pénurie, comme pour le choc pétrolier de 1973) implique automatiquement la hausse brutale des prix sur le marché, qui en conséquence fait baisser la demande, via les politiques d'économies d'énergie par exemple.

La problématique de la prévision du montant total d'hydrocarbures disponibles est donc absolument centrale pour l'économie de l'énergie en général. La peur de la pénurie est un problème presque aussi vieux que l'industrie pétrolière[1] qui cherche depuis toujours à savoir quelles sont nos réserves ultimes et combien de temps nous allons pouvoir en profiter. Au fur et à mesure que nous consommons ces réserves, ce problème devient source de travaux de plus en plus nombreux.

L'estimation des réserves est en réalité à la fois un problème géologique, économique et statistique :

– géologique car les objets d'étude que sont les gisements d'hydrocarbures relèvent des sciences de la terre que sont la sédimentologie, la tectonique, la sismique, le forage, etc...

[1] Voir notamment Maretheux [54] qui, en 1919, prévoyait l'épuisement imminent des gisements d'hydrocarbures des États-Unis!

– économique car l'extraction des hydrocarbures naturels à un coût. Ce coût détermine la fraction des quantités en place sous terre qui représentent effectivement des réserves. Par réserves nous entendons quantités qui pourront un jour être exploitées de façon rentable, c'est-à-dire à un coût moindre que ce qu'elles rapportent à l'exploitant ;
– et enfin statistique à plusieurs niveaux. D'abord à l'échelle "microscopique" du champ lui-même. En effet, les mesures physiques ou sismologiques nécessaires à la détermination de l'existence éventuelle d'un champ et à l'évaluation de ses caractéristiques géologiques, géophysiques et géochimiques sont réalisées par sondage. L'extrapolation des résultats à la totalité du champ relève de techniques de modélisation où les statistiques interviennent. En ce qui concerne l'estimation des réserves au niveau plus macroscopique du bassin sédimentaire, voire du monde, les statistiques jouent de nouveau un rôle important. En effet, il s'agit de déterminer certaines caractéristiques de la distribution des tailles des champs au vu seulement de ceux qui ont été découvert jusqu'alors, c'est-à-dire un sous-échantillon. En termes statistiques, nous sommes face à un problème d'inférence.

Le but de notre travail est de fournir un éclairage sur le dernier point, qui concerne l'estimation des réserves à l'échelle d'un bassin sédimentaire. Pour cela, nous voulons d'abord construire un modèle statistique de la distribution des tailles des gisements d'hydrocarbures. Nous cherchons ensuite à estimer les paramètres du modèle au moyen du sous-échantillon des observations. Puis, de ces estimations, nous déduirons des évaluations de réserves restant à découvrir.

L'Institut Français du Pétrole (IFP) a été à l'origine de ce travail de recherche. Il a souhaité que soit menée une étude approfondie des aspects mathématiques sous-jacents aux travaux de Jean Laherrère en la matière ([45] et suivants). Ce dernier crée différents modèles de la distribution des tailles des champs existant dans le sous-sol. En évaluant les paramètres de ces modèles souvent de manière très empirique, Laherrère fournit des estimations de réserves restant à découvrir que l'IFP et nombre d'industriels souhaitent mettre à l'épreuve.

Conjointement, l'IFP et Total (aujourd'hui TotalFinaElf) décident alors de construire leur propre modèle, basé sur les idées de Laherrère, mais qui en étend considérablement les fondements mathématiques dans le but de développer un outil simple et statistiquement fiable d'estimation des réserves.

Une des contraintes fortes quant au développement de ce modèle est la quantité en général faible de données disponibles. Celles-ci sont en effet soit inexistantes, soit extrêmement coûteuses à acquérir. Il était donc nécessaire de créer un modèle dont l'estimation des paramètres requiert le minimum de données en entrée, en l'occurrence une seule liste de tailles des champs déjà découverts dans une certaine zone du globe, éventuellement sans même

l'historique de ces découvertes. En sortie, l'utilisateur doit trouver une liste des tailles des champs restant à découvrir.

Notre thèse est divisée en cinq chapitres dont nous présentons les caractéristiques et résultats.

Le problème de l'estimation des réserves

Le premier chapitre présente la problématique générale de l'estimation des réserves et énonce les premières hypothèses naturalistes du modèle de distribution des tailles des champs que nous cherchons à créer.

Considérons le ratio R/P (pour réserves/production) qui mesure en années la durée de vie des réserves au rythme actuel de production. Depuis plus de 30 ans, ce ratio est stable et est égal à 30 ans pour le cas du pétrole, 60 pour celui du gaz naturel. Autrement dit, en 1970, on annonçait la fin du pétrole en 2000 ! Cela signifie donc que de nouvelles réserves voient le jour au cours du temps. Plusieurs facteurs technico-économiques influent sur ce renouvellement des réserves :
– le progrès technologique ;
– la meilleure connaissance des gisements au cours du temps ;
– le prix du pétrole brut ;
– la découverte de nouveaux champs.
En particulier, la découverte de nouveaux champs est le sujet sur lequel nous axons notre travail en considérant que tous les autres déterminants restent constants.

En nous fixant une définition de ce que représentent les réserves d'hydrocarbures d'un point de vue technico-économique, nous cherchons à modéliser la distribution des tailles des champs qui existent dans le sous-sol au moyen des observations que sont les champs déjà découverts. Ce dernier, l'échantillon observé, représente un sous-échantillon sans remise du premier, l'échantillon parent. Ainsi, l'exploration pétrolière est interprétée comme un tirage sans remise : les découvertes sont les tirages sans remise au sein de l'ensemble des champs qui existent dans la nature. Nous ferons l'hypothèse que la loi de l'échantillon parent, supposé i.i.d., est approximativement une loi de Lévy-Pareto (exponentielle d'une loi exponentielle), notamment pour sa propriété d'invariance par changement d'échelle. Il est en effet classique dans l'exploration pétrolière qu'à côté d'une grande découverte se trouve une moins grande, à proximité de laquelle se trouve encore une petite...

Si l'échantillon observé était tiré de manière équiprobable alors il serait simple de tirer des conclusions sur la loi de l'échantillon parent au moyen de la distribution de l'échantillon observé par des techniques d'inférence empruntées à la théorie des sondages par exemple. Or, l'échantillon observé est un échantillon dont le tirage est biaisé par un "effet taille". En effet, pour

l'explorateur, il est *a priori* plus facile (et plus rentable aussi) de découvrir une structure de taille importante. Ainsi, plus un champ de l'échantillon parent est gros et plus sa probabilité d'avoir été découvert, donc sa probabilité de figurer dans l'échantillon observé, est importante.

Nous commençons donc à entrevoir les prémices d'un modèle dit de "super-population" où un échantillon parent est supposé i.i.d. d'une certaine loi et l'échantillon observé est un sous-échantillon de ce dernier. Ce sous-échantillon sera pour notre part sans remise et biaisé par un effet taille, dont la seul caractéristique connue *a priori* est qu'il accorde une probabilité d'observation croissante avec la taille de l'individu.

Le chapitre 2 est consacré à la construction rigoureuse du modèle statistique correspondant à la problématique de type sondage brièvement exposée ci-dessus.

Le modèle

Après avoir spécifié l'échelle d'étude de nos échantillons, celle du système pétrolier, qui garantit l'homogénéité géologique d'un certain nombre de paramètres importants, nous montrons qu'en vue du traitement statistique à venir, il est plus adapté de travailler avec les logarithmes des tailles (LogTailles) des champs, qui sont supposés suivre une loi exponentielle.

Nous effectuons ensuite une petite revue de la théorie des tirages successifs biaisés par effet taille. Dans l'ensemble de ces travaux, une information *a priori* importante est disponible sur la forme précise du biais (probabilités d'inclusion du type fonction puissance par exemple). Nous construisons pour notre part un modèle dans lequel la seule information *a priori* sur les probabilités de découverte des champs est le fait que celle-ci est croissante en fonction de leur taille.

Appelons $\mathbb{X} = \{X_1, ..., X_N\}$ l'échantillon parent des LogTailles des champs qui existent dans le sous-sol, que l'on suppose i.i.d. de loi à densité f contre une mesure μ. L'échantillon observé $\mathbb{Y} = \{Y_1, ..., Y_n\}$ est un sous-échantillon sans remise, biaisé par effet taille. Nous interprétons ce modèle comme un modèle de censure dont les caractéristiques sont les suivantes :

- conditionnellement à la donnée de $\mathbb{X}$, on suppose qu'un objet X_i a la probabilité $\omega(X_i)$ d'avoir été tiré ;
- implicitement, on considère donc une fonction de pondération ω à valeurs dans $]\,0\,;1]$ et un échantillon de censure $\{\varepsilon_1, ..., \varepsilon_N\}$ tel que, $\mathbb{P}(\varepsilon_i \,|\, \mathbb{X}) = \omega(X_i)$ pour $1 \leq i \leq N$;
- de nouveau, conditionnellement au fait que $\varepsilon_1 + \cdots + \varepsilon_N = n$ est connu, l'échantillon des observés est alors $\{Y_1, ..., Y_n\} = \{X_i \,/\, \varepsilon_i = 1\}$.

Nous montrons que, conditionnellement au fait que le nombre d'observations n est connu, ce qui est bien sûr le cas en pratique, l'échantillon observé

$\mathbb{Y} = \{Y_1, ..., Y_n\}$ est un échantillon i.i.d. de densité

$$f_\omega \;=\; \frac{\omega \times f}{\mathbb{E}_f\left(\omega(X_1)\right)}. \tag{1}$$

La densité des observations est donc proportionnelle au produit de la fonction de pondération et de la densité de l'échantillon parent. Ce résultat est très intuitif dans le sens où il s'interprète en voyant que f_ω est un compromis direct entre ω et f. La fonction ω "tord" la distribution de densité f en favorisant l'apparition d'observations là où elle est proche de 1 et en les censurant sinon.

Dans notre cas, l'hypothèse que plus un champ est gros et plus sa probabilité de découverte est grande se traduit naturellement par une hypothèse de croissance sur ω.

Notons que si l'on multiplie ω par une constante $c \in \left]\,0\,;1\right[$ alors l'expression de f_ω reste inchangée. Pour que notre modèle à venir soit identifiable, nous sommes donc amenés à imposer la contrainte $\max \omega = 1$. Ceci se traduit, dans le cas de l'hypothèse de croissance de ω, par $\omega(\max \mathbb{Y}) = 1$ *i.e.* le plus gros champ a été trouvé de façon sûre. Du point de vue naturaliste, cela signifie que l'on travaille sur une zone géographique où l'exploration est suffisamment mature.

En pratique, f et ω sont inconnues. L'enjeu de notre travail est d'estimer à la fois f et ω au moyen de f_ω, bref, de séparer la partie "loi parente" de la partie "biais" en estimant la densité des observations.

Dans le chapitre 3, nous spécifions les formes des lois que nous manipulons ensuite. Nous résolvons les équations de vraisemblance associées au modèle et montrons que l'on peut effectivement séparer la partie biais de la partie loi parente. Grâce aux évaluations du biais, on construit des estimateurs du nombre de champs restant à découvrir par classe de taille.

Estimation dans un modèle

Nous supposons que la fonction ω est une fonction en escalier dont les marches sont situés sur une partition $m = \{I_1, ..., I_{|m|}\}$ de l'intervalle d'étude $[\min \mathbb{Y}\,; \max \mathbb{Y}]$, appelée modèle. La loi parente des LogTailles des champs du sous-sol est exponentielle de densité f_α :

$$\omega(x) \;=\; \sum_{I \in m} \omega_I \mathbb{1}_I(x) \quad \text{et} \quad f_\alpha(x) \;=\; \alpha e^{-\alpha x} \mathbb{1}_{\mathbb{R}^+}(x), \quad \forall\, x \in \mathbb{R}.$$

On peut alors trouver une suite $(\theta_I)_{I \in m}$ telle que la densité résultante par (1) s'écrit

$$f_{\alpha,\omega}(x) \;=\; \exp\left(\sum_{I \in m} (\theta_I + \log\alpha - \alpha x) \mathbb{1}_I(x)\right).$$

Elle s'interprète donc comme un modèle exponentiel de polynômes par morceaux de degré 1, dont tous les termes de degré 1 sont égaux. De plus, les termes de degré 0 sont croissants si l'hypothèse de monotonie est active. La relation entre les $(\omega_I)_{I\in m}$ et les $(e^{\theta_I})_{I\in m}$ est bijective via l'hypothèse $\omega(\max \mathbb{Y}) = 1$, ce qui rend équivalente l'estimation de e^θ ou de ω.

Sans la contrainte de monotonie en ω, l'estimateur du maximum de vraisemblance du couple (α, e^θ) est donné par la solution du système suivant :

$$\left\{ \begin{array}{rcl} e^\theta & = & \left(\dfrac{P_n(I)}{\mu_\alpha(I)}\right)_{I\in m} \\ \dfrac{1}{\alpha} & = & \bar{Y} + \displaystyle\sum_{I\in m} e^{\theta_I}\mu'_\alpha(I) \end{array} \right. \tag{2}$$

où l'on a :
- P_n la mesure empirique et $P_n(I)$ celle de l'intervalle I ;
- $\mu_\alpha(I)$ la mesure exponentielle de paramètre α de l'intervalle I ;
- $\mu'_\alpha(I)$ sa dérivée par rapport à α ;
- $\bar{Y}$ la moyenne empirique de l'échantillon $\mathbb{Y} = \{Y_1, ..., Y_n\}$.

La première équation du système (2) s'interprète comme un écart entre la mesure empirique de l'intervalle I et la mesure $\mu_\alpha(I)$, qu'il devrait posséder s'il n'y avait pas de biais dans le tirage. On retrouve donc ici clairement l'idée de probabilité d'inclusion.

La seconde équation du système (2) montre l'impact du biais dans le tirage sur l'estimation de α. En effet, si la somme qui se trouve dans le membre de droite était nulle, on retrouverait la formule classique $\hat{\alpha} = 1/\bar{Y}$ d'estimation par maximum de vraisemblance du paramètre d'un échantillon de loi exponentielle.

Nous résolvons ensuite les équations de vraisemblance sous contrainte de monotonie. Les solutions, dont la forme provient d'une technique de régression isotonique, sont données par la résolution du système

$$\left\{ \begin{array}{l} e^\theta_* = \left(\min\limits_{j\geq i} \max\limits_{h\leq i} \left[\dfrac{P_n\left(\bigcup_{k=h}^{j} I_k\right)}{\mu_\alpha\left(\bigcup_{k=h}^{j} I_k\right)} \right] \right)_{1\leq i\leq |m|} \\ \dfrac{1}{\alpha} = \bar{Y} + \displaystyle\sum_{I\in m} e^{\theta_I}_* \mu'_\alpha(I). \end{array} \right. \tag{3}$$

Les intervalles de type $\bigcup_{k=h}^{j} I_k$ qui apparaissent dans les solutions du système (3) sont ceux qui correspondent aux plages de décroissance de l'estimation sans contrainte de monotonie. La valeur correspondante de ω contrainte à la

monotonie sur ces intervalles est alors une valeur moyenne pondérée de celles qui intervenaient hors contrainte, de sorte que la résultante sur m entière soit croissante.

Nous montrons ensuite que la solution du problème contraint sur m est égale à la solution du problème non contraint sur la partition (dite isotonique) obtenue comme celle des plages de constance de ω contrainte, ce qui aura une conséquence importante pour la sélection de modèles que nous allons effectuer par la suite.

La fonction ω étant estimée par $\widehat{\omega}$ via l'estimateur du maximum de vraisemblance $\widehat{e^{\theta}}$ de e^{θ}, nous nous appuyons ensuite sur la théorie des sondages et les estimateurs de Horvitz-Thompson pour fournir des estimateurs du nombre total de champs, sur les classes de taille définies par m :

$$\widehat{N}_{HT} \;=\; \sum_{I \in m} \frac{n P_n(I)}{\widehat{\omega}_I}$$

et pour les réserves totales :

$$\widehat{R}_{HT} \;=\; \sum_{I \in m} \frac{\sum_{\{i\,/\,Y_i \in I\}} e^{Y_i}}{\widehat{\omega}_I}$$

Sélection de modèles

Nous faisons maintenant varier les modèles au sein de plusieurs familles finies et construisons un critère de pénalisation de la vraisemblance afin de sélectionner le meilleur modèle possible, en termes de risque d'estimation. Nous fournissons de plus une borne de risque non-asymptotique à l'estimateur sur le modèle sélectionné.

Notre travail s'appuie sur celui de Castellan [20] et [21], qui montre un théorème de sélection de modèles exponentiels de polynômes par morceaux généraux que nous souhaitons adapter à notre cas contraint :
– à l'égalité des termes de degré 1 ;
– éventuellement à la monotonie des termes de degré 0.

Pour deux densités p et q contre la mesure μ, on définit la distance de Kullback entre p et q par

$$K(p,q) = \int p \log \frac{p}{q}\, d\mu$$

si la mesure $P = p d\mu$ est dominée par $Q = q d\mu$, et $+\infty$ sinon. On définit aussi la distance de Hellinger par

$$h^2(p,q) = \frac{1}{2} \int (\sqrt{p} - \sqrt{q})^2 \; d\mu.$$

Ces deux distances sont reliées par l'inégalité

$$2h^2(p,q) \;\leq\; K(p,q).$$

Sous certaines hypothèses, la distance de Kullback a un comportement proche de celui d'une distance euclidienne, en particulier on peut définir la projection Kullback $\bar{f}$ d'une densité f sur un ensemble de densités E comme $\bar{f} = \text{argmin}\{K(f,g),\, g \in E\}$.

Soit f une densité sur $[\,0\,;1]$ à estimer. Si $\bar{f}_m$ (respectivement $\hat{f}_m$) désigne sa projection Kullback (respectivement son estimateur du maximum de vraisemblance) sur un modèle m alors le risque de $\hat{f}_m$ s'écrit

$$\mathbb{E}(K(f,\hat{f}_m)) \;=\; K(f,\bar{f}_m) + \mathbb{E}(K(\bar{f}_m,\hat{f}_m)). \tag{4}$$

Le premier terme de (4) est un terme de biais qui mesure la distance (non aléatoire mais inconnu) entre f et le modèle dans lequel on cherche à l'estimer. Le second terme, espérance d'une variable aléatoire, est un terme de variance à l'intérieur du modèle m.

Le but de la sélection de modèles est de construire un critère empirique qui permette de sélectionner le modèle qui fasse le meilleur compromis entre biais et variance en terme de risque d'estimation. Bien entendu, ce meilleur modèle, que nous appelons "oracle" existe, mais il est inconnu car il dépend de f. Notre critère vise donc à sélectionner un modèle qui fasse aussi bien que l'oracle, dans le sens où le risque de l'estimateur sur le modèle sélectionné est majoré par le risque de l'oracle, à constante multiplicative près.

Le critère que nous avons construit est un critère de maximum de vraisemblance pénalisé dans la mesure où il s'écrit

$$\text{crit}_n(m) \;=\; P_n(-\log \hat{f}_m) + \text{pen}_n(m)$$

où $P_n(-\log \hat{f}_m)$ est l'opposée de la log-vraisemblance et $\text{pen}_n(m)$ est une fonction de pénalité qui dépend du modèle m et du nombre d'observations n.

En pratique, nous supposons que les modèles m appartiennent à une famille $\mathcal{M}_n$ définie comme étant soit la famille des partitions à pas régulier, soit la famille des partitions à pas irrégulier mais dont chaque intervalle est réunion d'élément d'une partition fine à pas régulier. On notera $\hat{m}$ la partition qui minimise le critère : $\hat{m} = \text{argmin}\{\text{crit}_n(m),\, m \in \mathcal{M}_n\}$.

– Si l'on choisit la fonction pen_n telle que

$$\text{pen}_n(m) \;\geq\; \lambda\left(1+\sqrt{2(1+1/\lambda)L_m}\right)^2 \frac{|m|}{n}, \tag{5}$$

pour une constante $\lambda > 1/2$, les L_m étant un ensemble de poids mesurant la complexité de la famille $\mathcal{M}_n$;

– si le nombre de morceaux de la partition la plus fine de la famille $\mathcal{M}_n$ est en $\mathcal{O}(\sqrt{n}/\log^2 n)$;
– si $\tilde{f} = \hat{f}_{\hat{m}}$ désigne l'estimateur du maximum de vraisemblance pénalisé de f, qu'il soit contraint ou non à la monotonie en ω, alors on a la borne de risque suivante :

$$\mathbb{E}\left(h^2(f,\tilde{f})\mathbb{I}_{\Omega_n\cap\{\tilde{f}\geq\rho_n\}}\right) \leq \; c \inf_{m\in\mathcal{M}_n}\left\{2K(f,\bar{f}_m) + \mathrm{pen}_n(m) + c'\frac{2\Sigma_n+1}{n}\right\},$$

où c et c' sont deux constantes, ρ_n est une suite qui tend vers 0 et Ω_n est un ensemble de grande probabilité dans le sens où la probabilité de son complémentaire est négligeable devant toute puissance de n.

Cette borne de risque est la même que celle du théorème principal de Castellan [21], mais les hypothèses en sont légèrement différentes, essentiellement en raison de la restriction $\mathcal{O}(\sqrt{n}/\log^2 n)$ sur le cardinal de la partition la plus fine de $\mathcal{M}_n$. Le théorème de Castellan, en effet, pose une hypothèse en $\mathcal{O}(n/\log^2 n)$, ce qui est évidemment moins restrictif du point de vue théorique, mais n'a aucun impact pratique puisque ces bornes sont définies à constante multiplicative inconnue près. Cette restriction est due au fait que les coefficients des polynômes sur chaque intervalle du modèle subissent des contraintes telles qu'ils ne sont pas indépendants entre eux. Castellan, pour sa part, estime les densités exponentielles de polynômes par morceaux intervalle par intervalle en s'appuyant sur l'indépendance des coefficients entre chaque intervalle des partitions. Du point de vue théorique, notre restriction sur l'hypothèse $\mathcal{O}(n/\log^2 n)$ s'interprète en voyant que la prise en compte des dépendances entre les coefficients estimés tout au long d'une partition impose que chaque intervalle contienne plus d'information (donc plus d'observations) que dans le cas de Castellan. À nombre d'observation fixé, cela implique donc des estimations sur partitions moins fines.

Une fois la preuve de ce théorème achevée, nous menons une discussion sur les choix optimaux des poids L_m et de la constante λ qui intervient dans la définition (5) de la fonction de pénalité. Pour cette dernière, nous mettons en œuvre la récente méthode dite de la "pente", détaillée par Lebarbier dans sa thèse [50]. Très grossièrement, celle-ci consiste
– à sélectionner le meilleur modèle $\hat{m}_D$ de dimension D au sens du critère de la log-vraisemblance pour $1 \leq D \leq D_{\max}$. En pratique, $D_{\max} \leq 20$, ou 14, selon que les partitions sont régulières ou irrégulières ;
– construire le nuage de points de coordonnées $(\mathrm{pen}_n(\hat{m}_D), -\log P_n(\hat{f}_{\hat{m}_D})$;
– calculer la pente de la droite de régression du précédent nuage de points.
L'opposé du double de cette pente fournit alors une valeur performante de la constante λ, adaptée aux données.

Le cinquième et dernier chapitre de la thèse concerne la mise en pratique des protocoles d'estimation décrits dans les chapitres 3 et 4.

Applications

Du point de vue purement algorithmique, les traitements nécessaires aux estimation décrites ci-dessus sont lourds. Au sein d'un modèle, les solutions des équations de vraisemblance (sous contrainte de monotonie en ω ou non) ne sont pas définies de manière analytique. Une méthode par itération s'impose alors. Ensuite, la sélection de modèles dans le cas de partitions irrégulières impose un nombre exponentiel de modèles à visiter.

Il faut noter que nos applications ne rentrent pas exactement dans le cadre des hypothèses du théorème principal ci-dessus. En effet, les densités sur lesquelles nous travaillons ne sont pas bornées. En pratique, nous levons ce problème en travaillant conditionnellement aux données, l'estimateur du maximum de vraisembalnce ne charge alors que l'intervalle compact d'étude $[\min \mathbb{Y}; \max \mathbb{Y}]$, que nous pouvons ramener à $[0;1]$ par simple homothétie/translation.

Pourtant, les résultats de ces estimations peuvent être présentés de manière très simple. Nous avons donc choisi de créer un logiciel convivial permettant à tout utilisateur de mettre en pratique les protocoles d'estimation préprogrammés, soit sur des simulations de données, soit sur des données réelles issues de bassins sédimentaires.

Grâce à ce logiciel, nous avons pu tester nos protocoles d'estimation en simulation ainsi que sur des données réelles sur plusieurs bassin sédimentaires mondiaux.

Nous avons programmé 20 protocoles d'estimation différents. En effet, nous avons la possibilité de spécifier ou non la contrainte de monotonie en ω ainsi que la valeur du paramètre α de la loi de Pareto sous-jacente. De plus, nous proposons 5 modes de partition de l'intervalle d'étude (sur les LogTailles) :
- partitions de pas régulier ;
- partitions de pas irréguliers dont les intervalles sont basés sur une partition fine de pas régulier ;
- partitions dont les intervalles ont une fréquence d'observation régulière (blocs statistiques équivalents) ;
- partitions dont les intervalles sont basés sur une partition fine de blocs statistiquement équivalents ;
- partition complètement spécifiée par l'utilisateur.

Il est aussi possible de mettre en compétition les divers protocoles au sens de leurs critères pénalisés respectifs. Dans le cas des partitions 2 et 4, la propriété d'équivalence des modèles contraints et non contraints à la monotonie sur une partition moins fine est particulièrement intéressante. Elle implique en effet qu'il suffit de retenir dans la sélection de modèles ceux dont l'estimation sans contrainte conduit à un estimateur monotone de ω.

Lorsque la densité à estimer appartient exactement à un modèle exponentiel de polynômes par morceaux, nos simulations portent sur les protocoles d'estimation à α inconnu, avec ou sans contrainte de monotonie en ω et selon tous les modes de partition sauf celui spécifié par l'utilisateur, soit 8 protocoles au total. Lorsque la densité à estimer n'appartient pas à un modèle exponentiel de polynômes par morceaux, nos simulations portent sur les protocoles d'estimation à α inconnu, avec monotonie en ω et selon tous les modes de partition sauf celui spécifié par l'utilisateur, soit 4 protocoles au total.

Notons que lorsque les partitions sont basées sur des intervalles de fréquence d'observation régulière plutôt que de pas régulier, nous sortons assez largement du cadre du théorème principal puisque ces partitions deviennent aléatoires.

Les conclusions des simulations sont les suivantes :
- les protocoles basés sur des intervalles irréguliers construits sur ceux d'une partition plus fine (pas régulier ou blocs statistiques équivalents) sont plus performants que ceux basés sur sur une partition de pas régulier ou de blocs statistiques équivalents ;
- les protocoles basés sur les blocs statistiques équivalents sont plus performants que ceux basés sur des partitions régulières ;
- les estimations sont plus performantes lorsque la densité à estimer n'appartient pas aux modèles sur lesquels on l'estime ;
- dans ce dernier cas, les protocoles tenant compte de la contrainte de monotonie en ω sont plus performants que ceux qui relaxent la contrainte.

Enfin, nous fournissons quatre applications aux données réelles de quatre bassins sédimentaires du monde : mer du Nord, delta du Congo, offshore du golfe du Mexique et un exemple traité volontairement de façon non optimale sur la Sibérie. En effet, les trois premiers sont des cas d'école dans le sens ou sans imposer la contrainte en ω, le résultat est monotone. Nous souhaitions montrer, dans le cas de la Sibérie, le cas où un des protocole sélectionne un modèle sur lequel l'estimation de ω n'est pas monotone alors que le protocole optimal au sens de l'ensemble des critères pénalisés sélectionne bien un modèle avec monotonie.

Nous concluons ce chapitre et la thèse par un petit bémol sur nos résultats sur données réelles dans la mesure où notre méthode générale ne peut être validée en pratique. En effet, aucun bassin sédimentaire ne s'est encore vu exploré dans son intégralité, c'est-à-dire que l'on ne dispose pas de terrain d'étude sur lequel l'échantillon parent $\mathbb{X} = \{X_1, ..., X_N\}$ est totalement connu.

Nos seules méthodes de validation possibles sont donc les simulations et la rétroprévision. En se plaçant en arrière dans le temps, nous vérifions que

les estimations faites ne sont pas contradictoires avec les résultats connus aujourd'hui.

Perspectives

Il existe plusieurs champs d'investigation ouverts par cette thèse.

Du point de vue théorique, il serait intéressant d'énoncer un théorème analogue à celui du chapitre 4 dans le cadre des partitions basées sur les blocs statistiquement équivalents. Plus généralement, on pourrait énoncer un résultat similaire pour des partitions basées sur des sous-ensembles des statistiques d'ordre de l'échantillon observé $\mathbb{Y} = \{Y_1, ..., Y_n\}$. Il nous semble que cela impliquerait un important travail d'adaptation de la preuve.

Nous pensons qu'à moindre frais il serait envisageable d'adapter ce théorème aux densités exponentielles de polynômes par morceaux dont les coefficients suivent un ensemble de contraintes de type polyêdre convexe général. Il est probable qu'un tel résultat soit aussi valable pour un ensemble de contrainte de type convexe fermé plus général encore. Notre théorème serait alors un cas particulier. Un autre cas particulier intéressant serait celui de la sélection de modèles de logsplines.

Du point de vue algorithmique, les codes source du logiciel peuvent être largement optimisés. Les temps de calcul nécessaires aux estimations dans le cas des partitions irrégulières sont encore extrêmement (et beaucoup trop) longs. L'intégration de l'évaluation de bornes pour les intervalles de confiance dans les estimations reste aussi à mettre en œuvre.

Concernant les applications au monde pétrolier et gazier, le modèle devra peu à peu prendre en compte les autres facteurs technico-économiques qui rentrent en jeu dans le processus de renouvellement des réserves : prix du brut, meilleurs connaissance des gisements, progrès technologique. Chacun d'entre-eux nécessite une approche particulière en terme de modélisation.

Enfin, un des champs naturels d'investigation est maintenant celui de la production des réserves dont on a estimé les volumes. En particulier, la connaissance du processus temporel qui régit la découverte des champs (s'il existe) est une donnée fondamentale de la prévision de production pour les années à venir. Certains modèles basiques existent déjà sur ce thème (comme le modèle de Hubbert – voir les articles de Laherrère [45] et suivants – célèbre dans l'industrie pétrolière), mais ils ont déjà montré leur manque de fiabilité à mainte reprises. Si une telle prévision devient un jour possible, elle permettra notamment, en la mettant au regard d'une prévision de la demande en hydrocarbures naturels, de détecter le moment où l'offre risque de ne plus satisfaire la demande. Nous disposerions alors de modèles capables de prévoir d'éventuels chocs pétroliers, comme celui de 1973 que nous avons évoqué en ce début d'introduction.

Chapitre 1

Présentation du problème

Dans ce chapitre, nous présentons la problématique de l'estimation des réserves d'hydrocarbures à un niveau macroscopique. par macroscopique nous entendons non pas l'échelle d'un gisement, mais celle d'un ensemble cohérent de gisements, comme typiquement celle d'un bassin. Avant de définir de façon rigoureuse ce que nous entendons par "réserves", nous commençons par une section introductive motivant une approche probabiliste de notre sujet. Au cours de celle-ci, pour des raisons de clarté d'exposition, nous utilisons le terme de réserves dans son sens intuitif, c'est-à-dire peu ou prou celui des quantités d'hydrocarbures mesurées et disponibles aujourd'hui. Cette section montre notamment la nécessité de définir cette notion de manière aussi rigoureuse que possible, ce qui fait l'objet de la seconde section, où nous justifions notre emploi dans toute la suite du chiffre dit de *réserves prouvées*.

La troisième section de ce chapitre est conscrée à la description d'un premier modèle de la distribution des tailles des champs d'un bassin pétrolier et d'un embryon de méthode statistique d'évaluation de réserves à cette échelle. Cette dernière approche a fait l'objet d'une littérature relativement abondante et possède un certain crédit au sein de l'industrie pétrolière. En nous appuyant sur certains de nos travaux antérieurs, nous montrons pourtant son manque de robustesse et la nécessité subséquente de l'améliorer en en dérivant un modèle plus fin mais nécessairement plus complexe. Une ébauche de ce modèle clos ce chapitre. Son étude représente l'enjeu théorique de notre thèse,

1.1 Approche probabiliste de l'évaluation du potentiel de réserves d'une zone géographique

Depuis presque toujours, la vision déterministe de l'évaluation du potentiel de réserves d'une zone géographique (comme un bassin par exemple) est considérée comme peu pertinente. Celle-ci consiste à fournir un chiffre unique

ne tolérant pas de marge d'erreur. Il est aujourd'hui très communément admis et démontré, comme nous le verrons dans la suite, qu'une approche de type probabiliste (*i.e.* modélisation probabiliste et estimation statistique des paramètres du modèle) est tout à fait adaptée au problème.

L'aléa dans l'estimation existe en effet en au moins deux aspects de l'exploration pétrolière :
- dans l'évaluation du potentiel d'un gisement en premier lieu puisque les paramètres géologiques du champ sont eux-mêmes incertains ;
- dans l'estimation du montant des réserves à une échelle plus macroscopique puisqu'il s'agit de prévoir, au vu de ce qui a déjà été observé, ce qui peut rester encore à découvrir.

Avant de détailler chacun des deux thèmes énoncés ci-dessus, commençons par décrire le mode de génèse d'un gisement d'hydrocarbures pour tenter de comprendre ensuite où intervient l'aléa dans l'estimation du potentiel de réserves d'un gisement, puis d'une zone géographique entière contenant un ensemble de gisements.

1.1.1 Genèse d'un gisement d'hydrocarbures

Cette section présente les rudiments de géologie pétrolière permettant de comprendre la formation d'un gisement d'hydrocarbures. Pour une approche géologique plus détaillée, le lecteur peut se reporter à Lepez *et al.* [22] ou Perrodon [65].

Un bassin sédimentaire est une zone de dépression remplie de sédiments qui se sont accumulés, compactés puis solidifiés. Les différentes strates composées se superposent alors au cours du temps (plusieurs millions d'années). On appelle *subsidence* ce phénomène. Ces strates successives de sédiments peuvent se déformer en fonction de l'activité tectonique de l'écorce terrestre pour former des anticlinaux (dômes), synclinaux (cuvettes) ou encore des failles. Certaines de ces structures sont susceptibles de receler des hydrocarbures, décrivons-en les raisons.

Les sédiments organiques (composés de carbone, hydrogène, azote et oxygène) sont généralement détruits par les bactéries aérobies, c'est-à-dire vivant à l'air libre. En revanche, ceux qui se sont déposés dans des milieux aquatiques pauvres en oxygène se trouvent protégés de l'action des bactéries puis s'accumulent, s'enfouissent et subissent l'action de micro-organismes anaérobies (n'ayant pas besoin d'air pour vivre) pour donner naissance au *kérogène*, dont les molécules sont fixées au sein d'une *roche mère* argileuse.

Certains de ces sédiments dégradés sont entraînés par subsidence vers de grandes profondeurs où règnent des conditions de hautes pression et température. Le kérogène subit alors un *craquage* thermique au cours duquel l'azote et l'oxygène sont éliminés et les longues chaînes moléculaires se brisent pour

ne plus laisser que des molécules courtes contenant carbone et hydrogène : les *hydrocarbures*. Plus la température est élevée et plus les molécules obtenues seront courtes, donc les hydrocarbures légers. Le kérogène se dégrade d'abord en pétrole (hydrocarbures liquides dont les chaînes carbonées comportent au moins 6 atomes) à partir de 50°C. Au delà de 120°C, le pétrole est à son tour craqué en gaz, dont la molécule la plus élémentaire, qui ne possède qu'un atome de carbone, est la molécule de méthane.

Sous l'effet de la pression de subsidence, le pétrole et le gaz issus du kérogène sont expulsés de la roche mère (ce processus est appelé *migration primaire*). Étant plus légers que l'eau, ils ont ensuite tendance à remonter vers la surface de la croûte terrestre en circulant dans des drains perméables ou des fractures géologiques (c'est la *migration secondaire*). Ils peuvent alors soit être bloqués par une *couverture* imperméable, soit ne pas être arrêtés et s'échapper ou suinter à la surface. Pour constituer un gisement, il faut que la roche réservoir soit suffisamment poreuse pour que les hydrocarbures puissent s'y accumuler. Le couple (couverture + roche réservoir) est appelé *piège*. Par opposition aux pièges stratigraphiques (changements brutaux de porosité au sein d'une même strate de roche), les pièges les plus courants sont les pièges structuraux :
– anticlinaux (dôme imperméable au dessus d'une roche réservoir)
– failles (juxtaposition d'une couche réservoir et d'une couche imperméable).

En résumé, la naissance d'un gisement d'hydrocarbures dans un bassin sédimentaire a donc pour origine la combinaison de plusieurs facteurs : l'existence d'une roche mère, la possibilité de migration (primaire et secondaire), l'existence d'une roche réservoir et d'une couverture qui constituent un piège.

La figure 1.1 présente une coupe Nord-Ouest / Sud-Est d'une partie d'un bassin pétrolier Lybien au lieu-dit de Ghadames. On identifie clairement les failles le long desquelles les hydrocarbures ont été drainés pour s'accumuler en gisements d'huile ou de gaz.

Nous utiliserons, dans la suite, indifféremment le mot "gisement" et le mot "champ".

1.1.2 Premier pas vers un modèle probabiliste

Il est intéresant de noter que Kontorovitch *et al.* [44] considèrent que le processus spatio/temporel de création des bassins pétroliers dans leur globalité est par essence aléatoire. Il justifie alors à lui seul une approche probabiliste de l'évaluation des réserves d'hydrocarbures à cette échelle. Cette interprétation motive le fait que l'on peut regarder l'ensemble des gisements d'hydrocarbures présents dans le sous-sol, figés à la date t d'aujourd'hui, comme la

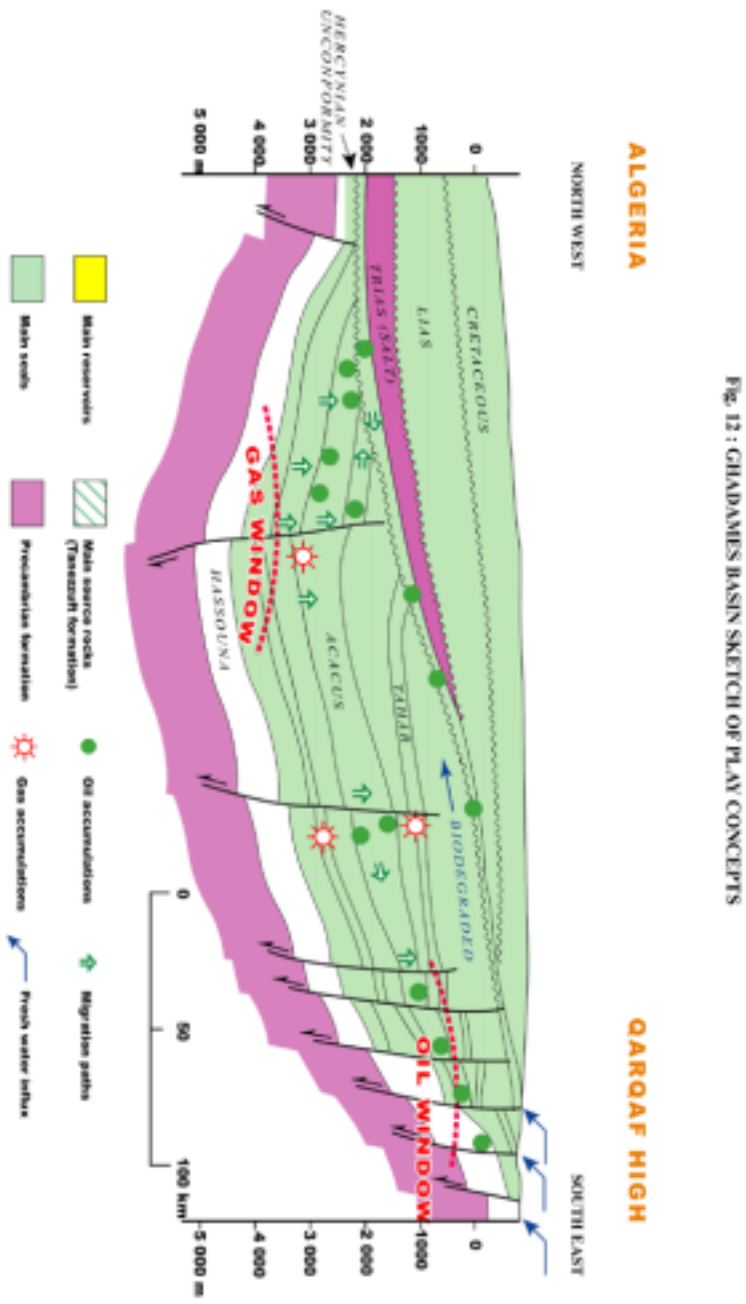

Fig. 12 : GHADAMES BASIN SKETCH OF PLAY CONCEPTS

FIG. 1.1 – Coupe géologique du système pétrolier de Ghadames (Lybie) – Source : IFP.

réalisation d'un processus aléatoire. En particulier, la taille d'un champ "mesurée"[1] aujourd'hui s'interprète comme la réalisation d'une variable aléatoire caractérisée par une famille de paramètres géologiques. Considérant ensuite une zone géographique où ces paramètres géologiques sont homogènes, l'ensemble des tailles des champs qui y sont inclus peut alors être vu comme la réalisation d'un échantillon de cette variable aléatoire[2]. Cette idée sous-tend l'ensemble de notre modélisation à venir mais, pour l'heure, intéressons-nous à l'évaluation des réserves au niveau "microscopique" du champ.

1.1.3 Estimation du potentiel d'un gisement au moyen de la loi LogNormale

Grace aux techniques de sismique 2D ou 3D (et même 4D depuis peu) qui permettent des cartographies 2D, 3D (ou spatio/temporelles) du sous-sol, les géologues d'exploration et les modélisateurs de gisements identifient un *prospect*, ou *objectif géologique* comme une zone géographique susceptible de receler une accumulation d'hydrocarbures. Essentiellement, les géologues cherchent à relever sur les cartes sismiques des anomalies prouvant l'existence de pièges potentiels (couple roche réservoir et couverture) et de drains de migration. Des carottages permettent ensuite de déceler la présence d'une roche mère en amont des drains, d'une roche réservoir effective en aval de ceux-ci, d'une couverture ainsi que d'évaluer les paramètres géologiques du prospect.

À l'échelle du gisement, la quantité de réserves en hydrocarbures apparaît alors comme le produit de plusieurs facteurs[3] :
- une probabilité de présence d'hydrocarbures (l'existence d'un piège et d'une roche mère ne garantit en effet pas la présence d'hydrocarbures[4]) ;
- des indicateurs géologiques comme la porosité, la saturation en hydrocarbures et le volume de roche réservoir qui permettent d'évaluer le volume d'hydrocarbures en place ;
- un indicateur de taux de récupération escompté.

L'approche probabiliste de l'évaluation du prospect consiste à interpréter chacun des caractères ci-dessus comme la réalisation d'une variable aléatoire. La valeur estimée étant généralement considérée comme un indicateur de tendance centrale pour la loi correspondante (mode, moyenne ou médiane

[1] ou plus précisément estimée par les géologues.

[2] petit "hic" : dans cette vision, l'effectif de cet échantillon est alors lui aussi nécessairement aléatoire... Nous reviendrons très sérieusement sur ce point dans le second chapitre.

[3] toutes les compagnies n'utilisent pas exactement les mêmes indicateurs, nous présentons ceux qui leur sont commun.

[4] ceux-ci ont en effet pu se dégrader en fonction de conditions de pression et température inadaptées à leur préservation ou tout simplement disparaître suite à des evènements sismiques créant des fuites dans la couverture.

suivant les approches). Les lois utilisées pour modéliser ces variables aléatoires sont le plus souvent des lois triangulaires positives[5]. En tout état de cause, ces lois sont toujours des lois de probabilité subjectives, dans le sens où elles sont toujours paramétriques et la valeur des paramètres associés est fixée par l'expert, de par sa propre expérience.

Le montant des réserves est donc lui aussi interprété comme la variable aléatoire produit des précédentes. La distribution potentielle des réserves du prospect est ensuite généralement modélisée par une loi LogNormale. En effet, même si le nombre de lois impliquées dans le produit est faible et que ces variables ne sont pas identiquement distribuées, il semble que "conformément" à la version multiplicative du Théorème Central Limite, en pratique, le produit de lois triangulaires dont le support est commun devient rapidement LogNormal. Les simulations intensives de type Monte-Carlo pratiquées par les compagnies à ce sujet sur les évaluations de prospects montrent qu'une loi LogNormale dont les paramètres sont évalués au moyen des formules de Tukey décrit suffisamment bien la distribution attendue de la vraie taille du prospect. Pour autant que nous savons, l'hypothèse LogNormale a été validée et est utilisée dans toutes les compagnies pétrolières.

1.1.4 Des diverses définitions de réserves d'hydrocarbures

Il existe de nombreuses définitions des réserves d'hydrocarbures. En premier lieu, il faut noter que le terme de réserves désigne un concept de nature technico-économique plutôt que géologique. Ainsi, de manière générale, on parlera de :

- *réserves*, pour les hydrocarbures qui sont, ou seront, récupérables ;
- *ressources*, pour les quantités d'hydrocarbures en place dans le gisement, sans faire référence aux contraintes d'accessibilité et/ou de prix de revient. Notons que cette notion est identique à celle de *volumes en place*, aussi couramment utilisée.

Mc Kelvey en 1972 [61], Brobst et Pratt en 1973 [17] ont défini les réserves d'énergies fossiles comme étant les accumulations identifiées qui peuvent être extraites de façon rentable avec les techniques d'aujourd'hui et sous les conditions économiques actuelles. Ainsi, le terme souvent employé de "réserves récupérables" est un pléonasme, puisque les hydrocarbures identifiés comme réserves sont (à quelques nuances près) destinés à être produits et économiquement rentables.

[5] on peut d'ailleurs s'interroger sur la légitimité de cette modélisation. L'utilisation de lois triangulaires peut sembler plausible pour la saturation et la porosité, mais cela l'est beaucoup moins pour le volume de roche. En effet, la loi triangulaire stipule que les petits gisements ont une probabilité d'existence qui décroit avec leur volume. Ceci peut sembler contraire à l'intuition dans la mesure où, nous le verrons, certains experts s'accordent à penser que les pièges les plus petits sont les plus nombreux, c'est-à-dire qu'il y a potentiellement bien plus de "gouttes d'huile" dans le sous-sol que de champs géants !

Concept politique et technico-économique

Sont désignées comme ressources toutes les quantités en place dans la croûte terrestre, identifiées ou non. La première étape est l'identification de ces ressources, donc l'exploration, afin de découvrir d'éventuels gisements. L'exploration impose deux limites. La première est d'ordre politique, car certaines zones géographiques ne sont que partiellement ouvertes à l'exploration par les États qui les contrôlent. La seconde est technique, car il existe des zones où les méthodes géologiques et géophysiques (pour l'offshore ultra-profond par exemple) demeurent encore insuffisantes. Mais pour passer du stade de ressources à celui de réserves, il existe une troisième barrière, technico-économique qui tient aux contraintes de production. En effet, il existe de nombreux gisements d'hydrocarbures que l'on ne sait pas aujourd'hui mettre techniquement en production. Ces gisements, bien que parfaitement identifiés, peuvent être trop profonds dans le cas de l'offshore, ou renfermer des pétroles bruts que nous ne sommes pas en mesure de récupérer totalement en raison de leur trop grande viscosité par exemple.

La technique ne constitue donc pas la seule barrière à la transformation des ressources en réserves. Que dire en effet d'un gisement pour lequel on dispose d'une technique de récupération, mais dont le coût est supérieur au produit de la vente des hydrocarbures qui en sont extraits ? Ou, ce qui est équivalent d'un point de vue bilan énergétique global, dont l'énergie nécessaire à la production se révèle supérieure à l'énergie qu'ils peuvent fournir ? Ce gisement ne sera pas rentable économiquement, et ne sera donc pas mis en production. Le passage des ressources aux réserves peut donc être résumé par la figure 1.2.

Par ailleurs, pour les compagnies comme pour les pays producteurs, les réserves présentent un important intérêt, autant stratégique que politique. Les chiffres de réserves font donc souvent l'objet d'effets d'annonce dont la rigueur peut parfois être sujette à caution :

- les grands gisements sont souvent sous-évaluées par les géologues à leur découvertes car ces derniers n'ont pas besoin d'estimations fines pour assurer la rentabilité du projet de développement. Ils se réservent ainsi la possibilité de réévaluations positives dans le futur, toujours accueillies favorablement par les compagnies ;
- à l'inverse, les gisements marginaux sont pour beaucoup surévalués au moment de leur découverte. Dans la majeure partie des cas, malgré sa bonne foi, l'optimisme du géologue découvreur est à mettre en cause. Il n'est cependant pas si rare de voir des petits champs surévalués afin d'en justifier plus aisément la mise en développement ;
- citons aussi le cas de l'OPEP qui en 1986 changea sa définition des réserves et augmenta ainsi très artificiellement de presque 20 % le montant des réserves mondiales !

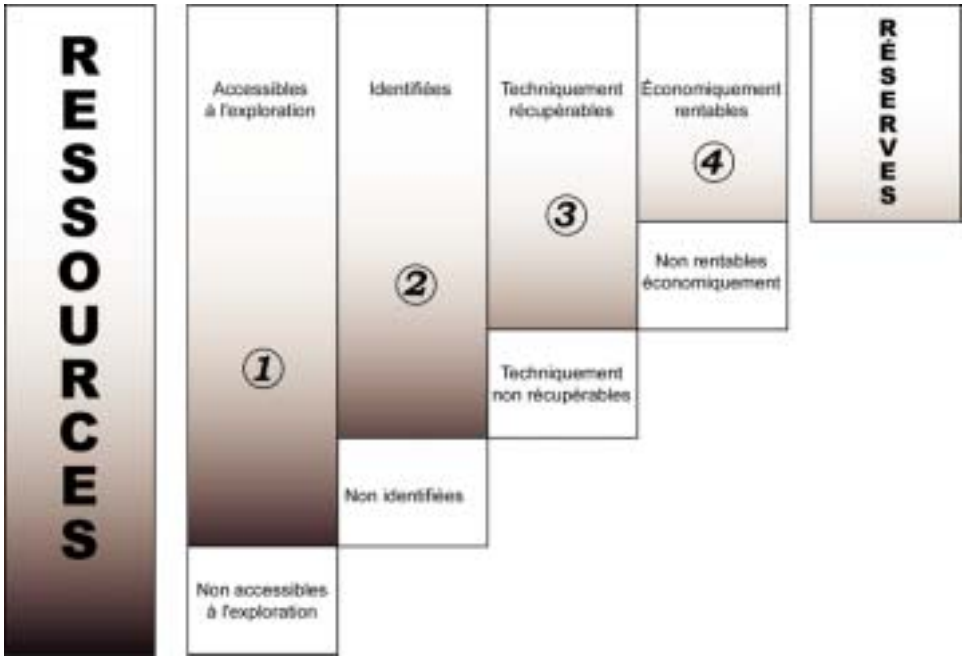

FIG. 1.2 – Des ressources aux réserves (avec l'aimable autorisation de Jean-Noël Boulard) [15].

Le flou règne donc et reste entretenu sur la définition quantitative et l'utilisation rigoureuse du terme de réserves. Un cadre précis de définition existe cependant, nous le détaillons dans la suite.

Réserves Déterministes et probabilistes

Comme cela vient d'être évoqué, on parle de réserves pour des hydrocarbures qui peuvent être mis en production à court ou moyen terme. Les réserves sont donc des volumes hypothétiques, puisque soumis aux incertitudes diverses des modifications technologiques, de la conjoncture économique, etc. Les seules réserves que l'on connaisse de façon certaine, ou déterministe, sont celles que l'on a déjà produites - on a d'ailleurs coutume de dire que l'on ne connaît les réserves d'un champ que lorsque l'on ferme son exploitation. Les approches déterministes reposent sur le principe que les valeurs de chaque paramètre nécessaire à l'estimation des réserves sont supposées certaines. Elles conduisent à une estimation supposée totalement fiable, ignorant toute marge d'erreur. Une approche probabiliste fournit, quant à elle, des indicateurs aléatoires, inscrits dans des fourchettes, ou en termes statistiques, d'intervalles de confiance ou, plus précisément, d'intervalles de prédiction.

La partie 1.1.3 a décrit les incertitudes auxquelles sont soumises les mesures des caractéristiques d'un gisement. Cette approche conduit à une probabilité

qu'un prospect, ou objectif géologique (c'est-à-dire une zone géographique susceptible de receler des réserves) contienne des hydrocarbures.

Un gisement étant déclaré contenir des hydrocarbures, il faut ensuite évaluer les quantités en place (ces données ne sont que rarement publiées), puis estimer les réserves associées. A ce titre, on cherchera donc à évaluer les quantités récupérables par rapport aux quantités en place, il s'agit de la notion de taux de récupération. Les techniques modernes de géosciences (géologie, géophysique, géochimie et géostatistique) permettent de décrire les réserves potentielles du champ comme une distribution de probabilité. Dans la pratique, on estimera les réserves en fournissant des paramètres (comme la moyenne ou certains fractiles de la distribution : 10%, 50%, 90%, etc.) de la loi LogNormale dont on suppose que taille du champ est une observation.

Réserves P90, P50, P10, etc.

On définit par Px la valeur de réserves que le champ a x% de chances de dépasser. Le Px est donc le fractile d'ordre $(1-x)$% de la distribution de la taille du champ. Par exemple, si un champ a un P10 de 1 Gbep, alors il y a 10% de chances que la taille réelle des réserves du champ dépasse 1 Gbep. Le P50 est par définition la médiane de la distribution. On a donc *a priori* une probabilité égale que les vraies réserves d'un champ soient supérieures ou inférieures au P50.

Les valeurs que l'on retrouve très fréquemment dans l'estimation des tailles des champs sont les P95, P90, P50, P10 et P5. On peut aussi trouver des estimations sous la forme "mini, mode, maxi", ou "mini, moyenne, maxi". Les mini et maxi sont en fait des P5 et P95, ou P10 et P90[6]. Le mode (ou maximum de vraisemblance) est la valeur théorique la plus probable de la distribution. La moyenne (ou espérance) est la valeur moyenne des réserves que l'on observerait sur un grand nombre de champs présentant a priori exactement la même distribution de taille des réserves. La figure 1.3 montre la fonction de queue de répartition LogNormale typique pour un champ dont le P50 est fixé à 500 Mbep. Pour une probabilité de x% donnée, la courbe fournit la taille telle que l'on a x% de chances que la taille réelle du champ soit supérieure à cette valeur.

Ces notations-définitions de réserves sont les plus rigoureuses et précises qui existent. Cependant, on en rencontre bien d'autres dans la littérature de l'industrie pétrolière. Détaillons les dans la suite.

[6] il s'agit là bien entendu de dénominations erronées car les minimum et maximum de la loi LogNormale sont 0 et $+\infty$.

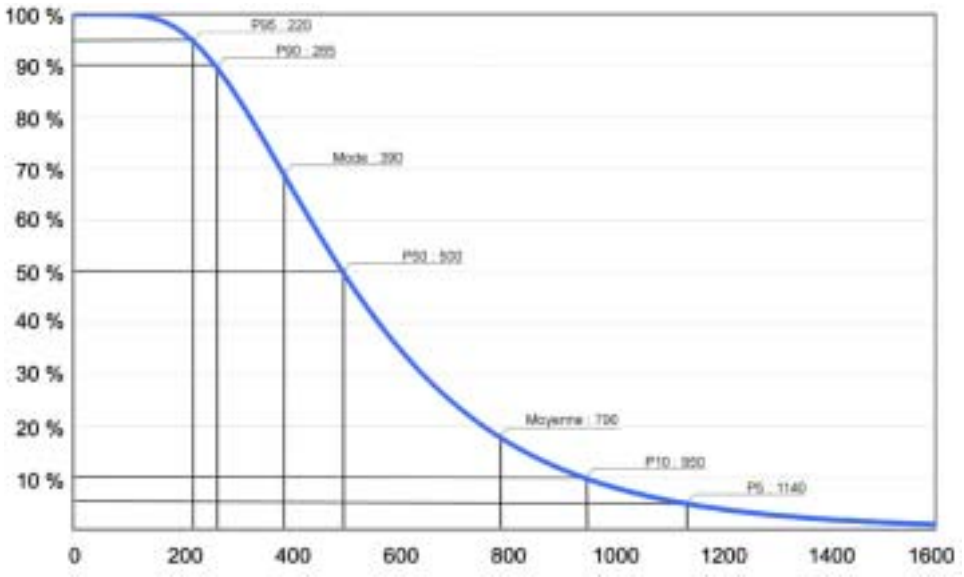

FIG. 1.3 – Fonction de queue de répartition d'une loi LogNormale de médiane 500 Mb.

Réserves 1P, 2P et 3P

Sont largement utilisées également les trois valeurs notées 1P, 2P et 3P, dérivées des précédentes et qui évaluent également de manière probabiliste le potentiel de réserves d'un champ. Ces valeurs correspondent à des Px, variables selon les compagnies ou les auteurs :
– le 1P est généralement égal au P90 ou au P95 précédemment défini ;
– le 2P est toujours égal au P50 ;
– le 3P est généralement égal au P10 ou au P5.

Enfin, nous présentons dans la section qui suit une autre terminologie, elle aussi très usitée équivalente à celle que nous venons d'exposer.

Réserves prouvées, probables et possibles ?

Les trois termes "prouvées", "probables" et "possibles" correspondent – le plus souvent, mais les exceptions sont légions, voir par exemple Masseron [60] dont les définitions sont légèrement différentes – aux valeurs respectives 1P, (2P - 1P) et (3P - 2P). Ce qui signifie, en inversant le système, que l'on a :
– 1P = prouvées ;
– 2P = prouvées + probables ;
– 3P = prouvées + probables + possibles.

La figure 1.4 résume schématiquement ces diverses définitions.

Il est à noter que ces définitions ont été établies et officialisées en 1997 par la SPE (Society of Petroleum Engineers) au cours du World Petroleum Congress [1]. Plus "précisément", on définit par prouvées des réserves qui ont une chance raisonnable d'être mises à jour ; puis, par "chance raisonnable", on considère en réalité un P90. Ces définitions sont pourtant loin d'être universellement adoptées et restent contestées par nombre d'acteurs du monde pétrolier : lorsque des chiffres de réserves sont avancés, ce sont, la plupart du temps, des chiffres de réserves dites prouvées. Cependant, s'agit-il d'un P95, d'un P90 ou autre ? Il est presque systématiquement impossible de répondre à cette question et parfois, le flou est volontaire de la part des auteurs. Encore aujourd'hui, on peut trouver des chiffres de réserves prouvées qui correspondent à des Px allant de P50 à P98 sans mention explicite ! Il convient donc de rappeler quelques règles de sécurité relatives à l'utilisation des divers chiffres de réserves que l'on peut rencontrer.

Mises en garde quant à l'utilisation des définitions

La vision probabiliste de l'estimation des réserves d'un champ est aujourd'hui courante. Cependant, elle n'est pas sans risque dans son utilisation. Par exemple, l'addition des réserves en vue de connaître le potentiel d'un bassin ou d'un pays n'est pas chose aussi aisée qu'il peut paraître. En effet, la somme des Px (ou des modes) des réserves de deux champs n'est pas, en général, le Px de la somme des réserves des deux champs. Pour mémoire, il convient de retenir que : sommer les réserves 1P (prouvées) des champs d'un bassin tend à sous-estimer les réserves 1P du bassin entier, et sommer les réserves 3P (prouvées + probables + possibles) des champs d'un bassin tend à surestimer les réserves 3P du bassin entier. En ce qui concerne les 2P, tous les cas, sous ou surestimation, sont possibles.

Les seules estimations qui peuvent être légitimement sommées (d'un point de vue statistique) sont les estimations d'espérances, car l'espérance est un opérateur linéaire. Cependant, il est important d'avoir à l'esprit que la moyenne ne s'avère un outil d'estimation efficace qu'à très grande échelle : dans l'exemple de la figure 1.3, l'espérance de la distribution n'est en réalité atteinte que dans moins de 20% des cas. Concrètement, cela signifie que si plusieurs champs possèdent cette distribution de taille, moins de 20% d'entre eux auront une taille réelle supérieure à la moyenne de la distribution ! Pourtant, il faut noter que la somme des réserves réelles des champs sera, pour sa part, proche du nombre de champs multiplié par l'espérance de la distribution. Ce faux paradoxe résulte de la Loi des Grands Nombres, qui stipule qu'asymptotiquement, les erreurs, ou écarts à la moyenne tendent à se compenser, c'est-à-dire que la moyenne des écarts à l'espérance tend vers zéro. Disposer d'un indicateur de dispersion de la distribution est donc aussi crucial que de savoir estimer l'espérance.

Il convient donc de conserver la plus grande prudence lorsque l'on effectue des calculs sur les réserves. Il est pourtant indispensable de sommer des estimations de réserves à l'échelle des champs pour obtenir des ordres de grandeur de réserves à un niveau plus macroscopique (région, pays ou encore ensemble des champs dans lesquels une compagnie possède des intérêts). La plupart du temps, seules les réserves prouvées sont publiées. Elles fournissent donc la seule matière dont on dispose pour la réalisation des études statistiques. Si les sommer ne s'avère pas une procédure correcte d'un point de vue mathématique, il n'est pourtant guère possible, bien souvent, de procéder autrement. Reste à être conscient du type d'erreur que l'on commet de façon systématique et à nuancer en conséquence le résultat final obtenu.

1.1.5 Caractéristiques des réserves

On distingue traditionnellement deux types d'hydrocarbures : *conventionnels* et *non conventionnels*[7].

Dans ce domaine également il n'existe pas de définition claire et précise de ce qui est conventionnel par opposition à ce qui ne l'est pas. On peut classer la qualité des pétroles grâce à plusieurs indices tels que la densité (mesurée en degrés API), la viscosité et leur teneur en soufre. En ce qui concerne le gaz, on parle moins de critères de qualité (capacité calorifique, teneur en soufre ou gaz inertes comme le CO_2, etc.) que de différence d'origine. Ainsi, on distingue le gaz associé au pétrole ou à des condensats, et les gaz dits secs (ces derniers représentant environ les deux tiers des réserves mondiales actuelles de gaz). Le fait qu'un gaz soit considéré comme conventionnel ou non dépend essentiellement de sa difficulté d'extraction et de mise en production.

Campbell, Perrodon et Laherrère [19], considèrent comme hydrocarbures conventionnels les hydrocarbures qui peuvent être produits dans les conditions techniques et économiques actuelles et prévisibles pour le futur. Cette définition, très proche dans sa formulation de celle des réserves prouvées de Mc Kelvey [61], intègre cependant les progrès technologiques et le contexte économique futur. Ainsi, par opposition, on peut dire que les hydrocarbures non conventionnels sont, de façon simpliste, ceux qui sont difficiles et coûteux à produire.

Cependant, il est extrêmement difficile d'appréhender ce que seront les conditions techniques et économiques dans l'avenir. En effet, on peut *a posteriori* mesurer l'impact d'une nouvelle technologie sur l'extraction des réserves, mais comment prévoir ce que sera la technologie dans 20 ans ?

L'exemple typique en la matière est l'offshore profond. À la fin des années soixante-dix, tous les gisements situés à une profondeur supérieure à 200 m étaient considérés comme non conventionnels (et donc, non comptabilisés

[7] Ces deux termes sont des anglicismes.

dans les estimations de réserves d'hydrocarbures conventionnels). La technologie de l'époque ne permettait pas de disposer pas de moyens suffisamment efficaces pour produire ces gisements de façon rentable. Aujourd'hui, il est courant d'exploiter des gisements cinq à dix fois plus profonds, c'est-à-dire, par 2000 m de profondeur d'eau. La barrière entre le conventionnel et le non conventionnel a donc nettement reculé au cours du temps.

Autre exemple : celui des pétroles lourds et extralourds. Le bassin de l'Orénoque au Venezuela, connu depuis les années trente, contient de l'huile extralourde (8 à 10 degrés API). En 1967, une première évaluation des quantités en place a conduit à une évaluation de 693 Gbep (soit l'équivalent de plus de la moitié des réserves prouvées mondiales de pétrole conventionnel). En 1967, on pouvait annoncer : ressources 693 Gbep et réserves 0 ! En 1983, une nouvelle évaluation a conduit à un volume de ressources de 1200 Gbep et des réserves (qui ne sont pas encore, à proprement parler, considérées comme réserves prouvées) de l'ordre de 100 à 300 Gbep, grâce notamment aux progrès accomplis en matière de forage horizontal.

Ici encore, la limite entre conventionnel et non conventionnel tend à se déplacer, au cours du temps, vers des hydrocarbures bruts de plus en plus difficiles à produire, en termes de conditions de production, de localisation et bien sûr de qualité. Il convient cependant de nuancer ce propos en se référant aux aspects géopolitiques de l'économie pétrolière. En effet, les pétroles du Moyen-Orient, par exemple, sont peu onéreux à produire et existent en très grande quantité. Une des conséquences des chocs pétroliers de 1973 et 1979 a été de permettre la découverte et la production rentable, dans le monde entier, de pétroles plus difficilement accessible. Le pétrole le moins cher à produire n'est donc plus nécessairement aujourd'hui le seul ou le premier à être exploité, comme évoqué dans [22], où est menée une étude rigoureuse des déterminants des coûts dans l'industrie pétrolière amont.

Les hydrocarbures non conventionnels sont donc les réserves du futur. Le déplacement de la limite traduit ce que certains auteurs appellent le continuum du carbone fossile (voir notamment Bauquis [11]). Lorsque les réserves d'un certain type d'hydrocarbures que l'on sait produire s'épuisent, on cherchera à développer la production de nouveaux types, dont les non conventionnels. Peu à peu, progrès technologiques et incitations politiques aidant, ces hydrocarbures deviendront plus courants à produire, donc conventionnels ou "conventionnalisés". Ainsi, mêmes s'ils coexistent évidemment, nous passons peu à peu des pétroles faciles à produire issus des États-Unis, d'Algérie ou du Moyen Orient, aux pétroles offshore ou lourds, pour aujourd'hui, nous tourner vers des pétroles extralourds et vers l'offshore ultraprofond. En guise d'illustration, citons les deux projets phares du groupe TotalFinaElf dans ces deux domaines pour ce tout début de XXI$^{\text{ème}}$ siècle :

– le développement du champ d'huile Girassol, à 150 km au large des côtes de l'Angola par 1350 m de fond ;

– le projet Sincore au Vénezuela qui vise au développement des huiles extralourdes de la ceinture de l'Orénoque.

1.1.6 Réserves ultimes

Le terme de réserves ultimes représente l'ensemble des hydrocarbures à produire passés, présents et à venir. La fraction "passée" des réserves ultimes est donc constituée de l'ensemble des hydrocarbures produits depuis les débuts de leur exploitation. La partie "présents" représente l'ensemble des réserves connues à ce jour. La fraction "à venir" représente l'ensemble des réserves restant à découvrir. Ces réserves restant à découvrir provenant :
– soit de réévaluations de champs déjà connus ;
– soit de l'amélioration technique des taux de récupération due aux progrès technologiques[8] ;
– soit de champs dont le développement est lié à économie favorable, et typiquement à un prix du brut élevé.
– ou enfin de nouvelles découvertes.
Sur ces quatre points, seul le dernier est le point d'intérêt de notre travail de thèse. Les paragraphes suivant illustrent cependant sommairement chacun d'eux car ils sont fondamentaux pour comprendre ce que recouvre la notion de réserves ultimes.

Réévaluation de champs connus, ou *field growth*

Comme on l'a montré précédemment, la taille d'un champ n'est pas connue mais seulement estimée, au moyen de probabilités *a priori*. Au fur et à mesure de l'exploitation du champ, les paramètres géologiques nécessaires à l'estimation de sa taille sont de mieux en mieux connus. L'intervalle de prédiction a donc tendance à se resserrer vers la vraie valeur, comme l'illustre le schéma de la figure 1.4.

Les chiffres publiés étant généralement ceux des réserves prouvées, conservatifs par définition, on voit ainsi qu'ils ont tendance à croître au cours du temps. L'impact d'une réévaluation est donc globalement positif sur le volume total des réserves ultimes, mais il peut être localement nul voire négatif. Il est difficile à prévoir autrement que par analogies et expérience passée.

Amélioration des taux de récupération et progrès technique

Les taux de récupération sont actuellement de l'ordre (et en moyenne mondiale) de 30% pour les gisements d'huile et de 80% pour les gisements de gaz. Il existe donc une marge de progression importante dans les quantités

[8] le taux de récupération est le ratio *réserves/ressources*, il exprime la quantité d'hydrocarbures que l'on va récupérer sur un gisement par rapport à la quantité en place initialement.

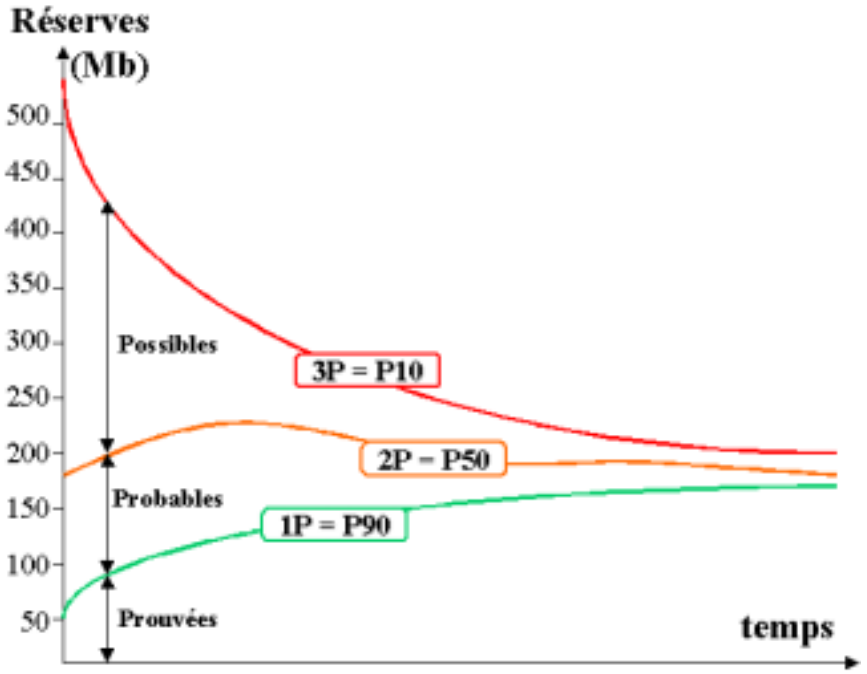

FIG. 1.4 – Variation schématique des paramètres de la distribution de la taille d'un champ d'environ 200 Mbep au cours du temps.

de pétrole récupérable. Les progrès technologiques permettent cette amélioration de la récupération par de nombreuses méthodes (puits multidrains, drains horizontaux, injection d'eau ou de gaz, etc.).

Dans de nombreux cas (le champ géant d'Allwyn en mer du Nord est à ce titre un cas d'école), la mise en place de ces nouvelles technologies contribue à faire passer des ressources au stade de réserves. Il se peut aussi que ces techniques ne fassent qu'accélérer la déplétion du champ sans pour autant créer de nouvelles réserves. Ces deux situations sont illustrées sur les profils de production de la figure 1.5 où les réserves sont représentées l'aire sous la courbe (à un facteur 360 près – nombre de jours de l'année comptable).

Il faut cependant noter que l'accélération de la production présente malgré tout un avantage pour la compagnie pétrolière qui anticipe ainsi ses revenus et augmente ainsi mécaniquement la valeur actuelle nette du projet. Pour une approche très détaillée des aspects rentabilité des projets d'investissement, on pourra se reporter avec profit à Babusiaux [8].

L'impact du progrès technique est donc positif ou nul. Il difficilement prévisible car les progrès se font souvent par sauts technologiques totalement irréguliers[9].

[9]Citons comme exceptions la résolution de la sismique 2 ou 3D qui augmente régu-

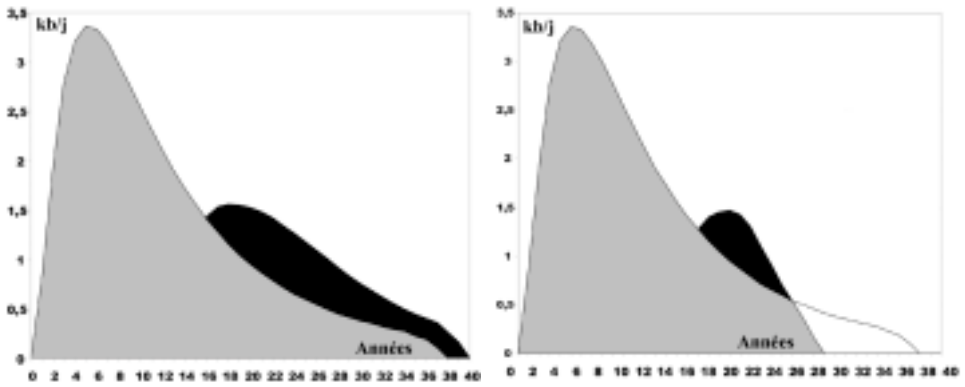

FIG. 1.5 – Variantes de l'effet de récupération assistée sur un profil de production (production journalière en fonction de l'année) : à gauche il y a création de réserves (surface noire), à droite il y a accélération de la déplétion (surface noire = surface blanche).

Impact de l'économie

Il est absolument déterminant (voir Masseron [60] pour une mise en évidence très détaillée). Le prix du baril de brut conditionne fortement l'accès à certaines ressources minières à la rentabilité marginale : un prix du pétrole élevé permet d'aller exploiter de façon rentable des gisements dont l'extraction des hydrocarbures est délicate, comme les extra-lourds ou l'offshore ultra-profond par exemple.

Entre les deux chocs pétroliers de 1973 et 1979 par exemple, le pétrole passe schématiquement de 10 à 30 $/b et permet d'aller explorer puis exploiter les gisements de mer du Nord qui devenaient rentables à partir d'un prix du brut de 15 à 20 $/b. Les progrès techniques et économies d'échelle permettent aujourd'hui d'avoir des coûts de production en mer du Nord d'environ 5 à 6 $/b. L'économie a donc eu un impact considérable sur l'augmentation des réserves puisque sans choc pétrolier, les ressources de mer du Nord n'auraient probablement été mises au jour que bien plus tard[10].

L'impact de l'économie, essentiellement au travers du prix du baril, est extrêmement difficile à prévoir. Il est clair qu'il peut être positif ou négatif suivant que les coûts d'extraction par baril de tel ou tel champ sont supérieurs ou inférieurs au prix du brut, qui est extrêmement volatil.

---

lièrement avec les capacités de calcul des ordinateurs ; la croissance de l'efficacité des catalyseurs ; ou encore la réduction des coûts dans la chaîne GNL.

[10] Même si son lancement est bien antérieur aux chocs pétroliers, il est clair que ces derniers ont permis d'asseoir la politique énergétique Française de recours à l'électricité nucléaire. L'économie a donc ici joué le rôle de catalyseur de substitution énergétique.

Nouvelles découvertes

C'est sur ce dernier point exclusivement que va se porter tout notre travail à venir. Nous allons chercher à estimer combien de champs restent encore à découvrir et quelles sont leurs tailles.

Il est clair que de nouveaux champs découverts créent des réserves qui doivent être incluses aux évaluations de réserves ultimes.

Réserves ultimes : conclusion

Les quatre points ci-dessus peuvent être résumés par le tableau suivant qui présente les impacts positifs ou négatifs de chacun d'entre eux, ainsi que leur prévisibilité (sur laquelle nous reviendrons partiellement en 1.1.8) :

Impact de...	réévaluations	progrès technologiques	l'économie	nouvelles découvertes
Positif	+ + +	+ +	+ +	+ + +
Négatif	–		– –	
Prévisibilité	faible	quasi-nulle	ardue	possible

TAB. 1.1 – Les déterminants de l'évolution des réserves ultimes.

Le chiffre véritable mais inconnu des réserves ultimes est sensé être un chiffre stable au cours du temps. En pratique, les réserves ultimes sont bien entendu seulement estimées et dépendent largement de la définition de réserves que l'on se donne [11]. Ces estimations sont, pour leur part, hautement instables au cours du temps, comme le sont celles des réserves prouvées et des réserves restant à découvrir. Ce chiffre fait, la plupart du temps, plutôt l'objet d'un consensus d'experts [67]. Au niveau mondial, il est communément admis que les réserves ultimes de pétrole sont de l'ordre de 2800 Gb et se répartissent comme sur la figure 1.6.

Nous discutons maintenant des chiffres que nous utilisons pour notre étude statistique concernant le nombre de champs restant à découvrir.

[11] On peut donc en particulier distinguer les réserves ultimes d'hydrocarbures conventionnels des non-conventionnels, mais cela nécessite de prendre des libertés avec les définitions données dans ce chapitre. Disons que les professionnels s'y retrouvent généralement !...

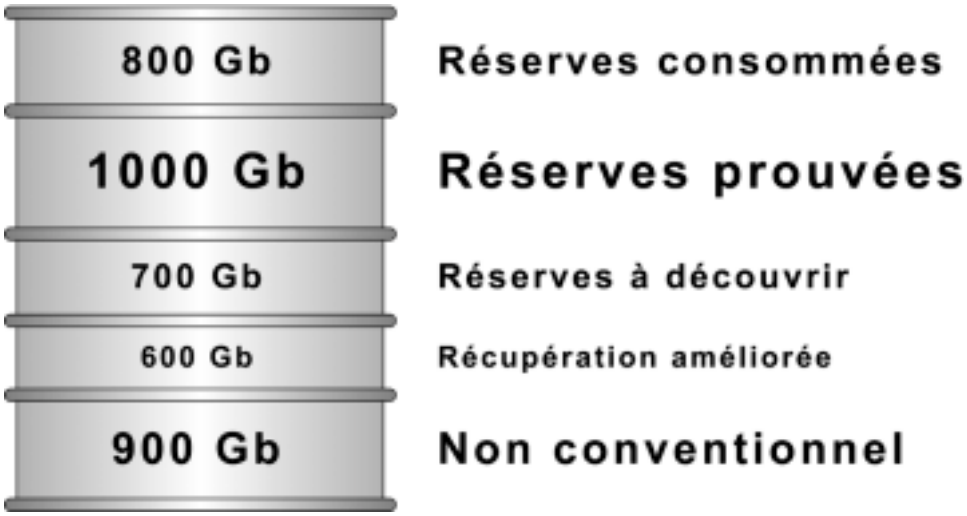

FIG. 1.6 – Constitution des réserves ultimes (consensus du WPC 1997 [1]).

1.1.7 Quels chiffres utilisons-nous ?

Dans toute notre thèse, nous nous focalisons exclusivement sur les *hydrocarbures conventionnels* et utilisons pour nos études les chiffres de *réserves prouvées* fournis par la base de données gisements Pétroconsultant 1998. Nous utilisons les chiffres du pétrole et du gaz ramenés à l'unité volumique homogène du "bep" pour "baril équivalent pétrole". Dans notre travail, il n'y a en effet pas lieu de distinguer pétrole et gaz dans un même champ puisque ce sont des molécules d'hydrocarbures qui sont issues du même kérogène, ont migré et ont été piégées à l'identique. Les conditions de pression et température seules ont fait que certaines de ces molécules se sont craquées pour donner lieu à du gaz plutôt que du pétrole.

Notre objectif est de présenter un protocole original d'estimation de réserves ultimes d'hydrocarbures présentant le même type de carcatéristiques que ceux qui sont de commercialisation courante et utilisés de façon pratique aujourd'hui. Ceci nous place donc d'emblée dans le cadre des hydrocarbures conventionnels[12]. Plus précisément, nous cherchons une estimation du nombre de champs restant à découvrir ainsi que de leurs tailles. Par ailleurs, les ressources d'hydrocarbures non-conventionnels sont à l'heure actuelle et à quelques exceptions près, beaucoup trop spéculatives et ne présentent, qu'un intérêt économique marginal, même si leur intérêt prospectif est évident.

Un autre point à discuter est celui du choix de la définition de réserves pertinente pour notre travail : celle des ressources (où, rappelons le, des

[12]Comme justifié précédemment, on ne peut d'ailleurs pas à proprement parler de réserves d'hydrocarbures non-conventionnels. Voir plus précisément la section 1.1.5 pour une discussion de ce sujet.

quantités en place dans les gisements), ou celle des réserves, c'est-à-dire celui d'une quantité technico-économique représentant les quantités destinées à être produites pour les besoins en énergie fossile des hommes à court ou moyen terme.

D'un point de vue physique, ou naturaliste, il est clair qu'il aurait été légitime de travailler avec les chiffres des quantités en place. En effet, comme on l'a expliqué précédemment, ils sont relatifs à des critères géologiques tangibles donc sensés être parfaitement stables au cours du temps puisqu'issus de processus physiques dont l'évolution ne se mesure que sur l'échelle des temps géologiques. Pourtant, l'expérience montre que les évaluations de quantités en place, lorsqu'elles sont publiées, ce qui est plutôt rare, sont très instables au cours du temps, subissant réévaluation sur réévaluation. De plus, ces chiffres n'ont pas de réalité économique puisqu'une fraction variable seulement de leur montant sera produite et valorisée. Les chiffres des ressources rencontrent donc les écueils de l'accessibilité très délicate aux données, d'une fiabilité douteuse suite aux nombreuses réévaluations dans le temps et enfin d'une interprétation économique impossible.

Pour notre étude, il est indispensable d'avoir recours à des données relativement fiables et suffisamment abondantes. On peut alors tabler directement sur les chiffres des réserves prouvées, conformément aux définitions données en 1.1.4. Ces chiffres présentent certains avantages :
- ils sont conservatifs, puisqu'ils représentent des valeurs auxquelles les experts estiment que les réserves véritables[13] ont 90 % de chances d'être supérieures ;
- ils ont une forte réalité économique puisqu'ils servent de base dans l'industrie pour l'évaluation des projets d'exploration (choix d'investissement pour la compagnie, accords de prêts pour les banques, calculs de rentabilité économique de la production, etc.)[14]. ;
- enfin, aspect plus qu'important, ce sont les chiffres les plus publiés par l'ensemble des organismes de collecte de données statistiques de l'industrie pétrolière.

Les chiffres correspondant aux autres définitions de réserves sont extrêmement rarement publiés et constituent surtout des données qui restent internes aux compagnies.

Le fait de travailler avec les réserves prouvées à une date donnée ne nous affranchit pas pour autant de certains biais dont on pourrait penser qu'il serait pertinent de les prendre en compte pour nos prévisions.

[13] C'est-à-dire les quantités qui auront été produites à la fin de l'exploitation du gisement.

[14] De plus, ce sont les valeurs que les sociétés pétrolières cotées à la bourse de New York doivent déclarer en fin d'année pour leur bilan d'exercice à la SEC (*Security Exchange Commission*, équivalent américain de la Commission des Opérations en Bourse française).

1.1.8 Peut-on tenir compte d'un biais systématique ?

Il faut être bien conscient des biais systématiques auxquels nous sommes confrontés lorsque l'on regarde l'évolution au cours du temps des chiffres avec lesquels nous travaillons. Il conviendra donc de nuancer les résultats que nous obtenons en travaillant avec le chiffre des réserves prouvées 1998 :

- en premier lieu et comme on l'a déjà fait remarquer en 1.1.4, les réserves prouvées sont des fractiles d'ordre 10 % des distributions *a priori* des tailles des champs. Or la somme des fractiles d'ordre 10 % tend à sous-estimer le fractile d'ordre 10 % de la distribution de la somme.
- les évaluations de réserves prouvées ont tendance à croître au cours du temps du fait de la meilleure connaissance géologique des gisements, comme vu en 1.1.6[15].

Le premier de ces biais pourrait être pris en compte si l'on connaissait les lois exactes des tailles de tous les champs de l'échantillon de travail. Mais les simulations de ces lois ne sont jamais accessibles. Il est donc statistiquement impossible d'en tenir compte. Le second n'est pas plus facile à intégrer... En effet, la meilleure connaissance des caractéristiques géologiques du sous-sol est principalement liée aux progrès technologiques. Il est extrêmement périlleux de tenter d'intégrer une composante de "dérive temporelle" des estimations due aux progrès techniques car comme nous l'avons vu en 1.1.6, un grand nombre de progrès techniques se font par sauts difficilement prévisibles. Citons malgré tout un exemple d'utilisation d'une composante de progrès technologique dans une approche économétrique de l'évolution du taux de succès dans l'exploration[16] dans Forbes *et al.* [31]. Ce travail présente, à notre connaissance, un des seuls modèles quantitatifs lié à l'industrie pétrolière incluant le progrès technologique comme variable explicative[17].

Notre choix s'étant porté sur le chiffre des réserves prouvées d'hydrocarbures, pétrole et gaz ramenés à l'unité commune du baril équivalent pétrole, le "bep", venons-en maintenant au problème de l'estimation des réserves ultimes à une l'échelle d'un bassin.

---

[15] Quelques auteurs, qualifiés de pessimistes, dont Campbell et Laherrère (voir par exemple [18] et [48]), ont mis en évidence le fait que sur de nombreux champs, la majeure partie de la "paternité" du renouvellement des réserves est due à ce phénomène de réévaluation.

[16] Ce taux s'exprime comme le rapport entre le nombre de puits d'explorations non-secs (c'est-à-dire montrant des traces d'hydrocarbures) forés et le nombre total de puits d'exploration forés.

[17] En pratique, le modèle économétrique présenté fait intervenir le progrès technologique comme une variable *dummy* pour chaque année sur laquelle court l'estimation à l'intérieur d'un modèle préétabli. Le principal doute que nous formulons face à cette approche est qu'une telle représentation fait apparaître le progrès technique de l'année i comme un évènement exceptionnel et non un évènement pérenne, ce qui est contraire à la réalité.

1.2 Principes d'estimation des réserves ultimes à un l'échelle d'un bassin

Dans cette section nous rentrons au cœur de notre sujet. Nous allons aborder l'estimation des réserves à une échelle macroscopique[18] suffisante pour comporter un grand nombre de champs (typiquement on peut penser à l'échelle du bassin sédimentaire), comparée à l'échelle microscopique du gisement. Plus précisément, par estimation des réserves, nous cherchons à estimer le nombre total de champs restant à découvrir et le montant de réserves que ces champs sont susceptibles de receler.

Historiquement, les diverses approches statistiques du problème ont, pour la grande majorité d'entre elles, consisté à tenter de trouver une loi de probabilité représentant correctement la distribution inconnue de l'ensemble des champs présents dans le sous-sol.

Nous commençons par présenter et commenter les diverses lois utilisées dans la littérature et l'industrie de façon récurrente. Nous nous intéressons ensuite à un protocole d'estimation particulier. Nous détaillons les problèmes que pose ce protocole, notamment vis-à-vis de la robustesse des estimations qu'il implique. Cependant, nous montrons comment en améliorant le modèle sous-jacent (prenant en compte le "biais d'exploration") nous construisons un modèle plus fidèle à la réalité et à la dynamique de l'exploration.

A partir de cet endroit, nous parlerons de "taille d'un champ" pour désigner le chiffre de ses réserves. En pratique, nos données sont les chiffres des réserves prouvées.

1.2.1 Les diverses lois utilisées

A une échelle macroscopique, on peut penser à deux façons de faire intervenir une loi de probabilité pour modéliser la distribution de l'ensemble des champs du sous-sol l'échantillon :

- on interprète les réserves des champs existant dans le sous-sol comme un ensemble de variables aléatoires non nécessairement identiquement distribuées. Cette vision est consistante avec celle qui consiste à modéliser la distribution des réserves d'un gisement par une loi de probabilité (LogNormale par exemple, comme dans la section 1.1.3) dépendant de paramètres géologiques locaux propres à chaque gisement ;
- on interprète les réserves des champs comme un ensemble de données déterministes sur lequel on va tenter d'ajuster une loi de probabilité. Cela revient à dire que l'on interprète cet ensemble de données comme la réalisation d'un échantillon d'une seule et même variable aléatoire de loi inconnue dépendant de paramètres géologiques macroscopiques.

[18] Le choix de l'échelle adéquate à considérer est un sujet que nous discuterons en détail au début du chapitre 2.

En pratique, on opte très clairement pour la seconde interprétation car les données relatives aux lois utilisées pour décrire la distribution de la taille de chaque champ ne sont pour ainsi dire jamais accessibles dans leur totalité, et quand bien même elles le seraient, le nombre de paramètres à estimer serait beaucoup trop important. Quelles sont alors les diverses lois à envisager pour la distribution de la taille des champs ?

La loi LogNormale

Historiquement (L'article de 1963 de Kaufman [42] est la référence la plus ancienne que nous ayons pu trouver à ce sujet), la première des lois considérée pour étudier la distribution des tailles des champs est, encore une fois, la loi LogNormale. L'argument "produit de variables aléatoires" que nous avions donné en 1.1.3 ne tient plus ici. De nombreuses références, à commencer par Kaufman [42], mais aussi de plus récentes comme Lee at Wang [51] ou encore Laherrère [45] et Sun et Woodroofe [77], ne mentionnent pour justification que la bonne adéquation de la loi LogNormale, pour des paramètres bien choisis, aux données observées sur des zones très matures, c'est-à-dire proches de l'exhaustion.

Au sein de l'industrie, on trouve une justification bien plus empirique, consistant à dire qu'en pratique, et en dépit du bon vieux Théorème Central Limite, la somme de lois LogNormales a tendance à être LogNormale. Bien entendu, ceci est tout à fait faux mais la simulation de la figure 1.7 illustre ce propos : on représente sur cette figure la distribution obtenue par la somme de 30 1000-échantillons d'une loi LogNormale de paramètres (2,2). Le résultat ressemble clairement à une loi LogNormale comme l'indiquent les histogrammes de la distribution et de la log-distribution. Cependant tout test, paramétrique ou non, rejette assez fortement l'hypothèse de LogNormalité. Il n'en reste pas moins que la distribution observée est bien plus proche d'une loi LogNormale que d'une loi Normale.

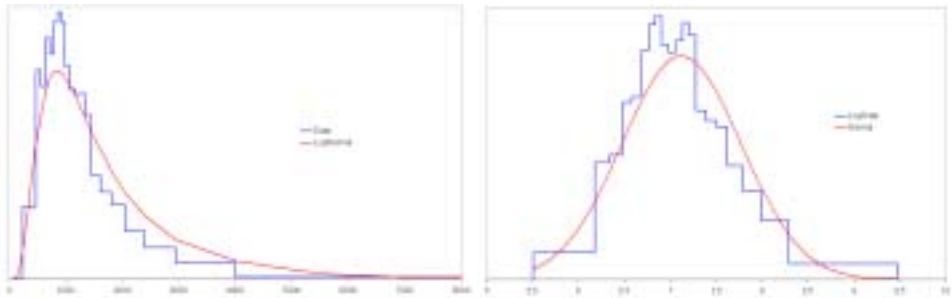

FIG. 1.7 – Distributions empiriques de la somme de 30 fois 1000-échantillons de loi LogNormales(2,2) et de son logarithme. Figurent les estimations par maximum de vraisemblance des lois LogNormales et Normales associées.

Ainsi, la loi LogNormale se trouve-t-elle "légitimée" pour modéliser les distributions des réserves tant au niveau microscopique que macroscopique. Les exemples qui suivent montrent en effet une certaine adéquation. Cependant, on peut voir qu'au cours du temps, la moyenne et le mode des lois LogNormales que l'on peut ajuster au mieux sur les données décroissent très significativement, rendant l'existence de champs de petite taille de plus en plus probable. À la limite de ce processus, on peut penser qu'au moment de l'exhaustion la densité de probabilité que l'on pourra ajuster au mieux aux données pourra être décroissante, autrement dit, la probabilité d'existence d'un champ est d'autant plus grande que ce champ est petit. Ceci est conforme au bon sens. Il est fort probable qu'il y ait plus de "gouttes d'huile" dans le sous-sol que de champs de plusieurs dizaines de millions de barils. Ce point, sur lequel nous revenons et dont nous donnons une justification en 1.2.3 est détaillé dans Lepez [52]. Ce travail montre que la loi LogNormale est bien adaptée à la modélisation de la distribution des tailles des champs déjà découverts dans une région non mature, mais qu'elle ne l'est pas pour modéliser la distribution des tailles de l'ensemble des champs du sous-sol.

Tenter de modéliser la distribution "ultime" de la taille des champs nécessite donc de bien représenter les petits champs. Une des voies possibles est donc d'utiliser une loi à densité décroissante susceptible de présenter certaines analogies de comportement avec la loi LogNormale.

La loi de Lévy-Pareto ou loi fractale-linéaire

Une autre loi très utilisée dans la modélisation des tailles des champs est la loi de Lévy-Pareto. Depuis quelques années, de nombreux auteurs s'y sont référés concernant les problèmes de modélisation ou d'estimation de réserves, citons notamment Houghton [41], Crovelli *et al.* [25], Perrodon [68] ou encore dernièrement Kontorovich *et al.* [44].

Définition 1.2.1. *on dit qu'une variable aléatoire possède la loi de Lévy-Pareto de paramètre (ou d'exposant) $\alpha > 0$ si elle est l'exponentielle d'une variable aléatoire de loi Exponentielle de paramètre α.*

Le principal intérêt de cette loi du point de vue modélisation est son invariance stochastique par changement d'échelle :

Définition 1.2.2. *On dit qu'une variable aléatoire X à valeurs dans $[\,1\,;+\infty[$ est invariante par changement d'échelle si :*

$$\forall\ a > 1, \quad \frac{1}{a}\, X_{|_{[\,a\,;+\infty[}} \overset{\mathcal{L}oi}{=} X$$

où $X_{|_{[\,a\,;+\infty[}}$ désigne la variable aléatoire X restreinte à l'intervalle $[\,a\,;+\infty[$.

Cette invariance, non pas géométrique mais en "forme de distribution", est un argument avancé par les géologues utilisateurs pionniers de cette loi dans le domaine pétrolier. Or, comme nous le montrons dans la Proposition 1.2.4, la loi de Lévy-Pareto est la seule loi de probabilité continue positive à être invariante par changement d'échelle.

Il est par ailleurs bien connu des explorateurs qu'à côté d'une découverte importante, on a toutes les chances de faire d'autres découvertes plus petites, au voisinage desquelles des structures encore moins importantes risquent de se trouver. C'est le concept de "satellite" bien connu dans toute l'industrie pétrolière et qui va dans le sens de notre idée d'invariance stochastique par changement d'échelle. Notons que les accumulations turbiditiques[19] en particulier formeraient des structures géologiques présentant de façon flagrante ce type d'invariance.

Alors comment justifier l'emploi courant de deux lois aussi différentes que la loi logNormale et la loi de Lévy-Pareto ? Pour mettre en exergue le fait que les deux lois sont empiriquement cousines, Mandelbrodt [56] cite plusieurs exemples d'utilisations inadéquates de la loi LogNormale là où une loi de Lévy-Pareto est mieux adaptée. Deux raisons croisées à cela. En premier lieu, la loi LogNormale se comporte bien souvent[20], en pratique et de façon purement empirique, comme une loi de Lévy-Pareto sur ses grandes observations. Parallèlement, les observations de phénomènes physiques suivant approximativement des lois de Lévy-Pareto sont souvent censurées de leurs petits objets, tout simplement parcequ'ils ne sont pas visibles, pas mesurables. Nous reviendrons sur cette idée fondamentale dans la section 1.2.3.

La loi de Lévy-Pareto fait partie de la classe des lois dites à queues épaisses, ou lois α-stables. Elles ont été largement étudiées en théorie des probabilités car elles apparaissent naturellement comme des lois limites des Théorèmes Centraux Limites de Lévy[21] pour les variables aléatoires ne possédant pas de moments d'ordre 2 ou 1, voir Feller [30] et Araujo et Giné [5].

Lemme 1.2.3. *La fonction de queue de répartition Q et la densité f d'une variable aléatoire X de loi de Lévy-Pareto de paramètre α sont données, pour x réel, par :*

$$\begin{aligned} Q(x) &= \mathbb{P}(X > x) = (x \vee 1)^{-\alpha} \\ f(x) &= \alpha x^{-(\alpha+1)} \mathbb{1}_{[1;+\infty[}. \end{aligned}$$

[19] Éboulements de roches sédimentaires alluviales ; aval d'un delta par exemple. L'Offshore ultra-profond de l'Angola est de ce type.

[20] Plus précisément, dès que la moyenne et la variance de la loi Normale sous-jacente sont suffisamment grandes, disons supérieures respectivement à 3 et 1 pour fixer les idées.

[21] D'où le nom de la loi... Nous justifierons la partie "Pareto" dans la section 2.1.2.

Preuve : soit Y une variable aléatoire de loi Exponentielle de paramètre α et $X = e^Y$. Soit $x > 1$:

$$Q(x) \;=\; \mathbb{P}(X > x) \;=\; \mathbb{P}(Y > \log x) \;=\; e^{-\alpha \log x} \;=\; x^{-\alpha}$$

Par ailleurs, si $x < 1$, il est clair que $Q(x) = 1$.

La forme de la densité f s'obtient évidemment en dérivant Q. ♣

Nous montrons maintenant que la loi de Lévy-Pareto est la seule loi invariante par changement d'échelle.

Proposition 1.2.4. *Soit X une variable aleatoire continue à support dans $[1; +\infty[$. Les deux assertions suivantes sont équivalentes :*

(i) X est invariante par changement d'échelle au sens de la définition 1.2.2 ;
(ii) il existe $\alpha > 0$ tel que X soit de loi de Lévy-Pareto de paramètre α.

Preuve : commençons par montrer le sens $(i) \Rightarrow (ii)$.

Soit $a > 1$ et $x \in \mathbb{R}$:

$$\mathbb{P}\left(X_{|_{[a\,;+\infty[}} > x\right) \;=\; \mathbb{P}(X > x \,|\, X > a) \;=\; \frac{Q(x \vee a)}{Q(a)} \tag{1.1}$$

or par hypothèse,

$$\mathbb{P}(X > x) \;=\; \mathbb{P}\left(\frac{1}{a}\, X_{|_{[a,+\infty[}} > x\right) \;=\; \mathbb{P}\left(X_{|_{[a,+\infty[}} > ax\right).$$

En combinant (1.1) à cette dernière égalité il vient

$$\begin{cases} Q(ax) \;=\; Q(a) \times Q(x) & \forall\, a, x > 1 \\ Q(1) \;=\; 1 \quad \text{et} \quad Q(+\infty) \;=\; 0. \end{cases} \tag{1.2}$$

Il reste ensuite à résoudre cette équation fonctionnelle pour trouver l'expression de Q.

De (1.2) on déduit successivement que pour tout n, p entiers naturels strictement positifs :

$$Q(a^n) \;=\; Q(a)^n$$

puis,

$$Q(a) \;=\; Q\big(a^{n \times \frac{1}{n}}\big) \;=\; Q\big(a^{\frac{1}{n}}\big)^n$$

donc

$$Q\big(a^{\frac{1}{n}}\big) \;=\; Q(a)^{\frac{1}{n}}$$

et enfin

$$Q\big(a^{\frac{p}{n}}\big) \;=\; Q(a)^{\frac{p}{n}}.$$

Puis, comme X est supposée continue alors Q est continue. On utilise ensuite la densité topologique de $\mathbb{Q}_*^+$ dans $\mathbb{R}_*^+$ pour conclure que pour tout λ de $\mathbb{R}_*^+$ on a $Q(a^\lambda) = Q(a)^\lambda$.

Ainsi, pour tout $x, a > 1$ on a :

$$Q(x) \;=\; Q\big(a^{\frac{\log x}{\log a}}\big) \;=\; Q(a)^{\frac{\log x}{\log a}} \;=\; x^{\frac{\log Q(a)}{\log a}}.$$

Il en résulte que la quantité $\log Q(a)/\log a$ ne dépend pas du point a. En posant alors $\alpha = -\log Q(a)/\log a > 0$ on retrouve alors bien l'expression de la fonction de queue de distribution de la loi de Lévy-Pareto de paramètre α donnée par le Lemme 2.1.

Dans le sens $(ii) \Rightarrow (i)$, si X suit une loi de Lévy-Pareto de paramètre α alors par (1.1) on a :

$$\mathbb{P}\left(X_{|[a,+\infty[} > x\right) \;=\; \left(\frac{x \vee a}{a}\right)^{-\alpha}.$$

Ainsi, en posant $Y = \dfrac{1}{a}\, X_{|[a,+\infty[}$ on a :

$$\mathbb{P}(Y > y) \;=\; \mathbb{P}\left(X_{|[a,+\infty[} > ay\right) \;=\; (y \vee 1)^{-\alpha}$$

donc finalement, $Y \overset{\mathcal{L}oi}{=} X$. ♣

L'adéquation à une loi de Lévy-Pareto se mesure visuellement de façon très pratique par ce que nous appelons dans la suite le diagramme LogLog.

Définition 1.2.5. *Soit $z = \{z_1, ..., z_p\}$ une série de données réelles strictement positives. Soit σ la permutation de l'ensemble $\{1, ..., p\}$ qui à $\{z_1, ..., z_p\}$ associe la série ordonnée décroissante $z_\sigma = z_{\sigma(1)} \geq z_{\sigma(2)} \geq \cdots \geq z_{\sigma(p)}$.*

On appelle diagramme LogLog de la série $\{z_1, ..., z_p\}$ le graphique à double échelle logarithmique où sont portés les points $(i, z_{\sigma(i)})_{1 \leq i \leq p}$ (Voir figure 1.8).

Remarque (1) : un échantillon qui suit une loi de lévy-Pareto de paramètre α possède un diagramme LogLog qui est à tendance linéaire de pente $-1/\alpha$. Pour une preuve, on peut se reporter au Lemme 2.1.3 du chapitre 2. En conséquence, ce diagramme est parfois appelé diagramme fractal car il est supposé mettre en évidence ce caractère sur les distributions testées. En effet, à supposer qu'un diagramme LogLog soit linéaire, effectuer un changement d'échelle au sens de la définition 1.2.4 revient à changer l'origine de l'axe des ordonnées et tronquer les données qui se trouvent sous le nouvel axe des abscisses. Cette opération ne change évidemment pas la caractère linéaire du diagramme ni sa pente, ce qui montre intuitivement l'invariance de la loi par changement d'échelle.

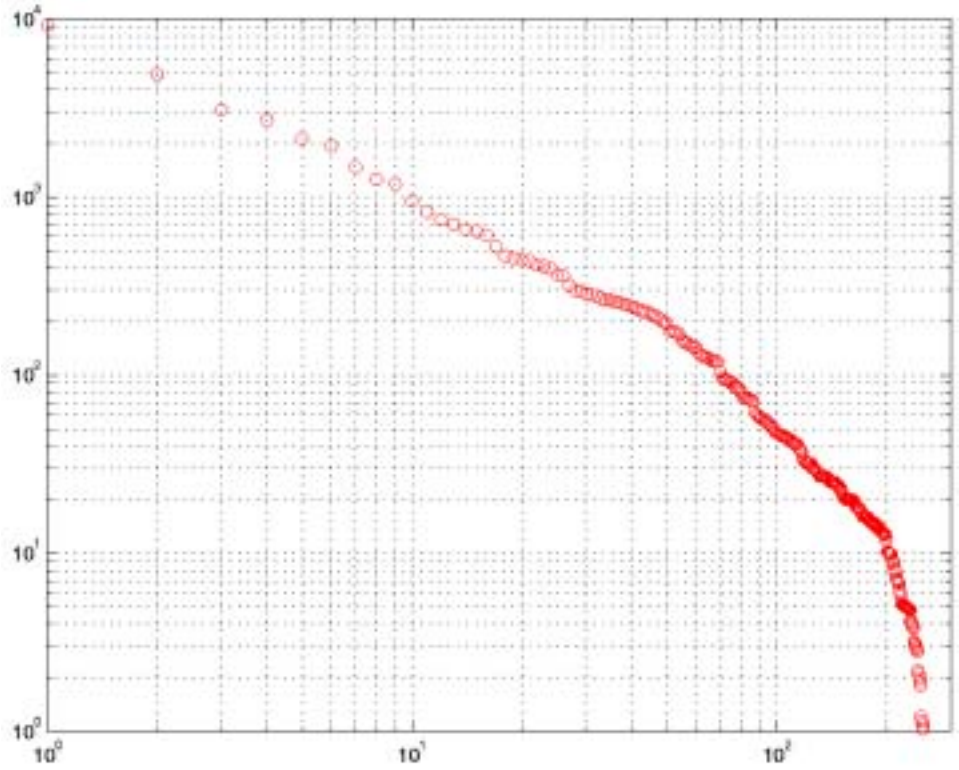

FIG. 1.8 – Diagramme LogLog des tailles des champs du Viking Graben de mer du Nord (1998, en Mb, seuls les champs de taille supérieure à 1 Mb ont été représentés).

Remarque (2) : le diagramme LogLog n'est, en fait, qu'une façon non orthodoxe de représenter la fonction de répartition empirique d'une distribution numérique.

À la vue de ces deux premiers exemples de distributions, l'une parfaitement Lévy-Pareto et l'autre représentant la distribution observée des tailles des champs d'une province pétrolière, on peut penser que la distribution observée semble être "sous" la tendance linéaire pour les gros objets, c'est-à-dire les premiers points du diagramme. Cette constatation est valide pour de nombreuses provinces pétrolières[22] et certains auteurs ont tenté d'en rendre compte en utilisant des lois de Lévy-Pareto légèrement perturbées : la loi "*stretched exponential*" et la loi dite "fractale parabolique".

La loi "stretched exponential"

C'est une loi que l'on définit au moyen de sa fonction de queue de répartition Q_s, pour λ et c strictement positifs donnés, par

$$Q_s(x) = \exp(-\lambda x^c) \mathbb{1}_{[\,0\,;+\infty[}\,(x).$$

On trouvera dans Laherrère et Sornette [49] de nombreux exemples d'application de cette loi, et en particulier à la modélisation de la taille des champs d'une province pétrolière. Notons que, suivant Frisch et Sornette [32], les auteurs justifient aussi le fait que cette loi apparaîtrait de façon "naturelle" lorsque le phénomène observé est issu du produit d'un nombre fini phénomènes aléatoires.

Cette loi a l'avantage de ne pas sous-estimer l'existence de petits objets et de posséder, comme la loi LogNormale, des moments de tous ordres. Elle est donc notamment dans le domaine d'attraction de la loi Normale. Cependant, dans son utilisation à des fins de modélisation de phénomènes hautement dispersifs, il est, en pratique et pour des tailles d'échantillons modérés (de l'ordre de la centaine), impossible de faire la différence entre cette loi et une loi de Lévy-Pareto. On aura alors tendance à lui préférer la première qui ne possède qu'un seul paramètre à estimer au lieu de deux.

La loi "fractale parabolique" de Laherrère

La loi dite "fractale parabolique" est née du besoin de prendre en compte la courbure observée sur les diagrammes LogLog des tailles des champs d'hydrocarbures. En conséquence, au lieu d'ajuster une droite aux données dans le diagramme LogLog, Laherrère [47] y ajuste une parabole.

[22] des exceptions existent cependant. Il n'est pas rare que le plus gros champ, ou les deux plus gros champs d'une région soient très nettement plus importants que les suivants et largement au dessus de la tendance linéaire de leurs successeurs. C'est ce que Jean Laherrère appelle "l'effet roi" (voir notamment [45], [46], [47], etc.).

Notons de suite que le terme "fractal" qualifiant cette distribution est fallacieux puisque nous avons montré en 1.2.4 que seule la loi de Lévy-Pareto présente le caractère d'invariance stochastique par changement d'échelle.

On peut montrer par une preuve tout-à-fait analogue à celle du lemme 2.1.3, cf. Lepez et Mandonnet [53], que la fonction de queue de répartition Q_p de cette loi s'écrit, pour α et β strictement positifs :

$$Q_p(x) \;=\; 1 \vee \exp\left(-\alpha + \sqrt{\alpha^2 - \beta \log x}\right) \mathbb{1}_{[\,1\,;\,e^{\alpha^2/\beta}]}\,.$$

La loi de Lévy-Pareto s'obtient comme cas limite lorsque α^2/β tend $+\infty$ sous la contrainte que α/β tende vers une constante strictement positive.

Un caractère surprenant de cette loi est qu'elle est à support compact et que sa densité tend vers $+\infty$ lorsque l'on se rapproche de la borne maximale de l'intervalle. D'un point de vue purement naturaliste, il est difficile de justifier l'emploi de cette loi qui est modale en les plus gros objets. D'un point de vue empirique, Lepez et Mandonnet [53] montrent que des tests d'adéquation à cette loi comparés à la loi de Lévy-Pareto, sur plusieurs jeux de données, ont tendance à être moins bons. Ceci pourrait paraître étrange compte tenu du fait que le modèle parabolique contient le modèle linéaire, mais il ne faut pas perdre de vue que le nombre de paramètres à estimer y est supérieur et le résultat alors moins fiable.

La figure 1.9 présente les diagrammes LogLog des quatre lois que nous avons présenté ci-dessus pour des valeurs de paramètres donnant les plus grandes observations comparables.

Venons-en maintenant aux premiers travaux d'estimation de réserves basés sur la modélisation de la distribution de la taille des champs qui ont suscité notre sujet de thèse.

1.2.2 Mise en évidence du manque de robustesse de certaines méthodes d'estimation

Les travaux de modélisation de Jean Laherrère ([45] et suivants) ont été précurseurs en Europe. Ils consistent essentiellement à ajuster une loi de probabilité comme celles décrites ci-avant aux données observées des tailles des champs dans une province pétrolière, ce au moyen de leur diagramme Log-Log. Une fois la "meilleure" loi estimée (dans la suite, nous notons Q sa fonction de queue de distribution) par des méthodes purement graphiques, des estimateurs sont construits pour déduire le nombre de champs restant à découvrir. Au début de notre travail de thèse, il a été très surprenant de constater qu'il existe outre-Atlantique depuis plus de vingt ans une littérature très abondante sur le sujet (se reporter à la bibliographie). Ces écrits semblent pourtant n'avoir reçu aucun écho sur le vieux continent, ou tout du

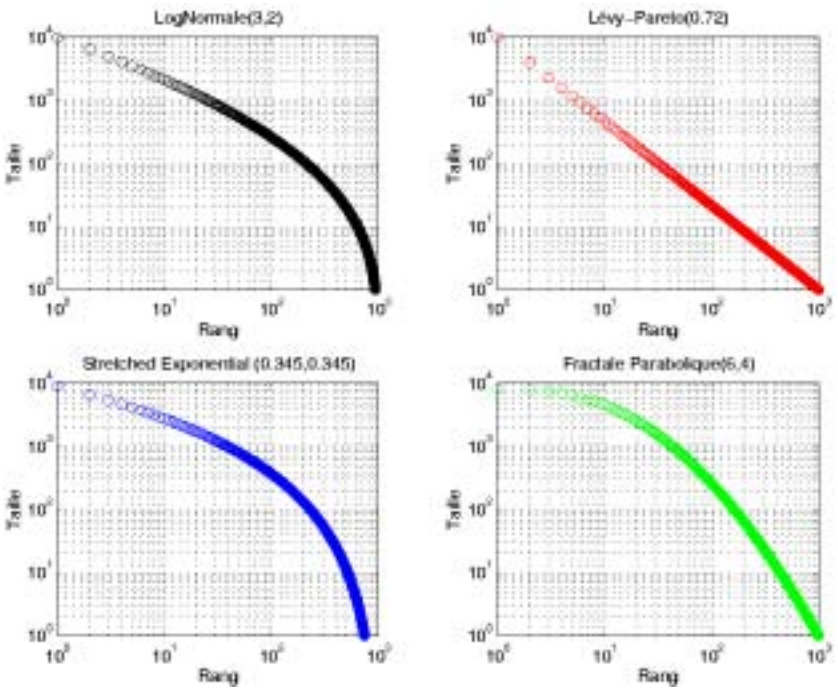

FIG. 1.9 – Diagrammes LogLog des quatre lois de probabilité LogNormale, Lévy-Pareto, *Stretched Exponential* et Fractale Parabolique.

moins en France, où Laherrère est considéré, à juste titre, comme pionnier local.

Nous ne développerons pas l'étude des formules d'estimation que nous avons extraites des divers travaux de Laherrère, travaux basés sur des lois de Lévy-Pareto essentiellement. Nous reportons simplement les conclusions et revoyons le lecteur intéressé à Lepez et Mandonnet [53].

Dans cette étude, nous avons cherché à inscrire les formules d'estimation ponctuelles exhibées par Laherrère dans des intervalles de confiance, compte tenu de l'ébauche de modèle probabiliste sous-jacent à ses travaux. Une province pétrolière étant donnée, on appelle N le nombre total de champs existant dans le sous-sol et de taille supérieure à une certaine unité, que nous qualifions d'unité de rentabilité économique[23]. Soit n le nombre de tels champs déjà découverts et $(T_i)_{1\leq i\leq n}$ la taille de ces champs. L'estimateur $\widehat{N}$ de N implicitement proposé par Laherrère est

$$\widehat{N} \;=\; \frac{1}{Q\big(\max\limits_{1\leq i\leq n} T_i\big)}.$$

[23]Cette unité représente donc une taille critique au dessous de laquelle il n'est pas rentable d'aller exploiter le champ.

L'utilisation de cet estimateur suppose évidemment que l'on ait découvert le plus gros objet de façon sûre, hypothèse de laquelle nous verrons que nous ne pouvons jamais nous affranchir, mais qui est peu contraignante, car, comme nous le verrons, les plus gros champs ont tendance à être trouvés les premiers dans les compagnes d'exploration. On peut alors facilement montrer que les intervalles de confiance associés à cet estimateur sont libres de loi[24] et qu'en particulier, pour a et b strictement positifs :

$$\mathbb{P}\left(N \in \left[\frac{\widehat{N}}{a}\,;\, b\widehat{N}\right]\right) \;=\; e^{-1/a} - e^{-b}.$$

Ce qui donne des intervalles de confiance absolument gigantesques : si par exemple $\widehat{N} = 1000$, ce qui est un ordre de grandeur raisonnable pour nos applications, alors un intervalle de confiance "raisonnablement centré" à 90 % seulement pour N est $[100\,;\,10000]$! De plus, cet intervalle de confiance ne prend en compte que l'information fournie par le maximum de la distribution observée et néglige le reste de l'échantillon. En particulier, rien n'empêche $\widehat{N}$ d'être plus petit que n...

En ce qui concerne le montant total des réserves, les estimateurs ponctuels proposés sont le produit de $\widehat{N}$ par $\mathbb{E}_Q(T)$ dans le cas intégrable, et de $\widehat{N}$ par $\zeta(1/\alpha)$ (où ζ désigne la fonction zeta de Riemann) dans le cas non intégrable de la loi de Lévy-Pareto de paramètre $\alpha < 1$.

Dans le cas intégrable, l'estimateur récupéré fluctue au moins autant que $\widehat{N}$. Dans le cas non-intégrable, on montre que l'estimateur proposé néglige totalement les fluctuations dues au fait que lorsque $\alpha < 1$ la loi de Pareto est dans le domaine d'attraction d'une loi stable qui est hautement dispersive. L'estimateur proposé est donc sûrement non valide et il est inutile de se hasarder ici à écrire des intervalles de confiance, qui seraient quoi qu'il arrive inexploitables.

Laherrère s'est parfaitement rendu compte de la haute instabilité de ses estimations, et l'a mis sur le compte du choix, selon lui mal adapté, de la loi de Lévy-Pareto. On observe en effet une courbure plus ou moins forte en pratique sur les données et il convenait d'en tenir compte, d'où l'apparition notamment de l'usage de la loi "fractale-parabolique" décrite ci-dessus.

Nous justifions pour notre part la courbure observée par la sous-représentation chronique des données de petite taille.

1.2.3 Mise en évidence d'une censure sur les petits champs dans le processus de découverte

Nous reprenons ici quelques un des arguments développés dans Lepez [52].

[24] et auront ainsi les mêmes "performances" quelle que soit la loi des observations.

Une première raison de la censure naturelle dans l'observation des petits champs est tout simplement liée aux techniques géophysiques de visualisation du sous-sol. Des pièges dont l'ordre de grandeur de taille est trop faible par rapport à la résolution de l'étude sismique ne peuvent évidemment pas être identifiés. Avec les progrès technologiques réalisés en la matière [25], les petites structures deviennent plus faciles à identifier, sans parler des nouveaux types de piégeage que l'on a pu mettre au jour.

Pour comprendre une autre cause de la censure des petits champs, supposons-nous un instant être responsable de l'exploration d'une zone géographique donnée du globe pour le compte d'une compagnie pétrolière. Pour chaque objectif géologique est établie une *fiche prospect* par les géologues de terrain faisant figurer toutes les valeurs des paramètres cités ci-dessus. Le but de cette fiche est d'évaluer le risque financier lié au développement du gisement potentiel au regard du montant de réserves qu'il est susceptible de receler. Le montant des réserves étant proportionnel à chacun de ces paramètres (cf. 1.1.3), toutes choses égales par ailleurs, une valeur faible de l'un d'entre-eux pénalisera le montant total estimé de réserves et le gisement potentiel risque de ne pas être développé puisque trop risqué d'un point de vue économique. Si le champ, même petit, existe effectivement, il est ignoré puisque non rentable en probabilité comparativement à des prospects *a priori* moins risquées. Bref, dans l'industrie : "petite probabilité d'existence = non-existence[26]" alors que la réalité est certainement du type "petite probabilité d'existence = (éventuellement très) petite accumulation".

Les pièges à faible potentiel sont donc censurés puisqu'il y a une grande probabilité *a priori* que leur développement ne se fasse pas dans des conditions acceptables de rentabilité.

D'un point de vue statistique, cela signifierait que la distribution des tailles des champs est simplement tronquée à gauche par une valeur que l'on pourrait appeler seuil de rentabilité économique. Ce seuil, en mer du Nord, était d'environ 50 Mbep dans les années 70 et descendu à 30 voire 20 Mbep de nos jours[27]. Certains petits champs ignorés alors sont finalement apparus dans la distribution car ils sont devenus, d'une part, visibles à la sismique et, d'autre part, rentables économiquement, pour peu qu'ils n'aient pas été éloignés d'une structure de production préexistante à proximité. Pour les petits champs en effet, le développement est très fortement conditionné au coût d'extraction. On ne peut construire une plate-forme pour développer un champ de 25 Mbep éloigné de tout alors que tirer un *pipe* depuis une

[25] Notamment le développement de la sismique 3D grâce aux améliorations des performances des traitements informatiques.

[26] Au moins d'un point de vue économique.

[27] Notons aussi que les petites structures sont aussi mieux identifiées que par le passé du fait des progrès des techniques de sismique, notamment dans la réduction du pas de maillage.

plate-forme d'exploitation préexistante à proximité a un coût quasiment nul. En mer du Nord, on est donc aujourd'hui amené à développer des champs de très petite taille pour peu qu'ils soient "satellites" de plus grosses structures[28]. Des champs d'une taille qui serait jugée absolument insignifiante au moyen-orient. A l'inverse, des champs encore plus petits sont exploités dans certaines zones du globe où ils sont très nombreux et proches comme dans le delta du Niger ou le golfe du Mexique.

En conclusion, progrès techniques aidant et coûts d'extraction baissant, des structures totalement ignorées voila 20 ans sont aujourd'hui bien identifiées et exploitées en nombre de plus en plus important. La faible probabilité d'existence des petits champs imposée par la loi LogNormale est donc très certainement à remettre en cause...

Le graphe de la figure 1.10 montre l'évolution dans le temps des découvertes d'une province pétrolière et le fait que les "petits" champs sont systématiquement sous-réprésentés. Ceci se traduit par un effondrement de la courbe, mais celle-ci a tendance à se "redresser" à mesure que l'exploration progresse. Une simulation d'un tel processus est donnée par la figure 2.3.

Notre point de vue est donc que la loi de Lévy-Pareto est adaptée à la description de la distribution de la taille des cahmps du sous-sol. L'effondrement du diagramme LogLog sur les petits champs (cf. figure 1.10) est à notre sens bien moins dû à une loi parente rendant improbable la présence de petits champs qu'à des probabilités d'inclusion très faibles sur ces derniers (à ce propos, Lepez [52] montre empiriquement et par simulation comment un tirage à probabilités inégales d'inclusion d'une loi de lévy-Pareto peut paraître très LogNormal du point de vue statistique).

1.3 Une problématique de type sondage

Nous présentons ici une ébauche du modèle que nous allons par la suite étudier en détail. Il constitue le cœur de notre thèse. Nous commençons dans les deux premières parties par replacer notre travail dans le contexte de la théorie des sondages puis nous formulons nos premières hypothèses sur le modèle que nous sommes amenés à construire puis à détailler dans le chapitre deux.

Commençons par définir rigoureusement ce que nous entendons par sondage :

Définition 1.3.1. *Soit un ensemble à $N \in \mathbb{N}$ éléments appelés respectivement "population parente" (ou juste "population") et "individus". On appelle*

[28] Le cas est fréquent. La notion de champ satellite est très familière des géologues qui ont l'habitude de faire des forages d'exploration et de trouver de nouveaux champs au voisinage d'un gisement déjà identifié. Nous reviendrons sur ce point crucial à la fin de ce chapitre.

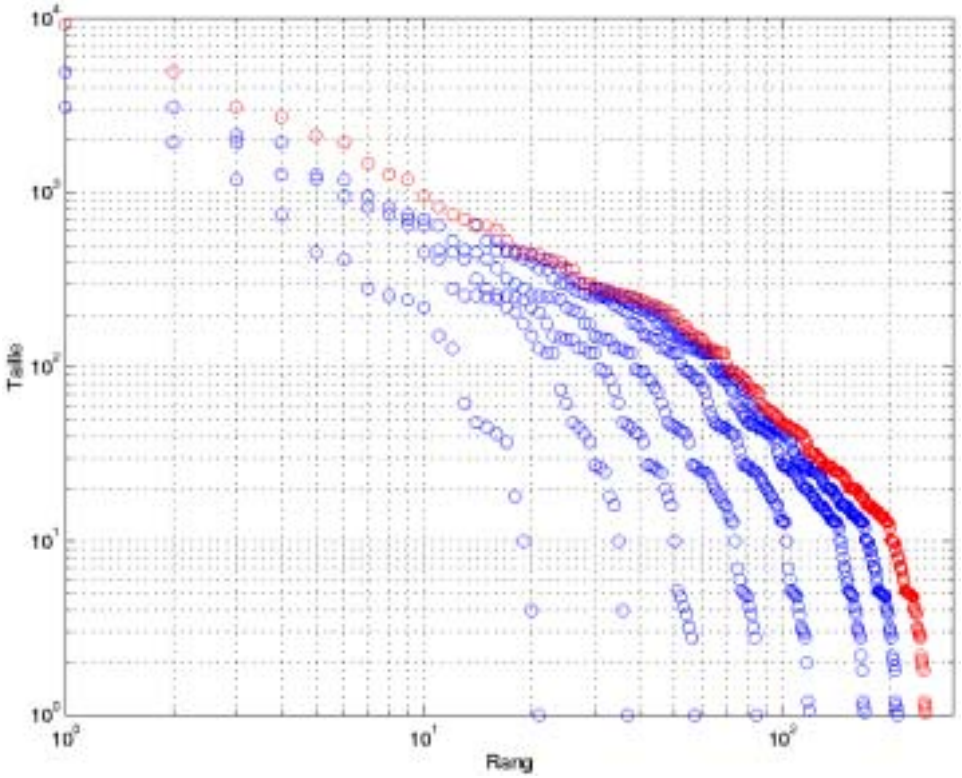

FIG. 1.10 – Diagrammes LogLog du Viking Graben de mer du Nord (1973, 1974, 1975, 1978, 1983, 1988, 1993 en bleu et 1998 en rouge)

sondage tout tirage, avec ou sans remise de $n \in \mathbb{N}$ individus parmi la population. L'ensemble des individus tirés s'appelle "échantillon observé" et le rapport $p = n/N$ est appelé taux de sondage.

Remarque (1) : de façon plus abstraite, on peut aussi voir un sondage comme la donnée d'un ensemble $\{i_1, ..., i_N\}$, d'un entier n et d'une application $j : \mathbb{N}_n \longrightarrow \mathbb{N}_N$. L'échantillon observé est alors l'ensemble $\{i_{j(1)}, ..., i_{j(n)}\}$. Le tirage est effectué sans remise si et seulement si l'application j est injective.

Remarque (2) : notons que si le tirage est effectué avec remise, rien n'empêche à n d'être plus grand que N[29]. S'il est effectué sans remise alors $p \in [0, 1]$.

Remarque (3) : en pratique, seule une caractéristique numérique (ou vectorielle) X des individus $\{i_1, ..., i_N\}$ de la population nous intéresse. Nous serons donc amenés, par abus de langage, à identifier l'ensemble $\{X_1, ..., X_N\} = \{X(i_1), ..., X(i_N)\}$ des valeurs des variables pertinentes sur les individus aux individus eux-mêmes. Un sondage étant effectué, nous identifions alors l'échantillon observé à l'ensemble $\{Y_1, ..., Y_n\} = \{X(i_{j(1)}), ..., X(i_{j(n)})\}$.

1.3.1 Tirage à probabilités d'inclusion inégales

Dans notre contexte, la population parente est l'ensemble des tailles des champs qui existent dans le sous-sol. La population observée est constituée de l'ensemble des champs découverts $\{Y_1, ..., Y_n\}$ *jusqu'à une date de référence* t. Ils constituent évidemment un tirage *sans remise* de n individus parmi les N éléments de la population parente.

Dans toute la suite, nous ne parlerons de sondage que pour des tirages sans remise.

Prendre en compte la censure sur les petits champs signifie que l'on attribue de faibles probabilités de figurer dans l'échantillon observé aux éléments de la population parente $\{X_1, ..., X_N\}$ qui sont effectivement de petite taille. Comparativement, on va avoir tendance à associer aux individus de grande taille des probabilités fortes de figurer dans la population observée. Les probabilités d'inclusion, comme nous allons les définir de suite, peuvent donc ne pas être identiques pour tous les individus de la population parente.

Définition 1.3.2. *Soit $\{Y_1, ..., Y_n\}$ un échantillon issu d'un sondage sur $\{X_1, ..., X_N\}$. Soit Z une variable nominale définie sur $\{X_1, ..., X_N\}$ et soit k une modalité de cette variable. La probabilité d'inclusion π_i d'un individu*

[29] Conformément à cette définition, le protocole de bootstrap apparaît alors comme un sondage particulier.

X_i de $\{X_1, ..., X_N\}$ présentant la modalité k est la fréquence d'observation de la modalité k dans $\{Y_1, ..., Y_n\}$:

$$\pi_i = \frac{|\{j \in \mathbb{N}_n \mid Z(Y_j) = k\}|}{|\{j \in \mathbb{N}_N \mid Z(X_j) = k\}|}.$$

Remarque (1) : tous les individus présentant la modalité k ont donc la même probabilité d'inclusion. Celle-ci peut alors s'interpréter comme un taux de sondage "conditionnel à la modalité k".

Remarque (2) : lorsque la variable d'intérêt Z est ordinale ou cardinale, il suffit de la considérer alors comme nominale pour obtenir une définitions de la probabilité d'inclusion conforme au "bon sens" et adaptée aux variables cardinales. Dans notre contexte par exemple, la variable d'intérêt "taille des champs" est cardinale. On peut la considérer comme nominale[30] puis considérer la probabilité d'inclusion d'individus d'une taille, ou plage de taille, fixée à l'avance.

1.3.2 Modèle de Superpopulation

Dans la l'approche traditionnelle de la théorie des sondages, on cherche à faire de l'inférence sur une variable numérique caractéristique d'une population entière. Pour cela, on dispose d'un sous-échantillon de cette population sur lequel on aura mesuré ce caractère. L'aléa n'intervient, dans ce type de sondage, que dans le tirage de l'échantillon considéré. La population de référence n'est pas supposée être elle même issue d'un échantillonnage (tirage) d'une variable aléatoire. Le caractère aléatoire des estimateurs des paramètres pertinents ne provient donc, lui aussi, que du mode de tirage de l'échantillon observé.

Cependant, lorsque l'on ne dispose pour seules données que d'un échantillon dont on sait qu'il est issu d'un tirage de type sondage il faut prendre en compte un deuxième type d'aléa lié au fait que la population parente est inconnue.

Bien entendu, il est absolument impossible de tester toutes les configurations possibles de la population parente c'est pourquoi le statisticien cherche à prendre en compte l'information *a priori* qu'il peut posséder[31]. Il en vient ainsi à "probabiliser" la population parente inconnue, typiquement en la considérant comme la réalisation d'un échantillon d'une variable aléatoire d'une loi paramétrique dont le vecteur des paramètres est de petite dimension.

[30] et par construction aussi ordinale...

[31] En attribuant par exemple des probabilités d'existence plus ou moins grandes à certaines configurations plutôt qu'à d'autres.

Dans cette approche, qualifiée d'approche modèle ou de modèle de superpopulation [6], les données sont considérées comme issues d'un "processus de tirage à deux étages" :

Définition 1.3.3. *On appelle modèle de superpopulation un sondage au sens de la remarque (3) de la définition 1.3.1 dans lequel la population parente* $\{X_1, ..., X_N\}$ *est un* N*-échantillon d'une variable aléatoire de loi elle aussi dite parente.*

Si le tirage est sans remise, la population observée ne peut donc pas être considérée comme un sous-échantillon de $\{X_1, ..., X_N\}$ possédant la même loi parente. Cependant, c'est grâce à cette sous-population représentant un échantillon d'une loi biaisée par rapport à la loi parente que l'on doit inférer sur des caractéristiques de la population parente. Ces dernières devront être estimées en prenant en compte les deux types d'aléa mentionnés ci-dessus :
– l'aléa lié à la loi parente sur l'échantillon parent $\{X_1, ..., X_N\}$;
– l'aléa dû au tirage de $\{Y_1, ..., Y_n\}$ au sein de $\{X_1, ..., X_N\}$.

Le problème d'estimation des paramètres de la loi parente devient encore plus épineux lorsque le tirage se fait non seulement sans remise mais de plus à probabilités inégales d'inclusion. En effet, si le taux de sondage p n'est pas proche de 0 ou de 1 le fait de tirer sans remise empêche de considérer les tirages comme indépendants. De plus, les probabilités inégales d'inclusion impliquent que l'échantillon observé ne peut plus alors être assimilé lui même à un échantillon de la loi parente, comme nous le verrons au chapitre 2. Dans notre cas, nous rencontrons ces deux écueils, puisque l'on a montré en 1.2.3 qu'il y a censure naturelle sur les petits champs et que notre processus de découverte des champs du sous-sol est par nature assimilé à un tirage sans remise.

Par ailleurs, définir correctement les probabilités d'inclusion au sens de la définition 1.3.2 lorsque l'on est dans le cadre d'un modèle de superpopulation peut être très délicat. En effet, la population parente étant inconnue, les effectifs d'individus présentant une modalité k d'une variable nominale définie sur la population parente sont alors eux-mêmes des variables aléatoires. On peut aussi adopter un autre point de vue, consistant à se donner à l'avance une fonction fixe ω définie sur le support de la loi parente, à valeurs dans $[0\,;\,1]$ et se donner alors N variables aléatoires $(\varepsilon_i)_{1\leq i\leq N}$ de lois de Bernoulli de paramètres $(\omega(X_i))_{1\leq i\leq N}$[32] représentant le fait que chaque $(X_i)_{1\leq i\leq N}$ a été sélectionné dans le tirage ou non. Le nombre total d'observations n devient alors aléatoire. Cette ébauche de protocole manque encore nettement de rigueur dans sa définition statistique. Nous résolvons totalement ce problème dans la partie 2.3.1.

Justifions maintenant notre choix de modélisation.

[32]Conditionnellement à la donnée de $\{X_1, ..., X_N\}$.

1.3.3 Premières hypothèses

Dans notre problématique, les individus de la population parente (c'est-à-dire l'ensemble des champs qui existent dans le sous-sol) qui sont non observés restent bien entendu inconnus. Cette dernière propriété justifie à elle seule, selon nous, une approche de type modèle de superpopulation au sens de la définition 1.3.3.

Dans ce modèle, nous distinguons les individus, ou objets, qui sont les gisements $\{X_1, ..., X_N\}$ de la population parente et les $\{Y_1, ..., Y_n\}$ qui sont ceux de l'échantillon observé. Étant donné que nous ne nous intéressons qu'à une et une seule caractéristique particulière de ces individus : leur taille T, nous identifierons les individus $\{X_1, ..., X_N\}$ et $\{Y_1, ..., Y_n\}$ et leurs caractéristiques $\{T(X_1), ..., T(X_N)\}$ ou $\{T(Y_1), ..., T(Y_n)\}$ selon le cas. Le fait de parler d'un individu de la population parente ou de l'échantillon étant généralement sans ambiguïté, on désignera donc indifféremment l'individu d'indice i de $\{X_1, ..., X_N\}$ par sa taille $T_i = T(X_i)$ ou l'individu d'indice j de $\{Y_1, ..., Y_n\}$ par sa taille $T_j = T(Y_j)$, .

La loi de probabilité au moyen de laquelle nous choisissons de modéliser la distribution de la population parente est la loi de Lévy-Pareto. Nous avons légitimé son emploi dans les sections précédentes, tant par des arguments descriptifs d'invariance stochastique par changement d'échelle que par des tests statistiques que l'on peut retrouver dans Lepez [52], Lepez et Mandonnet [53], ainsi qu'au chapitre 5 qui concerne les applications pratiques de nos travaux.

Par ailleurs, nous avons montré que nous devons prendre en compte des probabilités de découverte différentes selon la taille des individus de la population parente. En effet, et c'est là encore un "classique" de l'exploration pétrolière : "*The Big Stuff Gets Found First*", c'est-à-dire que plus la taille d'un champ est importante et plus il a une probabilité forte d'avoir été découvert tôt dans le processus d'exploration.

Enfin, comme nous l'avons mentionné dans la section 1.1.6, nous nous consacrons exclusivement à l'estimation du nombre de champs restant à découvrir, ainsi que leur taille *dans les conditions technologiques et économiques actuelles*. Nous ne prenons ainsi en compte ni les progrès technologiques à venir, ni les réévaluations de réserves au cours du temps sur les champs déjà connus, ni un scénario d'évolution du prix du brut.

Nous en venons maintenant à la modélisation de la distribution de la taille des champs ainsi que du processus de découverte proprement dits.

Chapitre 2

Modélisation

Dans ce chapitre, nous construisons rigoureusement notre modèle probabiliste de la distribution des tailles des champs du sous-sol.

Nous commençons par définir notre échelle d'étude. Celle-ci nous est dictée par les caractéristiques géologiques dont nous devons assurer l'homogénéité afin que notre échantillon soit représentatif. Nous formulons ensuite précisément les hypothèses de notre modèle. Nous terminons ce chapitre par la formalisation complète ainsi que diverses interprétations de ce dernier.

2.1 Caractéristiques géologiques de l'échantillonnage

Comme on l'a vu à la fin du chapitre 1, dans le modèle à construire, nous allons assimiler l'exploration pétrolière à un sondage à probabilités d'inclusion inégales au sein d'une approche de type modèle de superpopulation (cf. section 1.3.3).

La première question à se poser dans la démarche statistique de gestion d'un protocole de sondage est celle du choix et du mode de constitution de l'échantillon des observations. En effet, sur cet échantillon nous allons baser des estimations de paramètres inconnus qui sont censés lui être intrinsèques. Il est donc nécessaire de s'assurer que l'échantillon de travail est homogène en ces paramètres, c'est-à-dire que la population observée est bien représentative d'une seule et même valeur des paramètres et n'est pas constituée de plusieurs sous-familles associées à des valeurs diverses. Ceci éliminerait alors toute pertinence au protocole d'estimation. Or, ces paramètres étant inconnus, nous devons avoir recours à d'autres variables d'information dites auxiliaires, connues, qui vont garantir l'homogénéité souhaitée de notre échantillon.

Au moyen de la description de la génèse des hydrocarbures effectuées dans le chapitre 1, nous allons d'abord nous intéresser aux variables auxiliaires à prendre en compte, puis nous renforçons la justification donnée dans 1.3.3 du fait que la distribution de la taille des champs peut être correctement décrite

par une loi de Lévy-Pareto, ne dépendant que d'un seul paramètre. Nous identifions ce paramètre comme une mesure (tant statistique que géologique) de concentration et discutons en détail ce dernier point.

2.1.1 Choix de l'échelle du système pétrolier

Dans notre cadre, nous cherchons avant tout à garantir une certaine homogénéité géologique entre les champs que nous considérons dans notre échantillon. Les variables auxiliaires que nous allons prendre en compte pour s'assurer de cette homogénéité sont donc des variables qualitatives relatives à des paramètres géologiques de la zone géographique dans laquelle nous considérons notre échantillon d'observations. Le nombre de ces variables doit être choisi de façon opportune. En effet, si l'on devait garantir l'homogénéité d'un trop grand nombre de variables, on arriverait à une échelle d'étude trop petite (pouvant aller jusqu'au niveau du champ), et donc un nombre d'observations insuffisant pour obtenir des statistiques fiables. Si elles sont trop peu nombreuses en revanche, on risque d'amalgamer plusieurs sous-familles non homogènes en réalité.

Il est communément admis qu'une échelle d'étude appropriée doit avoir au plus la taille d'un bassin sédimentaire (voir Laherrère [45], Perrodon [67] ou Kontorovich *et al.* [44]). Comme nous venons de le voir, un bassin est essentiellement un corps géologique représentant une dépression dans la croûte terrestre. Un bassin est ensuite susceptible d'être le lieu de la superposition de plusieurs types de roche-mère, d'être non homogène dans les phénomènes de subsidence interne et de tectonique qui lui ont été appliqués et donc de présenter des structures de pièges très variées.

La classification géologique des bassins sédimentaires a donné lieu à de nombreux travaux, notamment Perrodon [64] et ce n'est qu'au début des années 80 que la géologie moderne a identifié le type d'entités géologiques susceptibles de présenter des relations étroites entre la distribution des tailles des champs et la structure tectonique dudit objet les contenant (voir Perrodon [66], Klemme [43] ou Demaison et Huizinga [28]).

Ces travaux ont montré que l'échelle d'étude adéquate pour répondre à ce problème est le système pétrolier. Cette structure respecte en effet la cohérence des paramètres qui sont à la base de la notion de gisement comme décrit dans la section 1.1.1. La définition qui suit est issue de Demaison et Huizinga [28] :

Définition 2.1.1. *On définit un système pétrolier comme une entité géologique incluse dans un bassin sédimentaire qui respecte la cohérence de certains éléments et évènements essentiels à la formation d'accumulations pétrolières. Ces éléments sont le type de roche mère, le mode de migration (primaire et secondaire) et le type de piégeage.*

Un bassin sédimentaire comporte donc en général plusieurs systèmes pétroliers, qui peuvent être entrelacés. Le bassin de mer du Nord par exemple comporte trois systèmes pétroliers dont l'un est à dominante de champs pétroliers (le Viking Graben, cf. section 5.3.1) et les deux autres sont à dominante gaz (Moray Firth et Central Graben). Il faut noter que certains systèmes sont très petits (moins d'une dizaine de champs, voire un seul champ) et ne peuvent donc pas faire l'objet d'une étude statistique.

L'identification géologique d'un système pétrolier peut être une tâche extrêmement difficile. Le bassin de Trinidad dans l'offshore du golfe du Mexique ainsi que les grands bassins du Moyen Orient par exemple, sont une superposition et un entrelacement de trois systèmes pétroliers différents ou plus. Comprendre leur structure a pris de nombreuses années. Une fois identifiés, les géologues établissent une sorte de carte d'identité du système, dite carte stratigraphique , qui décrit les caractéristiques des divers éléments de la définition 2.1.1. La figure 2.1 est un exemple de carte stratigraphique du système pétrolier de Ghadames en Libye déjà évoqué par la figure 1.1. À gauche du diagramme se trouve l'échelle des temps géologiques (correspondant à la profondeur de roche) et la nature des roches associées. Les plages de couleur sur les colonnes *seal*, *reservoir* et *source* indiquent respectivement la présence de roche couverture, roche réservoir et roche mère. La présence de champs d'huile ou de gaz (points noirs et étoilés de la colonne HC) est possible lorsque deux plages de couleur *seal* et *reservoir* sont contigües et qu'une plage de couleur *source* se trouve plus en profondeur (car les hydrocarbures migrent toujours *in fine* en direction de la surface).

Se placer à l'échelle d'un système pétrolier garantit que l'échantillon des champs est géologiquement homogène. On peut alors supposer que la distribution des tailles des champs à l'intérieur du système est homogène et que les tailles fournissent donc un échantillon d'une seule et même loi. Il reste alors à caractériser cette loi. Nous nous appuyons alors sur les observations des géologues décrites en 1.2.1 et sur le fait que les lois observées peuvent être prises dans une famille paramétrique ne dépendant que d'un unique paramètre qui s'interprète comme une mesure de concentration de la distribution.

2.1.2 Mesure(s) de concentration et notion d'habitat

Les géologues ont depuis longtemps constaté qu'il existe, trés schématiquement, deux types de distributions des tailles des champs à l'intérieur d'un système pétrolier donné. Sur des systèmes dits matures[1], les explorateurs se sont aperçus que l'on pouvait observer des distributions soit très inégalitaires, soit plutôt homogènes, dans le sens où la distribution des tailles

[1]Les systèmes matures sont des systèmes explorés depuis de nombreuses années sur lesquels on pense avoir mis au jour une fraction importante de l'ensemble des champs qui existent en réalité dans le sous-sol et notamment, bien sûr, les plus grosses structures.

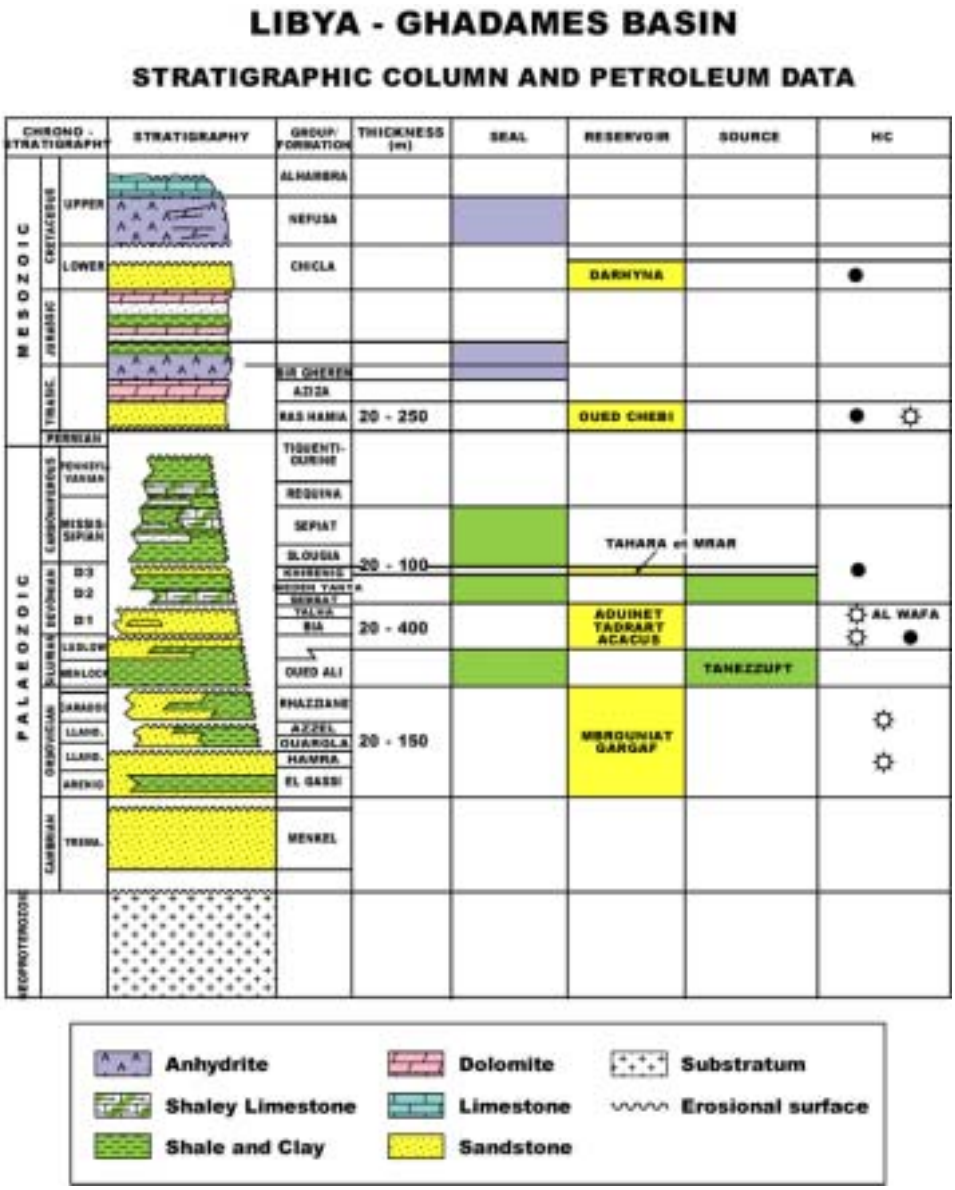

FIG. 2.1 – Carte stratigraphique du système pétrolier de Ghadames en Libye – Source : IFP.

observées n'est pas très dispersée, ce dernier cas étant relativement rare et semblant assez caractéristique des zones de delta (comme la région du delta du Congo par exemple, cf. section 5.3.2). La classification de ce caractère inégalitaire peut se faire au moyen d'un indice de concentration appelé *habitat* du système qui va représenter le paramètre de forme le plus important de notre modèle. Décrivons en premier lieu et au moyen de deux exemples ce que représente un indice de concentration.

Délaissons un instant les hydrocarbures et considérons la répartition des richesses dans une population donnée. Intuitivement, un indice de concentration (ou d'inégalités, selon que l'on est statisticien ou sociologue) est un instrument de mesure de la disparité de l'allocation dans la distribution du montant total de richesses disponibles parmi les individus. Aux deux extrêmes on peut trouver la distribution la plus inégalitaire qui soit : celle qui alloue à une seule personne la totalité de la richesse disponible ; et la plus égalitaire, qui divise de manière égale la richesse disponible entre tous les individus.

Dans le problème qui nous concerne, les individus sont les gisements d'hydrocarbures et leurs richesses sont représentées par leurs tailles.

Il existe de nombreuses façons de mesurer le caractère inégalitaire d'une distribution (voir Gourieroux [36]). Citons, par exemple, dans le cas des distributions finies, la courbe de Lorenz et l'indice de Gini qui peut lui être associé. Détaillons rapidement cette approche, car elle fournit une visualisation très claire du concept non trivial de mesure de concentration.

Fixons-nous une population de n individus. Selon les notations classiques en ce domaine, nous classons par ordre croissant leurs richesses $p_1 \leq p_2 \leq \cdots \leq p_n$. Considérons, pour tout $i = 1, ..., n$ la proportion q_i de richesses détenue par les i individus les plus pauvres :

$$q_i \;=\; \frac{\displaystyle\sum_{1\leq k\leq i} p_k}{\displaystyle\sum_{1\leq k\leq n} p_k}.$$

La courbe de Lorenz consiste à relier entre eux, dans le carré $[\,0\,;1\,]\times[\,0\,;1\,]$, les points $(i/n, q_i)_{i=1,\,...,\,n}$. Notons que cette courbe est nécessairement convexe et située sous la première bissectrice. La distribution est alors parfaitement égalitaire si la courbe de Lorenz est confondue avec la première bissectrice, et plus elle s'en éloigne, plus la distribution observée est inégalitaire. Il est même possible de définir une relation d'ordre entre les courbes de Lorenz, donc un ordre entre mesures d'inégalité dans différentes populations : nous dirons que la première distribution est plus inégalitaire que la seconde si sa courbe de Lorenz est en tout point située au dessous de la courbe de Lorenz

de la seconde distribution. Cet ordre n'est pas total puisque deux courbes de Lorenz peuvent se croiser.

La figure 2.2 présente deux courbes de Lorenz associées à deux lois de Pareto de paramètres respectifs 3 et 1,2 ainsi que deux échantillons de ces lois afin d'illustrer visuellement leur dispersion et leur caractère inégalitaire.

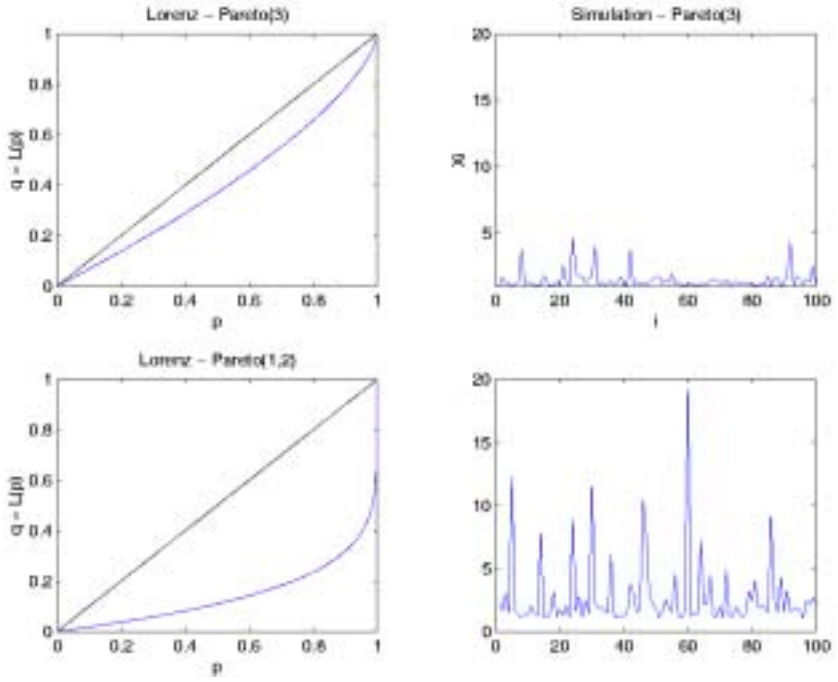

FIG. 2.2 – Courbes de Lorenz et échantillons de deux lois de Pareto de paramètres respectifs 3 et 1,2.

Il peut aussi être intéressant de mesurer le caractère inégalitaire d'une distribution au moyen d'un seul indice scalaire. Pour cela, on peut considérer l'indice de Gini G, égal à deux fois l'aire comprise entre la courbe de Lorenz et la première bissectrice. Cet indice, compris entre 0 et 1, indique que plus sa valeur est proche de 1 et plus la distribution est inégalitaire. Il vaut 0 si et seulement si la distribution est parfaitement uniforme. Notons que cet indice est compatible avec la relation d'ordre entre les courbes de Lorenz. On montre qu'il vaut

$$G = \frac{\sum_{1 \leq i,j \leq n} |p_i - p_j|}{2n^2 \sum_{1 \leq i \leq n} p_i}.$$

Il existe bien entendu des versions continues de ces deux outils. Pour une variable aléatoire X continue donnée à valeurs dans $\mathbb{R}^+$, de densité f et de

fonction de répartition F inversible, la courbe de Lorenz a pour équation

$$q \;=\; L(p) \;=\; \frac{1}{\mathbb{E}(X)} \int_0^p F^{-1}(t)\, dt$$

et l'indice de Gini vaut alors

$$G \;=\; \frac{1}{2\mathbb{E}(X)} \int_{[0;+\infty[^2} |u-v| f(u) f(v)\, du\, dv.$$

On voit alors sur ces deux expressions qu'elles ne sont définies que si la distribution considérée est intégrable. Ceci qui peut s'interprêter "physiquement" en réalisant qu'un échantillon d'une loi non intégrable est susceptible de receler des individus d'une taille si gigantesque ("suffisamment" grands pour amener la moyenne vers $+\infty$) pour qu'ils en deviennent incommensurables avec le reste de la distribution. La mesure de concentration au sens de Lorenz perd alors son sens. Ceci soulève donc le problème de la classification des mesures de concentration dans le cas des distributions non intégrables en général.

Une solution partielle de ce problème peut être trouvée dans l'approche de l'économiste Pareto [63], qui s'est intéressé, à la fin du XIX$^{\text{ème}}$ siècle, à la distribution des revenus des individus d'une population donnée. Pour les besoins de son étude, Pareto a modélisé cette distribution au moyen d'une famille de "lois puissances", c'est-à-dire des lois dont les fonctions de queue de répartition sont du type déjà vu en fin de chapitre 1 :

$$Q(t) \;=\; \mathbb{P}(T > t) \;=\; t^{-\alpha} \quad \text{pour} \quad t \geq 1 \quad \text{et} \quad \alpha > 0. \tag{2.1}$$

Lorsque $\alpha > 1$, la loi de Pareto est intégrable et sa courbe de Lorenz et son indice de Gini sont donnés par

$$L(p) \;=\; 1 - (1-p)^{\frac{\alpha-1}{\alpha}} \quad \text{et} \quad G \;=\; \frac{1}{2\alpha - 1}.$$

Sur la figure 2.2, les indices de Gini sont donc respectivement de 0,2 et 0,72. Il est à noter que pour des distributions assez concentrées, la valeur théorique de l'indice de Gini peut être assez éloignée de la valeur empirique calculée sur un échantillon simulé.

Le paramètre α, indépendamment de sa valeur s'interprète à lui seul comme une mesure de concentration. Il est en effet clair que plus α grandit et plus la probabilité de générer de larges objets devient faible et plus la distribution devient alors égalitaire, et vice-versa.

Il devient donc possible de quantifier le caractère inégalitaire d'une distribution, même non intégrable, pour peu qu'elle appartienne à la classe des lois puissance grâce, justement, à la valeur de son exposant caractéristique.

Pour aller un peu plus loin dans ce raisonnement,comment se donner une mesure de concentration d'une distribution non intégrable qui n'appartiendrait pas à la classe des fonctions puissance ? Par les Théorèmes Centraux Limite de Lévy relatifs aux lois stables (voir Araujo et Giné [5] et Feller [30]), une loi non intégrable (ou, plus généralement, ne possédant pas de moment d'ordre 2) se trouve toujours dans le domaine d'attraction d'une loi dite α-stable, dont la densité s'écrit comme le produit d'une fonction à variation lente (comme un logarithme) par une fonction puissance, comme la densité d'une loi de Lévy-Pareto. L'exposant apparaissant dans cette loi stable (ou, plus précisément, dans la fonction puissance de la loi stable) peut alors être pris comme indice de concentration même si la loi de départ à laquelle on s'intéresse n'est pas purement une loi puissance. Notons que les lois possédant des moments d'ordre supérieurs à 2 sont toutes dans le domaine d'attraction de la loi Normale par le Théorème Central Limite classique et leurs caractères inégalitaires deviennent donc non comparables si l'on suit le principe décrit ci-dessus. Mais elles sont alors en particulier intégrables et la courbe de Lorenz et l'indice de Gini font alors parfaitement l'affaire comme outils de mesure.

Nous avons ainsi présenté notre vision "unificatrice" des approches parallèles suivies par Pareto et Lévy qui ont tout deux, totalement indépendamment l'un de l'autre, mis en évidence l'intêret des lois puissances. C'est pourquoi nous appelons finalement ces distributions "lois de Lévy-Pareto".

Conformément au chapitre 1, où nous avons montré que les distributions des tailles de systèmes matures peuvent être approchées par des lois de Lévy-Pareto, nous avons alors une mesure toute naturelle de leur caractère inégalitaire, comme étant l'exposant α (à estimer en pratique) de la loi puissance approchant le mieux la distribution observée.

Or, les géologues utilisent depuis longtemps la valeur absolue h de la pente du diagramme Log-Log (qui, rappelons-le, est quasi-linéaire) comme indicateur de concentration.

Définition 2.1.2. *On appelle habitat d'un système pétrolier la pente h du diagramme Log-Log des tailles des champs du système. On dit que l'habitat est*

$$\begin{cases} \textit{concentré si } h > 1 \\ \textit{dispersé si } h \leq 1. \end{cases}$$

En effet, si la pente est forte, il n'existe que quelques gros champs qui "concentrent" toute la masse des réserves car la taille des champs décroît extrêmement vite. A l'inverse, si la pente est faible, la décroissance est bien plus lente et le montant total des réserves se "disperse" en un nombre important de champs de tailles comparables.

Nous présentons maintenant un Lemme qui permet de faire le lien avec la loi de Lévy-Pareto.

Lemme 2.1.3. *Si l'on suppose que la distribution observée suit exactement une loi de Lévy-Pareto d'exposant α alors l'habitat h vaut $1/\alpha$.*

Remarque : il en résulte que l'habitat est

$$\begin{cases} \text{concentré si } \alpha < 1 \\ \text{dispersé si } \alpha \geq 1. \end{cases}$$

On voit donc que la démarcation intuitive faite par les géologues à la valeur 1 qui distingue le comportement dispersé du comportement concentré n'est autre que la distinction entre intégrabilité (cas dispersé, rare en pratique et qui semble caractéristique des zones de delta) et non-intégrabilité (cas concentré, extrêmement fréquent) pour la loi de Lévy-Pareto sous-jacente.

Preuve : Nous ne donnons ici qu'une heuristique, pour fixer les idées.

Soit $\{T_1, ..., T_N\}$ un échantillon dont le comportement en diagramme LogLog est à tendance linéaire de pente $-k$, c'est-à-dire d'habitat k. Soit $T_{(1)} \leq \cdots \leq T_{(N)}$ l'échantillon ordonné associé. On a, pour $i = 1, ..., N$

$$\log T_{(i)} \approx \log T_{(N)} - k \log(N - i + 1). \tag{2.2}$$

Par ailleurs, en notant

$$Q_N(t) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{T_i \geq t}$$

la fonction de queue de répartition empirique associée à $\{T_1, ..., T_N\}$, par définition, on a $NQ_N(T_{(i)}) = N - i + 1$, donc en injectant dans 2.2 on obtient

$$\begin{aligned} \log T_{(i)} &\approx \log T_{(N)} - k \log N - k \log(Q_N(T_{(i)})) \\ &\approx \log T_{(1)} - k \log(Q_N(T_{(i)})). \end{aligned}$$

Ainsi, on a

$$Q_N(T_{(i)}) \approx \left(\frac{T_{(i)}}{T_{(1)}} \right)^{-\frac{1}{k}}.$$

On retrouve donc la forme de la fonction de queue de répartition de la loi de Lévy-Pareto donnée par 2.1 pour l'échantillon "normalisé" $(T_{(i)}/T_{(1)})_{1 \leq i \leq N}$ et l'on conclut par le Théorème de Glivenko-Cantelli. ♣

2.2 Présentation et hypothèses du modèle

Dans la section précédente, nous avons mis en place quelques bases de la construction de notre modèle. Nous le formulons de façon plus précise dans cette partie.

Nous supposons dans la suite que l'on se place dans un système pétrolier bien identifié donné.

2.2.1 Les tailles des champs et les logtailles

Nous faisons l'hypothèse que la distribution des tailles des champs peut être correctement approchée par une loi de Lévy-Pareto.

L'ensemble des tailles des champs (exprimées en Mbep) existant dans la nature est donc modélisé par un échantillon i.i.d. $\{T_1, ..., T_N\}$ (auquel on associe la statistique d'ordre $T_{(1)} \leq \cdots \leq T_{(N)}$) d'effectif N inconnu d'une loi de Lévy-Pareto donnée par 2.1 dont l'exposant α est lui aussi inconnu.

En vertu de la propriété d'invariance par changement d'échelle donnée par la Proposition 1.2.4, nous supposerons toujours que notre échantillon est normalisé en taille, c'est-à-dire que si $T_{(1)} \neq 1$ alors nous effectuerons la transformation $(T_i/T_{(1)})_{1\leq i\leq N} \to (T_i)_{1\leq i\leq N}$. Nous supposerons donc toujours que l'échantillon avec lequel nous travaillons est issu d'une loi de Lévy-Pareto minorée par 1 (nous dirons aussi basée en 1). Rappelons ici (cf. 1.2.1) que la loi du logarithme d'une variable aléatoire de loi de Lévy-Pareto est une loi exponentielle de même paramètre.

Définition 2.2.1. *Soit $\{T_1, ..., T_N\}$ un échantillon normalisé des tailles des champs d'un système pétrolier. On appelle échantillon des logtailles l'échantillon $\{X_1, ..., X_N\} = \{\log T_1, ..., \log T_N\}$*

Pour la suite, et notamment en ce qui concerne le travail d'estimation des paramètres inconnus du modèle, nous travaillerons sur l'échantillon des logtailles plutôt que sur l'échantillon des tailles directement. Nous utilisons en effet des techniques relatives aux modèles exponentiels et la manipulation de la loi Exponentielle dans ce contexte est clairement mieux adaptée que celle de la loi de Lévy-Pareto, bien que les deux approches soient parfaitement équivalentes.

2.2.2 Le sous-sol et les découvertes

L'ensemble des découvertes est l'ensemble des champs qui ont été mis au jour jusqu'à aujourd'hui. L'ensemble des logtailles $\{Y_1, ..., Y_n\}$[2] des découvertes est donc un sous-ensemble de l'échantillon des logtailles $\{X_1, ..., X_N\}$ des champs existant dans le sous-sol.

[2] le nombre de champs découverts n est donc bien sûr connu.

Nous interprétons ces découvertes comme un tirage sans remises de n individus dans $\{X_1, ..., X_N\}$. Si le tirage était équiprobable, la loi d'un Y_i serait alors clairement la même que celle d'un X_i et $\{Y_1, ..., Y_n\}$ serait un échantillon i.i.d. de la même loi exponentielle que les $(X_i)_{1 \leq i \leq n}$.

Mais cette hypothèse n'est pas tenable car elle signifierait que les champs du sous-sol sont découverts "au hasard", au sens populaire du terme, c'est-à-dire, correspondant au schéma de loto classique. Or, le bon sens seul montre que même si l'on forait au hasard dans le sous-sol à la recherche d'un champ, la probabilité de découverte d'un champ fixé serait proportionnelle à sa "taille", et plus précisément, son étendue cartographique par exemple. Mais ce n'est pas le seul biais dans le tirage à considérer.

2.2.3 Le biais par effet taille

Les techniques de type géosciences (géophysique, géochimie, géologie et géostatistique) interviennent de façon forte dans le tirage décrit ci-dessus. Elles le rendent hautement non-équiprobable, dans le sens où elles engendrent la tendance à associer aux objets des probabilités d'être découverts d'autant plus grandes qu'ils sont "gros" (dans une unité éventuellement à définir).

En effet, le métier de géologue d'exploration consiste justement à identifier en priorité les champs restants de taille la plus importante possible car ce sont eux qui sont le plus enclins à être rentables à exploiter. Nous appelons ce phénomène biais par effet taille.

Nous prenons donc en compte le biais dans le tirage (sous-entendu biais par rapport à un tirage sans-remise équiprobable) en nous donnant une fonction de pondération, inconnue, proportionnelle aux probabilités d'inclusion d'un élément du sous-sol dans l'ensemble des découvertes faites jusqu'à aujourd'hui. Cette fonction dépend donc du temps. La suite de fonctions de pondération obtenue en faisant varier le temps est nécessairement croissante. Le fait que "plus on est gros et plus on est visible" (*"big stuff gets found first"*, voir Campbell et Laherrère [18], Kaufman [4], Meisner et Demirmen [62], Wiorkowski [82] et de nombreux autres) se traduit donc par une hypothèse de croissance sur la fonction de pondération.

Notons que pour la plupart des auteurs (dont Kaufman [4] et Bickel *et al.* [13]), le biais par effet taille n'est relatif qu'à une fonction de pondération prise comme égale à l'identité à l'origine des temps, c'est-à-dire avant qu'aucun tirage, *i.e.* aucune découverte, ne soient effectués. Autrement dit, la probabilité de découverte d'un champ serait directement proportionnelle à sa taille. Nous commenterons cette hypothèse dans la section qui suit.

Venons-en maintenant à une première formalisation statistique du modèle.

2.3 Formalisation statistique

Dans cette partie, nous mettons rigoureusement en place les outils statistiques de description du modèle ébauché en 2.2. Nous commençons par détailler le mode d'échantillonnage sans-remise pondéré des champs découverts au sein de l'ensemble des champs existant dans le sous-sol. Nous montrons notamment que l'ensemble $\{Y_1, ..., Y_n\}$ des découvertes est, conditionnellement au fait que n est connu, un échantillon d'une loi qui s'exprime au moyen de la fonction de pondération et de la loi de l'échantillon $\{X_1, ..., X_N\}$. Nous formulons les hypothèses nécessaires pour lever un problème d'identifiabilité du modèle et présentons enfin le modèle exponentiel auquel nous aboutissons pour décrire la loi de l'échantillon $\{Y_1, ..., Y_n\}$.

2.3.1 Mode d'échantillonnage

Dans un système pétrolier donné, il existe un nombre N inconnu de champs de taille supérieure à une unité[3]. Comme justifié par la partie 2.2.1, on supposera dans la suite, l'ensemble des logtailles $\mathbb{X} = \{X_1, ..., X_N\}$ modélisé par un échantillon d'une loi Exponentielle de paramètre α inconnu. Seul un nombre fini n_t de ces champs ont été découverts à la date t. Dans la suite, nous supposerons t fixée et omettrons alors la dépendance en cette variable dans les notations. Nous reviendrons sur les aspects dynamiques et *backfitting* dans la dernière section de ce chapitre ainsi que dans le chapitre 5 pour les applications.

Notons $\mathbb{Y} = \{Y_1, ... Y_n\}$ l'ensemble des logtailles des champs observés. C'est un "sous-échantillon" de $\{X_1, ..., X_N\}$ (au sens ensembliste), que nous dirons biaisé par effet taille (nous donnerons un sens mathématique précis à ce terme un peu plus loin). C'est à dire que la probabilité d'observer un champ de logtaille X_i donnée est fonction de X_i. Pour décrire ce mode de tirage nous travaillons donc conditionnellement à l'échantillon $\mathbb{X}$. Nous modélisons le biais dans le tirage au moyen d'une pondération π définie sur $\mathbb{X}$. Il existe plusieurs façons de concevoir un tel tirage sans remises. L'un des plus étudié est le tirage successif dont nous détaillons le protocole. Le lecteur peut se référer à [38] pour une revue très complète, quoique peu récente, de la théorie des tirages sans remises.

On tire les objets dans $\{X_1, ..., X_N\}$ un par un. À chaque X_i, $i = 1, ..., N$ est associée un poids $\pi(X_i) \in \mathbb{R}^+$ proportionnellement auquel il risque d'être sélectionné. Posons

$$\Pi_{\mathbb{X}} \;=\; \sum_{i=1}^{N} \pi(X_i).$$

[3]Notons que la valeur de cette unité n'a pas réellement d'importance étant donnée la propriété d'invariance par changement d'échelle de la loi de Lévy-Pareto (voir Proposition 1.2.4). En pratique, notre unité sera 1 ou éventuellement 10 Mbep.

Au premier tirage, un X_i, $i \in \mathbb{N}_N$ a pour probabilité $\pi^1(X_i)$ d'être tiré qui vaut

$$\pi^1(X_i) \;=\; \frac{\pi(X_i)}{\Pi_{\mathbb{X}}}.$$

Notons X_{i_1} l'élément tiré et posons $\check{\mathbb{X}}_i = \{X_1, ..., X_{i-1}, X_{i+1}, ..., X_N\}$. Pour le second tirage, chaque X_i pour $i \in \mathbb{N}_N \setminus \{i_1\}$ a une probabilité $\pi^2(X_i)$ d'être sélectionné qui s'exprime

$$\pi^2(X_i) \;=\; \frac{\pi(X_i)}{\sum_{j=1}^N \pi(X_j) - \pi(X_{i_1})} \;=\; \frac{\pi(X_i)}{\Pi_{\check{\mathbb{X}}_{i_1}}}.$$

Et ainsi de suite... Au $n^{\text{ième}}$ tirage successif, pour $i \in \mathbb{N}_N \setminus \{i_1, ..., i_{n-1}\}$ on obtient

$$\pi^n(X_i) \;=\; \frac{\pi(X_i)}{\sum_{j=1}^N \pi(X_j) - \sum_{k=1}^{n-1} \pi(X_{i_k})} \;=\; \frac{\pi(X_i)}{\Pi_{\check{\mathbb{X}}_{i_1,...,i_{n-1}}}}.$$

On peut ensuite s'intéresser, le nombre de tirages n étant fixé, à la probabilité que l'objet X_i ait été tiré à un rang inférieur ou égal à n. Dans le cas général non-équiprobable qui nous concerne, la Proposition suivante fournit une formule de calcul élémentaire et explicite de cette probabilité au moyen des seules pondérations initiales $(\pi(X_i))_{1 \le i \le N}$.

Proposition 2.3.1. *Soient $\mathbb{X} = \{X_1, ..., X_N\}$ et $\pi : \mathbb{X} \longrightarrow \mathbb{R}_*^+$ une fonction. Pour $n \in \mathbb{N}_N$, on note $\omega_{\mathbb{X}}^n(X_i)$ la probabilité qu'un X_i appartienne à un sous-ensemble de taille n issu d'un tirage successif dans $\mathbb{X}$ dont les probabilités d'inclusion au premier tirage sont proportionnelles aux $(\pi(X_i))_{i \in \mathbb{N}_N}$.*

On pose

$$\beta_{\mathbb{X}}(X_i) \;=\; \frac{\pi(X_i)}{\Pi_{\check{\mathbb{X}}_i}} \quad et \quad \delta_{\mathbb{X}}^n(X_i) \;=\; \frac{\omega_{\mathbb{X}}^n(X_i)}{\omega_{\mathbb{X}}^1(X_i)}.$$

On a alors

$$\delta_{\mathbb{X}}^n(X_i) \;=\; 1 + \sum_{\substack{j=1 \\ j \neq i}}^N \beta_{\mathbb{X}}(X_j) \delta_{\check{\mathbb{X}}_j}^{n-1}(X_i) \tag{2.3}$$

avec la convention $\delta_{\mathbb{X}}^0(X_i) = 0$. De plus,

$$\sum_{i=1}^N \omega_{\mathbb{X}}^n(X_i) \;=\; n. \tag{2.4}$$

Remarque (1) : la spécification $\omega^1_{\mathbb{X}}(X_i) = \pi(X_i)/\Pi_{\mathbb{X}}$ permet ensuite d'obtenir les $\omega^n_{\mathbb{X}}(X_i)$ pour tout $i = 1, ..., N$.

Remarque (2) : malgré le côté relativement élémentaire de la formule (2.3), nous n'avons pas trouvé de références qui en mentionnent l'existence. Nous verrons que sa forme récursive est pourtant, de notre point de vue, à la base d'importants travaux visant à fournir des estimations de ces probabilités d'inclusion au lieu de calculs exacts.

Preuve : L'équation (2.3) se montre par une récurrence basée sur un calcul de dénombrement. Notons (HR_n) cette équation qui représente notre hypothèse de récurrence au rang n.

(HR_1) est triviale : $\delta^1_{\mathbb{X}}(X_i) = 1$.

Pour éviter les lourdeurs de calcul, nous montrons (HR_2) $\Rightarrow$ (HR_3). Le cas général (HR_{n-1}) $\Rightarrow$ (HR_n) se démontrant grâce à un raisonnement analogue.

En notant $\tilde{\omega}^3_{\mathbb{X}}(X_i)$ la probabilité que X_i ait été tiré au troisième tirage exactement, on a

$$\omega^3_{\mathbb{X}}(X_i) \;=\; \omega^2_{\mathbb{X}}(X_i) + \tilde{\omega}^3_{\mathbb{X}}(X_i).$$

Or, la probabilité de tirer les éléments du triplet (X_j, X_k, X_i) dans cet ordre exactement est

$$\frac{\pi(X_j)}{\Pi_{\mathbb{X}}}\frac{\pi(X_k)}{\Pi_{\check{\mathbb{X}}_j}}\frac{\pi(X_i)}{\Pi_{\check{\mathbb{X}}_{j,k}}} \;=\; \omega^1_{\mathbb{X}}(X_i)\beta_{\mathbb{X}}(X_j)\beta_{\check{\mathbb{X}}_j}(X_k).$$

Donc

$$\tilde{\omega}^3_{\mathbb{X}}(X_i) \;=\; \omega^1_{\mathbb{X}}(X_i)\sum_{\substack{j=1\\ j\neq i}}^{N}\sum_{\substack{k=1\\ k\neq i,j}}^{N}\beta_{\mathbb{X}}(X_j)\beta_{\check{\mathbb{X}}_j}(X_k).$$

Puis, par (HR_2) on a

$$\begin{aligned}
\delta^3_{\mathbb{X}}(X_i) &= 1 + \sum_{\substack{j=1\\ j\neq i}}^{N}\beta_{\mathbb{X}}(X_j) + \sum_{\substack{j=1\\ j\neq i}}^{N}\beta_{\mathbb{X}}(X_j)\sum_{\substack{k=1\\ k\neq i,j}}^{N}\beta_{\check{\mathbb{X}}_j}(X_k)\\
&= 1 + \sum_{\substack{j=1\\ j\neq i}}^{N}\beta_{\mathbb{X}}(X_j)\left[1 + \sum_{\substack{k=1\\ k\neq i,j}}^{N}\beta_{\check{\mathbb{X}}_j}(X_k)\right]\\
&= 1 + \sum_{\substack{j=1\\ j\neq i}}^{N}\beta_{\mathbb{X}}(X_j)\delta^2_{\check{\mathbb{X}}_j}(X_i),
\end{aligned}$$

ce qui est exactement HR_3.

La preuve de la formule (2.4) peut se concevoir de plusieurs façons, notamment par des considérations élémentaires de dénombrement. Nous choisissons plutôt de la montrer par une récurrence qui fait pleinement appel à l'équation (2.3). Notons de nouveau (HR_n) l'hypothèse de récurrence qui consiste à postuler que (2.4) est vraie.

(HR_1) est évidente :

$$\sum_{i=1}^{N} \omega_{\mathbb{X}}^{1}(X_i) \;=\; \sum_{i=1}^{N} \frac{\pi(X_i)}{\Pi_{\mathbb{X}}} \;=\; 1.$$

Montrons (HR_{n-1}) $\Rightarrow$ (HR_n). Par (2.3) on a

$$\begin{aligned}
\sum_{i=1}^{N} \omega_{\mathbb{X}}^{n}(X_i) &= \sum_{i=1}^{N} \omega_{\mathbb{X}}^{1}(X_i) + \sum_{i=1}^{N} \omega_{\mathbb{X}}^{1}(X_i) \sum_{\substack{j=1 \\ j\neq i}}^{N} \beta_{\mathbb{X}}(X_j) \frac{\omega_{\check{\mathbb{X}}_j}^{n-1}(X_i)}{\omega_{\check{\mathbb{X}}_j}^{1}(X_i)} \\
&= 1 + \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j\neq i}}^{N} \frac{\pi(X_i)}{\Pi_{\mathbb{X}}} \frac{\pi(X_j)}{\Pi_{\check{\mathbb{X}}_j}} \frac{\Pi_{\check{\mathbb{X}}_j}}{\pi(X_i)} \omega_{\check{\mathbb{X}}_j}^{n-1}(X_i) \\
&= 1 + \sum_{j=1}^{N} \frac{\pi(X_j)}{\Pi_{\mathbb{X}}} \sum_{\substack{i=1 \\ i\neq j}}^{N} \omega_{\check{\mathbb{X}}_j}^{n-1}(X_i).
\end{aligned}$$

Puis, en appliquant(HR_{n-1}) à chaque $\omega_{\check{\mathbb{X}}_j}^{n-1}(X_i)$ il vient que

$$\sum_{i=1}^{N} \omega_{\mathbb{X}}^{n}(X_i) \;=\; 1 + (n-1) \sum_{j=1}^{N} \frac{\pi(X_j)}{\Pi_{\mathbb{X}}} \;=\; n.$$

♣

La figure 2.3 montre les résultats (en rouge) d'un tirage successif biaisé par effet taille pour diverses valeurs de n sur un 500-échantillon d'une loi de Lévy-Pareto de paramètre 0,7 (en bleu) avec une probabilité d'inclusion au premier tirage $\pi(x) = \sqrt{x}$. La figure complémentaire 2.4 montre l'évolution des probabilités d'inclusion empiriques $(\omega^n)_{1\leq n\leq 500}$ par classe de taille.

La formule récursive (2.3) est pourtant inutilisable en pratique car elle requiert un temps de calcul ordinateur et une capacité mémoire gigantesques. Pour le montrer, nous allons évaluer un ordre de grandeur du nombre d'opérations élémentaires (additions et multiplications) nécessaires à un ordinateur pour l'implémenter.

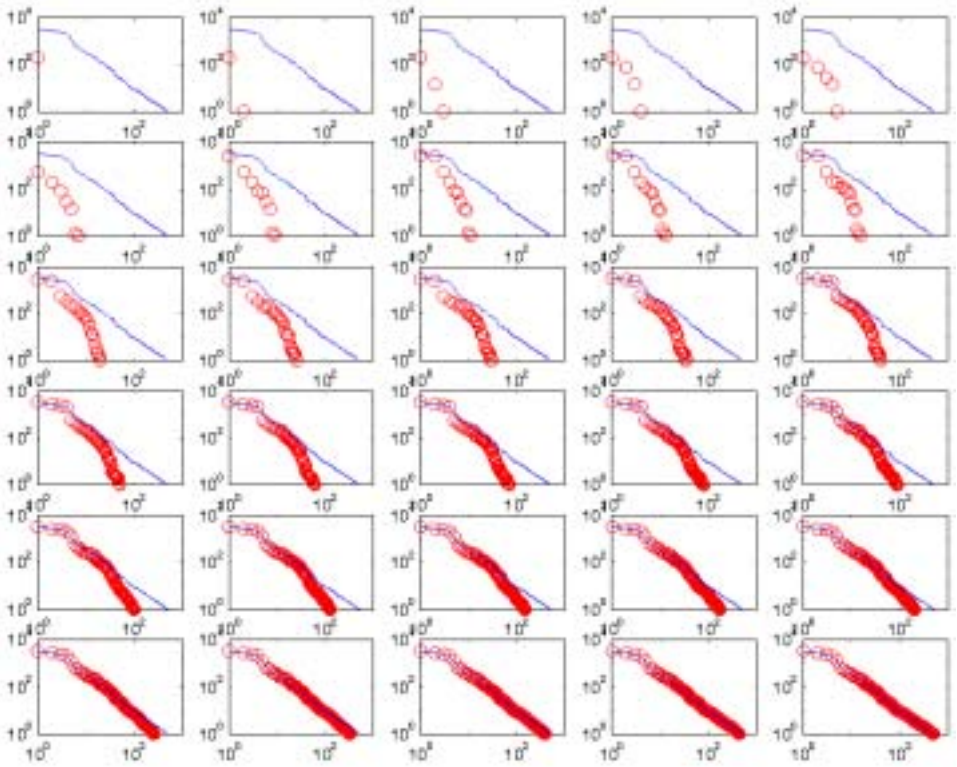

FIG. 2.3 – Simulation d'un tirage successif biaisé par effet taille. En abscisse se trouve le rang dans la statistique d'ordre et en ordonnée, la taille.

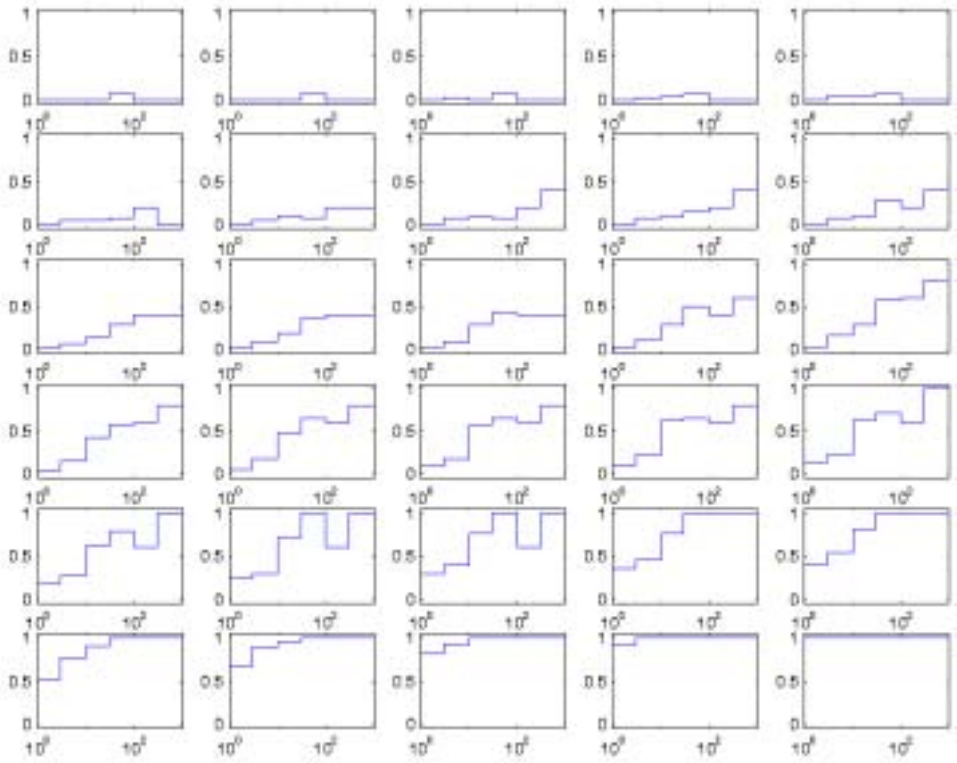

FIG. 2.4 – Probabilités d'inclusion empiriques associées aux tirages de la figure 2.3. En abscisse se trouve la taille et en ordonnée, la probabilité d'inclusion associée.

Soit $C(N,n)$ le nombre d'opérations élémentaires nécessaire au calcul d'un $\delta^n_{\mathbb{X}}(X_i)$, pour $i = 1, ..., N$. L'équation (2.3) implique la relation de récurrence suivante :

$$\begin{aligned} C(N,n) &= (N-1)(N+1+C(N-1,n-1)) \\ &\geq (N-1)^2 + (N-1)C(N-1,n-1) \\ &\vdots \\ &\geq \sum_{i=1}^{n-1}(N-i)^i + \frac{(N-1)!}{(N-n)!}\,C(N-n+1,1) \\ &\geq \sum_{i=1}^{n-1}(N-i)^i + \frac{(N-1)!}{(N-n-1)!}. \end{aligned}$$

Le deuxième terme du membre de droite de la dernière inégalité est nettement prépondérant. On en déduit que la complexité algorithmique croit au moins "factoriellement vite" en n et N. À titre d'exemple, un simple programme Matlab® (cf. 6.1) d'application de la formule (2.3) pour $N = 25$ et $n = 10$ a une complexité algorithmique supérieure à 10 millions de milliards d'opérations et dépasse les capacités de calcul d'un ordinateur actuel, même très puissant. Or, ne perdons pas de vue que les ordres de grandeur de nos applications pratiques sont de $N = 1000$ et $n = 100$... Cette difficulté justifie de nombreux travaux qui permettent de contourner, autant que possible, de fastidieux (voire inenvisageables) calculs pour évaluer les $(\omega^n(X_i))_{1\leq i\leq N}$.

Le travail qui semble fondateur en ce domaine est celui de Bengt Rosén [74] et [75]. Il consiste à fournir des estimateurs simples des $\omega^n(X_i)$ connaissant les $\pi(X_i)$ pour i dans $\mathbb{N}_N$. Ces estimateurs ont été repris et appliqués dans de nombreux domaines où les tirages à probabilités d'inclusion inégales sont utilisés (biométrie, astronomie, estimations de type capture-recapture, sondages évidemment, etc.). Concernant l'évaluation des réserves pétrolières, citons entre autres Holst [39], Gordon [35] et Bickel *et al.* [13].

Ces derniers se placent dans une situation où la population dans laquelle le tirage est effectué est finie, d'effectif éventuellement inconnu, ce qui est notre cas. La connaissance (ou l'estimation) des probabilités d'inclusion $(\omega^n(X_i))_{i\in\mathbb{N}_N}$ est la clef de l'estimation des réserves ultimes via notamment les estimateurs de type Horvitz-Thompson [40] comme nous le verrons dans le chapitre 3. Cependant, tous les travaux cités ci-dessus rencontrent tous l'écueil de la connaissance *a priori* des $(\pi(X_i))_{i\in\mathbb{N}_N}$, qui est très délicate à justifier en pratique.

Grâce aux estimateurs de Rosén [74], elle est équivalente (à estimation près) à la connaissance *a priori* des $(\omega^n(X_i))_{i\in\mathbb{N}_N}$. Dans leur article, Bickel *et al* [13] supposent de plus que la fonction π de pondération représentant les probabilités d'inclusion au premier tirage est l'identité. On parle alors de tirage

directement proportionnel à la taille. D'autres auteurs, dont Lee et Wang [51], ont considéré pour leur part que π appartient à une classe paramétrée de fonctions (typiquement la classe des fonctions puissance). Ce choix est arbitraire et donc largement discutable. Pourquoi ne pas considérer plutôt une fonction logarithmique, qui aurait l'avantage de fournir une pondération globalement indépendante de la dimension des observations dans le sens suivant : le poids des champs pris en compte est généralement une unité de volume, le baril. On pourrait concevoir que l'unité soit plutôt surfacique, comme l'étendue cartographique du champ ; ou encore qu'elle soit de dimension 1 comme le diamètre du champ, vu comme un ensemble compact. Un champ qui aurait une taille T pour le poids "de dimension 1" aurait une taille de l'ordre T^2 ou T^3 pour les poids surfaciques ou volumiques. Il est alors douteux de considérer une pondération π qui dépende si fortement de la dimension de la mesure de taille choisie alors qu'il s'agit toujours du même champ. Prendre un log élimine cette dépendance puisque $\log(T)$, $\log(T^2)$ et $\log(T^3)$ sont proportionnels et définissent donc des probabilités d'inclusion égales.

Cependant, prenant en considération ces difficultés de justification *a priori* de l'appartenance de π (donc de ω^n) à une famille paramétrée, nous choisissons une voie non-paramétrique pour évaluer les $(\omega^n(X_i))_{i\in\mathbb{N}_N}$ directement. En ce sens, notre travail se rapproche de celui de Sun et Woodroofe [77], qui considèrent que la loi sous-jacente des tailles des champs appartient à un modèle exponentiel et que la pondération ω^n est croissante et atteint son maximum (égal à 1) en la plus grande observation. Ils estiment ensuite ω^n de façon semi-paramétrique avec hypothèse de monotonie par maximum de vraisemblance pénalisé au sens de Good et Gaskins [34][4] pour corriger le manque de régularité de leur estimateur. Le problème du "réglage" optimal de cette pénalité est soulevé mais cependant pas abordé.

Soit $X_{(1)} \leq \cdots \leq X_{(N)}$ la statistique d'ordre associée à l'échantillon $\mathbb{X}$. La loi du *"big stuff gets found first"* s'interprète comme le fait que les probabilités d'inclusion $\omega^n(X_{(i)})$ sont croissantes en fonction de i (plus un objet est gros, et plus il a une probabilité d'avoir été découvert importante). Autrement dit, en déconditionnant par rapport à $\mathbb{X}$ et en voyant la probabilité d'inclusion $\omega^n : \text{Support}(X) \longrightarrow]\,0\,;1]$ comme une fonction de la taille de X, cela revient à considérer ω^n comme croissante. Nous serons amenés à imposer cette contrainte de forme dans certaines de nos estimations. Il pourra notamment être intéressant de comparer les estimateurs contraints et non contraints à la monotonie. Dans la suite, nous allons justifier de la nécessité de travailler conditionnellement au fait que n est fixé (donc t aussi) et nous omettrons

[4] Un terme de pénalité est ajouté à la log-vraisemblance du modèle pour rendre plus petites que la normale les vraisemblances associées à certaines formes indésirables d'estimateurs. En l'occurence ici, il s'agit de pénaliser les estimateurs de ω^n à la croissance trop forte.

alors la dépendance de ω en cette variable dans l'écriture.

2.3.2 Modèle de censure

Nous formalisons ici notre modèle de tirage de type superpopulation comme un modèle de censure de certaines observations (les champs non encore découverts). Les résultats que nous montrons dans cette section sont manifestement considérés comme classiques ou naturels par les auteurs qui nous les ont inspirés (essentiellement Bickel *et al.* [13] et Sun et Woodroofe [77]). Ils sont effectivement très intuitifs et nous n'avons pas été en mesure de trouver des références qui mettent en évidence le véritable problème statistique que pose cette modélisation. Un travail rigoureux à ce sujet se trouve ainsi pleinement justifié.

Pour rendre compte du tirage sans remise, nous considérons les champs non encore découverts comme censurés : à tout X_i, pour $i = 1, ..., N$, on associe une variable de censure ε_i de loi de Bernoulli, conditionnellement à X_i, de paramètre $\omega(X_i)$:

$$\left\{ \begin{array}{rcl} \mathbb{P}(\varepsilon_i = 1 \,|\, X_i) & = & \omega(X_i) \\ \mathbb{P}(\varepsilon_i = 0 \,|\, X_i) & = & 1 - \omega(X_i). \end{array} \right. \tag{2.5}$$

Un X_i est observé si la variable ε_i associée vaut 1 (donc avec une probabilité $\omega(X_i)$) et non observé sinon. Dans la suite et d'une façon générale, nous désignerons par la lettre Y les objets observés. Définir correctement d'un point de vue probabiliste ces individus observés demande quelques préalables :

Définition 2.3.2. *Soit $\mathbb{X} = \{X_1, ..., X_N\}$ un échantillon de loi à densité f et une fonction ω : Support$(X) \longrightarrow]\,0\,;1]$. On y associe un échantillon de censure $\varepsilon = \{\varepsilon_1, ..., \varepsilon_N\}$ comme ci-dessus donné par (2.5).*

Soit $\tilde{\varepsilon}$ la variable aléatoire à valeurs dans $\mathcal{P}(\mathbb{N}_N)$ définie par $\tilde{\varepsilon} = \{i \in \mathbb{N}_N \,/\, \varepsilon_i = 1\}$.

Soit $n \geq 1$, en se plaçant conditionnellement au fait que $|\tilde{\varepsilon}| = n$, on définit l'ensemble des observés par $\mathbb{Y} = \{Y_1, ..., Y_n\} = \{X_{\tilde{\varepsilon}_1}, ..., X_{\tilde{\varepsilon}_n}\} = \{X_j,\, j \in \tilde{\varepsilon}\}$.

Nous dirons alors que $\mathbb{Y}$ est un ω-sous-échantillon de $\mathbb{X}$.

Remarque : sans conditionnement particulier, le nombre d'observations $|\tilde{\varepsilon}| = \varepsilon_1 + \cdots + \varepsilon_N$ est aléatoire. Nous montrons par la suite qu'il est de loi Binômiale de paramètres N et $\mathbb{E}_f(\omega(X)))$.

Cherchons maintenant la loi des observés et montrons en particulier leur indépendance, conditionnellement au fait que leur nombre n est fixé.

Proposition 2.3.3. *L'ensemble* $\{Y_1, ..., Y_n\}$ *des observés défini en 2.3.2 forme un n-échantillon de loi à densité* f_ω *donnée par*

$$f_\omega(x) = \frac{\omega(x)f(x)}{\mathbb{E}_f(\omega(X))} \quad \forall\, x \in \text{Support}(X) \tag{2.6}$$

En particulier, les $(Y_i)_{1\leq i\leq n}$ *sont conditionnellement indépendantes.*

Remarque (1) : il est inutile de préciser que $\mathbb{E}(\omega(X))$ doit exister puisque, par définition, $\omega(X)$ est bornée.

Remarque (2) : la Proposition 2.3.3 montre que la pondération par ω perturbe la densité f en la rendant petite là où ω est petite et plus importante là où elle est élevée. La quantité $\mathbb{E}_f(\omega(X))$ apparaît simplement comme constante de normalisation pour que la fonction f_ω soit effectivement une densité.

Preuve : nous cherchons la loi jointe de $\{Y_1, ..., Y_n\}$ conditionnellement à $|\tilde{\varepsilon}| = n$. Soit $\{A_1, ..., A_n\} \in \mathcal{P}(\text{Support}(X))^n$, par la formule de Bayes on a :

$$\begin{aligned}\mathbb{P}(Y_1 \in A_1, ..., Y_n \in A_n \,\big|\, |\tilde{\varepsilon}| = n) \\ = \frac{\mathbb{P}(\{Y_1 \in A_1, ..., Y_n \in A_n\} \cap \{|\tilde{\varepsilon}| = n\})}{\mathbb{P}(|\tilde{\varepsilon}| = n)}.\end{aligned} \tag{2.7}$$

Cherchons d'abord la loi de $|\tilde{\varepsilon}|$. Les $(\varepsilon_i)_{1\leq i\leq N}$ sont de loi de Bernoulli indépendantes conditionnellement aux $(X_i)_{1\leq i\leq N}$, par propriété de l'espérance conditionnelle et des lois de Bernoulli, elles sont donc aussi inconditionnellement de loi de Bernoulli de paramètre $\mathbb{E}_f(\omega(X))$:

$$\begin{aligned}\mathbb{P}(\varepsilon_i = 1) &= \mathbb{E}(\varepsilon_i) \\ &= \mathbb{E}_f(\mathbb{E}(\varepsilon_i \,|\, X_i) \\ &= \mathbb{E}_f(\mathbb{P}(\varepsilon_i = 1 \,|\, X_i)) \\ &= \mathbb{E}_f(\omega(X_i)).\end{aligned}$$

Montrons de plus qu'elles sont indépendantes.

Soit $\tau \in \{0,1\}^N$, on définit $\tilde{\tau} = \{i \in \mathbb{N}_N \,/\, \tau = 1\}$. On a

$$\begin{aligned}\mathbb{P}(\varepsilon = \tau) &= \int_{\mathbb{R}^N} \mathbb{P}(\varepsilon = \tau \,|\, X_1 = x_1, ..., X_N = x_N) \prod_{i=1}^N f(x_i)\, dx_1 ... dx_N \\ &= \int_{\mathbb{R}^N} \prod_{i=1}^N \mathbb{P}(\varepsilon_i = \tau_i \,|\, X_1 = x_1, ..., X_N = x_N) f(x_i)\, dx_1 ... dx_N \\ &= \int_{\mathbb{R}^N} \prod_{i\in\tilde{\tau}} \omega(x_i) f(x_i) \prod_{j\notin\tilde{\tau}} (1-\omega(x_j)) f(x_j)\, dx_1 ... dx_N \\ &= \mathbb{E}_f(\omega(X))^{|\tilde{\tau}|} (1 - \mathbb{E}_f(\omega(X)))^{N-|\tilde{\tau}|}.\end{aligned}$$

Cette dernière égalité assure l'indépendance des $(\varepsilon_i)_{1\leq i\leq N}$.

Remarquons ensuite que la probabilité (non conditionnelle) de l'évènement $\{\varepsilon = \tau\}$ ne dépend que de $|\tilde{\tau}| = \tau_1 + \cdots + \tau_N$ et que

$$\{|\tilde{\varepsilon}| = n\} \;=\; \bigcup_{\substack{\tau\in\{0,1\}^N \\ |\tilde{\tau}|=n}} \{\varepsilon = \tau\}. \tag{2.8}$$

Comme l'ensemble des C_N^n évènements de la réunion de droite est un ensemble d'évènements incompatibles, il en résulte que pour tout n de $\mathbb{N}_N$ il vient

$$\mathbb{P}(|\tilde{\varepsilon}| = n) \;=\; \mathrm{C}_N^n \mathbb{E}_f\left(\omega(X)\right)^{|\tilde{\tau}|}(1 - \mathbb{E}_f\left(\omega(X)\right))^{N-|\tilde{\tau}|} \tag{2.9}$$

et $|\tilde{\varepsilon}|$ est donc de loi Binômiale de paramètres N et $\mathbb{E}_f\left(\omega(X)\right)$.

Nous allons maintenant appliquer une méthode analogue pour calculer le terme $\mathcal{P} = \mathbb{P}(\{Y_1 \in A_1, ..., Y_n \in A_n\} \cap \{|\tilde{\varepsilon}| = n\})$. Par (2.8) on a :

$$\mathcal{P} = \sum_{\substack{\tau\in\{0,1\}^N \\ |\tilde{\tau}|=n}} \mathbb{P}(\{Y_1 \in A_1, ..., Y_n \in A_n\} \cap \{|\tilde{\varepsilon}| = n\} \cap \{\varepsilon = \tau\}).$$

Mais pour tout τ tel que $|\tilde{\tau}| = n$ on a $\{\varepsilon = \tau\} \subset \{|\tilde{\varepsilon}| = n\}$. L'intersection est donc réduite au premier évènement. D'où

$$\begin{aligned}
\mathcal{P} &= \sum_{\substack{\tau\in\{0,1\}^N \\ |\tilde{\tau}|=n}} \mathbb{P}(\{Y_1 \in A_1, ..., Y_n \in A_n\} \cap \{\varepsilon = \tau\}) \\
&= \sum_{\substack{\tau\in\{0,1\}^N \\ |\tilde{\tau}|=n}} \mathbb{P}(\{X_{\tilde{\tau}_1} \in A_1, ..., X_{\tilde{\tau}_n} \in A_n\} \cap \{\varepsilon = \tau\}) \\
&= \sum_{\substack{\tau\in\{0,1\}^N \\ |\tilde{\tau}|=n}} \int_{\mathbb{R}^N} \prod_{i=1}^{n} w(x_{\tilde{\tau}_i}) \mathbb{1}_{A_i}\left(x_{\tilde{\tau}_i}\right) f(x_{\tilde{\tau}_i}) \prod_{j\notin\tilde{\tau}} (1 - w(x_j)) f(x_j)\, dx_1 ... dx_N \\
&= \sum_{\substack{\tau\in\{0,1\}^N \\ |\tilde{\tau}|=n}} (1 - \mathbb{E}_f\left(\omega(X)\right))^{N-n} \int_{A_1\times\cdots\times A_n} \prod_{i=1}^{n} w(x_{\tilde{\tau}_i}) f(x_{\tilde{\tau}_i})\, dx_{\tilde{\tau}_1} ... dx_{\tilde{\tau}_n}.
\end{aligned}$$

Les variables d'intégration étant muettes, l'intégrale dans la somme ne dépend que de n et l'on en déduit que

$$\mathcal{P} \;=\; \mathrm{C}_N^n (1 - \mathbb{E}_f\left(\omega(X)\right))^{N-n} \int_{A_1\times\cdots\times A_n} \prod_{i=1}^{n} w(x_i) f(x_i)\, dx_1 ... dx_n. \tag{2.10}$$

Nous concluons ensuite la preuve en injectant (2.10) et (2.9) dans (2.7) :

$$\mathbb{P}(Y_1 \in A_1, ..., Y_n \in A_n \,\big|\, |\tilde{\varepsilon}| = n) = \frac{1}{\mathbb{E}_f\left(\omega(X)\right)^n} \int_{A_1 \times \cdots \times A_n} \prod_{i=1}^{n} w(x_i) f(x_i)\, dx_1 ... \, dx_n.$$

ce qui donne l'indépendance conditionnelle des $(Y_i)_{1 \leq i \leq n}$ et la forme voulue de la densité de leur loi. ♣

2.3.3 Problème d'identifiabilité

On peut voir sur la formule (2.6) que pour tout $c \neq 0$, $f_\omega = f_{c\omega}$. Le changement $\omega \to c\omega$ laisse donc invariante la densité des observés[5]. Dans nos estimations futures de la fonction de pondération inconnue ω, sans hypothèse complémentaire, nous nous heurterons à un problème d'identifiabilité du modèle. C'est pourquoi nous formulons les hypothèses qui suivent, distinguant le cas ω libre de contrainte de monotonie du cas ω contrainte à la croissance :

- **Hypothèse (Hnc) :** ω est supposée telle que $\max\{\omega(Y_i),\, i = 1, ..., n\} = 1$, *i.e.* le champ qui réalise le max a été trouvé de façon sûre : il n'existe aucun autre champ de même taille dans $\{X_1, ..., X_N\}$.

Cette hypothèse, qui semble très formelle, prend tout son sens lorsque l'on impose la contrainte de monotonie à ω :

- **Hypothèse (Hc) :** ω est supposée croissante et $\omega(X_{(N)}) = \omega(Y_{(n)}) = 1$, *i.e.* le plus gros champ a été trouvé de façon sûre.

Comme nous l'avons développé dans le chapitre 1, cette dernière hypothèse est naturelle lorsque l'on veut effectuer *in fine* des estimations de réserves, compte tenu du caractère hautement dispersif de la loi de Lévy-Pareto (voir Lepez et Mandonnet [53] pour quelques mises en évidence empiriques).

Définition 2.3.4. *Dans le cas particulier de la densité f_α de la loi exponentielle de paramètre α, nous noterons $f_{\alpha,\omega}$ la densité des observés donnée par (2.6).*

En pratique, nous allons estimer cette densité des variables observées afin d'en déduire des estimations de α et surtout ω dans un premier temps. Ces dernières nous permettrons ensuite de construire des estimateurs de l'effectif global N et du cumul des réserves S pour finir.

[5] Ceci est compatible avec le fait que la fonction ω n'est définie qu'à une constante multiplicative près par rapport à π, la fonction de pondération au premier tirage.

2.4 Rétroprévision : le modèle comme outil d'analyse de l'efficacité d'exploration

Notre modèle étant construit, quel type d'information peut-on récupérer de l'estimation à venir de la fonction ω ?

Rappelons que la fonction ω dépend du temps. Fixons nous une date t d'étude et notons alors la dépendance en cette variable : ω^t. Cette fonction décrit la pondération affectée à chaque champ de la population parente[6] de taille x donnée proportionnellement à laquelle il a été découvert avant la date t. Au tout début de l'exploration pétrolière, très peu de champs ont été découvert, les probabilités d'inclusion étaient donc faibles pour tous les éléments de la population parente, excepté pour les éléments de grande taille, peu nombreux et trouvés rapidement. À l'inverse, lorsque l'exhaustion est proche, les probabilités d'avoir été découvert sont élevées pour tous les individus.

L'évolution au cours du temps de la fonction ω traduit donc le *processus de découverte* des champs du système pétrolier considéré.

À supposer que l'on se soit fixé une origine des temps à la date $t = 0$, il est clair que le processus $(\omega^t)_{t\in\mathbb{R}^+}$ est un processus croissant en t. C'est même un processus croissant de fonctions croissantes si, qui plus est, on se place sous l'hypothèse (Hc). Il est alors intéressant de regarder comment croissent les fonctions ω^t en fonction de t.

Grâce à cette information, on peut analyser l'efficacité *a posteriori* de l'exploration pétrolière sur un système donné. En effet, détaillons deux cas. Dans le premier cas illustré par la figure 2.5, les fonctions ω^t sont "lentement croissantes". Ceci traduit le fait que même si les plus gros objets ont eu tendance à avoir été découverts avec une plus forte probabilité, nombre de champs de taille plus modeste ont eux aussi été mis au jour.

A l'inverse, sur le cas présenté par la figure 2.6, l'efficacité maximale d'exploration aurait supposé que l'on ne découvre exclusivement que les plus grands champs restant. Ceci implique alors des formes de courbes type courbes en S, voire en marche d'escalier de 0 à 1 pour les ω^t.

L'étude "*backfit*", ou rétroprévision, des courbes ω^t fournit donc aussi un outil historique d'étude de l'efficacité d'exploration qu'il peut-être particulièrement pertinent de mettre au regard d'indices classiques d'efforts d'exploration[7] dans une région.

Mais avant toutes choses, il faut pouvoir effectivement être en mesure d'estimer à t fixée, la fonction ω^t sous l'une ou l'autre des hypothèses (Hnc) ou (Hc), ce qui est objet du chapitre suivant.

[6] Au sens de la définition 1.3.1.

[7] Comme le nombre de puits d'explorations forés au cours du temps par exemple.

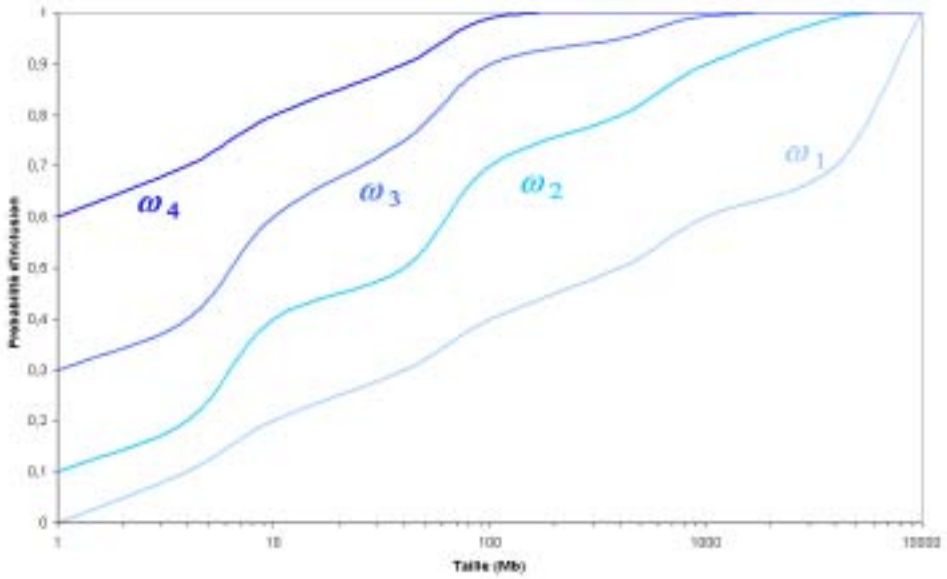

FIG. 2.5 – Courbes représentatives de la fonction ω_t pour $t = 1, 2, 3, 4$ dans le cas d'une exploration peu efficace.

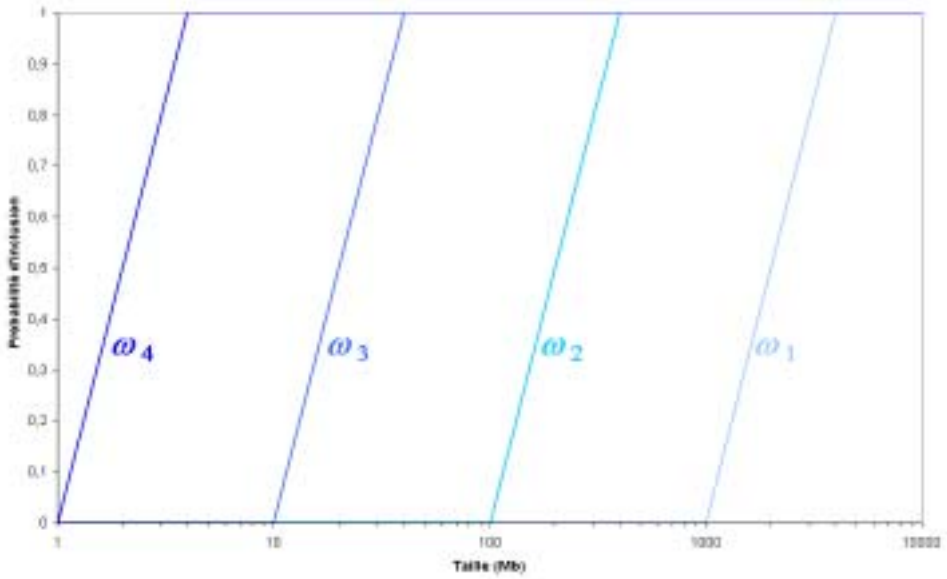

FIG. 2.6 – Courbes représentatives de la fonction ω_t pour $t = 1, 2, 3, 4$ dans le cas d'une exploration très efficace.

Chapitre 3

Estimation dans un modèle

À partir de cet instant, $n \geq 1$ étant donné, conformément à la section 2.3.2, nous supposons que nous travaillons conditionnellement au fait que $|\tilde{\varepsilon}| = n$, *i.e.* que n logtailles ont été effectivement observées.

Le but de ce chapitre est de présenter les estimateurs de la fonction de pondération ω ainsi que du nombre de champs restant à découvrir en fonction de classes de tailles fixées à l'avance.

Nous modélisons la loi des LogTailles de la population parente (*i.e.* des champs du sous-sol) par une loi exponentielle de paramètre α inconnu et le biais dans le tirage ω (interprété comme les probabilités d'inclusion des observations) par une fonction de pondération constante par morceaux inconnue, à valeurs dans $[\,0\,;1\,]$ et définie sur une partition fixée. Le choix de la partition la mieux adaptée fait l'objet du chapitre 4.

Dans la première partie de ce chapitre, nous nous appuyons sur la proposition 2.3.3 pour construire la forme précise de la densité $f_{\alpha,\omega}$ (cf. définition 2.3.4) que nous allons chercher à estimer par maximum de vraisemblance. Nous résolvons ensuite les équations de vraisemblance du modèle, d'abord sans intégrer la contrainte de monotonie en ω, puis en la rendant active. Nous montrons notamment un résultat peu intuitif sur le fait que le second cas peut facilement se déduire du premier. Nous concluons ce chapitre en déduisant de l'estimation de ω et de la théorie des sondages les estimateurs finaux du nombre de champs restant à découvrir dans chaque classe de taille.

3.1 Présentation du modèle exponentiel

Nous cherchons à estimer le couple (α, ω), de façon non-paramétrique en ω, par maximum de vraisemblance. De manière classique, nous verrons que la résolution des équations de vraisemblance montre que l'estimateur du maximum de vraisemblance $\hat{\omega}$ de ω est nécessairement constant entre deux observations et nul en dehors de l'intervalle $[\min_{1\leq i\leq n} Y_i\,;\,\max_{1\leq i\leq n} Y_i\,]$. Nous pourrons

donc, dès lors, considérer que ω est une fonction constante par morceaux, dont les sauts se situent sur les observations $\{Y_1, ..., Y_n\}$. La localisation de ces sauts est sans importance pour le travail théorique d'optimisation qui suit.

Nous supposerons simplement donnée une partition finie m quelconque de l'intervalle $I = [\min_{1\leq i\leq n} Y_i \,;\, \max_{1\leq i\leq n} Y_i\,]$, définissant des classes de taille, que l'on écrit

$$m \;=\; \{I_1, ..., I_{|m|}\} \quad \text{avec} \quad I \;=\; I_1 \cup \cdots \cup I_{|m|}.$$

3.1.1 Densité des observations

Nous donnons dans cette section deux écritures de la densité des observations $f_{\alpha,\omega}$ (voir la définition2.3.4).

Suivant les notations ci-dessus, la fonction ω s'écrit

$$\omega \;=\; \sum_{i=1}^{|m|} \omega_{I_i} \mathbb{1}_{I_i} \quad \text{ou encore} \quad \omega \;=\; \sum_{I\in m} \omega_I \mathbb{1}_I\,. \tag{3.1}$$

Pour obtenir une forme adéquate de la densité des observations $f_{\alpha,\omega}$ (voir la définition2.3.4), nous définissons la mesure exponentielle et la densité associée :

Définition 3.1.1. *On appelle mesure exponentielle de paramètre α et l'on note μ_α la mesure sur $\mathbb{R}^+$ définie par*

$$d\mu_\alpha(x) \;=\; f_\alpha(x)\, d\mu(x) \;=\; \alpha e^{-\alpha x} \mathbb{1}_{\mathbb{R}^+}(x)\, d\mu(x)$$

où μ désigne la mesure de Lebesgue sur $\mathbb{R}^+$.

Par la formule (3.1), la définition 3.1.1 et la Proposition 2.3.3, il vient que la fonction $f_{\alpha,\omega}$ s'écrit, pour $x \geq 0$

$$f_{\alpha,\omega}(x) \;=\; \sum_{I\in m} \frac{\omega_I}{\mathbb{E}_{f_\alpha}(\omega(X))}\, \alpha e^{-\alpha x} \mathbb{1}_I\,(x)$$

avec

$$\mathbb{E}_{f_\alpha}(\omega(X)) \;=\; \sum_{I\in m} \omega_I \mu_\alpha(I). \tag{3.2}$$

Nous allons maintenant donner un paramétrage de ω qui nous permettra d'interpréter plus facilement la densité $f_{\alpha,\omega}$ en termes de modèle exponentiel.

Lemme 3.1.2. *Soit* $\Omega^m = \left\{(\omega_I)_{I\in m} \in]\,0\,;1]^{|m|} \,/\, \max_{I\in m} \omega_I = 1\right\}$.

L'application

$$\begin{array}{rcl} \Omega^m & \longrightarrow & \left(\mathbb{R}_*^+\right)^{|m|} \\ \omega = (\omega_I)_{I\in m} & \longmapsto & e^\theta = (e^{\theta_I})_{I\in m} = \left(\dfrac{\omega_I}{\mathbb{E}_{f_\alpha}(\omega(X))}\right)_{I\in m} \end{array}$$

est une bijection ensembliste dont la réciproque est donnée par

$$\begin{array}{rcl} \left(\mathbb{R}_*^+\right)^{|m|} & \longrightarrow & \Omega^m \\ e^\theta = (e^{\theta_I})_{I\in m} & \longmapsto & \omega = (\omega_I)_{I\in m} = \left(\dfrac{e^{\theta_I}}{\max\limits_{J\in m} e^{\theta_J}}\right)_{I\in m} . \end{array}$$

Remarque : l'équation (3.2) montre qu'en particulier on a

$$\sum_{I\in m} e^{\theta_I}\mu_\alpha(I) \;=\; 1. \tag{3.3}$$

Preuve : Pour tout I de m, par (3.2) et par définition de e^{θ_I}, on a

$$e^{\theta_I}\sum_{J\in m}\omega_J\mu_\alpha(J) - \omega_I \;=\; 0,$$

ce qui se traduit matriciellement par

$$\left(\begin{bmatrix} e^{\theta_{I_1}} \\ \vdots \\ e^{\theta_{I_{|m|}}} \end{bmatrix} \begin{bmatrix} \mu_\alpha(I_1) & \cdots & \mu_\alpha\left(I_{|m|}\right) \end{bmatrix} - \mathrm{Id}_{|m|}\right) \begin{bmatrix} \omega_{I_1} \\ \vdots \\ \omega_{I_{|m|}} \end{bmatrix} \;=\; \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}. \tag{3.4}$$

Posons

$$V = \begin{bmatrix} e^{\theta_1} \\ \vdots \\ e^{\theta_{|m|}} \end{bmatrix} \begin{bmatrix} \mu_\alpha(I_1) & \cdots & \mu_\alpha\left(I_{|m|}\right) \end{bmatrix}.$$

L'équation (3.4) implique donc que le vecteur $(\omega_{I_1}, ..., \omega_{I_{|m|}})$ appartient à $\mathrm{Ker}(V - \mathrm{Id}_{|m|})$. Or, V est de rang 1 et sa trace vaut 1 par (3.3). La matrice $V - \mathrm{Id}_{|m|}$ possède donc 0 et 1 pour valeurs propres, de multiplicités respectives 1 et $|m| - 1$. L'espace $\mathrm{Ker}(V - \mathrm{Id}_{|m|})$ est donc de dimension 1.

Par ailleurs le vecteur $(e^{\theta_{I_1}}, ..., e^{\theta_{I_{|m|}}})$ appartient à ce noyau et l'engendre donc. Ainsi, il existe $c \in \mathbb{R}_*$ tel que pour tout I de m on ait $\omega_I = ce^{\theta_I}$. Puis, on interprète la contrainte de majoration de ω :

$$\max_{I\in m} \omega_I = 1 \quad \Rightarrow \quad c = \frac{1}{\max_{I\in m} e^{\theta_I}},$$

ce qui achève la preuve. ♣

Utilisant les $(e^{\theta_I})_{I\in m}$, on peut alors écrire

$$f_{\alpha,\omega}(x) \;=\; \sum_{I\in m} e^{\theta_I} \alpha e^{-\alpha x} \mathbb{1}_I(x).$$

qui s'interprète comme élément d'un modèle exponentiel que nous allons commenter.

3.1.2 Interprétation statistique

Grâce au Lemme 3.1.2, il est légitime de reparamétrer la fonction $f_{\alpha,\omega}$ en une fonction f_{α,e^θ} en posant naturellement

$$f_{\alpha,e^\theta} \;=\; f_{\alpha,\omega}.$$

Il vient alors

$$f_{\alpha,e^\theta}(x) \;=\; \sum_{I\in m} \exp(\theta_I + \log\alpha - \alpha x)\mathbb{1}_I(x) \tag{3.5}$$

$$\;=\; \left(\sum_{I\in m} e^{\theta_I}\mathbb{1}_I(x)\right) f_\alpha(x). \tag{3.6}$$

Ainsi, par l'équation (3.5), f_{α,e^θ} se réalise comme une exponentielle de polynômes par morceaux contre la mesure de Lebesgue. Les coefficients des termes de degré 0 sont les $(\theta_I + \log\alpha)_{I\in m}$ et les coefficients des termes de degré 1 sont tous égaux à α.

Une première façon d'interpréter notre modèle est effectivement de le voir comme un modèle exponentiel de polynômes par morceaux de degré 1 contraint, en particulier à l'égalité des termes de degré 1. De plus, Sous l'hypothèse (Hc), on réclame la croissance des termes de degré 0.

Par l'équation (3.6), on peut aussi interpréter la fonction f_{α,e^θ} comme étant un élément d'un modèle exponentiel de polynômes par morceaux, de degré 0 cette fois, mais contre la mesure exponentielle de paramètre α (inconnu) ; autrement dit un histogramme contre la mesure μ_α. Dans le cadre de l'hypothèse (Hc), les termes de degré 0 sont aussi contraints à la monotonie.

Ces deux visions complémentaires vont nous permettre d'interpréter de façon très intuitive les résultats des résolutions des équations de vraisemblance sur $f_{\alpha,e^\theta} = f_{\alpha,\omega}$.

3.2 Résolution des équations de vraisemblance

Dans cette partie, nous montrons que le problème de maximisation de la vraisemblance, qu'il soit contraint ou non à la monotonie en la suite $e^\theta = \left(e^{\theta_I}\right)_{I\in m}$ admet une unique solution. Autrement dit, les modèles exponentiels de polynômes par morceaux basés sur la partition m sont identifiables.

Nous commençons par montrer cette propriété pour le modèle non contraint relatif à l'hypothèse (Hnc), puis pour le modèle contraint à la monotonie en e^θ par (Hc), ces hypothèses étant définies en 2.3.1. Enfin, nous montrons que la solution du problème contraint est égale à celle d'un problème non contraint dans lequel on aurait remplacé m par m_*, où m_* est la partition construite à partir de m et définie par la régression isotonique du vecteur e^θ pour une pondération adaptée.

Dans cette section on se donne un n-échantillon $\mathbb{Y} = \{Y_1, ..., Y_n\}$ de variables aléatoires à valeurs dans $\mathbb{R}_*^+$ indépendantes et identiquement distribuées de loi $dP = f_{\alpha,\omega}d\mu$, où μ désigne la mesure de Lebesgue sur $\mathbb{R}_*^+$.

3.2.1 Résolution sans contrainte de monotonie

Pour écrire la vraisemblance du problème avons besoin d'une définition préliminaire d'ordre général concernant la mesure empirique associée à l'échantillon $\mathbb{Y}$ et d'objets statistiques qui en sont dérivés.

Définition 3.2.1. *On note P_n la mesure empirique associée à l'échantillon $\mathbb{Y}$*

$$P_n \;=\; \frac{1}{n}\sum_{i=1}^{n}\delta_{Y_i}$$

où δ_{Y_i} est la mesure de Dirac au point Y_i.

On notera de plus

$$\nu_n \;=\; P_n - P$$

la mesure empirique recentrée, définie comme opérateur sur les fonctions P-intégrables, de sorte que pour toute fonction $g : \mathbb{R} \longrightarrow \mathbb{R}$ on a

$$P_n(g) \;=\; \int g\, dP_n \;=\; \frac{1}{n}\sum_{i=1}^{n} g(Y_i)$$

et

$$\nu_n(g) \;=\; \frac{1}{n}\sum_{i=1}^{n} g(Y_i) - \mathbb{E}_{f_{\alpha,\omega}}(g).$$

En particulier, pour tout intervalle I de $\mathbb{R}$ on note

$$P_n(I) \;=\; P_n(\mathbb{1}_I) \;=\; \frac{|\{Y_i \in I, i = 1, ..., n\}|}{n}$$

la fréquence empirique d'observation de l'intervalle I.

La moyenne empirique $\bar{Y}$ est aussi donnée par

$$\bar{Y} \;=\; P_n(\mathrm{Id}) \;=\; \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

Enfin, pour f une densité de probabilité, on définit

$$\gamma_n(f) \;=\; P_n(-\log(f)) \;=\; -\frac{1}{n}\sum_{i=1}^{n} \log f(Y_i)$$

l'opposée de la log-vraisemblance de f.

Remarque : de cette dernière définition, on voit que classiquement, il est équivalent de maximiser la vraisemblance ou de minimiser la quantité $\gamma_n(f)$, appelée contraste. Le concept de minimisation du contraste est très générale. Elle inclut une large variété de méthodes classiques d'estimation, suivant la forme du contraste. En ce sens, l'estimation par maximum de vraisemblance en est un cas particulier, tout comme l'estimation par régression.

Utilisant les notations de la définition 3.2.1, nous voyons que la vraisemblance du problème s'écrit

$$L(\alpha, e^{\theta}) \;=\; \prod_{i=1}^{n} f_{\alpha,\omega}(Y_i) \;=\; \alpha^n e^{-\alpha n \bar{Y}} \prod_{I\in m} \left(e^{\theta_I}\right)^{nP_n(I)}.$$

Pour maximiser la vraisemblance, nous cherchons à résoudre le problème suivant

$$\left\{ \begin{array}{ll} \max\limits_{\alpha>0,\, e^{\theta}\in\mathbb{R}_*^{+|m|}} & \alpha^n e^{-\alpha n \bar{Y}} \prod\limits_{I\in m} \left(e^{\theta_I}\right)^{nP_n(I)} \\ \text{s. c.} & \sum\limits_{I\in m} e^{\theta_I}\mu_\alpha(I) = 1. \end{array} \right. \tag{3.7}$$

où μ_α a été définie par (3.1.1). La contrainte qui apparaît dans ce problème exprime simplement le fait que la solution $(\hat{\alpha}, \hat{\omega})$ doit être telle que $f_{\hat{\alpha},\hat{\omega}}$ soit une densité.

Avant de résoudre le problème (3.7), nous devons donner quelques notations dont le but est de faciliter l'écriture et l'interprétation des solutions.

Définition 3.2.2. *Pour $i = 1, ..., |m|$ et I_i un intervalle de $m \in \mathcal{M}$, posant $b_{I_0} = \min \mathbb{Y}$ et $b_{I_{|m|}} = \max \mathbb{Y}$, nous adoptons l'écriture suivante des bornes de I_i :*

$$I_i = [\, b_{I_{i-1}} ; b_{I_i}[\quad \textit{pour } i = 1, ..., |m| - 1, \;\; \textit{et} \quad I_{|m|} = [b_{I_{|m|-1}} ; \, b_{I_{|m|}}].$$

Nous noterons aussi, pour $I = [\,b\,;b'\,]$ un intervalle[1] *de $\mathbb{R}^+$*

$$\mu'_\alpha(I) \;=\; \frac{d}{d\alpha}\,\mu_\alpha(I) \;=\; \frac{d}{d\alpha}\left(e^{-\alpha b} - e^{-\alpha b'}\right) \;=\; -\left(be^{-\alpha b} - b'e^{-\alpha b'}\right). \quad (3.8)$$

la dérivée par rapport à α de la mesure exponentielle de l'intervalle I.

On note N_I (respectivement n_I) le nombre d'individus de $\mathbb{X}$ (respectivement $\mathbb{Y}$) qui sont dans I :

$$N_I \;=\; |\{i \in \mathbb{N}_N \,/\, X_i \in I\}| \quad et \quad n_I \;=\; |\{i \in \mathbb{N}_n \,/\, Y_i \in I\}|. \quad (3.9)$$

Enfin, on note p le taux de sondage global de $\mathbb{Y}$ par rapport à $\mathbb{X}$ et p_I le taux de sondage à l'intérieur de l'intervalle I :

$$p \;=\; \frac{n}{N} \quad et \quad p_I \;=\; \frac{n_I}{N_I}. \quad (3.10)$$

La solution du problème (3.7) est donnée par la Proposition suivante.

Proposition 3.2.3. *L'estimateur du maximum de vraisemblance $\left(\hat{\alpha}, \widehat{e^\theta}\right)$ du couple $\left(\alpha, e^\theta\right)$ de $\mathbb{R}_*^+ \times \mathbb{R}_*^{+|m|}$ est donné par l'unique solution du système suivant :*

$$\begin{cases} e^\theta \;=\; \left(\dfrac{P_n(I)}{\mu_\alpha(I)}\right)_{I\in m} \\ \dfrac{1}{\alpha} \;=\; \bar{Y} + \displaystyle\sum_{I\in m} e^{\theta_I}\mu'_\alpha(I), \end{cases} \quad (3.11)$$

où μ'_α est définie par (3.8).

Remarque (1) : la première équation de (3.11) peut être vue comme la définition d'un histogramme contre la mesure $d\mu_\alpha$. Il s'agit ici de faire l'analogie avec l'histogramme vu comme estimateur du maximum de vraisemblance d'une densité (sous-entendu contre la mesure μ de Lebesgue). En effet, celui est égal à $P_n(I)/\mu(I)$ pour tout I de m et le parallèle à mener est alors évident.

Remarque (2) : supposons α fixé. Chaque $\widehat{e^{\theta_I}}$ mesure un "écart" relatif à la mesure μ_α sur l'intervalle I. On interprète cet écart comme un biais dans le tirage par rapport à l'échantillon initial i.i.d. de loi $d\mu_\alpha$, c'est-à-dire dans l'échantillon des logtailles des champs qui existent dans la nature. Par la loi des grands nombres, la mesure empirique (inconnue) de l'intervalle I

[1] La mesure exponentielle étant continue, la nature ouverte ou fermée des bornes n'a aucune importance.

converge presque sûrement vers la mesure exponentielle de I lorsque N tend vers l'infini :

$$N_I/N \xrightarrow[N\to+\infty]{p.s.} \mu_\alpha(I)$$

où N_i est défini par la formule (3.9).

Par ailleurs, le taux de sondage $p_I = n_I/N_I$ donné par la formule (3.10) est la vraie probabilité d'inclusion des éléments de $\mathbb{X}$ observés dans l'intervalle I conditionnellement au fait que n est connu. Si l'on se place dans des conditions "raisonnables" (cf. Rosén [74] et [75] ou Bickel *et al.* [13]) du style

$$\forall I \in m, \lim_{N,\, n\to+\infty} p_I \in \,]\,0\,;1[$$

on obtient

$$\frac{p_I}{p} \;=\; \frac{n_I}{N_I}\frac{N}{n} \underset{N,\,n\to+\infty}{\sim} \frac{P_n(I)}{\mu_\alpha(I)} \;=\; \widehat{e^{\theta_I}}.$$

Puis, l'estimateur $\widehat{\omega}_I$ de ω_I préconisé par le Lemme 3.1.2 est alors

$$\widehat{\omega}_I \;=\; \frac{\widehat{e^\theta_I}}{\max\limits_{J\in m}\widehat{e^{\theta_J}}} \underset{N\,n\to+\infty}{\sim} \frac{p_I}{p}\frac{p}{\max\limits_{J\in m} p_J} \;=\; \frac{p_I}{\max\limits_{J\in m} p_J}.$$

Or, les hypothèses (Hc) ou (Hnc) précisent justement que

$$\max_{J\in m} p_J \;=\; 1,$$

ainsi, cet estimateur doit bien être simultanément proche asymptotiquement de p_I et de ω_I. On a donc ici esquissé une preuve heuristique de la consistance de l'estimateur $\widehat{\omega}_I$ pour tout I, au moins à α fixé.

Remarque (3) : la seconde équation de (3.11) qui concerne l'estimation de α montre l'impact du biais dans le tirage au moyen du terme $\sum_{I\in m} e^{\theta_I}\mu'_\alpha(I)$. En effet, dans l'estimation par maximum de vraisemblance du paramètre d'une loi exponentielle basée sur un échantillon i.i.d. classique, ce terme n'existe pas et l'équation s'écrit simplement $1/\alpha = \bar{Y}$.

Preuve : On peut ici utiliser une méthode standard de maximisation du Lagrangien du problème (3.7). Cependant, par la proposition 4.6.7, l'estimateur du maximum de vraisemblance $f_{\hat{\alpha},\widehat{e^\theta}}$ de f_{α,e^θ} est caractérisé par le fait que pour tout s, polynôme par morceaux de degré 1 dont les termes de degré 1 sont égaux on a

$$\int_0^{+\infty} s f_{\hat{\alpha},\widehat{e^\theta}}\, d\mu = P_n(s).$$

Il suffit d'appliquer cette dernière égalité aux fonctions $\mathbb{1}_I$ pour tout I de m pour obtenir la première équation de (3.11) et à la fonction identité pour obtenir la seconde.

Résolvons maintenant le système (3.11) et montrons que la solution est unique.

Pour I un intervalle de $\mathbb{R}_*^+$ posons

$$\varphi_I'(\alpha) \;=\; \frac{d}{d\alpha}\left(\log\frac{\alpha}{\mu_\alpha(I)}\right) \;=\; \frac{1}{\alpha} - \frac{\mu_\alpha'(I)}{\mu_\alpha(I)} \tag{3.12}$$

et

$$\varphi'(\alpha) \;=\; -\bar{Y} + \sum_{I\in m} P_n(I)\varphi_I'(\alpha). \tag{3.13}$$

Pour résoudre (3.11), on voit qu'il suffit de substituer l'équation (3.12) dans l'équation (3.13), donc de montrer que φ' admet une unique racine. Or, pour $i = 1, ..., |m|$ on a

$$\varphi_{I_i}'(\alpha) \;=\; \frac{b_{I_{i-1}}e^{-\alpha b_{I_{i-1}}} - b_{I_i}e^{-\alpha b_{I_i}}}{e^{-\alpha b_{I_{i-1}}} - e^{-\alpha b_{I_i}}} + \frac{1}{\alpha} \;=\; \frac{b_{I_{i-1}} - b_{I_i}e^{-\alpha\mu(I_i)}}{1 - e^{-\alpha\mu(I_i)}} + \frac{1}{\alpha}.$$

Donc en effectuant un développement limité et un développement asymptotique de φ' en 0^+ et en $+\infty$ au moyen du membre de droite de l'équation ci-dessus, on obtient

$$\lim_{\alpha\to 0^+}\varphi'(\alpha) \;=\; \sum_{i=1}^{|m|} P_n(I_i)b_{I_i} - \bar{Y} \;>\; 0$$

et

$$\lim_{\alpha\to +\infty}\varphi'(\alpha) \;=\; \sum_{i=1}^{|m|} P_n(I_i)b_{I_{i-1}} - \bar{Y} \;<\; 0.$$

Il reste donc à voir que pour tout I de m, φ_I' est strictement décroissante. Or

$$\varphi_I''(\alpha) \;=\; \frac{d}{d\alpha}\varphi_I'(\alpha) \;=\; \frac{(\alpha\mu(I))^2e^{\alpha\mu(I)} - e^{2\alpha\mu(I)} + 2e^{\alpha\mu(I)} - 1}{\alpha^2(e^{\alpha\mu(I)} - 1)^2}$$

donc φ_I'' est du signe de la fonction $\psi(x) = x^2e^x - e^{2x} + 2e^x - 1$ pour x strictement positif. Mais, $\psi(0) = 0$ et

$$\psi'(x) \;=\; -2e^x(e^x - 1 - x - \frac{x^2}{2}) \;=\; -2e^x\sum_{k=3}^{+\infty}\frac{x^k}{k!} \;<\; 0.$$

Ainsi, ψ puis φ_I'' pour tout I de m sont négatives. Il en résulte que φ' est continue, strictement décroissante, strictement positive en 0^+ et strictement

négative en $+\infty$. Elle admet alors, par le Théorème des valeurs intermédiaires, une unique racine $\hat{\alpha}$ strictement positive.

On définit ensuite $\widehat{e^{\theta}}$ en posant

$$\widehat{e^{\theta_I}} \;=\; P_n(I)/\mu_{\hat{\alpha}}(I)$$

pour tout I de m, comme dans (3.11). ♣

D'un point de vue pratique, la racine de la fonction φ' peut être approchée par une simple dichotomie, qui fournit une vitesse de convergence exponentielle, ce qui est pour nous parfaitement suffisant. L'utilisation d'un algorithme de type Newton-Raphson serait ici intéressante pour le gain de temps éventuellement non-négligeable qu'elle apporterait (convergence quadratique contre convergence linéaire, voir Vignes [81]). Cependant, dans le cas de la résolution du problème de maximisation de la vraisemblance avec contrainte de monotonie, nous allons voir de suite que la fonction qui joue un rôle semblable à φ' dans la preuve de la Proposition 3.2.3, et que nous allons chercher à annuler, n'est pas dérivable. La méthode de Newton-Raphson n'est alors plus utilisable. Afin de conserver l'homogénéité algorithmique de la résolution de nos équations avec ou sans contraintes, nous en restons donc au choix de la dichotomie.

3.2.2 Résolution avec contrainte de monotonie

Ajoutant la contrainte de monotonie en e^{θ} au problème (3.7), nous cherchons à résoudre le problème de maximisation suivant

$$\left\{\begin{array}{ll} \max\limits_{(\alpha,e^{\theta})\in\mathbb{R}_*^+\times\mathbb{R}_*^{+|m|}} & \alpha^n e^{-\alpha n\bar{Y}} \prod\limits_{I\in m}\left(e^{\theta_I}\right)^{nP_n(I)} \\ \text{s. c.} & \\ (i) & \sum\limits_{I\in m} e^{\theta_I}\mu_\alpha(I) = 1 \\ (ii) & e^{\theta_{I_1}} \le e^{\theta_{I_2}} \le \cdots \le e^{\theta_{I_{|m|}}}. \end{array}\right. \tag{3.14}$$

Nous allons montrer que les solutions de ce programme ont une allure tout-à-fait semblable à celles du problème non contraint à la monotonie. En effet,

la Proposition 3.2.8 montre que celles-ci sont uniques solutions du système

$$\begin{cases} e_*^\theta = \left(\min\limits_{j \geq i} \max\limits_{h \leq i} \left[\dfrac{P_n\left(\bigcup_{k=h}^{j} I_k\right)}{\mu_\alpha\left(\bigcup_{k=h}^{j} I_k\right)} \right] \right)_{1 \leq i \leq |m|} \\ \dfrac{1}{\alpha} = \bar{Y} + \sum\limits_{I \in m} e_*^{\theta_I} \mu'_\alpha(I). \end{cases}$$

Les réunions d'intervalles de m qui apparaissent dans la première de ces deux équations sont bien entendu construites de telle façon que la suite e_*^θ soit croissante. Tout se passe donc ici comme si l'on remplaçait m par une nouvelle partition formée de réunions des intervalles qui la constituent, de façon à ce que le résultat final respecte la contrainte de monotonie.

Il était aussi possible d'envisager une technique similaire à celle employée par Sun et Woodroofe [77] proche, dans les idées et la mise en œuvre, d'un algorithme EM[2] : on part d'une valeur quelconque de α et l'on maximise la vraisemblance du modèle en le vecteur e^θ à α fixé sous contraintes de monotonie. On maximise ensuite la vraisemblance du modèle en α, le vecteur e^θ étant fixé à la valeur précédente et ainsi de suite. Les auteurs ont montré la convergence de cet algorithme vers le maximum de vraisemblance[3] du couple (α, e^θ) sous des conditions assez générales. La vitesse de convergence de leur algorithme n'est cependant pas évaluée et il y a toutes les chances pour que celle-ci dépende fortement de la forme de la pénalité que les auteurs ajoutent à leur vraisemblance.

Comme dans le cas non contraint, notre approche consiste à exprimer la solution générale en (α, e^θ) du problème de maximisation de la vraisemblance sous contrainte comme solution d'un système d'équations implicites. Nous montrons de nouveau que la résolution de ce système se résume à la recherche d'une racine d'une fonction strictement décroissante. En conséquence, la solution peut être approchée par une simple dichotomie, ce qui fournit une vitesse de convergence linéaire à notre approximation (voir Vignes [81]).

L'ajout de la contrainte de monotonie (ii) par rapport à (3.7) rend ce programme plus délicat à résoudre. Comme précédemment, nous allons procéder par substitution. Détaillons les étapes : commençons par supposer $\alpha > 0$ fixé

[2]Les auteurs nomment leur technique MM pour Maximization Maximization au lieu de Expectation Maximization dans le cas de l'algorithme EM.

[3]Vraisemblance pénalisée en l'occurrence, mais l'algorithme fonctionne dans notre cas.

et résolvons le sous-problème

$$\begin{cases} \max\limits_{e^\theta \in \mathbb{R}_*^{+|m|}} & \prod\limits_{I\in m} \left(e^{\theta_I}\right)^{nP_n(I)} \\ \text{s. c.} & \\ (i) & \sum\limits_{I\in m} e^{\theta_I}\mu_\alpha(I) = 1 \\ (ii) & e^{\theta_{I_1}} \leq e^{\theta_{I_2}} \leq \cdots \leq e^{\theta_{I_{|m|}}}. \end{cases} \tag{3.15}$$

Cela fait, par le Lemme 3.2.5, nous aurons une fonction $\alpha \longmapsto e_*^\theta(\alpha)$ qui à α associe la solution unique du problème (3.15). Nous résolvons ensuite le problème principal (3.14) en cherchant la solution du second sous-problème

$$\begin{cases} \max\limits_{\alpha>0} & \alpha^n e^{-\alpha n\bar{Y}} \prod\limits_{I\in m} \left(e_*^{\theta_I}(\alpha)\right)^{nP_n(I)} \\ \text{s. c.} & \sum\limits_{I\in m} e_*^{\theta_I}(\alpha)\mu_\alpha(I) = 1 \end{cases} \tag{3.16}$$

dont la résolution relève des mêmes idées que dans la section 3.2.1 à ceci près que la non-différentiabilité de la fonction $\alpha \longmapsto e_*^\theta(\alpha)$, contrairement à celle de la fonction définie par (3.11), complique la tâche.

La résolution du sous-problème (3.15) fait appel aux techniques de régression isotoniques (voir Robertson *et al.* [72]).

Définition 3.2.4. *On appelle classe des fonctions isotoniques sur une partition m l'ensemble des fonctions croissantes $f : m \longrightarrow \mathbb{R}$.*

Remarque : pour $\mathcal{I}$ un ensemble d'indices ordonné, nous identifions, sauf en cas d'ambiguïté, un vecteur $(u_i)_{i\in\mathcal{I}}$ à la suite de ses coordonnées, ainsi qu'à l'image de la fonction

$$\begin{array}{rccl} u & : & \mathcal{I} & \longrightarrow & \mathbb{R} \\ & & i & \longmapsto & u(i) = u_i. \end{array}$$

Nous parlerons donc indifféremment de régression isotonique d'une fonction, d'une suite ou d'un vecteur selon le point de vue que nous adoptons.

Lemme 3.2.5. *La solution $e_*^\theta(\alpha)$ du problème (3.15) est unique. Elle est donnée par la régression isotonique associée à la pondération $(\mu_\alpha(I))_{I\in m}$ de la fonction*

$$\begin{array}{rccl} e^\theta(\alpha) & : & m & \longrightarrow & \mathbb{R}^{|m|} \\ & & I & \longmapsto & \widehat{e^{\theta_I}}(\alpha) = \dfrac{P_n(I)}{\mu_\alpha(I)}. \end{array}$$

Preuve : Le programme (3.15) équivaut à maximiser en f la quantité

$$\sum_{I \in m} \widehat{e^{\theta_I}}(\alpha) \log(f(I)) \mu_\alpha(I) \tag{3.17}$$

dans la classe des fonctions isotoniques (interprétation de la contrainte (ii)) satisfaisant la contrainte (i) :

$$\sum_{I \in m} \left(\widehat{e^{\theta_I}}(\alpha) - f(I) \right) \mu_\alpha(I) \ = \ 0. \tag{3.18}$$

Par le Théorème (1.5.1) de Robertson *et al.* [72], on sait que la régression isotonique $e_*^\theta(\alpha)$ de $\widehat{e^\theta}(\alpha)$ maximise en f la quantité

$$\sum_{I \in m} \left[\Phi\left(f(I)\right) + \left(\widehat{e^{\theta_I}}(\alpha) - f(I) \right) \phi\left(f(I)\right) \right] \mu_\alpha(I) \tag{3.19}$$

pour toute fonction convexe Φ de dérivée ϕ. De plus, la solution est unique si Φ est strictement convexe.

Choisissons $\Phi(x) = x \log x$, pour $x > 0$ qui est strictement convexe. L'équation (3.19) devient

$$\sum_{I \in m} \left[\widehat{e^{\theta_I}}(\alpha) \log f(I) + \left(\widehat{e^{\theta_I}}(\alpha) - f(I) \right) \right] \mu_\alpha(I).$$

Par le Théorème (1.3.6) de Robertson *et al.* [72], on a

$$\sum_{I \in m} \left(\widehat{e^{\theta_I}}(\alpha) - e_*^\theta(\alpha) \right) \mu_\alpha(I) = 0.$$

Donc $e_*^\theta(\alpha)$ vérifie (3.18) et maximise (3.17) via (3.19). ♣

La figure 3.1 donne trois exemples de régression isotonique d'une fonction en escalier. En premier lieu, chaque marche est pondérée de la même façon (pondération par la mesure empirique) : les zones de décroissance sont "remplacées" par la valeur moyenne des marches qui y sont incluses. La seconde figure présente une pondération par la mesure de Lebesgue, c'est-à-dire que les zones de décroissance sont remplacées par la moyenne des valeurs des marches pondérée par la longueur des marches y sont incluses. Enfin, la dernière figure présente une pondération exponentielle de paramètre 1, *i.e.* les zones de décroissance sont remplacées par la moyenne des valeurs des marches pondérée par la mesure exponentielle de paramètre 1 des marches y sont incluses.

Donnons maintenant la forme analytique de la régression isotonique.

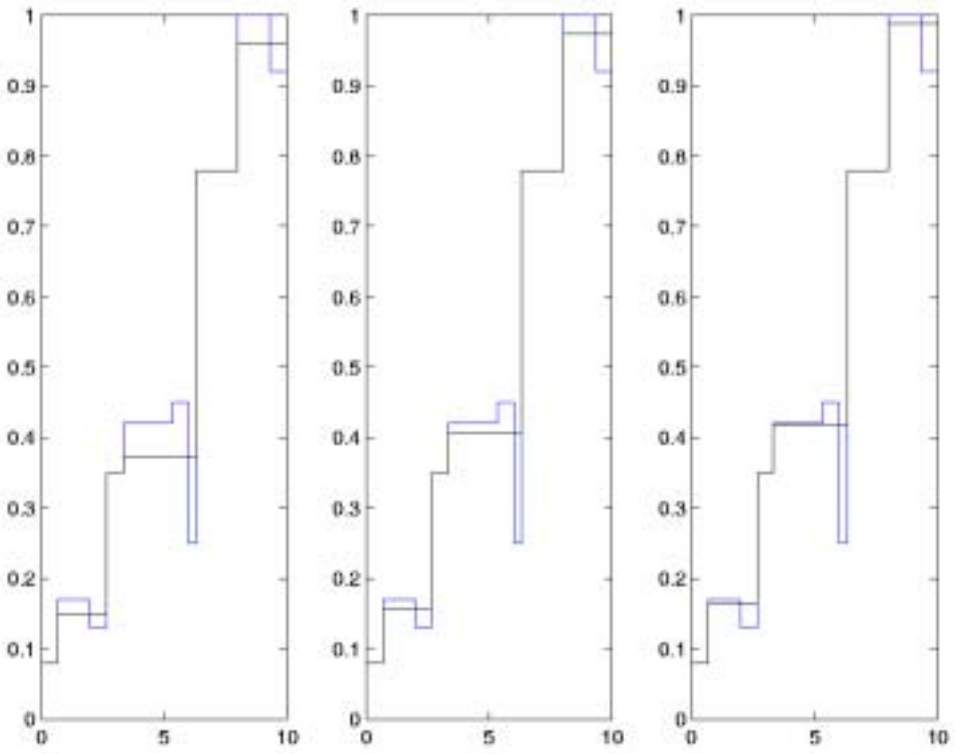

FIG. 3.1 – Régressions isotoniques (en noir) d'une fonction en escalier (en bleu) contre les pondérations empiriques (à gauche), Lebesgue (au milieu) et exponentielle de paramètre 1 (à droite).

Lemme 3.2.6. *La régression isotonique $e_*^{\theta}(\alpha)$ de la suite $\widehat{e^{\theta}}(\alpha)$ est donnée, pour tout $i = 1, ..., |m|$ par*

$$e_*^{\theta_{I_i}}(\alpha) = \min_{j \geq i} \max_{h \leq i} \left[\frac{P_n\left(\bigcup_{k=h}^{j} I_k\right)}{\mu_\alpha\left(\bigcup_{k=h}^{j} I_k\right)} \right]. \tag{3.20}$$

Remarque (1) : notons $I_*(\alpha) = \bigcup_{k=h_*(\alpha)}^{j_*(\alpha)} I_k$, où $h_*(\alpha)$ et $j_*(\alpha)$ réalisent les $(\min, \max)$ dans (3.20). Pour éviter les notations lourdes, nous écrirons I_* pour $I_*(\alpha)$, sauf lorsque montrer la dépendance en α s'avère nécessaire. L'ensemble $m_* = \{I_*, I \in m\}$, où les I_* ne sont pas répétés, définit une nouvelle partition de $[Y_{(1)}, Y_{(n)}]$ dont chaque élément est réunion d'éléments de m. Ceci nous amène à définir la notion de sur-partition ci-dessous. Comme précédemment nous omettons la dépendance en α de m_* (et donc des I_*), sauf lorsque cela est nécessaire.

Définition 3.2.7. *Pour une partition m donnée, on appelle sur-partition de m toute partition de $[Y_{(1)}, Y_{(n)}]$ formée d'intervalles issus de la réunion d'intervalles de m.*

Remarque (2) : pour tout I de m on a, par la remarque (1) et le Lemme 3.2.6 on a :

$$e_*^{\theta_I}(\alpha) = \widehat{e^{\theta_{I_*}}}(\alpha) = \frac{P_n(I_*)}{\mu_\alpha(I_*)}. \tag{3.21}$$

En conséquence, pour tout J de m contenu dans un I_* de la sur-partition m_* de m définie par régression isotonique :

$$e_*^{\theta_J}(\alpha) \;=\; \widehat{e^{\theta_{I_*}}}(\alpha)$$

où $\widehat{e^{\theta}}$ est défini dans la Proposition 3.2.3.

Preuve : Elle est une conséquence directe des formules classiques dites "minmax" du théorème (1.4.4) de Robertson *et al.* [72]. ♣

Nous pouvons maintenant passer à la résolution complète du problème (3.14).

Proposition 3.2.8. *L'estimateur du maximum de vraisemblance $\left(\hat{\alpha}_*, \widehat{e_*^{\theta}}\right)$ du couple $\left(\alpha, e^{\theta}\right)$ de $\mathbb{R}_*^+ \times \mathbb{R}_*^{+|m|}$ sous contrainte de monotonie de $\widehat{e_*^{\theta}}$ (i.e. sous la contrainte (Hc)) est donné par l'unique solution du système suivant :*

$$\begin{cases} e_*^{\theta} = \left(\min_{j \geq i} \max_{h \leq i} \left[\dfrac{P_n\left(\bigcup_{k=h}^{j} I_k\right)}{\mu_\alpha\left(\bigcup_{k=h}^{j} I_k\right)} \right] \right)_{1 \leq i \leq |m|} \\ \dfrac{1}{\alpha} = \bar{Y} + \sum_{I \in m} e_*^{\theta_I} \mu'_\alpha(I). \end{cases} \tag{3.22}$$

Preuve : La première équation de (3.22) résulte du Lemme (3.2.6). Comme dans le cas de la preuve de la Proposition (3.2.1), il suffit de montrer que la fonction

$$\varphi'_*(\alpha) = \frac{1}{\alpha} - \bar{Y} - \sum_{I\in m} e_*^{\theta_I}(\alpha)\mu'_\alpha(I) \tag{3.23}$$

admet une unique racine $\hat{\alpha}_*$ strictement positive. Pour ce faire, nous allons montrer que φ'_* est continue sur $\mathbb{R}_*^+$ et strictement décroissante entre une valeur strictement positive en 0^+ et strictement négative en $+\infty$.

En premier lieu, la continuité résulte du fait que, par Robertson *et al.* [72] page 24, l'application qui à un vecteur associe sa régression isotonique est continue (mais non différentiable en général, ce qui nous empêche de raisonner comme dans la preuve de (3.2.3)[4]). Ensuite, l'application $\alpha \longmapsto P_n(I)/\mu_\alpha(I)$ est, elle aussi, continue. Par composition on en déduit que φ'_* est continue.

Montrons qu'elle est strictement décroissante. Pour cela, on cherche une écriture de φ'_* proche de celle de φ' donnée par l'équation (3.13).

En utilisant les notations de la remarque (1) du Lemme 3.2.6 on écrit

$$\begin{aligned}
\varphi'_*(\alpha) &= \frac{1}{\alpha} - \bar{Y} - \sum_{I_*\in m_*}\sum_{\substack{I\subset I_*\\ I\in m}} \frac{P_n(I_*)}{\mu_\alpha(I_*)}\mu'_\alpha(I) \\
&= \frac{1}{\alpha} - \bar{Y} - \sum_{I_*\in m_*} \frac{P_n(I_*)}{\mu_\alpha(I_*)} \sum_{\substack{I\subset I_*\\ I\in m}} \mu'_\alpha(I) \\
&= \frac{1}{\alpha} - \bar{Y} - \sum_{I_*\in m_*} \frac{P_n(I_*)}{\mu_\alpha(I_*)}\mu'_\alpha(I_*) \\
&= \frac{1}{\alpha} - \bar{Y} - \sum_{I_*\in m_*} \Big(\sum_{\substack{I\subset I_*\\ I\in m}} P_n(I)\Big)\frac{\mu'_\alpha(I_*)}{\mu_\alpha(I_*)} \\
&= -\bar{Y} + \sum_{I\in m} P_n(I)\varphi'_{I_*}(\alpha),
\end{aligned}$$

où φ'_I a été définie par (3.12).

La fonction φ'_* s'exprime donc comme la translation par $-\bar{Y}$ du barycentre des fonctions $\varphi'_{*,I} : \alpha \longmapsto \varphi'_{I_*(\alpha)}(\alpha)$, pondéré par les $(P_n(I))_{I\in m}$. Mais bien que φ'_* et les φ'_I pour I dans m soient continues, les $\varphi'_{*,I}$ ne le sont pas en général. Cependant, leur "tendance" décroissante va tout de même nous permettre de conclure.

[4] notons ici, comme on l'a déjà mentionné, qu'en particulier il serait impossible d'appliquer un algorithme de type Newton-Raphson pour approcher la valeur de la racine de la fonction φ'_* puisque celle-ci n'est pas dérivable.

Soit J un intervalle fixé, formé de la réunion d'intervalles de m. On sait, par la preuve de la Proposition 3.2.3, que φ'_J est strictement décroissante. La restriction de $\varphi'_{*,I}$ à l'ensemble $\{\alpha / I_*(\alpha) = J\}$ est donc, elle aussi, strictement décroissante. Ainsi, si $\alpha < \alpha'$ sont tels que $\varphi'_{*,I}(\alpha) \leq \varphi'_{*,I}(\alpha')$ alors $I_*(\alpha) \neq I_*(\alpha')$.

Raisonnons par l'absurde et prenons $\alpha < \alpha'$ tels que $\varphi'_*(\alpha) \leq \varphi'_*(\alpha')$. Les $(P_n(I))_{I \in m}$ étant fixes par rapport à α, il existe alors nécessairement un I de m tel que $\varphi'_{*,I}(\alpha) \leq \varphi'_{*,I}(\alpha')$. On en conclut alors que $m_*(\alpha) \neq m_*(\alpha')$.

La continuité de φ'_* implique, via le Théorème des valeurs intermédiaires, qu'il existe alors une suite strictement croissante $(\alpha_n)_{n \in \mathbb{N}}$ de l'intervalle $[\alpha, \alpha']$ dont l'image par φ'_* est croissante. En particulier, les partitions $(m_*(\alpha_n))_{n \in \mathbb{N}}$ associées sont alors deux-à-deux distinctes, ce qui est impossible puisqu'il n'existe qu'un nombre fini de sur-partitions basées sur m. La fonction φ'_* est donc bien strictement décroissante.

Par ailleurs, pour toute sur-partition m' de m, on a

$$\begin{aligned} \lim_{\alpha \to 0^+} \sum_{J \in m'} P_n(J) \varphi'_J(\alpha) &= \sum_{J \in m'} P_n(J) b_J \\ &\geq \sum_{I \in m} P_n(I) b_I \quad > \quad \bar{Y} \end{aligned}$$

et

$$\begin{aligned} \lim_{\alpha \to +\infty} \sum_{J \in m'} P_n(J) \varphi'_J(\alpha) &= \sum_{J \in m'} P_n(J) b_{J-1} \\ &\leq \sum_{I \in m} P_n(I) b_{I-1} \quad < \quad \bar{Y}, \end{aligned}$$

où b_{I-1} désigne la borne de gauche de l'intervalle I.

Donc, φ'_* est continue, décroissante, bornée, donc elle admet des limites en 0^+ et $+\infty$ qui vérifient

$$\lim_{\alpha \to 0^+} \varphi'_*(\alpha) \; > \; 0$$

et

$$\lim_{\alpha \to +\infty} \varphi'_*(\alpha) \; < \; 0,$$

ce qui achève la preuve. ♣

L'objectif de la section suivante est de mettre en évidence le lien qui existe entre les modèles contraints et les modèles non contraints à la monotonie.

3.2.3 Équivalence des modèles contraints et des modèles non-contraints sur une sur-partition

Nous montrons dans cette section que la solution du modèle contraint à la monotonie est égale à la solution d'un modèle équivalent sans contraintes de monotonie sur une sur-partition de la partition de départ. La Proposition suivante est un Corollaire de la Proposition 3.2.8.

Corollaire 3.2.9. *Notons $\hat{m}_* = m_*(\hat{\alpha}_*)$ la sur-partition de m où m_* est définie par la remarque (2) du Lemme (3.2.6) et $\hat{\alpha}_*$ par la Proposition (3.2.8). Les estimateurs du maximum de vraisemblance $\left(\hat{\alpha}_*(m), \widehat{e^{\theta}_*}(m)\right)$ associés à la partition m sous (Hc) et $\left(\hat{\alpha}(\hat{m}_*), \widehat{e^{\theta}}(\hat{m}_*)\right)$ associés à la partition $\hat{m}_*$ sous (Hnc) sont égaux.*

Remarque (1) : par abus de langage sur les notations, on peut interpréter ce Lemme en disant que "$*$" (étoile) commute avec "$\wedge$" (chapeau).

Remarque (2) : d'un point de vue algébrique, on constate grâce au Corollaire 3.2.9 que la solution au problème contraint s'obtient comme solution du problème non contraint équivalent sur un espace "plus petit" (correspondant à une sur-partition de la partition de travail initiale). Cette dernière propriété est caractéristique des problèmes d'optimisation sous contraintes. Lorsque celles-ci sont affines, trouver la solution d'un problème contraint revient en fait à trouver le "bon" sous-espace affine (*i.e.* les bonnes liaisons entre variables du problème) de l'espace de travail sur lequel la solution va être obtenue en résolvant le problème correspondant, sans contraintes. Plus généralement dans le cadre de contraintes différentiables transverses, la solution au problème contraint correspond à la solution d'un problème non contraint équivalent, sur une sous-variété différentiable de l'espace de travail (les variables du problèmes étant alors définies localement par des fonctions implicites). Ceci correspond, en fait, à une interprétation géométrique des multiplicateurs de Kuhn et Tucker. En effet, la sous-variété contenant la solution du problème d'optimisation est celle définie par les équations correspondant aux contraintes saturées (multiplicateurs non nuls). Nous reviendrons sur ce dernier point dans le chapitre 4 (Lemme 4.6.6).

Preuve : Pour I dans m et I_* associé dans $\hat{m}_*$ on a

$$\widehat{e^{\theta_I}_*} = \widehat{e^{\theta_{I_*}}} = \frac{P_n(I_*)}{\mu_\alpha(I_*)}$$

où α est solution de

$$\frac{1}{\alpha} = \bar{Y} + \sum_{I_* \in \hat{m}_*} \widehat{e^{\theta_{I_*}}} \mu'_\alpha(I_*) \tag{3.24}$$

$$\Leftrightarrow \quad \frac{1}{\alpha} = \bar{Y} + \sum_{I_* \in \hat{m}_*} \widehat{e^{\theta_{I_*}}} \Big(\sum_{\substack{I \subset I_* \\ I \in m}} \mu'_\alpha(I) \Big) \tag{3.25}$$

$$\Leftrightarrow \quad \frac{1}{\alpha} = \bar{Y} + \sum_{I_* \in \hat{m}_*} \sum_{\substack{I \subset I_* \\ I \in m}} \widehat{e^{\theta_{I_*}}} \mu'_\alpha(I) \tag{3.26}$$

$$\Leftrightarrow \quad \frac{1}{\alpha} = \bar{Y} + \sum_{I \in m} \widehat{e_*^{\theta_I}} \mu'_\alpha(I). \tag{3.27}$$

La solution de ces équations étant unique, on en déduit que la solution $\hat{\alpha}$ de l'équation (3.24) associée au maximum de vraisemblance non contraint basé sur $\hat{m}_*$ et la solution $\hat{\alpha}_*$ de l'équation (3.27) associée au maximum de vraisemblance contraint basé sur m sont bien égales. ♣

3.3 Estimateurs de type Horvitz-Thompson

Dans cette section, comme suite à l'estimation des probabilités d'inclusion des champs dans l'échantillon des découvertes, nous construisons des estimateurs du nombre total de champs ainsi que des réserves ultimes.

3.3.1 Estimation du nombre de total de champs

À partition m de l'intervalle d'observation fixée, on dispose d'un estimateur $\widehat{\omega}$ de ω donné d'abord par le lemme 3.1.2 puis les Propositons 3.2.3 ou 3.2.8 selon que l'on contraigne $\widehat{\omega}$ à la monotonie ou non. On dispose donc, pour tout I de m d'estimations des probabilités d'inclusion de chacune des observations en voyant l'exploration conformément au modèle décrit dans le chapitre 2.

Un estimateur naturel du nombre total de champ peut alors être construit comme l'analogue de l'estimateur d'Horvitz-Thompson [40] lorsque les probabilités d'inclusion sont connues :

$$\widehat{N}_{HT} = \sum_{i=1}^{n} \frac{1}{\widehat{\omega}(Y_i)} = \sum_{I \in m} \frac{nP_n(I)}{\widehat{\omega}_I} = \sum_{I \in m} \frac{n_I}{\widehat{\omega}_I}. \tag{3.28}$$

En particulier, on peut "découper" cet estimateur afin de récupérer des estimateurs du nombre total de champs par classe de taille où $\widehat{\omega}$ est constant, c'est-à-dire sur chaque I de m :

$$\widehat{N}_I = \frac{nP_n(I)}{\widehat{\omega}_I} = \frac{n_I}{\widehat{\omega}_I}, \tag{3.29}$$

de sorte que

$$\sum_{I \in m} \widehat{N}_I \;=\; \widehat{N}_{HT}.$$

Ces estimateurs sont naturels dans la mesure où l'on a vu dans la remarque (2) de la Proposition 3.2.3 que $\widehat{\omega}_I$ est proche de n_I/N_I. De plus, comme nous le verrons en 3.3.4, lorsque la fonction ω est supposée connue, ces estimateurs possèdent d'excellentes qualités d'approximation.

En pratique, pour estimer le nombre de champs restant à découvrir par classe de taille, nous retrancherons aux estimateurs $\widehat{N}_I$ le nombre de champs déjà trouvés dans ladite classe :

$$\widehat{N}_I^r \;=\; \widehat{N}_I - n_I. \tag{3.30}$$

3.3.2 Une alternative ?

On aurait pu cependant utiliser le simple fait que si l'on déconditionne l'ensemble de notre approche par rapport au fait que n est fixé, alors n est une variable aléatoire de loi Binômiale de paramètre N et $\mathbb{E}_f(\omega(X))$, comme vu dans la remarque de la définition 2.3.2. En conséquence :

$$\mathbb{E}(n) \;=\; N\mathbb{E}_f(\omega(X)).$$

Un estimateur concurrent de (3.28) pourrait donc être

$$\widehat{N}_{\mathcal{B}} \;=\; \frac{n}{\mathbb{E}_f(\widehat{\omega}(X))}. \tag{3.31}$$

Cependant, et reprenant la définition des $(\varepsilon_i)_{1 \leq i \leq N}$ en (2.5), on peut montrer qu'à ω connue (cf. Delfiner [27]) les deux estimateurs sont liés par la relation :

$$\mathbb{E}(\widehat{N}_{HT} \,|\, \varepsilon_1, \, ..., \, \varepsilon_N) \;=\; \widehat{N}_{\mathcal{B}}.$$

Ces estimateurs concurrents sont donc de proches parents et il est clair que l'emploi des estimateurs de Horvitz-Thompson est bien plus informatif, puisqu'ils permettent d'obtenir des estimations du nombre total de champs classe par classe de taille et non sur la globalité comme $\widehat{N}_{\mathcal{B}}$.

3.3.3 Estimation des réserves ultimes

Comme pour l'estimation du nombre total de champs, nous allons nous appuyer sur les estimateurs de Horvitz-Thompson. Nous allons tenter de les introduire heuristiquement de façon naturelle. Auparavant, nous donnons quelques notations sur les réserves par classe de taille.

Définition 3.3.1. *Pour tout I de m, on note R_I (respectivement r_I) le montant de réserves des champs de l'échantillon $\mathbb{X}$ (respectivement $\mathbb{Y}$) contenues dans l'intervalle I :*

$$R_I = \sum_{X \in \mathbb{X} \cap I} \exp(X) \quad et \quad r_I = \sum_{Y \in \mathbb{Y} \cap I} \exp(Y).$$

Considérons une classe de taille I sur laquelle la probabilité d'inclusion ω est constante. Dans notre cas, nous remplacerons ω par son estimateur $\widehat{\omega}$. Le tirage au sein de cette classe est donc un tirage à probabilités égales d'inclusion ω_I et, conditionnellement à $\mathbb{X}$, la moyenne des tailles du tirage est un estimateur sans biais de la moyenne des tailles de la population parente sur l'intervalle I : $r_I/n_I \simeq R_I/N_I$. Ainsi,

$$R_I \ \simeq \ \frac{N_I}{n_I} \times r_I \ = \frac{r_I}{p_I}.$$

Ceci légitime l'emploi des estimateurs, pour tout I de m

$$\widehat{R}_I = \frac{r_I}{\widehat{\omega}_I} \quad \text{et} \quad \widehat{R}_{HT} = \sum_{I \in m} \widehat{R}_I \tag{3.32}$$

comme estimateurs des réserves ultimes par classe de taille ainsi que de réserves ultimes globales.

L'estimateur $\widehat{\omega}$ du biais dans le tirage étant lui même difficile à appréhender d'un point de vue théorique, nous ne sommes pas en mesure de fournir une étude théorique de la qualité de nos estimateurs de type Horvitz-Thompson $(\widehat{N}_I, \widehat{N}_{HT})$ ou encore $(\widehat{R}_I, \widehat{R}_{HT})$. Nous donnons cependant maintenant quelques pistes.

3.3.4 Commentaires sur la qualité d'estimation à la Horvitz-Thompson

Hormis l'article original [40], la littérature concernant les propriétés asymptotiques des estimateurs de Horvitz-Thompson est extrêmement dense : voir par exemple Hanif [38] pour de nombreuses références, Cordy [23] pour une approche un peu plus générale qui s'inscrit bien dans les modèles de superpopulation ou encore Dol *et al.* [29] pour une interprétation algébrique récente, simple et originale. La presque totalité de cette littérature concerne des modèles assimilables à des modèles de sondage ou de superpopulation présupposant que les probabilités d'inclusion sont soit connues, soit estimables grâce à des techniques de type Rosén [74] et [75] que nous avons déjà évoquées au chapitre 2[5]. C'est le cas notamment dans Bickel *et al.* [13]. La majorité de ces articles montrent que les estimateurs de Horvitz-Thompson, ou

[5] Rappelons qu'il s'agit de techniques d'estimation des probabilités d'inclusion au $n^{\text{ième}}$ tirage connaissant les probabilités d'inclusion au premier tirage.

assimilés Horvitz-Thompson lorsque les probabilités d'inclusion sont seulement estimées, sont consistants et asymptotiquement normaux sous des hypothèses "raisonnables" du type de celles présentées dans la remarque (2) de la Proposition 3.2.3. Essentiellement, on réclame que pour une classe de taille $I \in m$ fixée, la fréquence d'observation $p_I = n_I/N_I$ tende vers une limite finie lorsque n_I et N_I tendent vers l'infini. On pourra se reporter aux articles cités ci-dessus pour plus de précisions concernant chacun des cas particuliers évoqués, sachant que les techniques utilisées sont toutes dans la veine de celles crées par Rosén [74] et [75]. Pour notre part, nous utilisons simulation et méthodes de Monte-Carlo dans le chapitre 5 pour quantifier la qualité d'approximation de nos estimateurs de type Horvitz-Thompson.

Jusqu'alors, nous avons calculé les estimateurs du maximum de vraisemblance de α et de ω ainsi que des estimateurs du nombre total de champs et des réserves ultimes compte tenu du fait que la dimension de notre modèle était fixée. Maintenant, comment choisir "le meilleur modèle possible[6]", autrement dit, le modèle dans lequel notre estimateur sera le plus proche possible de la vraie densité ? C'est l'objectif des procédures dites de "sélection de modèles".

[6] Au travers de sa dimension, c'est-à-dire le cardinal de la partition

Chapitre 4

Sélection de modèles

Nous nous intéressons dans ce chapitre à l'estimation d'une densité f par sélection de modèles exponentiels de polynômes par morceaux de degré 1 comme ceux décrits à la fin du chapitre 2. Dans le chapitre 3 nous montrons comment estimer les paramètres du modèle associés à une partition fixée de l'intervalle d'étude. Notre but dans ce chapitre est de sélectionner la meilleure partition possible, ou le meilleur modèle possible selon la terminologie usuelle, au sens d'un certain critère statistique.

Notre travail vise à adapter à notre cas particulier le Théorème 2.3.2 de sélection de modèles par maximum de vraisemblance pénalisé de Castellan [20] sur les modèles exponentiels de polynômes par morceaux généraux. Mais notre modèle n'en satisfait pas les hypothèses. En effet, celui-ci s'interprète bien aussi en termes de polynômes par morceaux, mais les coefficients de ces polynômes sont liés par des contraintes, ce que n'inclut pas le résultat de Castellan.

Nous allons montrer que sous certaines hypothèses – essentiellement que le nombre de morceaux maximal des partitions considérées soit en $\mathcal{O}\left(\sqrt{n}/\log^2 n\right)$ où n est le nombre d'observations, mais d'autres plus techniques comme le fait que la densité f soit définie sur un compact où elle est minorée par une constante strictement positive[1] – le risque de Kullback de l'estimateur choisi par l'algorithme de sélection de modèles est dominé par le risque (Hellinger) de "l'oracle". Ce dernier est défini comme l'estimateur du maximum de vraisemblance (associé à un modèle) tel que son risque soit le plus petit possible au sein de la collection de modèles parmi lesquels se fait la sélection. Cet oracle dépend donc de la densité inconnue et il est donc lui-même inconnu, en particulier, ce n'est évidemment pas un estimateur.

[1] En pratique, nous nous affranchirons de ces hypothèses en travaillant consitionnellement aux données et en supposant donc que la densité à estimer est nulle en dehors de l'intervalle défini par celles-ci et minorée à l'intérieur, ce qui n'est pas restrictif lorsque l'on travaille à l'estimation d'une densité par maximum de vraisemblance.

Plus précisément, si l'on note $\gamma_n(\hat{f}_m) = P_n(-\log \hat{f}_m)$ la log-vraisemblance de l'estimateur $\hat{f}_m$ sur le modèle défini au moyen d'une partition m d'une collection $\mathcal{M}$, nous allons donner la forme de la fonction de pénalité $\mathrm{pen}_n(m)$ (qui dépend de la collection $\mathcal{M}$, voir 4.4.1) pour que la minimisation du critère $\mathrm{crit}_n(m) = \gamma_n(\hat{f}_m) + \mathrm{pen}_n(m)$ fournisse un modèle $\hat{m}$ tel que $\hat{f}_{\hat{m}}$ ait la propriété voulue, sur un ensemble de grande probabilité. Nous donnons alors une borne de risque associée à cet estimateur qui permet d'en évaluer les performances de façon non asymptotique.

Nous commençons ce chapitre en construisant un paramétrage de notre modèle adapté aux techniques que nous souhaitons mettre en œuvre. Nous introduisons ensuite la notion de risque d'un estimateur au moyen de laquelle nous allons évaluer les performances de notre méthode de sélection de modèles. La troisième section de ce chapitre est consacrée à la description heuristique du protocole de sélection de modèles par maximum de vraisemblance pénalisé. Nous énonçons et démontrons ensuite notre résultat principal puis commentons le choix de la fonction de pénalité en fonction de la richesse de la collection au sein de laquelle se pratique la sélection de modèles. Nous terminons ce chapitre par une section annexe contenant quelques Lemmes techniques nécessaires à la preuve du Théorème 4.4.1.

Dans tout ce chapitre, la densité f à estimer est supposée définie sur $[\,0\,;1]$ et l'on se donne un échantillon i.i.d. $\mathbb{Y} = \{Y_1, ..., Y_n\}$ de variables aléatoires de densité f. En pratique, ces observations sont les log-tailles des champs d'un système pétrolier.

4.1 Définition d'un modèle

Dans toute cette section, on suppose donnée une partition $m = \{I_1, ..., I_{|m|}\}$ de l'intervalle $[\,0\,;1]$.

4.1.1 Construction d'une base adaptée

Conformément aux développements des chapitres 2 et 3, nous allons travailler sur des modèles exponentiels de polynômes par morceaux de degré 1 sur $[\,0\,;1]$. De plus, ces polynômes sont tels que leurs coefficients de degré 1 sont tous égaux, et nous pourrons être amenés à imposer une contrainte de monotonie sur les termes de degré 0.

Nous commençons ce chapitre en définissant un premier paramétrage adapté à notre problème.

Définition 4.1.1. *Soit $\mathbb{R}^1_m[X]$ l'espace des polynômes de degré 1 par morceaux basés sur m. L'espace $\mathbb{R}^1_m[X]$ est un sous-espace de $\mathbb{L}^2([0\,;1],\mu)$ muni du produit scalaire*

$$< s \,|\, t > = \int_0^1 st\, d\mu.$$

Pour tout $I \in m$ on pose

$$\begin{cases} \varphi_I^0 &= \dfrac{\mathbb{1}_I}{\sqrt{\mu(I)}} \\ \varphi_I^1 &= \dfrac{2\sqrt{3}}{\mu(I)^{3/2}}(\text{Id-m}(I))\mathbb{1}_I\,, \end{cases}$$

où m(I) désigne le milieu du segment I, de sorte que $\left\{\varphi_I^0,\, \varphi_I^1\right\}_{I\in m}$ est une base orthonormée de $\mathbb{R}_m^1[X]$.

De plus on pose

$$\varphi^1 \;=\; \frac{2\sqrt{3}}{\sqrt{\sum_{I\in m}\mu(I)^3}}\sum_{I\in m}(\text{Id-m}(I))\mathbb{1}_I \;=\; \frac{1}{\sqrt{\sum_{I\in m}\mu(I)^3}}\sum_{I\in m}\mu(I)^{3/2}\varphi_I^1,$$

de sorte que la famille $B_m = \left\{(\varphi_I^0)_{I\in m},\, \varphi^1\right\}$ est orthonormée pour $<\cdot|\cdot>$.

Enfin, on définit l'espace vectoriel E_m engendré par B_m des polynômes de degré 1 par morceaux basés sur m dont tous les termes de degré 1 ont même coefficient : $E_m = \text{Vect}(B_m)$.

Remarque : notons que l'espace E_m est de dimension $|m|+1$.

Nous sommes maintenant en mesure de construire un modèle exponentiel de polynômes par morceaux compatible avec notre problématique au moyen du paramétrage défini ci-dessus.

Définition 4.1.2. *Soit $\mathcal{E}_m$ le modèle exponentiel défini par*

$$\mathcal{E}_m \;=\; \left\{ h \in \mathbb{L}^1([0\,;1],\mu) \,\middle/\, h>0,\, \log h \in E_m,\, \int_0^1 h\,d\mu = 1 \right\}.$$

Le modèle $\mathcal{E}_m$ est donc construit comme un espace de fonctions définies à partir d'un paramétrage représenté par un espace vectoriel de dimension $|m|+1$. On impose de plus la contrainte que ces fonctions sont des densités de probabilité. Il semble alors intuitif que le "bon" paramétrage sous-jacent au modèle $\mathcal{E}_m$ soit un espace vectoriel de paramètres de dimension $|m|$ et non $|m|+1$. C'est-à-dire que $\mathcal{E}_m$ est un espace défini par $|m|$ paramètres.

L'objectif des parties suivantes est de construire ce bon paramétrage. L'espace $\mathcal{E}_m$ sera identifié comme hyperplan de E_m, ou de $\mathbb{R}^{|m|}\times\mathbb{R}$ selon que l'on interprète les éléments de $\mathcal{E}_m$ comme images d'éléments de l'espace sous-jacent E_m ou de l'espace $\mathbb{R}^{|m|}\times\mathbb{R}$ des coordonnées des éléments de E_m.

4.1.2 Identifications des espaces de paramètres

Un vecteur de E_m est repéré par ses coordonnées dans $\mathbb{R}^{|m|} \times \mathbb{R}$. Dans toute la suite, nous regarderons l'espace $\mathbb{R}^{|m|} \times \mathbb{R}$ comme un espace euclidien muni de son produit scalaire canonique noté $(\cdot \,|\, \cdot)$.

Pour $a \in \mathbb{R}^{|m|} \times \mathbb{R}$, nous écrirons indifféremment

$$\begin{cases} a &= (a^0, a^1) \text{ avec } a^0 \in \mathbb{R}^{|m|} \text{ et } a^1 \in \mathbb{R}, \text{ ou} \\ a &= \left((a^0_I)_{I \in m}, a^1\right) \text{ avec } (a^0_I)_{I \in m} \in \mathbb{R}^{|m|} \text{ et } a^1 \in \mathbb{R}. \end{cases}$$

Ainsi,

$$(a \,|\, b) \;=\; \sum_{I \in m} a^0_I b^0_I + a^1 b^1.$$

Dans toute la suite de ce chapitre, nous aurons besoin d'identifier canoniquement les polynômes de E_m aux vecteurs de leurs coordonnées dans une base quelconque de $\mathbb{R}^{|m|} \times \mathbb{R}$ (et notamment B_m). La définition suivante d'un isomorphisme canonique entre $\mathbb{R}^{|m|} \times \mathbb{R}$ et E_m permet de s'affranchir de notations lourdes pour la suite.

Définition 4.1.3. *Soit $\beta_m = \left\{(\beta^0_I)_{I \in m}, \beta^1\right\}$ une base quelconque de E_m.*

On définit l'application linéaire $[\cdot \,|\, \beta_m]$ par

$$\begin{array}{rccl} [\cdot \,|\, \beta_m] \quad : & \mathbb{R}^{|m|} \times \mathbb{R} & \longrightarrow & E_m \\ & a & \longmapsto & [a \,|\, \beta_m] = \displaystyle\sum_{I \in m} a^0_I \beta^0_I + a^1 \beta^1. \end{array}$$

Remarque : en particulier, pour $a \in \mathbb{R}^{|m|} \times \mathbb{R}$ on a

$$[a \,|\, B_m] \;=\; \sum_{I \in m} a^0_I \varphi^0_I + a^1 \varphi^1$$

donc

$$\int_0^1 [a \,|\, B_m]^2 \; d\mu \;=\; \sum_{I \in m} {a^0_I}^2 + {a^1}^2 \;=\; \|a\|_2^2.$$

Franchissons une petite étape supplémentaire en définissant cette fois une application surjective entre $\mathbb{R}^{|m|} \times \mathbb{R}$ et $\mathcal{E}_m$ directement.

Définition 4.1.4. *Soit t la fonction définie par*

$$\begin{array}{rccl} t \quad : & \mathbb{R}^{|m|} \times \mathbb{R} & \longrightarrow & \mathcal{E}_m \\ & a & \longmapsto & \exp([a \,|\, B_m] - \psi(a)) \end{array}$$

où

$$\begin{array}{rccl} \psi & : & \mathbb{R}^{|m|} \times \mathbb{R} & \longrightarrow & \mathbb{R} \\ & & a & \longmapsto & \log \displaystyle\int_0^1 \exp([\,a \,|\, B_m])\, d\mu. \end{array}$$

On note $\nabla\psi$ le gradient de ψ. Il s'écrit

$$\begin{array}{rccl} \nabla\psi & : & \mathbb{R}^{|m|} \times \mathbb{R} & \longrightarrow & \mathbb{R}^{|m|} \times \mathbb{R} \\ & & a & \longmapsto & \mathbb{E}_{t(a)}(B_m) \;=\; \left(\mathbb{E}_{t(a)}(\varphi_I^0)_{I\in m}, \mathbb{E}_{t(a)}(\varphi^1)\right) \\ & & & & \;=\; \left(\left(\displaystyle\int_0^1 \varphi_I^0 t(a)\, d\mu\right)_{I\in m}, \int_0^1 \varphi^1 t(a)\, d\mu\right). \end{array}$$

Remarque : l'application t est non injective.

Grâce à ces quelques outils, nous allons construire le bon paramétrage de $\mathcal{E}_m$ par un sous-espace de $\mathbb{R}^{|m|} \times \mathbb{R}$ au moyen de son image par $[\,\cdot \,|\, B_m]$.

4.1.3 Paramétrage final du modèle

Il est clair que l'application t est surjective. Intuitivement, si l'on quotiente $\mathbb{R}^{|m|} \times \mathbb{R}$ par la relation d'équivalence selon laquelle a est équivalent à b si $t(a) = t(b)$ alors t sera bijective sur l'espace quotient et cet espace sera le bon espace de paramètres de $\mathcal{E}_m$. Il suffit donc de trouver à quelles conditions sur a et b on a $t(a) = t(b)$.

Définition 4.1.5. *Soit u le vecteur unitaire de $\mathbb{R}^{|m|} \times \mathbb{R}$ défini par*

$$u = \left(\sqrt{\mu(I_1)},\, ...,\, \sqrt{\mu\left(I_{|m|}\right)},\, 0\right).$$

On pose alors $G_m = u^\perp = \left\{ a \in \mathbb{R}^{|m|} \times \mathbb{R} \;\middle/\; \displaystyle\sum_{I\in m} a_I^0 \sqrt{\mu(I)} = 0 \right\}.$

Nous allons montrer que G_m est le bon espace de paramètres dans $\mathbb{R}^{|m|} \times \mathbb{R}$. Il nous faut donc montrer que t restreinte à G_m est une bijection.

Proposition 4.1.6. *La restriction $t_{|_{G_m}}$ de t à G_m définit une bijection entre G_m et $\mathcal{E}_m$.*

Preuve : • Montrons d'abord que $t_{|_{G_m}}$ est injective :

Soient a et b dans G tels que $t(a) = t(b)$, en passant au log on a

$$[\,a \,|\, B_m] - \psi(a) \;=\; [\,b \,|\, B_m] - \psi(b)$$

donc, appliquant la propriété (i) du Lemme 4.6.1 on obtient

$$[\, a - \psi(a)u \,|\, B_m] \;=\; [\, b - \psi(b)u \,|\, B_m].$$

Mais, B_m est une base de E_m donc

$$a - \psi(a)u \;=\; b - \psi(b)u,$$

soit encore

$$b - a \;=\; (\psi(b) - \psi(a))u. \tag{4.1}$$

Puis, comme a et b sont dans $G_m = u^\perp$ on a

$$(b - a \,|\, u) \;=\; (\psi(b) - \psi(a))\|u\|_2^2 \;=\; \psi(b) - \psi(a) \;=\; 0.$$

Ainsi, $\psi(b) = \psi(a)$, puis en injectant dans (4.1) on obtient $a = b$ et $t_{|_{G_m}}$ est injective.

- Montrons maintenant que $t_{|_{G_m}}$ est surjective.

Soit $t \in \mathcal{E}_m$ on veut montrer qu'il existe $a \in G_m$ tel que $t = t(a)$. Soit alors $t \in \mathcal{E}_m$. Par définition de $\mathcal{E}_m$, il existe $b \in \mathbb{R}^{|m|} \times \mathbb{R}$ tel que

$$t \;=\; t(b) \;=\; \exp([\, b \,|\, B_m])$$

et en particulier, $\psi(b) = 0$.

Soit alors a la projection orthogonale de b sur G_m. Comme u est un vecteur unitaire on a

$$a = b - (b \,|\, u)\, u$$

Utilisant les propriétés du Lemme 4.6.1, il vient

$$\begin{aligned}
\log t(a) &= [\, a \,|\, B_m] - \psi(a) \\
&= [\, b - (b \,|\, u)\, u \,|\, B_m] - \psi(b - (b \,|\, u)\, u) \\
&= [\, b \,|\, B_m] - (b \,|\, u)\, [\, u \,|\, B_m] - \psi(b) + (b \,|\, u) \\
&= [\, b \,|\, B_m] - (b \,|\, u) + (b \,|\, u) \\
&= [\, b \,|\, B_m] \\
&= \log t(b)
\end{aligned}$$

d'où la surjectivité de $t_{|_{G_m}}$. ♣

On en déduit alors que le bon sous-espace de polynômes de E_m de dimension $|m|$ sous-jacent à $\mathcal{E}_m$ est caractérisé par la définition qui suit :

Définition 4.1.7. *Soit S_m le sous-espace de E_m défini par*

$$S_m = \{\, [\, a \,|\, B_m] / \; a \in G_m \}\,.$$

Remarque : par définition, S_m est aussi l'orthogonal à la fonction constante égale à 1 dans $\mathbb{L}^2$:

$$S_m = \mathbb{1}^{\perp} = \left\{ s \in E_m \;\middle/\; < s \,|\, \mathbb{1} > = \int_0^1 s \, d\mu = 0 \right\}.$$

Il en résulte alors que

$$\mathcal{E}_m = \left\{ h \in \mathbb{L}^1([0\,;1],\mu) \;\middle/\; h > 0,\, \log h \in S_m,\, \int_0^1 h \, d\mu = 1 \right\} \tag{4.2}$$

avec S_m est de dimension $|m|$. Par abus de langage, nous dirons aussi que le modèle $\mathcal{E}_m$ est de dimension $|m|$ (en tout état de cause, il s'agit d'une variété de dimension $|m|$).

Remarque : la nécessité de se restreindre au sous-espace G_m de $\mathbb{R}^{|m|} \times \mathbb{R}$ pour la définition d'un paramétrage bijectif de $\mathcal{E}_m$ est le pendant de la contrainte $\max\{\omega_I, I \in m\} = 1$ que l'on impose dans les hypothèses (Hc) et (Hnc) présentées dans la section 2.3.3. Sans cette dernière contrainte, le modèle n'est pas identifiable car ω ne peut alors être définie qu'à constante multiplicative près, tout comme un élément de $\mathcal{E}_m$ n'est défini qu'à translation par u près par son paramétrage dans $\mathbb{R}^{|m|} \times \mathbb{R}$. Imposer à la classe de fonctions de poids ω le fait que $\max\{\omega_I, I \in m\} = 1$ est ainsi équivalent à imposer de paramétrer les éléments de $\mathcal{E}_m$ par les éléments de G_m.

Conformément à l'hypothèse (Hc) que nous avons utilisée aux chapitres 2 et 3 nous devons aussi définir, pour toute partition m de $[\,0\,;1]$, un modèle exponentiel de polynômes par morceaux de degré 1 dont tous les termes de degré 1 sont égaux et dont les termes de degré 0 sont croissants. Nous procédons alors par analogie avec ce qui a été vu ci-dessus.

Définition 4.1.8. *On note $\dot{G}_m$ l'intersection de demi-sous-espaces fermés de G_m définie par*

$$\dot{G}_m = \left\{ a \in G_m \;\middle/\; \frac{a_1^0}{\sqrt{\mu(I_1)}} - \frac{2\sqrt{3}\,\mathrm{m}(I_1)a^1}{\sqrt{\sum_{I\in m}\mu(I)^3}} \leq \cdots \right.$$

$$\left. \cdots \leq \frac{a_{|m|}^0}{\sqrt{\mu(I_{|m|})}} - \frac{2\sqrt{3}\,\mathrm{m}(I_{|m|})a^1}{\sqrt{\sum_{I\in m}\mu(I)^3}} \right\}$$

où $m(I)$ désigne le milieu du segment I.

On construit alors, pour tout $i = 1, ..., |m| - 1$, le vecteur $\dot{u}_i$ de $\mathbb{R}^{|m|} \times \mathbb{R}$:

$$\dot{u}_i = \left(\left(0, ..., 0, \frac{1}{\sqrt{\mu(I_i)}}, \frac{-1}{\sqrt{\mu(I_{i+1})}}, 0, ..., 0 \right), \frac{2\sqrt{3}(\mathrm{m}(I_i) - \mathrm{m}(I_{i+1}))}{\sqrt{\sum_{I \in m} \mu(I)^3}} \right),$$

de sorte que

$$\dot{G}_m \;=\; \{a \in G_m \,/\, \forall\; i = 1, ..., |m| - 1,\; (a \,|\, \dot{u}_i) \leq 0\}$$

On note alors

$$\dot{S}_m = \left\{ [\, a \,|\, B_m] / \; a \in \dot{G}_m \right\}$$

et $\dot{\mathcal{E}}_m$ le sous-ensemble convexe et fermé de $\mathcal{E}_m$ défini par $\dot{\mathcal{E}}_m = t(\dot{G}_m)$.

Les sous-ensembles convexes et fermés $\dot{G}_m$ et $\dot{S}_m$ de G_m et S_m sont des variétés de dimension $|m|$. La dimension de $\dot{\mathcal{E}}_m$ est donc elle aussi $|m|$, comme celle de $\mathcal{E}_m$.

Nos modèles étant maintenant clairement définis et paramétrés pour toute partition m de l'intervalle $[0\,;1]$, il nous faut nous munir de moyens d'évaluer la qualité d'approximation de la densité f au moyen de chacun de ces modèles en vue de pouvoir sélectionner ensuite le "meilleur".

4.2 Distance de Kullback et log-vraisemblance

Nous cherchons à mesurer l'écart entre la vraie densité inconnue f et son estimateur du maximum de vraisemblance $\hat{f}$. Dans cette section, nous allons définir les notions de distances de Kullback et de Hellinger entre densités de probabilité ainsi que les fonctions de risques associées. Nous montrerons comment la première intervient de façon naturelle dans l'évaluation du risque de l'estimateur du maximum de vraisemblance d'une densité inconnue.

4.2.1 Distances entre densités de probabilités

La distance (ou quantité d'information) de Kullback-Leibler est un outil classique de la théorie de l'information. Son intérêt est tout aussi important dans la théorie de l'estimation par maximum de vraisemblance, comme nous allons le voir.

Définition 4.2.1. *Soient P et Q deux mesures de densités respectives p et q définies sur $\mathbb{R}$ contre une mesure μ. On appelle distance de Kullback-Leibler,*

ou de façon plus rapide, distance de Kullback la quantité (non symétrique)

$$K(p,q) = \begin{cases} \displaystyle\int_0^1 p \log \frac{p}{q}\, d\mu & \text{si } P \text{ est dominée par } Q \\ +\infty & \text{sinon.} \end{cases}$$

Considérons que la densité q représente une certaine information (en l'occurrence, la "forme" d'une distribution). La mesure P dominée par Q est susceptible présenter des analogies de comportement vis-à-vis de Q (elle doit notamment ne pas charger les ensembles de $q\,d\mu$-mesure nulle). On peut interpréter ces analogies comme une quantité d'information que p "détient" sur q. La quantité $K(p,q)$ représente alors un défaut d'information de p vis-à-vis de q. En effet, et à titre d'illustration, si $K(p,q) = 0$ alors $p = q$ presque partout et p détient la totalité de l'information contenue dans q. On peut alors dire qu'il n'y a aucun défaut d'information. À l'inverse, si p n'est pas dominée par q alors p ne possède aucune information sur q et leur distance de Kullback est ainsi infinie.

Nous utiliserons aussi dans la suite la distance de Hellinger qui présente de fortes analogies de comportement avec la distance de Kullback lorsque $\log(p/q)$ n'est pas trop grand (voir Birgé [14]).

Définition 4.2.2. *Soient p et q deux densités de probabilités définies contre la mesure μ. On définit la carré de la distance de Hellinger h^2 par*

$$h^2(p,q) = \frac{1}{2} \int_0^1 (\sqrt{p} - \sqrt{q})^2\, d\mu.$$

Un intérêt de la distance de Hellinger est qu'à l'inverse de la distance de Kullback, elle est bornée par 1. Par ailleurs, nous utiliserons le fait que (cf. Birgé [14])

$$2h^2(p,q) \;\leq\; K(p,q). \tag{4.3}$$

Les distances entre densités de probabilité permettent d'évaluer la proximité entre deux distributions. Plus généralement, nous verrons que la distance de Kullback permet aussi de mesurer la proximité entre une distribution et une classe, ou modèle, de distributions. Ainsi, si plusieurs estimateurs appartenant à plusieurs modèles sont en compétition, la distance de Kullback fournit alors un critère permettant d'arbitrer entre les modèles concurrents.

Classiquement, on parle du risque d'un estimateur comme l'espérance de sa distance à la densité inconnue. On distingue donc risque Kullback et risque Hellinger.

Établissons maintenant le lien entre la distance de Kullback et la procédure d'estimation par maximum de vraisemblance.

4.2.2 Lien avec la log-vraisemblance

Fixons nous une partition m et soit alors $\mathcal{E}_m$ (ou respectivement $\dot{\mathcal{E}}_m$) le modèle exponentiel associé.

Définition 4.2.3. *La log-vraisemblance d'une densité t est donnée par*

$$\gamma_n(t) \;=\; P_n(-\log t) \;=\; -\frac{1}{n}\sum_{i=1}^{n}\log t(Y_i).$$

On définit de plus le critère empirique k_n par

$$k_n(f,t) \;=\; \gamma_n(t) - \gamma_n(f).$$

L'estimateur du maximum de vraisemblance de f est donc la densite $\hat{f}$ qui minimise la log-vraisemblance. Elle minimise donc aussi la quantité $k_n(f,t)$ pour t variant dans $\mathcal{E}_m$, puisque $\gamma_n(f)$ est fixée. Comme pour la distance de Kullback, ce critère empirique s'interprète comme une mesure du défaut "d'information empirique" (inconnu puisque f est inconnue) que $\hat{f}$ présente vis-à-vis de f. L'estimation par maximum de vraisemblance vise donc à minimiser ce défaut d'information empirique.

Par ailleurs, pour une densité t fixée on a

$$\mathbb{E}_f\left(\gamma_n(t)\right) = -\int_0^1 \log(t) f \, d\mu$$

donc

$$K(f,t) = \mathbb{E}_f\left(\gamma_n(t) - \gamma_n(f)\right) = \mathbb{E}_f\left(k_n(f,t)\right).$$

La Proposition 4.6.7 montre qu'il existe une unique densité $\bar{f}_m$ inconnue qui minimise $K(f,t)$ lorsque t varie dans $\mathcal{E}_m$. On l'appelle projection Kullback de f sur $\mathcal{E}_m$ et représente l'élément de $\mathcal{E}_m$ qui contient le plus d'information sur f. Le terme de projection est ici parfaitement adapté puisque Csiszàr [26] a montré que lorsque deux densités p et q sont peu éloignées, la quantité $K(p,q)$ possède les mêmes propriétés que le carré d'une distance euclidienne.

Remarque : l'ensemble des descriptions faites ci-dessus restent valables si l'on remplace $\mathcal{E}_m$ par $\dot{\mathcal{E}}_m$. On peut donc définir de façon analogue la projection Kullback $\bar{\dot{f}}_m$ et l'estimateur du maximum de vraisemblance $\hat{\dot{f}}$ de f sur $\dot{\mathcal{E}}_m$.

4.3 Heuristique de la sélection de modèles

Dans toute la suite, et implicitement, lorsque le modèle de référence est le modèle contraint par (Hc), toutes les définitions ci-dessus et à venir deviennent relatives à (Hc), c'est-à-dire qu'elles se munissent d'un point " $\cdot$ " dans les

notations. Pour éviter les lourdeurs, dans la suite et sauf mention contraire explicite, nous utilisons les notations relatives au modèle $\mathcal{E}_m$.

Il est clair que la quantité inconnue $K(f, \hat{f}_m)$ dépend, à n fixé, de f et de la dimension $|m|$ du modèle $\mathcal{E}_m$. Le but de la sélection de modèles est de choisir $\mathcal{E}_m$, via sa dimension, de façon à minimiser le risque $\mathbb{E}_f(K(f, \hat{f}_m))$ pour un ensemble fini $\mathcal{M}$ de partitions m de l'intervalle $[0\,;1]$.

Le modèle optimal (donc "l'estimateur" optimal) et en particulier sa dimension sont inconnus puisqu'ils dépendent de f.

Définition 4.3.1. *Nous appelons oracle et notons $\check{f} \in \mathcal{E}_m$ la densité optimale :*

$$\check{f} \;=\; \operatorname{argmin}\{\mathbb{E}_f(K(f, \hat{f}_m)) \,/\, m \in \mathcal{M}\}$$

Notre travail consiste donc à construire un critère empirique capable de "faire aussi bien que l'oracle", c'est-à-dire sélectionner la partition $\hat{m}$ en fonction des observations, donc le modèle $\mathcal{E}_{\hat{m}}$, de façon à ce que le risque $\mathbb{E}_f(K(f, \hat{f}_{\hat{m}}))$ de l'estimateur $\hat{f}$ soit majoré (à une constante près) par le risque de l'oracle : $\mathbb{E}_f(K(f, \check{f})) = \inf_{m \in \mathcal{M}} \{\mathbb{E}_f(K(f, \hat{f}_m))\}$.

4.3.1 Description

Pour construire le critère empirique envisagé ci-dessus, nous évaluons le risque d'un estimateur du maximum de vraisemblance de f à m fixé.

En premier lieu, nous décomposons la distance de Kullback de f à $\hat{f}_m$ en faisant intervenir la projection Kullback $\bar{f}_m$ de f :

$$\begin{aligned} K(f, \hat{f}_m) &= \mathbb{E}_f(\gamma_n(\hat{f}_m) - \gamma_n(f)) \\ &= \mathbb{E}_f(\gamma_n(\hat{f}_m) - \gamma_n(\bar{f}_m)) + \mathbb{E}_f(\gamma_n(\bar{f}_m) - \gamma_n(f)) \\ &= \int_0^1 f \log \bar{f}_m \, d\mu - \int_0^1 f \log \hat{f}_m \, d\mu + K(f, \bar{f}_m). \end{aligned}$$

Ensuite, nous utilisons la propriété de la Proposition 4.6.7, en notant que $\log \bar{f}_m$ et $\log \hat{f}_m$ appartiennent à l'espace vectoriel S_m sous-jacent au modèle $\mathcal{E}_m$ pour obtenir

$$K(f, \hat{f}_m) \;=\; K(f, \bar{f}_m) + \int_0^1 \bar{f}_m \log \bar{f}_m \, d\mu - \int_0^1 \bar{f}_m \log \hat{f}_m \, d\mu$$

et finalement

$$K(f, \hat{f}_m) \;=\; K(f, \bar{f}_m) + K(\bar{f}_m, \hat{f}_m). \tag{4.4}$$

En prenant ensuite l'espérance sous f de cette égalité de type Pythagore on obtient

$$\mathbb{E}_f\left(K(f,\hat{f}_m)\right) \;=\; K(f,\bar{f}_m) + \mathbb{E}_f\left(K(\bar{f}_m,\hat{f}_m)\right). \tag{4.5}$$

Le risque $\mathbb{E}_f\left(K(f,\hat{f}_m)\right)$ de $\hat{f}_m$ s'exprime donc comme la somme de deux termes de nature différente.

Le premier terme du membre de droite $K(f,\bar{f}_m)$ de (4.5), non aléatoire et inconnu, représente au choix :
– la distance de f au modèle $\mathcal{E}_m$;
– le défaut d'information du modèle $\mathcal{E}_m$ par rapport à f ;
– le biais dans l'estimation de f par des éléments du modèle $\mathcal{E}_m$.

Le second terme du membre de droite $\mathbb{E}_f\left(K(\bar{f}_m,\hat{f}_m)\right)$ de (4.5) représente l'espérance de la distance entre $\hat{f}_m$ et $\bar{f}_m$ à l'intérieur du modèle $\mathcal{E}_m$. De nouveau, ce terme peut s'interpréter comme :
– la qualité d'estimation de $\bar{f}_m$ par $\hat{f}_m$ dans le modèle $\mathcal{E}_m$;
– un terme de variance dans l'estimation de f à l'intérieur du modèle $\mathcal{E}_m$.

Nous souhaitons minimiser l'expression (4.5), pour m appartenant à la famille finie de partitions $\mathcal{M}$. Or il est clair que les deux termes varient en sens inverses l'un de l'autre lorsque la dimension du modèle augmente.

En effet, plus la dimension du modèle est grande, et plus la quantité d'information sur f portée par le modèle est susceptible d'être grande. Donc le terme $K(f,\hat{f}_m)$ décroit. Mais à l'inverse, en augmentant le nombre de degrés de liberté, on autorise une plus grande variabilité de $\hat{f}_m$ à l'intérieur du modèle $\mathcal{E}_m$. Ainsi, le terme $\mathbb{E}_f\left(K(\bar{f}_m,\hat{f}_m)\right)$ a-t-il, pour sa part, tendance à croître.

Nous recherchons donc un critère statistique basé uniquement sur les observations qui fasse le meilleur compromis possible entre biais et variance pour minimiser en m le risque $\mathbb{E}_f\left(K(f,\hat{f}_m)\right)$. Pour cela, continuons à décortiquer ce terme...

Toujours par la propriété de la Proposition 4.6.7 on a

$$\mathbb{E}_f\left(K(f,\hat{f}_m)\right) \;=\; \int_0^1 f\log f\,d\mu - \int_0^1 \bar{f}_m\log\bar{f}_m\,d\mu + \mathbb{E}_f\left(K(\bar{f}_m,\hat{f}_m)\right).$$

Le terme $\int_0^1 f\log f\,d\mu$ étant constant, la minimisation de $\mathbb{E}_f\left(K(f,\hat{f}_m)\right)$ est donc équivalente à celle de

$$Q(m) \;=\; -\int_0^1 \bar{f}_m\log\bar{f}_m\,d\mu + \mathbb{E}_f\left(K(\bar{f}_m,\hat{f}_m)\right). \tag{4.6}$$

Or, sous certaines hypothèses, la variable aléatoire $K(\bar{f}_m, \hat{f}_m)$ se concentre autour de la valeur $|m|/2n$. Ce phénomène (dans lequel les inégalités de concentration de Talagrand [78] interviennent de façon centrale) est à rapprocher fondamentalement du phénomène de concentration du χ^2 multinômial[2] autour de son espérance (égale à $(|m|-1)/n$) décrit par Massart [58] et Castellan [20] dans sa thèse.

Ainsi, dans la quantité $Q(m)$ donnée par (4.6), la dépendance en l'inconnue f intervient principalement dans le terme $-\int_0^1 \bar{f}_m \log \bar{f}_m \, d\mu$ puisque $\mathbb{E}_f\left(K(\bar{f}_m, \hat{f}_m)\right)$ est de l'ordre de $|m|/2n$, ce de façon relativement indépendante de f.

Notre parti pris est alors d'estimer $-\int_0^1 \bar{f}_m \log \bar{f}_m \, d\mu$ de manière naturelle par $-\int_0^1 \hat{f}_m \log \hat{f}_m \, d\mu$. Le terme $Q(m)$ s'exprime donc de la façon suivante :

$$Q(m) \;=\; -\int_0^1 \hat{f}_m \log \hat{f}_m \, d\mu + \mathbb{E}_f\left[K(\bar{f}_m, \hat{f}_m) + K(\hat{f}_m, \bar{f}_m)\right] + \left[\int_0^1 \hat{f}_m \log \hat{f}_m \, d\mu - \mathbb{E}_f\left(\int_0^1 \hat{f}_m \log \hat{f}_m \, d\mu\right)\right].$$

Or, le terme $\mathbb{E}_f\left[K(\bar{f}_m, \hat{f}_m) + K(\hat{f}_m, \bar{f}_m)\right]$ du membre de droite de cette équation est de l'ordre de $|m|/n$ et le dernier terme est, par définition, d'espérance nulle.

Heuristiquement, on obtient alors

$$Q(m) \;\simeq\; -\int_0^1 \hat{f}_m \log \hat{f}_m \, d\mu + \frac{|m|}{n}.$$

On peut donc espérer que la partition $\hat{m}$ qui minimise le critère

$$\mathrm{crit}_n(m) \;=\; -\int_0^1 \hat{f}_m \log \hat{f}_m \, d\mu + \frac{|m|}{n} \;=\; \gamma_n(\hat{f}_m) + \frac{|m|}{n}. \qquad (4.7)$$

fournira un "bon" estimateur $\tilde{f} = \hat{f}_{\hat{m}}$ de f et l'on tombe ainsi sur le critère d'Akaïke [2], qui l'avait introduit dans un contexte plus général.

Le terme en $|m|/n$ intervient donc pour compenser deux effets :
– la tendance décroissante avec $|m|$ de $\gamma_n(\hat{f}_m)$
– la tendance croissante avec n de $\gamma_n(\hat{f}_m)$.

Il joue donc un rôle de pénalité vis-à-vis de ces monotonies. D'autres critères pénalisés comparables ont été introduits par divers auteurs comme Mallows [55] dans le contexte de la régression ou voir Grasa [37] pour une revue de

[2]Rappelons que $\chi^2_n(m)$ est défini par $\chi^2_n(m) = \sum_{I \in m} \frac{(P_n(I) - P(I))^2}{P(I)}$

diverses méthodes de sélection de modèles par pénalisation d'un contraste utilisées en économétrie. Sur ce dernier thème, les critères les plus connus sont ceux de :
– Akaïke encore avec pénalité en $|m|/n$;
– Schwarz avec pénalité en $|m| \log n/2n$;
– Hannan-Quinn avec pénalité en $|m| \log\log n/n$.

De façon plus générale, nous considèrerons un critère de séléction de modèles du type

$$\mathrm{crit}_n(m) = \gamma_n(\hat{f}_m) + \mathrm{pen}_n(m) \tag{4.8}$$

pour une fonction dite de pénalité pen_n dont nous déterminerons la forme adéquate par la suite. La procédure de minimisation du critère s'appelle sélection de modèles par maximum de vraisemblance pénalisé.

Définition 4.3.2. *Pour toute la suite, on se donne une famille finie $\mathcal{M}_n$ de partitions de $[0\,;1]$, qui peut dépendre du nombre d'observations n, et dont nous préciserons au besoin certaines propriétés.*

Pour tout m de $\mathcal{M}_n$ on note $\hat{f}_m$ l'estimateur du maximum de vraisemblance de f sur le modèle $\mathcal{E}_m$. Soit $\hat{m}$ tel que $\hat{m} = \arg\min\{\mathrm{crit}_n(m),\, m \in \mathcal{M}\}$ où $\mathrm{crit}_n(m)$ est défini par 4.8. On appelle estimateur du maximum de vraisemblance pénalisé l'estimateur $\tilde{f}$ défini par

$$\tilde{f} \;=\; \hat{f}_{\hat{m}}.$$

De façon totalement analogue, on définit l'estimateur $\widetilde{\dot{f}}$ du maximum de vraisemblance pénalisé sous les contraintes définies par l'hypothèse (Hc) par $\widetilde{\dot{f}} = \widehat{\dot{f}}_{\hat{m}}$.

L'un des objectifs de notre travail est de montrer que la fonction de pénalité à choisir peut être prise comme étant la même que l'on soit sous l'hypothèse (Hc), sous l'hypothèse (Hnc) ou encore sans aucune contrainte sur les coefficients des polynômes, comme dans le cadre du Théorème 2.3.2 de Castellan [20].

4.3.2 Commentaire

Il est primordial de noter que, même si f appartient à un modèle, le modèle idéal d'estimation au sens de la minimisation du risque *n'est pas nécessairement le "vrai" modèle auquel appartient f*.

Supposons par exemple que la vraie densité f appartienne exactement à un modèle de grande dimension $|m|$ et que le nombre d'observations n ne soit pas très grand devant $|m|$. Si le terme de biais à l'intérieur du vrai modèle

de dimension $|m|$ est évidemment nul, la variance, à l'inverse risque d'être extrêmement grande compte tenu du nombre élevé de degrés de liberté dans l'estimation. Un compromis efficace biais/variance sur un modèle de plus petite dimension a toutes les chances d'être plus intéressant en terme de risque.

Nous passons maintenant à notre principal résultat concernant la sélection de modèles exponentiels de polynômes par morceaux.

4.4 Théorème principal

4.4.1 Familles de partitions *ad hoc*

Nous verrons dans la suite que la forme de la fonction de pénalité qui apparaît dans (4.8) et dans le Théorème 4.4.1 dépend largement de la "richesse" de la collection de partitions $\mathcal{M}_n$ de $[0\,;1]$ parmi lesquelles nous devrons effectuer la sélection de modèles. Nous sommes donc amenés à définir deux types de classes de partitions.

(H1) : *il existe un entier k tel que pour tout entier $D \leq n$*

$$|\{m \in \mathcal{M}_n \,/\, |m| = D\}| = \mathcal{O}(D^k).$$

Autrement dit, le nombre de partitions ayant le même nombre de morceaux est polynômial. C'est le cas en particulier des partitions régulières, qui n'ont qu'un seul modèle m par dimension D choisie. L'alternative est celle des partitions irrégulières.

Si de plus on note

$$\Gamma_n = \inf_{m \in \mathcal{M}_n} \inf_{I \in m} \mu(I),$$

alors on suppose qu'il existe une suite $(\theta_n)_{n\in\mathbb{N}}$ qui tend vers l'infini et une constante Γ telles que $\Gamma_n \geq \Gamma\theta_n/\sqrt{n}$.

(H2) : *soit une partition régulière m_n de $[0\,;1]$. Pour un I quelconque de m_n on note $\Gamma_n = \mu(I)$ et l'on suppose qu'il existe une suite $(\theta_n)_{n\in\mathbb{N}}$ et une constante Γ telles que $\Gamma_n \geq \Gamma\theta_n/\sqrt{n}$.*

On considère alors la famille $\mathcal{M}_n$ de toutes les partitions de $[0\,;1]$ dont les intervalles sont formés de réunions d'intervalles de m_n.

Dans ce cas, le nombre de partitions ayant le même nombre de morceaux est exponentiel : le nombre de partitions à D morceaux de $\mathcal{M}_n$ est $\mathrm{C}^D_{1/\Gamma_n}$ (Γ_n est nécessairement l'inverse d'un entier).

Remarque : il est une autre classe de partitions auxquelles nous aurions pu nous intéresser. Il s'agit de la classe des blocs statistiquement équivalents

(cf. [76] et [79]) où les partitions sont construites de façon à ce que les intervalles de $[0\,;1]$ aient tous même mesure empirique à $1/n$ près. On peut alors considérer ces partitions de façon régulière ou irrégulière comme dans les hypothèses (H1) et (H2). Nous qualifions ces partitions d'homogènes en fréquence d'observation par opposition à celles définies par (H1) et (H2) qui sont, pour leur part, homogènes en mesure de Lebesgue. Il est aussi fréquent d'employer le terme de blocs statistiquement équivalents.

Nous ne traiterons pas le cas de ces partitions d'un point de vue théorique, car alors, les points sur lesquels se basent ces partitions sont aléatoires. Cet aléa aurait bien-sûr tendance à compliquer de façon notable les résultats qui suivent. Il peut cependant s'agir d'une piste de recherches futures.

4.4.2 Énoncé du Théorème

Le Théorème qui suit donne la forme de la pénalité à appliquer au protocole de sélection de modèle décrit dans la section 4.3.1. Il montre en particulier que la pénalité peut être choisie de la même forme que le modèle de référence soit non contraint (hypothèse (Hnc)) ou contraint à la monotonie des coefficients de degré 0 (hypothèse (Hc)).

Théorème 4.4.1. *Soit $\{Y_1, ..., Y_n\}$ un n-échantillon de variables aléatoires i.i.d. à valeurs dans $[0\,;1]$ de densité f par rapport à la mesure μ. On suppose que la famille de partitions $\mathcal{M}_n$ satisfait (H_1) ou (H_2). Soit $(L_m)_{m\in\mathcal{M}_n}$ une suite de poids positifs. On pose*

$$\Sigma_n \;=\; \sum_{m\in\mathcal{M}_n} \exp(-|m|L_m). \tag{4.9}$$

Pour tout m de $\mathcal{M}_n$ on note $\mathcal{E}_m$ le modèle exponentiel de polynômes par morceaux donné par la Définition 4.2.

On se donne enfin une suite ρ_n de réels positifs qui tend vers 0 telle que $\lim_{n\to+\infty} \theta_n\rho_n^2/\log n = 0$ et une constante $\lambda > 1/2$. On choisit alors une fonction de pénalité pen_n telle que pour tout m de $\mathcal{M}_n$ on a

$$\mathrm{pen}_n(m) \;\geq\; \lambda\left(1+\sqrt{2(1+1/\lambda)L_m}\right)^2\frac{|m|}{n}.$$

On suppose que f est positive et que $\log f \in \mathbb{L}^\infty([0\,;1],\mu)$.

Il existe alors un évènement Ω_n et deux constantes explicites c et c' telles que l'estimateur du maximum de vraisemblance pénalisé $\tilde{f}$ dans $\mathcal{E}_{\hat{m}}$ (respectivement $\widetilde{\dot{f}}$ dans $\mathcal{E}_{\widehat{\dot{m}}}$) donné par la Définition 4.3.2 vérifie

$$\mathbb{E}\left(h^2(f,\tilde{f})\mathbb{1}_{\Omega_n\cap\{\tilde{f}\geq\rho_n\}}\right) \leq\; c\inf_{m\in\mathcal{M}_n}\left\{2K(f,\bar{f}_m)+\mathrm{pen}_n(m)+c'\frac{2\Sigma_n+1}{n}\right\}.$$

De plus, pour tout $\alpha > 0$ on a

$$\mathbb{P}({}^c\Omega_n) \;=\; \mathcal{O}(n^{-\alpha}).$$

Remarque (1) : dans l'énoncé du Théorème ci-dessus, la constante c ne dépend que de λ. De plus, on peut montrer que si λ tend vers $1/2$ alors c tend vers l'infini. L'évènement Ω_n et la constante c' dépendent de f et de la constante λ.

Remarque (2) : les suites $\theta_n = \log^4 n$ et $\rho_n = 1/\log n$ satisfont les hypothèses du Théorème.

4.4.3 Commentaires

Le fait que la densité f doive être minorée apparaît comme une condition essentiellement technique qui doit pouvoir être relaxée en pratique. En particulier, s'affranchir de cette hypothèse serait un pas intéressant pour permettre de passer à l'estimation de densités non bornées, ce qui est notre cas en pratique.

Notre énoncé paraît peut-être un peu moins fort que celui du Théorème 2.3.2 de Castellan [20] en raison de la limitation du nombre de morceaux de la partition qui doit être en $\mathcal{O}\left(\sqrt{n}/\log^2 n\right)$ au lieu de $\mathcal{O}\left(n/\log^2 n\right)$. Ceci est dû au fait que dans ce dernier cas l'estimation de la densité inconnue f se fait intervalle par intervalle alors que nous devons introduire des dépendances sur les coefficients à estimer entre tous les intervalles des partitions sur lesquelles nous "testons" l'estimation. Ces dépendances s'interprètent comme des contraintes sur les estimateurs, conformément à notre travail du chapitre 3. Ce sont ces dépendances "tout au long de l'intervalle $[\,0\,;1]$" qui impliquent de disposer de plus de données par intervalle pour obtenir que les estimateurs sous contraintes vérifient les mêmes inégalités que les estimateurs libres de toute contrainte comme dans le cas traité par Castellan [20]. Notons enfin que cette "restriction" n'a pas d'impact pratique puisque le contrôle du nombre de morceaux de la partition ne se fait ici qu'à constante multiplicative près.

Nous pensons qu'il doit-être délicat de retourner à la seule hypothèse sur la dimension des modèles en $\mathcal{O}\left(n/\log^2 n\right)$ et de conserver la même conclusion dans le cas où les contraintes sur les coefficients des polynômes du modèle exponentiel impliquent des liaisons sur les coefficients de plusieurs intervalles en même temps – comme par exemple dans le cas des modèles exponentiels de splines où les liaisons portent sur les coefficients de polynômes de degré 3 sur des intervalles deux-à-deux contigus.

Nous pensons en revanche que le résultat ci-dessus avec l'hypothèse de type $\mathcal{O}\left(\sqrt{n}/\log^2 n\right)$ devrait se généraliser à tout modèle exponentiel de polynôme

par morceau sous contraintes de type polyèdre convexe fermé et probablement même sous contraintes convexes fermées générales. Un tel résultat engloberait en particuliers les modèles exponentiels de splines et, en dehors de notre cadre de thèse, fournirait un résultat intéressant sur la sélection de tels modèles. Là encore, nous sommes devant une piste de recherche pour des travaux futurs.

4.4.4 Choix de la fonction de pénalité

Le Théorème 4.4.1 fournit la forme de la fonction de pénalité, en fonction d'une suite de poids dépendant des familles de modèles considérés, ce à constante multiplicative près.

Dans cette section, nous étudions d'abord l'influence des poids $(L_m)_{m\in\mathcal{M}_n}$ sur la fonction de pénalité puis nous nous intéressons à la constante multiplicative à affecter à la forme de la pénalité "générique" que nous aurons alors choisie.

Choix des poids

Nous reprenons ici l'argumentaire de Castellan [20].

Celle-ci a montré que dans le cas (H1) des partitions régulières, l'application de poids constants $L_m = L$ pour tout $m \in \mathcal{M}_n$ conduit à une fonction de pénalité du type

$$\mathrm{pen}_n(m) = \lambda \frac{|m|}{n},$$

pour tout $\lambda > 1/2$. Le cas $\lambda = 1$ correspondant au critère d'Akaïke.

Le choix de poids variables, comme $L_m = 1/\sqrt{|m|}$ semble cependant améliorer le risque non-asymptotiquement, comme cela est démontré en ce qui concerne les histogrammes. Castellan recommande l'utilisation d'une pénalité du type

$$\mathrm{pen}_n(m) = \frac{|m|}{n} + \lambda \frac{|m|^{3/4}}{n},$$

pour une constante λ "raisonnable".

Dans le cas des partitions irrégulières, la complexité de la famille, *i.e.* le fait qu'il existe un nombre exponentiel de modèles de même dimension rend plus difficile la convergence de la série servant à définir la constante Σ_n, cf. (4.9). Afin de rendre Σ_n indépendante de n, un choix de poids constants[3] de l'ordre de $\log n$ est nécessaire. Cela mène alors à une fonction de pénalité de la forme suivante :

$$\mathrm{pen}_n(m) = \lambda \frac{|m| \log n}{n},$$

[3]Constants par rapport à la dimension du modèle, mais non par rapport à n.

pour un certain λ positif.

De nouveau, il est recommandé un choix de poids variables conduisant à une pénalité du type

$$\mathrm{pen}_n(m) = \lambda \frac{|m|}{n} \log \frac{\lambda'}{\Gamma_n |m|},$$

pour des constantes λ et λ' correctement calibrées. En pratique, nous choisirons $\lambda' = e^{\frac{5}{2}}$, comme suggéré par le travail de Lebarbier [50].

La calibration de la constante λ est un épineux problème auquel l'heuristique de pente que nous développons dans la suite semble apporter une réponse adéquate.

Heuristique de pente

Le but de cette section est de fournir une méthode qui permet de déterminer empiriquement la bonne constante par laquelle multiplier la fonction "générique" de pénalité déterminée grâce au théoreme 4.4.1. Nous commençons par décrire cette heuristique, puis, la façon dont nous l'appliquons en pratique. Il est à noter qu'une étude de simulation fournie dans Castellan [20] avait conduit au calibrage de certaines constantes pour une classe assez variée de densités. La méthode dite "de la pente" permet d'avoir une adaptation des constantes à l'échantillon de données que l'on traite.

L'heuristique de pente et son utilisation pratique est décrite en détail dans Lebarbier [50], nous ne donnons ici qu'un résumé sommaire et pratique de la méthode.

Nous cherchons à comprendre comment bien pénaliser la log-vraisemblance $\gamma_n(\hat{f}_m)$ afin de minimiser en m le risque inconnu $\mathbb{E}_f\left(K(f, \hat{f}_m)\right)$. On cherche donc une fonction de pénalité pen_n qui soit telle qu'en minimisant en m le critère $\mathrm{crit}_n(m) = \gamma_n(\hat{f}_m) + \mathrm{pen}_n(m)$, on récupère une partition $\hat{m}$ telle que $\mathbb{E}_f\left(K(f, \hat{f}_{\hat{m}})\right)$ soit du même ordre (ou, plus précisément, contrôlé) par le risque de l'oracle $\min_{m \in \mathcal{M}_n} \mathbb{E}_f\left(K(f, \hat{f}_m)\right)$ comme dans le Théorème 4.4.1. Le problème reste donc le choix de la constante que l'on fait intervenir devant le terme de pénalité.

L'idée de l'heuristique de pente résulte du travail de Massart [59], qui l'a formulée et validée dans le contexte gaussien. Une étude de simulation très complète montrant son efficacité dans ce même contexte a récemment vu le jour (voir Lebarbier [50]). Nous renvoyons le lecteur à ces références pour toute précision théorique ou algorithmique.

Cette heuristique repose sur les observations suivantes :

- pour un entier D fixé, appelons $\hat{f}_D$ l'estimateur du maximum de vraisemblance qui possède la plus grand vraisemblance parmi tous les estimateurs associés à des modèles de dimension exactement D ;

- la décroissance de $\gamma_n(\hat{f}_D)$ en fonction de D est approximativement linéaire ;
- l'opposé de la pente de la droite associée fournit une constante minimale pour le fonctionnement du critère pénalisé ;
- le double de cette dernière valeur fournit la constante optimale.

D'un point de vue pratique, il suffirait donc d'effectuer une régression au sens des moindres carrés du nuage des couples $(D, \gamma_n(\hat{f}_D))_{1 \leq D \leq D_{\max}}$[4] pour en estimer la pente. Le double de l'opposé de cette valeur estimée nous fournirait alors la constante cherchée.

En pratique pourtant, ces considérations sont difficilement vérifiées. Nous adoptons alors la méthodologie calibrée par Lebarbier [50], qui constate qu'en associant à la suite des pentes de l'enveloppe convexe du nuage de points $(D, \gamma_n(\hat{f}_D))_{1 \leq D \leq D_{\max}}$ la suite des dimensions associées dans le nuage, on observe des sauts de dimension qui peuvent être importants.

La pente minimale est alors celle qui conduit au plus grand écart de dimension[5] et la constante optimale est prise comme le double de cette dernière. C'est celle que nous retenons en pratique pour les simulations et applications du chapitre 5.

4.5 Démonstration du Théorème

La preuve que nous donnons s'appuie fidèlement sur celle du Théorème 2.3.2 de Castellan [20]. Nous ne redémontrons que les points qui s'en écartent ou que nous estimons fondamentaux pour la démarche de la preuve.

La principale différence entre la preuve de Castellan et la nôtre se situe au niveau du contrôle de l'écart entre la projection Kullback et l'estimateur du maximum de vraisemblance d'une densité sur un modèle. Dans la preuve originale, ce contrôle pouvait se faire intervalle par intervalle. Dans notre cas, l'égalité des termes de degré 1 sur l'ensemble des intervalles ne nous permet pas autant de liberté. C'est pourquoi, nous le verrons dans le détail de la preuve, l'ensemble sur lequel a lieu l'inégalité de contrôle du risque du Théorème 4.4.1 est plus petit dans notre cas que dans celui de Castellan. Il en possède cependant les mêmes caractéristiques, pourvu que le nombre d'observations dans chaque intervalle soit suffisamment grand, et plus précisément en $\mathcal{O}(\sqrt{n}/\log^2 n)$, contre $\mathcal{O}(n/\log^2 n)$ dans son cas.

Passons maintenant à la démonstration du Théorème. Nous utilisons les notations relatives aux modèles non contraints (sans les points " $\cdot$"), l'ensemble de la preuve développée ci-dessous s'adapte sans aucun changement au cas

[4] La valeur $D_{\max}$ peut, *a priori* être pris égal à n, mais les capacités de calcul machines imposent des valeurs nettement plus faibles, moins de 15 dans notre cas, pour n =100 à 1000.

[5] Si plusieurs pentes réalisent ce critère, la plus forte valeur absolue de pente est retenue.

contraint. Seuls les Lemmes techniques utiles en cours de preuve distinguent clairement les deux les cas lorsque cela est nécessaire.

Soit $m \in \mathcal{M}_n$, $M > 0$ et $f_m \in \mathcal{E}_m$ telle que $\|\log(f/f_m)\|_\infty \leq M$. Par la définition 4.3.2 de $\tilde{f}$ on a

$$\gamma_n(\tilde{f}) + \mathrm{pen}_n(\hat{m}) \;\leq\; \gamma_n(\hat{f}_m) + \mathrm{pen}_n(m) \;\leq\; \gamma_n(f_m) + \mathrm{pen}_n(m).$$

Puis, par la définition 3.2.1 de l'opérateur de mesure empirique recentrée ν_n ($\nu_n = P_n - P$), pour toute densité t telle que $\int_0^1 f \log t \, d\mu < +\infty$ on a

$$\gamma_n(t) \;=\; K(f,t) - \int_0^1 f \log f \, d\mu - \nu_n(\log t).$$

On en déduit ainsi que

$$K(f,\tilde{f}) \;\leq\; K(f,f_m) + \nu_n(\log \tilde{f} - \log f_m) + \mathrm{pen}_n(m) - \mathrm{pen}_n(\hat{m}). \quad (4.10)$$

La preuve du Théorème 4.4.1 réside essentiellement dans le contrôle du terme $\nu_n(\log \tilde{f} - \log f_m)$ uniformément en m. Pour ce faire, nous le décomposons en trois éléments que nous allons traiter séparément :

$$\nu_n\left(\log \frac{\tilde{f}}{f_m}\right) \;=\; \nu_n\left(\log \frac{\hat{f}_{\hat{m}}}{\bar{f}_{\hat{m}}}\right) + \nu_n\left(\log \frac{\bar{f}_{\hat{m}}}{f}\right) + \nu_n\left(\log \frac{f}{f_m}\right).$$

Nous cherchons d'abord à majorer $\nu_n\left(\log \dfrac{\hat{f}_{m'}}{\bar{f}_{m'}}\right)$ et $\nu_n\left(\log \dfrac{\bar{f}_{m'}}{f}\right)$ uniformément en m'. Pour cela, on pose

$$V_f^2(p,q) = \int_0^1 \left(\log \frac{p}{q}\right)^2 f \, d\mu, \quad (4.11)$$

pour toutes fonctions p et q telles que $\log \dfrac{p}{q} \in \mathbb{L}^2(f\,d\mu)$.

• **Contrôle de $\nu_n(\log \hat{f}_{m'} - \log \bar{f}_{m'})$:**

En premier lieu, on a

$$\left|\nu_n\left(\log \frac{\hat{f}_{m'}}{\bar{f}_{m'}}\right)\right| \;\leq\; \sup_{t \in \mathcal{E}_{m'}} \left|\nu_n\left(\frac{\log t - \log \bar{f}_{m'}}{V_f(t, \bar{f}_{m'})}\right)\right| V_f^2(\hat{f}_{m'}, \bar{f}_{m'}).$$

Ainsi, en posant

$$Z_{m'} \;=\; \sup\left\{|\nu_n(g)|,\; g \in S_{m'} \text{ et } \int_0^1 g^2 f \, d\mu = 1\right\}$$

on obtient

$$\left|\nu_n\left(\log\frac{\hat{f}_{m'}}{\bar{f}_{m'}}\right)\right| \;\leq\; Z_{m'}V_f^2(\hat{f}_{m'},\bar{f}_{m'}) \tag{4.12}$$

Soit $\varepsilon\in\,]\,0\,;1[$ et $(x_m)_{m\in\mathcal{M}_n}$ une suite de poids positifs. On introduit l'ensemble

$$\Omega_1 \;=\; \left\{Z_m\mathbb{1}_{\Omega_n[A]} \leq (1+\varepsilon)\left(\sqrt{\frac{|m|}{n}}+\sqrt{\frac{2x_m}{n}}\right),\ \forall m\in\mathcal{M}_n\right\} \tag{4.13}$$

où l'évènement $\Omega_n[A]$ est donné par la définition 4.6.9.

Puis, si A est tel que $A \;\leq\; \dfrac{4\varepsilon^2}{5\varepsilon+32}$, on déduit du Lemme 4.6.12 que

$$\mathbb{P}({}^c\Omega_1) \;\leq \sum_{m\in\mathcal{M}_n}\exp(-x_m). \tag{4.14}$$

On obtient donc que sur l'évènement Ω_1

$$\left|\nu_n\left(\log\frac{\hat{f}_{m'}}{\bar{f}_{m'}}\right)\right|\mathbb{1}_{\Omega_n[A]} \;\leq\; (1+\varepsilon)\left(\sqrt{\frac{|m'|}{n}}+\sqrt{\frac{2x_{m'}}{n}}\right)V_f^2(\hat{f}_{m'},\bar{f}_{m'}).$$

Utilisant maintenant le fait que pour tout θ strictement positif et tout couple (a,b) de réels on a

$$2ab \;\leq \theta a^2+\frac{b^2}{\theta},$$

il résulte qu'il existe deux constantes (λ_1,λ_2) telles que

$$\left|\nu_n\left(\log\frac{\hat{f}_{m'}}{\bar{f}_{m'}}\right)\right|\mathbb{1}_{\Omega_n[A]} \;\leq\; \lambda_1\left(\sqrt{\frac{|m'|}{n}}+\sqrt{\frac{2x_{m'}}{n}}\right)^2+\lambda_2V_f^2(\hat{f}_{m'},\bar{f}_{m'}).$$

Si l'on écrit ensuite $x'_m=|m'|L_{m'}+x_0$ et si l'on utilise le fait que pour tout θ strictement positif et tout couple (a,b) de réels on a

$$(a+b)^2 \;\leq\; (1+\theta)\,a^2+\left(1+\frac{1}{\theta}\right)b^2$$

on obtient finalement qu'il existe trois *nouvelles* constantes $\lambda_1>1/2$, λ_2 et λ_3 telles que

$$\left|\nu_n\left(\log\frac{\hat{f}_{m'}}{\bar{f}_{m'}}\right)\right|\mathbb{1}_{\Omega_n[A]} \;\leq\; \lambda_1\left(1+\sqrt{2L_{m'}}\right)\frac{|m'|}{n}+\lambda_2V_f^2(\hat{f}_{m'},\bar{f}_{m'})+\lambda_3\frac{x_0}{n}, \tag{4.15}$$

ce qui nous donne le contrôle souhaité.

• **Contrôle de $\nu_n(\log \bar{f}_{m'} - \log f)$:**

Toujours suivant Castellan [20], on applique la Proposition 1.4.5 page 61 qui nous montre qu'en posant

$$\Omega_2 \;=\; \left\{ \nu_n \left(\log \frac{\bar{f}_m}{f} \right) \leq K(f, \bar{f}_m) - 2h^2(f, \bar{f}_m) + 2\frac{x_m}{n}\,,\; \forall m \in \mathcal{M}_n \right\} \tag{4.16}$$

on a

$$\mathbb{P}({}^c\Omega_2) \;\leq \sum_{m \in \mathcal{M}_n} \exp(-x_m) \tag{4.17}$$

et qu'alors sur cet ensemble,

$$\nu_n \left(\log \frac{\bar{f}_{m'}}{f} \right) \;\leq\; K(f, \bar{f}_{m'}) - 2h^2(f, \bar{f}_{m'}) + 2\frac{|m'|L_{m'}}{n} + 2\frac{x_0}{n}. \tag{4.18}$$

ce qui nous donne le second contrôle voulu.

• **Contrôle de $\nu_n(\log f - \log f_m)$:**

Par l'inégalité de Bernstein (voir Pollard [70]) on a

$$\mathbb{P}\left[\left| \nu_n \left(\log \frac{f}{f_m} \right) \right| \geq V_f^2(f, f_m) \sqrt{\frac{2x_0}{n}} + \frac{Mx_0}{3n} \right] \;\leq\; \exp(-x_0).$$

Posant alors

$$\Omega_3 = \left\{ \left| \nu_n \left(\log \frac{f}{f_m} \right) \right| \leq V_f^2(f, f_m) \sqrt{\frac{2x_0}{n}} + \frac{Mx_0}{3n} \right\} \tag{4.19}$$

il vient

$$\mathbb{P}({}^c\Omega_3) \;\leq\; \exp(-x_0). \tag{4.20}$$

Par ailleurs, sur Ω_3, et de façon analogue à ci-dessus, on a que pour tout θ strictement positif :

$$\begin{aligned} \left| \nu_n \left(\log \frac{f}{f_m} \right) \right| \;&\leq\; V_f^2(f, f_m) \sqrt{\frac{2x_0}{n}} + \frac{Mx_0}{3n} \\ &\leq\; \theta V_f^2(f, f_m) + \left(\frac{2}{\theta} + \frac{M}{3} \right) \frac{x_0}{n}. \end{aligned}$$

Puis appliquant le Lemme 4.6.3 il vient que

$$K(f, f_m) \ \geq \ \Phi(-M) V_f^2(f, f_m)$$

où Φ est définie par (4.23).

Le choix $\theta = \Phi(-M)$ nous donne alors que sur l'ensemble Ω_3 il existe une constante λ (qui dépend de M) telle que

$$\left| \nu_n \left(\log \frac{f}{f_m} \right) \right| \ \leq \ K(f, f_m) + \lambda \frac{x_0}{n}, \tag{4.21}$$

d'où le troisième et dernier contrôle recherché.

• **Conclusion :**

Sur l'ensemble $\Omega_1 \cap \Omega_2 \cap \Omega_3 \cap \Omega_n[A]$, en appliquant (4.18) et (4.15) à $m' = \hat{m}$ on obtient qu'il existe deux constantes $\lambda_1 > 1/2$, λ_2 et une constante λ_3 qui dépend de M telles que

$$\begin{aligned} K(f, \tilde{f}) \ &\leq \ 2K(f, f_m) + \mathrm{pen}_n(m) - \mathrm{pen}_n(\hat{m}) \\ &+ \ \left(\lambda_1 \left(1 + \sqrt{2L_{\hat{m}}} \right)^2 + 2L_{\hat{m}} \right) \frac{|\hat{m}|}{n} + \lambda_2 V_f^2(\bar{f}_{\hat{m}}, \hat{f}_{\hat{m}}) \\ &+ \ \left(K(f, \bar{f}_{\hat{m}}) - 2h^2(f, \bar{f}_{\hat{m}}) \right) + \lambda_3 \frac{x_0}{n}. \end{aligned}$$

Par ailleurs, sur l'ensemble $\{\hat{f}_{\hat{m}} \geq \rho_n\} \cap \Omega_n[A]$, le Lemme 4.6.11 nous donne

$$\left\| \log \frac{\hat{f}_{\hat{m}}}{\bar{f}_{\hat{m}}} \right\|_\infty \ \leq \ \varepsilon.$$

On suppose donc à partir de maintenant que l'on se place sur l'ensemble $\Omega_1 \cap \Omega_2 \cap \Omega_3 \cap \Omega_n[A] \cap \{\hat{f}_{\hat{m}} \geq \rho_n\}$. Appliquant alors le Lemme 4.6.4 il vient

$$K(\bar{f}_{\hat{m}}, \hat{f}_{\hat{m}}) \ \geq \ \Phi(-\varepsilon)(1-\varepsilon) \int_0^1 \left(\log \frac{\hat{f}_{\hat{m}}}{\bar{f}_{\hat{m}}} \right)^2 f \, d\mu - 2\varepsilon \Phi(-\varepsilon) h^2(f, \bar{f}_{\hat{m}})$$

donc

$$V_f^2(\bar{f}_{\hat{m}}, \hat{f}_{\hat{m}}) \ \leq \ \frac{1}{\Phi(-\varepsilon)(1-\varepsilon)} K(\bar{f}_{\hat{m}}, \hat{f}_{\hat{m}}) + \frac{2\varepsilon}{1-\varepsilon} h^2(\bar{f}_{\hat{m}}, \hat{f}_{\hat{m}}).$$

Il en résulte cette fois-ci qu'il existe quatre constantes λ_1, λ_2, λ_3 et λ_4 telles que

$$\begin{aligned} K(f, \tilde{f}) \ &\leq \ 2K(f, f_m) + \mathrm{pen}_n(m) + \lambda_3 \frac{x_0}{n} \\ &+ \ \left(\lambda_1 \left(1 + \sqrt{2L_{\hat{m}}} \right)^2 + 2L_{\hat{m}} \right) \frac{|\hat{m}|}{n} - \mathrm{pen}_n(\hat{m}) \\ &+ \ K(f, \bar{f}_{\hat{m}}) - \lambda_2 h^2(f, \bar{f}_{\hat{m}}) + \lambda_4 K(\bar{f}_{\hat{m}}, \hat{f}_{\hat{m}}). \end{aligned}$$

Nous pouvons ensuite écrire que

$$\begin{aligned}\left(\lambda_1\left(1+\sqrt{8L_{\hat m}}\right)^2+2L_{\hat m}\right)\frac{|\hat m|}{n} &\leq \lambda_1\left(1+\sqrt{\left(2+\frac{2}{\lambda}\right)L_{\hat m}}\right)^2\frac{|\hat m|}{n}\\ &\leq \mathrm{pen}_n(\hat m)\end{aligned}$$

par définition de pen_n.

Il vient alors

$$\begin{aligned}K(f,\tilde f) &\leq 2K(f,f_m)+\mathrm{pen}_n(m)+\lambda_3\frac{x_0}{n}\\ &+ K(f,\bar f_{\hat m})-\lambda_2 h^2(f,\bar f_{\hat m})+\lambda_4 K(\bar f_{\hat m},\hat f_{\hat m}).\end{aligned}$$

On utilise ensuite l'identité de type Pythagore (4.4) pour déduire que l'on a

$$\lambda_2 h^2(f,\bar f_{\hat m})+(1-\lambda_4)K(\bar f_{\hat m},\hat f_{\hat m}) \leq 2K(f,f_m)+\mathrm{pen}_n(m)+\lambda_3\frac{x_0}{n}.$$

Par l'inégalité (4.3) et le fait que $h^2(f,\hat f_{\hat m})\leq 2h^2(f,\bar f_{\hat m})+2h^2(\bar f_{\hat m},\hat f_{\hat m})$, il existe alors deux nouvelles constantes λ_1 et λ_2 telles que sur l'évènement $\Omega_1\cap\Omega_2\cap\Omega_3$ on a

$$\lambda_1 h^2(f,\hat f_{\hat m})\mathbb{1}_{\Omega_n[A]\cap\{\hat f_{\hat m}\geq\rho_n\}} \leq 2K(f,f_m)+\mathrm{pen}_n(m)+\lambda_2\frac{x_0}{n}. \qquad (4.22)$$

Par ailleurs, les inégalités (4.14), (4.17) et (4.20) montrent que

$$\mathbb{P}({}^c\Omega_1\cup{}^c\Omega_2\cup{}^c\Omega_3) \leq (2\Sigma_n+1)\exp(-x_0).$$

On en conclut alors en intégrant (4.22) qu'il existe deux constantes c et c' (qui peuvent être explicitées au fil de la démonstration) telles que

$$\mathbb{E}\left(h^2(f,\tilde f)\mathbb{1}_{\Omega_n[A]\cap\{\tilde f\geq\rho_n\}}\right) \leq c\left\{2K(f,\bar f_m)+\mathrm{pen}_n(m)+c'\frac{2\Sigma_n+1}{n}\right\}$$

ce qui achève la preuve de la première partie du Théorème 4.4.1.

Pour la seconde partie, il suffit ensuite de prendre $\Omega_n=\Omega_n[A]$, alors le fait que pour tout $\alpha>0$ on ait $\mathbb{P}({}^c\Omega_n) = \mathcal{O}(n^{-\alpha})$ est assuré par le Lemme 4.6.10. ♣

4.6 Lemmes techniques

Cette section est dévolue aux preuves des Lemmes techniques qui interviennent dans la démonstration du théorème 4.4.1.

4.6.1 Un Lemme de nature algébrique

Les propriétés qui suivent sont évidentes mais d'utilité fréquente dans le chapitre 4 ainsi que pour la preuve de certains Lemmes nécessaires à la démonstration du théorème 4.4.1.

Lemme 4.6.1. *On considère le vecteur u et les applications $[\,\cdot\,|\,B_m]$, t et ψ respectivement introduits dans les définitions 4.1.5, 4.1.3 et 4.1.4.*

Les propriétés suivantes sont vérifiées

$$(i) \quad [\,u\,|\,B_m] = \mathbb{1}\!\!1\ ;$$

$$(ii) \quad \psi(u) = 1\,;$$

$$(iii) \quad t(u) = \mathbb{1}\!\!1\ ;$$

$$(iv) \quad \psi(a+\lambda u) = \psi(a) + \lambda\,;$$

$$(v) \quad t(a+\lambda u) = t(a).$$

Preuve : (i)

$$[\,u\,|\,B_m] \;=\; \sum_{I\in m} \sqrt{\mu(I)}\varphi_I^0 \;=\; \sum_{I\in m} \sqrt{\mu(I)}\frac{\mathbb{1}\!\!1_I}{\sqrt{\mu(I)}} \;=\; \mathbb{1}\!\!1\,.$$

(ii)

$$\psi(u) \;=\; \log\int_0^1 \exp([\,u\,|\,B_m])\,d\mu \;=\; \log\int_0^1 \exp(\mathbb{1}\!\!1\,)\,d\mu \;=\; \log(e) \;=\; 1.$$

(iii)

$$t(u) \;=\; \exp([\,u\,|\,B_m] - \psi(u)) \;=\; \exp(\mathbb{1}\!\!1 - 1) \;=\; \mathbb{1}\!\!1\,.$$

(iv)

$$\begin{aligned}
\psi(a+\lambda u) &= \log\int_0^1 \exp([\,a+\lambda u\,|\,B_m])\,d\mu \\
&= \log\int_0^1 \exp([\,a\,|\,B_m] + \lambda\mathbb{1}\!\!1\,)\,d\mu \\
&= \log\left(\exp(\lambda)\times\int_0^1 \exp([\,a\,|\,B_m])\,d\mu\right) \\
&= \psi(a) + \lambda.
\end{aligned}$$

(v)

$$\begin{aligned} t(a+\lambda u) &= \exp([\,a+\lambda u\,|\,B_m] - \psi(a+\lambda u)) \\ &= \exp([\,a\,|\,B_m] + \lambda\mathbb{1} - \psi(a) - \lambda)) \\ &= \exp([\,a\,|\,B_m] - \psi(a)) \\ &= t(a). \end{aligned}$$

♣

4.6.2 Un Lemme de nature géométrique

Le Lemme suivant, permet de déduire des propriétés d'orthogonalité d'éléments d'un modèle $\mathcal{E}_m$ dans $\mathbb{L}^2([\,0\,;1],\mu)$ à partir de propriétés d'orthogonalité de leurs paramétrages dans $G_m \subset \mathbb{R}^{|m|} \times \mathbb{R}$.

Lemme 4.6.2. *Soit E un espace vectoriel euclidien muni du produit scalaire $(\cdot\,|\,\cdot)$ et soit $\|\cdot\|_2$ la norme associée. Soit C un sous-ensemble convexe fermé de E. Les propriétés suivantes sont vérifiées :*

(i) Pour tout x de E il existe un unique élément (noté px) de C, appelé projection orthogonale de x sur C tel que

$$\|x - px\|_2 = \min_{y\in C} \|x - y\|_2.$$

Si, de plus $x \notin C$ alors $px \in \partial C$, où ∂C désigne la frontière de C.

(ii) De plus, px est caractérisé par le fait que pour tout y de C

$$(x - px\,|\,y - px) \;\leq\; 0.$$

(iii) L'application

$$\begin{array}{rcccc} p & : & E & \longrightarrow & C \\ & & x & \longmapsto & px \end{array}$$

est 1-lipschitzienne (i.e. elle réduit les distances).

(iv) Soit $\phi : E \longrightarrow \mathbb{R}$ une application strictement convexe qui atteint son minimum en un point $\bar{x}$ de E. Alors la restriction $\phi_{|C}$ de ϕ à C admet un unique minimum $\bar{x}_C$.

Remarque : il est clair que si C est un sous-espace affine de E alors on peut remplacer l'inégalité dans (ii) par une égalité. Cette égalité définit même la projection d'un point sur un sous-espace affine.

Preuve : **(*i*)** Soit $r > 0$ tel que $\bar{B}(x,r) \cap C \neq \varnothing$. La restriction à $\bar{B}(x,r)\cap C$ de l'application

$$\begin{array}{rcccc} d & : & C & \longrightarrow & \mathbb{R}^+ \\ & & y & \longmapsto & \|x - y\|_2^2 \end{array}$$

est continue sur un compact, elle admet donc un minimum. Par définition, d est strictement convexe, ce minimum est donc unique. On le note px. Il est clair que si $x \in C$ alors $px = x$ et que si $x \notin C$ alors la distance de px à C est nulle par définition de px, donc px est adhérent à C.

(ii) Il s'agit de montrer que

$$(\|x - px\|_2 \leq \|x - y\|_2 \ \forall\, y \in C) \ \Leftrightarrow \ (< x - px \,|\, y - px > \leq 0 \ \forall\, y \in C).$$

Nous supposons $x \notin C$, sinon il n'y a rien à démontrer.

Montrons d'abord le sens ($\Leftarrow$) :

$$\begin{array}{rrcl} & (x - px \,|\, y - px) & \leq & 0 \\ \Rightarrow & (x - px \,|\, x - px) & \leq & (x - px \,|\, x - y) \\ \Rightarrow & \|x - px\|_2^2 & \leq & \|x - px\|_2 \|x - y\|_2 \\ \Rightarrow & \|x - px\|_2 & \leq & \|x - y\|_2. \end{array}$$

Pour le sens ($\Rightarrow$), il s'agit de voir que sous l'hypothèse que px réalise le minimum de la fonction d de distance de x à C définie ci-dessus, alors l'hyperplan affine Π d'équation $\{y \in E \,/\, (x - px \,|\, y - px) = 0\}$ est un hyperplan d'appui de C.

Pour cela, montrons qu'il ne contient aucun point intérieur à C.

Si tel était le cas, alors il existerait une boule ouverte autour d'un point de $\Pi \cap C$ incluse dans C. Puisque ce point appartiendrait à Π, cette boule traverserait Π. En particulier, il existerait y dans C tel que $(x - px \,|\, y - px) > 0$. Par suite, comme C est convexe, le segment $]\, y\,;px[$ serait inclus dans C et tel que pour tout z de $]\, y\,;px[$ on ait $(x - px \,|\, z - px) > 0$.

Ensuite, comme Π est perpendiculaire au segment $[\, x\,;px]$ en le point px, Π est tangent à $\bar{B}(x, \|x - px\|_2)$ en px. De plus, cette boule fermée est située du "côté" du demi-espace $\{z \,/\, (x - px \,|\, z - px) > 0\}$ de Π puisqu'en particulier, pour $z = x$, $(x - px \,|\, x - px) = \|x - px\|_2^2 > 0$. Donc, comme le segment $]\, y\,;px[$ est du même côté de Π et non tangent à $\bar{B}(x, \|x - px\|_2)$, puisque non inclus dans Π, il traverse nécessairement $\bar{B}(x, \|x - px\|_2)$ en px et un un autre point $\tilde{y}$. Comme $\bar{B}(x, \|x - px\|_2)$ est convexe, on en déduit que le segment $]\, \tilde{y}\,;px[$ est inclus dans l'intérieur de C ainsi que dans la boule ouverte $B(x, \|x - px\|_2)$. En particulier, le point $\bar{y} = (\tilde{y} + px)/2$ est dans C ainsi que dans $B(x, \|x - px\|_2)$, donc $\|x - \bar{y}\|_2 < \|x - px\|_2$. Or, cette dernière inégalité est impossible car, par hypothèse, px réalise le minimum de d.

Ainsi, Π est un hyperplan d'appui de C, et comme $(x - px \,|\, x - px) = \|x - px\|_2^2 > 0$, on en déduit que pour tout y de C : $(x - px \,|\, y - px) \leq 0$.

(iii) Nous voulons montrer que pour tout (x, y) de E^2 :

$$\|px - py\| \leq \|x - y\|.$$

Comme px et py sont dans C, par la propriété (ii) on a

$$\begin{cases} (x - px \,|\, py - px) & \leq & 0\,; \\ (y - py \,|\, px - py) & \leq & 0. \end{cases}$$

En sommant ces deux inégalités on obtient :

$$\begin{array}{rcll} & (x - px + py - y \,|\, py - px) & \leq & 0 \\ \Rightarrow & \|py - px\|_2^2 & \leq & (x - y \,|\, px - py) \\ \Rightarrow & \|py - px\|_2 & \leq & \|x - y\| \end{array}$$

grâce à l'inégalité de Cauchy-Schwarz si $px \neq py$, ce dernier cas ne présentant pas d'intérêt.

(iv) De nouveau, nous supposons que $\bar{x} \notin C$. Comme ϕ est convexe et admet un unique minimum, elle est propre. C'est-à-dire que

$$\lim_{\|x\| \to +\infty} \phi(x) = +\infty.$$

Le lecteur peut s'en convaincre en voyant que les taux d'accroissements de ϕ sont strictement croissants dans toutes les directions en s'éloignant de $\bar{x}$. Pour une preuve détaillée, voir Rockafellar [73].

En conséquence, l'image réciproque d'un compact est compacte. Donc, pour $\lambda > \phi(\bar{x})$, l'ensemble

$$E_\lambda \;=\; \phi^{-1}([\phi(\bar{x}), \lambda]) \;=\; \{x \in E \,/\, \phi(\bar{x}) \leq \phi(x) \leq \lambda\}$$

est compact. Mais comme $\phi(\bar{x}) = \min\{\phi(x),\, x \in E\}$ on a aussi $E_\lambda = \{x \in E \,/\, \phi(x) \leq \lambda\}$. Or, cet ensemble est convexe car pour $(x, y) \in E_\lambda^2$ et $t \in [0, 1]$, comme ϕ est convexe on a

$$\phi(tx + (1-t)y) \;\leq\; t\phi(x) + (1-t)\phi(y) \;\leq\; t\lambda + (1-t)\lambda \;=\; \lambda.$$

Donc E_λ est un ensemble compact convexe. Posons maintenant $C_\lambda = E_\lambda \cap C$. C_λ est compact convexe comme intersection de deux compacts convexes. De plus, on a

$$C_\lambda \;=\; \{x \in C \,/\, \phi(x) \leq \lambda\} \;=\; \{x \in C \,/\, \phi_{|C}(x) \leq \lambda\}.$$

Donc $\phi_{|\mathcal{C}}$, qui est continue sur le compact C_λ (en tant que fonction convexe, elle est continue sur l'intérieur de son domaine), y admet un minimum en un point $\bar{x}_C$. Par stricte convexité de $\phi_{|C}$, ce minimum est unique : si $y \in C_\lambda$ est tel que $\phi_{|C}(y) = \phi_{|C}(\bar{x}_C)$ alors pour tout $z \in \,]y, \bar{x}_C[\, \subset C_\lambda$ on aurait $\phi_{|C}(z) < \phi_{|C}(\bar{x}_C)$, ce qui est impossible. Le point $\bar{x}_C$ est donc minimum global unique de $\phi_{|C}$ puisque pour tout $x \in C \smallsetminus C_\lambda$ on a $\phi_{|C}(x) > \lambda \geq \phi_{|C}(\bar{x}_C)$. ♣

4.6.3 Deux Lemmes techniques sur la distance de Kullback

On peut trouver les démonstrations des deux Lemmes qui suivent dans Castellan [20] page 115. Tout deux sont utiles soit pour la démonstration du théorème 4.4.1, soit d'autres lemmes qui lui sont préparatoires.

Lemme 4.6.3. *Soient p et q deux densités par rapport à la mesure μ alors on a*

$$K(p,q) \;\geq\; \Phi\left(-\left\|\log\frac{p}{q}\right\|_\infty\right)\int_0^1 p\left(\log\frac{p}{q}\right)^2 d\mu,$$

et pour tout $c\in\mathbb{R}$

$$K(p,q) \;\leq\; \Phi\left(\left\|\log\frac{p}{q}-c\right\|_\infty\right)\int_0^1 p\left(\log\frac{p}{q}-c\right)^2 d\mu$$

où Φ est la fonction définie pour tout $x\in\mathbb{R}$ par

$$\Phi(x) \;=\; \frac{e^x-1-x}{x^2}. \tag{4.23}$$

Lemme 4.6.4. *Soient p, q et s trois densités par rapport à la mesure μ telles qu'il existe $\eta>0$ tel que*

$$\left\|\log\frac{p}{q}\right\|_\infty \leq\; \eta,$$

alors pour tout $\theta>0$ on a

$$K(p,q) \;\geq\; \Phi(-\eta)\left(1-\theta\eta\sqrt{2}\right)\int_0^1 s\left(\log\frac{p}{q}\right)^2 d\mu - \Phi(-\eta)\frac{\eta}{\theta}\sqrt{2}h^2(s,p).$$

où Φ est la fonction définie par (4.23) et h est la distance de Hellinger donnée par la définition 4.2.2.

4.6.4 Caractérisation de l'EMV et de la projection Kullback

Définition 4.6.5. *Soit $\delta\in\mathbb{R}^{|m|}\times\mathbb{R}$, on pose*

$$\begin{array}{rccc} F_\delta & : & \mathbb{R}^{|m|}\times\mathbb{R} & \longrightarrow & \mathbb{R} \\ & & a & \longmapsto & \psi(a)-(a\,|\,\delta) \end{array}$$

où ψ est donnée par la définition 4.1.4.

On définit de plus deux valeurs particulières $\bar\delta$ et $\hat\delta$ de δ :

$$\bar\delta = E_f(B_m) = \left(\left(\int_0^1 \varphi_I^0 f\,d\mu\right)_{I\in m}, \int_0^1 \varphi^1 f\,d\mu\right)$$

et

$$\hat\delta = P_n(B_m) = \left(\left(\frac{1}{n}\sum_{i=1}^n \varphi_I^0(Y_i)\right)_{I\in m}, \frac{1}{n}\sum_{i=1}^n \varphi^1(Y_i)\right),$$

où B_m est la base donnée par la définition 4.1.1.

Le Lemme suivant montre donne les conditions d'existence d'un minimum de la fonction F_δ et localise ce minimum.

Lemme 4.6.6. *La fonction F_δ est convexe sur $\mathbb{R}^{|m|}\times\mathbb{R}$ et strictement convexe sur G_m (respectivement sur $\dot{G}_m$) (cf. définitions 4.1.5). Elle admet un unique minimum a sur G_m (respectivement $\dot{a}$ sur $\dot{G}_m$, cf. définitions 4.1.8) si et seulement si pour tout $c \in G_m$ on a*

$$(c\,|\,\delta) < \|[\,c\,|\,B_m]\|_\infty\,. \tag{4.24}$$

Si tel est le cas il existe alors un unique $\lambda \in \mathbb{R}$ (respectivement $\dot{\lambda} \in \mathbb{R}$ et $(\dot{\lambda}_1, ..., \dot{\lambda}_{|m|-1}) \in \mathbb{R}^{|m|-1}$) tel que

$$\begin{cases} \delta^0_I = \displaystyle\int_0^1 \varphi^0_I t(a)\,d\mu + \lambda\sqrt{\mu(I)} \quad \textit{pour tout } I \in m \\ \delta^1 = \displaystyle\int_0^1 \varphi^1 t(a)\,d\mu, \end{cases}$$

ou, de façon équivalente,

$$\delta = \mathbb{E}_{t(a)}(B_m) + \lambda u, \tag{4.25}$$

(respectivement

$$\delta = \mathbb{E}_{t(\dot{a})}(B_m) + \dot{\lambda}u + \sum_{i=1}^{|m|-1} \dot{\lambda}_i \dot{u}_i), \tag{4.26}$$

où u est le vecteur unitaire donné par la définition 4.1.5 (respectivement $(\dot{u}_i)_{1\le i\le |m|-1}$ sont les vecteurs donnés par la définition 4.1.8).

De plus, ces conditions caractérisent le minimum a (respectivement $\dot{a}$).

Preuve : Nous commençons par la preuve du lemme sur G_m.

Soit $\mathrm{Hess}_a F_\delta$ la matrice hessienne de F_δ en a et soit Y une variable aléatoire de densité $t(a)$. Les coefficients de $\mathrm{Hess}_a F_\delta$ s'écrivent

$$\begin{aligned} \frac{\partial^2 F_\delta}{\partial a^0_I \partial a^0_J}(a) &= \int_0^1 \varphi^0_I \varphi^0_J t(a)\,d\mu \;-\; \int_0^1 \varphi^0_I t(a)\,d\mu \int_0^1 \varphi^0_J t(a)\,d\mu \\ &= \mathrm{Cov}(\varphi^0_I(Y), \varphi^0_J(Y)) \end{aligned}$$

et

$$\begin{aligned} \frac{\partial^2 F_\delta}{\partial a^0_I \partial a^1}(a) &= \int_0^1 \varphi^0_I \varphi^1 t(a)\,d\mu - \int_0^1 \varphi^0_I t(a)\,d\mu \int_0^1 \varphi^1 t(a)\,d\mu \\ &= \mathrm{Cov}(\varphi^0_I(Y), \varphi^1(Y)). \end{aligned}$$

Donc, pour $b \in \mathbb{R}^{|m|} \times \mathbb{R}$,

$$ {}^t b \times \mathrm{Hess}_a F_\delta \times b \;=\; \mathrm{Var}\left(\sum_{I \in m} b^0_I \varphi^0_I(Y) + b^1 \varphi^1(Y) \right) \;=\; \mathrm{Var}([\, b \,|\, B_m \,](Y)). $$

Ainsi, $\mathrm{Hess}_a F_\delta$ est positive donc F_δ est convexe.

De plus, la forme quadratique $\mathrm{Hess}_a F_\delta$ s'annule si et seulement si la variable aléatoire $[\, b \,|\, B_m \,](Y)$ est constante presque sûrement. Or, par la propriété (i) du Lemme 4.6.1, $[\, b \,|\, B_m \,]$ est constante si et seulement si $b \in \mathrm{Vect}(u) = G_m^\perp$.

Il en résulte donc que F_δ est strictement convexe sur G_m.

La condition (4.24) s'obtient grâce au théorème 3.1 de [24] (voir [20] pour une preuve détaillée).

Explicitons maintenant la caractérisation du minimum de F_δ sous la contrainte définie par G_m.

Supposons la condition (4.24) remplie, F_δ admet alors un minimum sur G_m. Ce minimum est unique par stricte convexité. Soit LF_δ le lagrangien du problème d'optimisation :

$$ \begin{array}{rcccl} LF_\delta & : & \left(\mathbb{R}^{|m|} \times \mathbb{R}\right) \times \mathbb{R} & \longrightarrow & \mathbb{R} \\ & & (a, \lambda) & \longmapsto & \psi(a) - (a \,|\, \delta) + \lambda\, (a \,|\, u)\,. \end{array} $$

La condition du premier ordre s'écrit $\nabla LF_\delta = 0$ soit

$$ \left\{ \begin{array}{rcl} \dfrac{\partial LF_\delta}{\partial a^0_I}(a, \lambda) & = & \displaystyle\int_0^1 \varphi^0_I t(a)\, d\mu - \delta^0_I + \lambda \sqrt{\mu(I)} \;=\; 0 \\[2ex] \dfrac{\partial LF_\delta}{\partial a_1}(a, \lambda) & = & \displaystyle\int_0^1 \varphi^1 t(a)\, d\mu \;- \delta^1 = \; 0 \\[2ex] \dfrac{\partial LF_\delta}{\partial \lambda}(a, \lambda) & = & \displaystyle\sum_{I \in m} \sqrt{\mu(I)}\, a^0_I \;=\; 0. \end{array} \right. $$

ce qui est exactement le système recherché pour G_m.

Pour adapter la preuve au cas contraint $\dot{G}_m$, on commence par appliquer l'alinéa (iv) du Lemme 4.6.2 à la fonction F_δ et à l'ensemble convexe fermé $\dot{G}_m$ qui montre l'existence et l'unicité du minimum $\dot{a}$.

On applique ensuite le Théorème de Kuhn et Tucker à l'optimisation de F_δ sous les contraintes d'inégalité données par la définition 4.1.8 :

$$ \forall\; i = 1, \ldots, |m| - 1,\; (a \,|\, \dot{u}_i) \leq 0 \quad \text{et} \quad (a \,|\, u) = 0. $$

A l'optimum $\dot{a}$, il existe une unique suite de multiplicateurs $(\dot{\lambda}, (\dot{\lambda}_i)_{1\leq i\leq |m|-1})$ dans $\mathbb{R}\times\mathbb{R}^{|m|-1}$ (nuls si et seulement si les contraintes sont insaturées) tels qu'en le point $(\dot{a}, (\dot{\lambda}, (\dot{\lambda}_i)_{1\leq i\leq |m|-1}))$, le gradient du lagrangien

$$\begin{array}{rccl} L_{KT}F_\delta \ : & \left(\mathbb{R}^{|m|}\times\mathbb{R}\right)\times\left(\mathbb{R}\times\mathbb{R}^{|m|-1}\right) & \longrightarrow & \mathbb{R} \\ & (a,(\lambda,(\dot{\lambda}_i)_{1\leq i\leq |m|-1})) & \longmapsto & \psi(a)-(a\,|\,\delta) \\ & & & +\lambda\,(a\,|\,u)+\displaystyle\sum_{i=1}^{|m|-1}\lambda_i\,(a\,|\,\dot{u}_i) \end{array}$$

s'annule.

De façon analogue au cas G_m ci-dessus, l'annulation du gradient de $L_{KT}F_\delta$ à l'optimum conduit à l'équation voulue :

$$\delta = \mathbb{E}_{t(\dot{a})}(B_m) + \dot{\lambda}u + \sum_{i=1}^{|m|-1}\dot{\lambda}_i\dot{u}_i.$$

♣

La Proposition suivante est adaptée des Propositions 2.2.4 et 2.5.5. de Castellan [20] qui caractérisent la projection Kullback et l'estimateur du maximum de vraisemblance d'une densité f dans un modèle exponentiel de polynômes par morceaux général.

Proposition 4.6.7. *Soit f une densité sur $([0\,;1],\mu)$ telle que $\int_0^1 f\log f\,d\mu$ est finie. Il existe une unique densité $\bar{f}$ (respectivement $\overline{\dot{f}}$) de $\mathcal{E}_m$ (respectivement $\dot{\mathcal{E}}_m$) qui minimise la distance de Kullback $K(f,g)$ lorsque g varie dans $\mathcal{E}_m$ (respectivement $\dot{\mathcal{E}}_m$). La densité $\bar{f}$ (respectivement $\overline{\dot{f}}$) est appelée projection Kullback de f sur $\mathcal{E}_m$ (respectivement $\dot{\mathcal{E}}_m$).*

De plus $\bar{f}$ est caractérisée par

$$\int_0^1 s\bar{f}\,d\mu \;=\; \int_0^1 sf\,d\mu \;\;(=\;\mathbb{E}_f(s)) \quad \forall s\in S_m.$$

On notera alors $\bar{a}$ (respectivement $\overline{\dot{a}}$) l'unique vecteur de G_m (respectivement $\dot{G}_m$) tel que $\bar{f}=t(\bar{a})$ (respectivement $\overline{\dot{f}}=t(\overline{\dot{a}})$).

Par ailleurs, la même Proposition est valable si l'on remplace la distance de Kullback $K(f,g)$ par le critère empirique $k_n(f,g)$ (voir la définition 4.2.3). La densité $\hat{f}$ (respectivement $\widehat{\dot{f}}$) , unique solution du problème de minimisation est l'estimateur du maximum de vraisemblance de f sur $\mathcal{E}_m$ (respectivement $\dot{\mathcal{E}}_m$).

Il est, de façon analogue, caractérisé par

$$\int_0^1 s\hat{f}\, d\mu \;=\; \frac{1}{n}\sum_{i=1}^{n} s(Y_i) \;\; (=\; P_n(s)) \quad \forall s \in S_m.$$

On notera alors $\hat{a}$ (respectivement $\widehat{\hat{a}}$) l'unique vecteur de G_m (respectivement $\dot{G}_m$) tel que $\hat{f} = t(\hat{a})$ (respectivement $\widehat{\hat{f}} = t(\widehat{\hat{a}})$).

Preuve : Soit κ la fonction définie par

$$\begin{array}{rccc} \kappa & : & G_m & \longrightarrow & \mathbb{R} \\ & & a & \longmapsto & K(f, t(a)). \end{array}$$

On veut montrer qu'il existe un unique $\overline{\bar{a}} \in \dot{G}_m$ qui minimise $\kappa_{|_{\dot{G}_m}}$ la restriction de κ à l'ensemble convexe fermé $\dot{G}_m$.

Par la définition 4.6.5, on voit que $K(f, t(a)) = \int_0^1 f \log f\, d\mu + F_{\bar{\delta}}(a)$. Minimiser $\kappa_{|_{G_m}}$ (respectivement $\kappa_{|_{\dot{G}_m}}$) équivaut donc à minimiser $F_{\delta|_{G_m}}$ (respectivement $F_{\bar{\delta}|_{\dot{G}_m}}$). Or d'après la preuve du Lemme 4.6.6, $F_{\bar{\delta}|_{G_m}}$ est strictement convexe et atteint son minimum (unique) en un point $\bar{a}$.

Comme $\dot{G}_m$ est un sous-ensemble convexe fermé de G_m, par l'alinéa (iv) de la Proposition 4.6.2 la fonction $F_{\bar{\delta}|_{\dot{G}_m}}$ atteint son unique minimum en un point $\overline{\bar{a}}$. Il en va de même de $\kappa_{|_{\dot{G}_m}}$.

La seconde partie du Corollaire se démontre de façon strictement analogue en remplaçant la distance de Kullback K par le contraste empirique k_n et donc $\bar{\delta}$ par $\hat{\delta}$ dans le corps de la démonstration. ♣

4.6.5 Contrôle de l'écart entre EMV projection Kullback

Le Lemme suivant exprime et quantifie une idée de continuité des solutions du problème de minimisation de la fonction F_δ par rapport au paramètre δ. Il est adapté du Lemme 2.5.8 de [20].

Lemme 4.6.8. *Soit*

$$B \;=\; \sup_{\|c\|_2=1} \|[\,c\,|\,B_m]\|_\infty\,.$$

Soient de plus $\bar{\delta} \in \mathbb{R}^{|m|} \times \mathbb{R}$ et $\bar{a} \in G_m$ associé de sorte que $\bar{a}$ est minimal pour $F_{\bar{\delta}|_{G_m}}$. On suppose de plus que $\inf t(\bar{a}) = b > 0$.

Si $\delta \in \mathbb{R}^{|m|} \times \mathbb{R}$ est tel que

$$\|\bar{\delta} - \delta\|_2 \;<\; \frac{b}{2B}, \tag{4.27}$$

alors il existe $a \in G_m$ *minimal pour* $F_{\delta|_{G_m}}$ *et l'on a*

$$\|\bar{a} - a\|_2 \;\leq\; \frac{1}{2B}\, g^{-1}\left(\frac{2B\|\bar{\delta} - \delta\|_2}{b}\right)$$

et

$$\left\|\log \frac{t(\bar{a})}{t(a)}\right\|_\infty \;\leq\; g^{-1}\left(\frac{2B\|\bar{\delta} - \delta\|_2}{b}\right), \tag{4.28}$$

où g^{-1} *désigne l'inverse de la fonction*

$$\begin{array}{rccc} g & : & \mathbb{R}_*^+ & \longrightarrow & \mathbb{R} \\ & & x & \longmapsto & \dfrac{e^{-x} - 1 + x}{x}. \end{array}$$

Enfin, toutes les propriétés ci-dessus restent vraies si l'on remplace l'espace G_m *par son sous-ensemble* $\dot{G}_m$.

Remarque : si $\varepsilon \in]0\,;1[$ et $\|\bar{\delta} - \delta\|_2 \leq \dfrac{g(\varepsilon)b}{2B}$ alors on a

$$\left\|\log \frac{t(\bar{a})}{t(a)}\right\|_\infty \leq g^{-1}\left(\frac{2B\|\bar{\delta} - \delta\|_2}{b}\right) \leq \varepsilon < 1.$$

Preuve : Une fois encore, nous commençons par le cas non contraint.

Soit $\delta \neq \bar{\delta}$ et vérifiant (4.27). Pour minimiser $F_{\delta|_{G_m}}$, on étudie la quantité $F_{\delta|_{G_m}}(a) - F_{\delta|_{G_m}}(\bar{a})$ dont on montre qu'elle est strictement positive à l'exterieur d'une boule euclidienne fermée de G_m, de centre $\bar{a}$ et de rayon à préciser par la suite. Ceci impliquera alors que $F_{\delta|_{G_m}}$ atteint son minimum a sur cette boule. On a

$$F_{\delta|_{G_m}}(a) - F_{\delta|_{G_m}}(\bar{a}) \;=\; \psi(a) - \psi(\bar{a}) - (a - \bar{a}\,|\,\delta)\,. \tag{4.29}$$

Par ailleurs,

$$\begin{aligned} K(t(\bar{a}), t(a)) &= \int_0^1 t(\bar{a}) \log \frac{t(\bar{a})}{t(a)}\, d\mu \\ &= \int_0^1 [\,\bar{a} - a\,|\,B_m]\, t(\bar{a})\, d\mu - \psi(\bar{a}) + \psi(a), \end{aligned}$$

donc

$$K(t(\bar{a}), t(a)) \;=\; \big(\bar{a} - a\,|\,\mathbb{E}_{t(\bar{a})}(B_m)\big) - \psi(\bar{a}) + \psi(a). \tag{4.30}$$

Or, par l'équation (4.25) du Lemme 4.6.6, il existe $\lambda \in \mathbb{R}$ tel que $\mathbb{E}_{t(\bar{a})}(B_m) = \bar{\delta} - \lambda u$. Ainsi,

$$K(t(\bar{a}), t(a)) \;=\; \left(\bar{a} - a \,|\, \bar{\delta}\right) - \lambda \left(\bar{a} - a \,|\, u\right) - \psi(\bar{a}) + \psi(a).$$

Puis, comme a et $\bar{a}$ sont dans $G = u^{\perp}$,

$$K(t(\bar{a}), t(a)) \;=\; \left(\bar{a} - a \,|\, \bar{\delta}\right) - \psi(\bar{a}) + \psi(a). \tag{4.31}$$

Ainsi, en combinant (4.29) et (4.31), on obtient

$$F_{\delta|_G}(a) - F_{\delta|_G}(\bar{a}) \;=\; K(t(\bar{a}), t(a)) - \left(a - \bar{a} \,|\, \delta - \bar{\delta}\right). \tag{4.32}$$

Soit Φ la fonction définie par 4.23,par le Lemme 4.6.3 on a

$$F_{\delta|_G}(a) - F_{\delta|_G}(\bar{a}) \geq \Phi\left(-\left\|\log\frac{t(\bar{a})}{t(a)}\right\|_{\infty}\right) b \int_0^1 \log^2 \frac{t(\bar{a})}{t(a)}\, d\mu - \left(a - \bar{a} \,|\, \delta - \bar{\delta}\right).$$

Or,

$$\left\|\log\frac{t(\bar{a})}{t(a)}\right\|_{\infty} \;\leq\; \|\,[\,\bar{a} - a \,|\, B_m]\,\|_{\infty} + |\psi(\bar{a}) - \psi(a)|$$

et

$$\begin{aligned}
|\psi(\bar{a}) - \psi(a)| &= \left|\log \int_0^1 \exp([\,\bar{a} \,|\, B_m])\, d\mu + \log(\exp(-\psi(a)))\right| \\
&= \left|\log \int_0^1 \exp([\,\bar{a} \,|\, B_m] - \psi(a))\, d\mu\right| \\
&= \left|\log \int_0^1 \exp([\,\bar{a} \,|\, B_m] - [\,a \,|\, B_m])t(a)\, d\mu\right| \\
&= \left|\log \int_0^1 \exp([\,\bar{a} - a \,|\, B_m])t(a)\, d\mu\right| \\
&\leq \|[\,\bar{a} - a \,|\, B_m]\|_{\infty}.
\end{aligned}$$

Donc finalement,

$$\left\|\log\frac{t(\bar{a})}{t(a)}\right\|_{\infty} \;\leq\; 2\|\,[\,\bar{a} - a \,|\, B_m]\,\|_{\infty}.$$

Puis, par définition de B on a

$$\left\|\log\frac{t(\bar{a})}{t(a)}\right\|_{\infty} \;\leq\; 2B\|\bar{a} - a\|_2. \tag{4.33}$$

Par ailleurs,

$$\begin{aligned}\int_0^1 \log^2 \frac{t(\bar a)}{t(a)}\, d\mu &= \int_0^1 \left([\,\bar a - a \,|\, B_m] + \psi(a) - \psi(\bar a)\right)^2 d\mu \\ &= \|\bar a - a\|_2^2 + (\psi(\bar a) - \psi(a))^2 - 2(\psi(\bar a) - \psi(a)) \times \\ &\qquad \left(\sum_{I\in m} (\bar a_I^0 - a_I^0) \int_0^1 \varphi_I^0\, d\mu + (\bar a^1 - a^1) \int_0^1 \varphi^1\, d\mu\right).\end{aligned}$$

Or, $\displaystyle\int_0^1 \varphi^1\, d\mu = 0$ et pour tout I de m on a $\displaystyle\int_0^1 \varphi_I^0\, d\mu = \sqrt{\mu(I)}$ donc

$$\begin{aligned}\int_0^1 \log^2 \frac{t(\bar a)}{t(a)}\, d\mu &= \|\bar a - a\|_2^2 + (\psi(\bar a) - \psi(a))^2 - 2(\psi(\bar a) - \psi(a))\,(\bar a - a \,|\, u) \\ &= \|\bar a - a\|_2^2 + (\psi(\bar a) - \psi(a))^2 \\ &\geq \|\bar a - a\|_2^2.\end{aligned}$$

Appliquant l'inégalité de Cauchy-Schwarz on obtient alors

$$F_{\delta|_G}(a) - F_{\delta|_G}(\bar a) \;\geq\; b\Phi(-2B\|\bar a - a\|_2)\|\bar a - a\|_2^2 - \|\bar\delta - \delta\|_2 \|\bar a - a\|_2.$$

Posons $c = \|\bar a - a\|_2 / \|\bar\delta - \delta\|_2$, on a

$$F_{\delta|_G}(a) - F_{\delta|_G}(\bar a) \;\geq\; c\|\bar\delta - \delta\|_2^2 \left(bc\Phi(-2Bc\|\bar\delta - \delta\|_2) - 1\right).$$

Posons maintenant $\tau = 2B\|\bar\delta - \delta\|_2 / b$, de sorte que $\tau < 1$, on a

$$F_{\delta|_G}(a) - F_{\delta|_G}(\bar a) \;\geq\; \frac{c}{\tau}\|\bar\delta - \delta\|_2^2 (g(\tau bc) - \tau),$$

où la fonction g définie par $g(x) = x\Phi(-x)$ est strictement croissante sur $\mathbb{R}_*^+$. Elle définit donc une bijection de $\mathbb{R}_*^+$ sur $]\,0\,;1[$. Ainsi, $g(\tau bc) - \tau > 0$ si et seulement si $\tau bc > g^{-1}(\tau)$.

Ainsi, pour tout $c > g^{-1}(\tau)/b\tau$ et tout $a \in G_m$ tel que $\|\bar a - a\|_2 = c\|\bar\delta - \delta\|_2$, on a $F_{\delta|_G}(a) - F_{\delta|_G}(\bar a) > 0$.

Il en résulte que $F_{\delta|_G}$ possède un minimum unique a qui vérifie

$$\|\bar a - a\|_2 \;\leq\; \frac{g^{-1}(\tau)}{b\tau}\|\bar\delta - \delta\|_2 \;=\; \frac{1}{2B}\, g^{-1}\left(\frac{2B\|\bar\delta - \delta\|_2}{b}\right),$$

ce qui, en injectant dans (4.33) donne aussi

$$\left\|\log \frac{t(\bar a)}{t(a)}\right\|_\infty \;\leq\; g^{-1}\left(\frac{2B\|\bar\delta - \delta\|_2}{b}\right).$$

Pour adapter la preuve à $\dot{G}_m$, on repart de l'équation (4.30) avec les notations *ad hoc*. Par la caractérisation (4.26) du Lemme 4.6.6, il existe $(\dot{\lambda}, (\dot{\lambda}_i)_{1\leq i\leq |m|-1})$ dans $\mathbb{R}\times\mathbb{R}^{|m|-1}$ tels que

$$\mathbb{E}_{t(\overline{\dot{a}})}(B_m) = \overline{\dot{\delta}} - \dot{\lambda}u - \sum_{i=1}^{|m|-1} \dot{\lambda}_i \dot{u}_i.$$

Ainsi,

$$K(t(\overline{\dot{a}}), t(\dot{a})) \;=\; \left(\overline{\dot{a}} - \dot{a} \,|\, \overline{\dot{\delta}}\right) - \dot{\lambda}\left(\overline{\dot{a}} - \dot{a} \,|\, u\right) - \sum_{i=1}^{|m|-1} \dot{\lambda}_i \left(\overline{\dot{a}} - \dot{a} \,|\, \dot{u}_i\right) - \psi(\overline{\dot{a}}) + \psi(\dot{a}).$$

En premier lieu, $\dot{a}$ et $\overline{\dot{a}}$ sont dans $G_m = u^\perp$ donc $\left(\overline{\dot{a}} - \dot{a} \,|\, u\right) = 0$.

Puis, pour $i = 1, \ldots |m| - 1$:
- soit la contrainte associée à $\dot{u}_i$ est insaturée alors $\left(\overline{\dot{a}} - \dot{a} \,|\, \dot{u}_i\right) \neq 0$ mais $\dot{\lambda}_i = 0$;
- soit elle est saturée, alors $\dot{\lambda}_i \neq 0$ en général mais, par définition de $\dot{G}_m$, $\dot{a}$ et $\overline{\dot{a}}$ sont alors dans $\dot{u}_i^\perp$ et $\left(\overline{\dot{a}} - \dot{a} \,|\, \dot{u}_i\right) = 0$.

On en déduit alors, comme dans l'équation (4.32) du cas G_m, que

$$K(t(\overline{\dot{a}}), t(\dot{a})) \;=\; \left(\overline{\dot{a}} - \dot{a} \,|\, \overline{\dot{\delta}}\right) - \psi(\overline{\dot{a}}) + \psi(\dot{a}). \tag{4.34}$$

Le reste de la preuve est ensuite analogue à celui du cas non contraint. ♣

Le Lemme 4.6.11 qui suit utilise le Lemme 4.6.8 pour permettre de contrôler l'écart entre les densités $\hat{f}_m$ et $\bar{f}_m$ (respectivement $\widehat{\dot{f}}$ et $\overline{\dot{f}}$ dans le modèle $\mathcal{E}_m$ (respectivement $\dot{\mathcal{E}}_m$). Ce contrôle ne peut se faire que sur un évènement restreint, de grande probabilité.

La définition de cet évènement vise à contrôler les fluctuations de la mesure empirique des éléments d'une base de S_m autour de leur moyenne pour en déduire ensuite un contrôle des fluctuations de $\hat{f}_m$ autour de $\bar{f}_m$.

Définition 4.6.9. *Soit ν_n la mesure empirique recentrée donnée par la définition 3.2.1. Soit $\rho = \operatorname{ess\,inf} f$ et soit $\Upsilon_n = \rho \wedge \rho_n$ où $(\rho_n)_{n\in\mathbb{N}}$ est une suite de réels positifs qui tend vers 0.*

On définit l'ensemble $\Omega_n[A]$ avec les deux variantes suivantes :

$$\Omega_n^1[A] = \bigcap_{m\in\mathcal{M}_n} \left\{\forall I \in m,\, |\nu_n(\varphi_I^0)| \leq A\Upsilon_n\Gamma_n \text{ et } |\nu_n(\varphi^1)| \leq A\Upsilon_n\Gamma_n\right\}$$

dans le cas de partitions régulières (H1) (cf. 4.4.1) et

$$\Omega_n^2[A] = \left\{\forall I \in m_n,\, |\nu_n(\varphi_I^0)| \leq A\Upsilon_n\Gamma_n \text{ et } |\nu_n(\varphi^1)| \leq A\Upsilon_n\Gamma_n\right\}$$

dans le cas de partitions irrégulières (H2) (cf. 4.4.1).

Le Lemme qui suit montre, sous certaines hypothèses, que la probabilité de l'évènement ${}^c\Omega_n[A]$ est à décroissance rapide en fonction du nombre n d'observations, *i.e.* négligeable devant toute puissance de n.

Lemme 4.6.10. *On suppose que* $\Gamma_n = \mathcal{O}\left(\dfrac{\log n}{\Upsilon_n\sqrt{n}}\right)$.

Alors pour tout $\alpha > 0$ *on a*

$$\mathbb{P}\left({}^c\Omega_n^1[A]\right) \;=\; \mathcal{O}\left(n^{-\alpha}\right) \quad \textit{et} \quad \mathbb{P}\left({}^c\Omega_n^2[A]\right) \;=\; \mathcal{O}\left(n^{-\alpha}\right).$$

Remarque : les conditions imposées sur l'espace $\Omega_n[A]$ sont ici assez restrictives par rapport au contrôle dont nous avons réellement besoin dans la suite. Dans le cas des partitions irrégulières, travailler sur un ensemble plus important

$$\tilde{\Omega}_n^2[A] = \left\{\forall I \in m_n,\, |\nu_n(\varphi_I^0)| \leq A\Upsilon_n\sqrt{\mu(I)\Gamma_n} \text{ et } |\nu_n(\varphi^1)| \leq A\Upsilon_n\sqrt{\Gamma_n}\right\}$$

serait plus pertinent et donnerait strictement les mêmes résultats. Il en va de même pour $\Omega_n^1[A]$. Cependant, les espaces définis par 4.6.9 sont plus simples à manipuler et sont suffisants pour énoncer nos résultats sans aucune perte de généralité.

Preuve : Dans le cas des partitions régulières on a

$$\mathbb{P}\left({}^c\Omega_n^1[A]\right) \;\leq\; \sum_{m\in\mathcal{M}_n}\sum_{I\in m}\mathbb{P}\left(|\nu_n(\varphi_I^0)| > A\Upsilon_n\Gamma_n\right) + \mathbb{P}\left(|\nu_n(\varphi^1)| > A\Upsilon_n\Gamma_n\right).$$

Appliquant l'inégalité de Bernstein (*cf.* [70]) aux processus $\nu_n(\varphi_I^0)$ et $\nu_n(\varphi^1)$, et en notant que

$$\begin{cases} \mathrm{Var}(\varphi_I^0(Y)) \;\leq\; \|f\|_\infty \quad \text{et} \quad \|\varphi_I^0\|_\infty = \dfrac{1}{\sqrt{\mu(I)}} \leq \dfrac{1}{\sqrt{\Gamma_n}} \quad \forall I \in m \\ \mathrm{Var}(\varphi^1(Y)) \;\leq\; \|f\|_\infty \quad \text{et} \quad \|\varphi^1\|_\infty = \dfrac{\sqrt{3}}{\sqrt{\min\limits_{I\in m}\mu(I)}} \leq \sqrt{\dfrac{3}{\Gamma_n}}, \end{cases}$$

on obtient

$$\begin{cases} \mathbb{P}\left(|\nu_n(\varphi_I^0)| > A\Upsilon_n\Gamma_n\right) \;\leq\; 2\exp\left(-n\,\dfrac{A^2\Upsilon_n^2\Gamma_n^2}{2\left(\|f\|_\infty + A\Upsilon_n\sqrt{\Gamma_n}/3\right)}\right) \\ \mathbb{P}\left(|\nu_n(\varphi^1)| > A\Upsilon_n\Gamma_n\right) \;\leq\; 2\exp\left(-n\,\dfrac{A^2\Upsilon_n^2\Gamma_n^2}{2\left(\|f\|_\infty + A\Upsilon_n\sqrt{\Gamma_n/3}\right)}\right). \end{cases}$$

Ainsi, il existe une constante C telle que

$$\begin{cases} \mathbb{P}\left(|\nu_n(\varphi_I^0)| > A\Upsilon_n\Gamma_n\right) \;\leq\; 2\exp\left(-Cn\Upsilon_n^2\Gamma_n^2\right) \\ \mathbb{P}\left(|\nu_n(\varphi^1)| > A\Upsilon_n\Gamma_n\right) \;\leq\; 2\exp\left(-Cn\Upsilon_n^2\Gamma_n^2\right). \end{cases}$$

Finalement,

$$\mathbb{P}\left({}^c\Omega_n^1[A]\right) \;\leq\; \frac{4}{\Gamma_n^2}\exp\left(-Cn\Upsilon_n^2\Gamma_n^2\right).$$

Dans le cas des partitions irrégulières,

$$\mathbb{P}\left({}^c\Omega_n^2[A]\right) \;\leq\; \sum_{I\in m}\mathbb{P}\left(|\nu_n(\varphi_I^0)|>A\Upsilon_n\Gamma_n\right)+\mathbb{P}\left(|\nu_n(\varphi^1)|>A\Upsilon_n\Gamma_n\right),$$

puis, de nouveau par application de l'inégalité de Bernstein il existe une constante C telle que,

$$\mathbb{P}\left({}^c\Omega_n^2[A]\right) \;\leq\; \frac{4}{\Gamma_n}\exp\left(-Cn\Upsilon_n^2\Gamma_n^2\right).$$

Ainsi, dans les deux cas, si $\Gamma_n = \mathcal{O}\left(\dfrac{\log n}{\Upsilon_n\sqrt{n}}\right)$, alors pour tout $\alpha>0$ on obtient bien

$$\mathbb{P}\left({}^c\Omega_n^1[A]\right) \;=\; \mathcal{O}\left(n^{-\alpha}\right) \quad \text{et} \quad \mathbb{P}\left({}^c\Omega_n^2[A]\right) \;=\; \mathcal{O}\left(n^{-\alpha}\right).$$

♣

Nous sommes maintenant en mesure de contrôler l'écart entre la projection Kullback de f et son estimateur du maximum de vraisemblance sur l'évènement $\Omega_n[A]$.

Lemme 4.6.11. *Soit $\{Y_1, ..., Y_n\}$ une famille de variables aléatoires indépendantes à valeurs dans $[\,0\,;1\,]$ de loi commune $dP = f\,d\mu$. Soit m un modèle de $\mathcal{M}_n$ et $\hat{f}_m$ (respectivement $\bar{f}_m$) l'estimateur du maximum de vraisemblance (respectivement la projection de Kullback) de f sur $\mathcal{E}_m$. On suppose que* $\operatorname{essinf}\hat{f}_m \geq b_n$.

Soient $\varepsilon\in\,]\,0\,;1[$ et $A\leq\dfrac{g(\varepsilon)}{4\sqrt{2}}$. Alors, sur l'ensemble $\Omega_n[A]$ on a

$$\left\|\log\frac{\bar{f}_m}{\hat{f}_m}\right\|_\infty \;\leq\; \varepsilon$$

De plus, la Proposition ci-dessus reste valable si l'on remplace $\hat{f}_m$ (respectivement $\bar{f}_m$) par $\hat{\dot{f}}_m$ (respectivement $\overline{\dot{f}}_m$) et $\mathcal{E}_m$ par $\dot{\mathcal{E}}_m$.

Preuve : Soit m une partition de $[\,0\,;1]$. On cherche à appliquer la remarque du Lemme 4.6.8 aux vecteurs $\bar{\delta}$ et $\hat{\delta}$ de la définition 4.6.5. Rappelons que les vecteurs $\bar{a}$ et $\hat{a}$ de G_m associés qui minimisent respectivement $F_{\bar{\delta}}$ et $F_{\hat{\delta}}$ sur G_m sont tels que $t(\bar{a})=\bar{f}_m$ et $t(\hat{a})=\hat{f}_m$.

Évaluons en premier lieu la quantité

$$B \;=\; \sup_{\|c\|_2=1} \|[\,c\,|\,B_m]\|_\infty$$

Utilisant la définition 4.1.1 on a

$$\begin{aligned}
\sup_{\|c\|_2=1} \|[\,c\,|\,B_m]\|_\infty &\leq \max_{I\in m} \sup_{c^{0\,2}+c^{1\,2}=1} \left\| c^0\varphi_I^0 + c^1\varphi^1_{|_I} \right\|_\infty \\
&\leq \max_{I\in m} \sup_{c^{0\,2}+c^{1\,2}=1} |c^0|\|\varphi_I^0\|_\infty + |c^1| \left\| \varphi^1_{|_I} \right\|_\infty \\
&\leq \max_{I\in m} \sup_{c^{0\,2}+c^{1\,2}=1} \frac{|c^0|}{\sqrt{\mu(I)}} + \frac{|c^1|\sqrt{3}\mu(I)}{\sqrt{\sum_{J\in m}\mu(J)^3}} \\
&\leq \max_{I\in m} \sup_{c^{0\,2}+c^{1\,2}=1} \frac{|c^0|}{\sqrt{\mu(I)}} + \frac{|c^1|\sqrt{3}}{\sqrt{\mu(I)}} \\
&\leq \max_{I\in m} \frac{2}{\sqrt{\mu(I)}} \\
&\leq \frac{2}{\sqrt{\Gamma_n}}.
\end{aligned}$$

Évaluons maintenant $\|\bar\delta - \hat\delta\|_2 \;=\; \sqrt{\sum_{I\in m} \nu_n^2(\varphi_I^0) + \nu_n^2(\varphi^1)} = \|\nu_n(B_m)\|_2$ sur l'ensemble $\Omega_n[A]$. On a

$$\begin{cases} |\nu_n(\varphi_I^0)| \;\leq\; A\Upsilon_n\Gamma_n \quad \forall I \in m \\ |\nu_n(\varphi^1)| \;\leq\; A\Upsilon_n\Gamma_n. \end{cases}$$

Il en résulte que

$$\|\bar\delta - \hat\delta\|_2 \leq\; A\Upsilon_n\sqrt{|m|\Gamma_n^2 + \Gamma_n^2} \;\leq\; A\Upsilon_n\sqrt{2\Gamma_n} \tag{4.35}$$

car $\Gamma_n \leq 1$ et $|m|\Gamma_n \leq 1$.

Par suite,

$$\|\bar\delta - \hat\delta\|_2 \;\leq\; \frac{g(\varepsilon)b_n}{2B}.$$

Appliquons maintenant la remarque du Lemme 4.6.8 pour obtenir

$$\left\| \log\frac{\bar f}{\hat f} \right\|_\infty \;=\; \left\| \log\frac{t(\bar a)}{t(\hat a)} \right\|_\infty \;\leq\; \varepsilon.$$

Enfin, l'ensemble des arguments ci-dessus valent aussi pour $\overline{\dot f}$ et $\widehat{\dot f}$ en adaptant simplement les notations. ♣

4.6.6 Inégalité de concentration

Le Lemme suivant permet le contrôle de la quantité $\nu_n(\log \hat{f}_m - \log \bar{f}_m)$ via l'inégalité (4.12). Il est basé sur une inégalité de concentration de type Talagrand [78], améliorée successivement en termes de constantes par Massart [57] puis Rio [71] et enfin très récemment par Bousquet [16].

Lemme 4.6.12. *Soit $\{Y_1, ..., Y_n\}$ un échantillon de variables aléatoires indépendantes et de loi commune $dP = f\,d\mu$, distribuées sur $[\,0\,;1\,]$ et telles que $\rho = \operatorname{ess\,inf} f > 0$. Soit $m \in \mathcal{M}_n$ et Z_m la variable aléatoire définie par*

$$Z_m = \sup\left\{ |\nu_n(g)|,\ g \in S_m \ et \int_0^1 g^2 f\,d\mu = 1 \right\}.$$

Soit $\varepsilon \in]\,0\,;1[$ et $A \leq \dfrac{4\varepsilon^2}{5\varepsilon + 32}$. Alors pour tout $x > 0$ on a

$$\mathbb{P}\left[Z_m \mathbb{1}_{\Omega_n[A]} \geq (1+\varepsilon)\left(\sqrt{\frac{|m|}{n}} + \sqrt{\frac{2x}{n}} \right) \right] \ \leq\ \exp(-x)$$

où l'ensemble $\Omega_n[A]$ est donné par la définition 4.6.9.

De plus, la même conclusion vaut pour

$$\dot{Z}_m = \sup\left\{ |\nu_n(g)|,\ g \in \dot{S}_m \ et \int_0^1 g^2 f\,d\mu = 1 \right\}.$$

Preuve : Soit $B_m^P = \left\{ \left(\phi_I^0\right)_{I \in m},\ \phi^1 \right\}$ une base orthonormée de $\mathbb{L}^2([\,0\,;1\,], P)$ telle que pour tout $I \in m$, $\phi_I^0 = \mathbb{1}_I / P(I)$. Il existe alors $c \in \mathbb{R}^{|m|} \times \mathbb{R}$, qui dépend de f, tel que $\phi^1 = [\,c\,|\,B_m]$.

On peut alors écrire en utilisant cette base, et conformément à la définition 4.1.3

$$Z_m = \sup_{\|a\|_2\,=\,1} \left| \nu_n\left([\,a\,|\,B_m^P] \right) \right| = \sup_{\|a\|_2 = 1} \left| \left(a\,|\,\nu_n\left(B_m^P\right) \right) \right| = \|\nu_n\left(B_m^P\right)\|_2. \tag{4.36}$$

Or, pour $I \in m$,

$$|\nu_n(\phi_I^0)| \ =\ \left| \nu_n\left(\sqrt{\frac{\mu(I)}{P(I)}}\,\varphi_I^0 \right) \right| \ \leq\ \frac{|\nu_n(\varphi_I^0)|}{\sqrt{\rho}}$$

et par l'inégalité de Cauchy-Schwarz,

$$|\nu_n(\phi^1)| \ =\ |\nu_n([\,c\,|\,B_m])| \ \leq\ \|c\|_2 \|\nu_n(B_m)\|_2.$$

Par ailleurs, par la remarque de la définition 4.1.3 on a

$$\|c\|_2^2 \;=\; \int_0^1 \left[\, c \,|\, B_m \right]^2 \, d\mu \;\leq\; \frac{1}{\rho} \int_0^1 \left(\phi^1\right)^2 f \, d\mu \;=\; \frac{1}{\rho}.$$

Ainsi, sur l'ensemble $\Omega_n[A]$, il résulte de (4.35) que

$$|\nu_n(\phi^1)| \;\leq\; \frac{\|\nu_n(B_m)\|_2}{\sqrt{\rho}} \;\leq\; A\sqrt{2\rho\Gamma_n}.$$

Enfin, sur l'ensemble $\Omega_n[A]$ il vient ainsi

$$\begin{cases} |\nu_n(\phi_I^0)| \;\leq\; A\sqrt{2\rho\Gamma_n} & \text{pour tout } I \in m \\ |\nu_n(\phi^1)| \;\leq\; A\sqrt{2\rho\Gamma_n}. \end{cases}$$

Le sup dans (4.36) est atteint pour $a = \nu_n\left(B_m^P\right)/Z_m$. Ainsi, pour $z > 0$, si l'on définit

$$A_m(z) = \left\{ a \in G \,/\, \|a\|_2 = 1,\, |a_I^0| \leq \frac{A}{z}\sqrt{2\rho\Gamma_n} \text{ et } |a^1| \leq \frac{A}{z}\sqrt{2\rho\Gamma_n} \right\}$$

et si l'on considère un sous ensemble A'_m dénombrable et dense dans A_m on obtient alors sur l'ensemble $\Omega_n[A] \cap \{Z_m \geq z\}$ que

$$Z_m = \sup_{a \in A'_m} \left| \nu_n\left(\left[\, a \,|\, B_m^P \right]\right) \right|.$$

Nous appliquons maintenant une inégalité de concentration de type Talagrand [78] à la variable aléatoire Z_m. Plus précisément, nous appliquons une version améliorée par Rio [71] pour obtenir

$$\mathbb{P}\left[\sup_{a \in A'_m} \left| \nu_n\left(\left[\, a \,|\, B_m^P \right]\right) \right| \geq (1+\varepsilon)\epsilon_m + \sqrt{\frac{2\sigma^2 x}{n}} + \frac{\kappa(\varepsilon) b x}{n} \right] \;\leq\; \exp(-x),$$

avec $\kappa(\varepsilon) = 2,5 + 32/\varepsilon$ et

$$\begin{aligned} \epsilon_m \;=\; \mathbb{E}(Z_m) \;&\leq\; \mathbb{E}\left(\left\| \nu_n\left(B_m^P\right) \right\|_2 \right) \\ &\leq\; \sqrt{\mathbb{E}\left(\left\| \nu_n\left(B_m^P\right) \right\|_2^2 \right)} \\ &\leq\; \sqrt{\sum_{I \in m} \frac{1}{n} \mathrm{Var}(\phi_I^0(Y)) + \frac{1}{n}\mathrm{Var}(\phi^1(Y))} \\ &\leq\; \sqrt{\sum_{I \in m} \frac{1}{n}(1 - P(I)) + \frac{1}{n}} \\ &\leq\; \sqrt{\frac{|m|}{n}}, \end{aligned}$$

puis

$$\sigma^2 \;=\; \sup_{a\in A'_m} \mathrm{Var}\left(\left[\,a\,|\,B_m^P\right](Y)\right) \;\le\; \sup_{a\in A'_m} \|a\|_2^2 \;\le\; 1,$$

et enfin

$$\begin{aligned} b \;=\; \sup_{a\in A'_m} \left\|\left[\,a\,|\,B_m^P\right]\right\|_\infty \;&=\; \sup_{a\in A'_m} \max_{I\in m} \left\| a_I^0\phi_I^0 + a^1\phi^1_{|_I} \right\|_\infty \\ &\le\; \sup_{a\in A'_m} \max_{I\in m} \frac{2}{\sqrt{\mu(I)}} \left\| a_I^0\phi_I^0 + a^1\phi^1_{|_I} \right\|_2 \\ &\le\; 2\times \sup_{a\in A'_m} \max_{I\in m} \sqrt{\frac{1}{\mu(I)}\int_0^1 \left(a_I^0\phi_I^0 + a^1\phi^1_{|_I}\right)^2 \frac{f}{\rho}\,d\mu} \\ &\le\; 2\times \sup_{a\in A'_m} \max_{I\in m} \sqrt{\frac{{a_I^0}^2 + {a^1}^2}{\rho\mu(I)}} \\ &\le\; \frac{4A}{z}. \end{aligned}$$

Ainsi,

$$\mathbb{P}\left[Z_m \mathbb{1}_{\Omega_n[A]\cap\{Z_m\ge z\}} \;\ge (1+\varepsilon)\sqrt{\frac{|m|}{n}} + \sqrt{\frac{2x}{n}} + \frac{4A\kappa(\varepsilon)x}{zn} \right] \;\le\; \exp(-x).$$

En choisissant maintenant $z = \sqrt{2x/n}$ et tenant compte du fait que par définition $A \le 2\varepsilon/\kappa(\varepsilon)$ on obtient finalement

$$\mathbb{P}\left[Z_m \mathbb{1}_{\Omega_n[A]} \;\ge (1+\varepsilon)\left(\sqrt{\frac{|m|}{n}} + \sqrt{\frac{2x}{n}}\right) \right] \;\le\; \exp(-x)$$

Du fait de la définition de Z_m en tant que supremum, l'adaptation de la conclusion à $\dot{Z}_m$ est immédiate car dans ce cas, le sup est pris sur un ensemble plus petit que dans le cas de Z_m. ♣

Chapitre 5

Applications

L'objectif final de ce chapitre est de présenter les applications de nos travaux décrits dans l'ensemble des chapitres précédents à certains systèmes pétroliers matures de la planète.

Il comprend trois parties. Nous allons commencer par décrire en détail la programmation du logiciel interfacé sous Matlab®, que nous avons appelé "select", qui nous a notamment permis de pratiquer les simulations permettant d'évaluer l'efficacité de notre méthode. Bien entendu, l'objectif de l'utilisation de ce logiciel est l'application aux cas concrets. Quelques codes source de ces programmes figurent en Annexe 1. Le mode d'emploi, l'interface et un exemple d'application pratique du logiciel sont fournis en annexe 2. Les simulations sont ensuite décrites et la construction d'intervalles de confiance autour des estimations données est envisagée. Les résultats de ces simulations figurent en Annexe 3. La dernière partie, quant à elle, présente les résultats d'application concrètes à certains systèmes pétroliers des bassins suivants :
- le Viking Graben de mer du Nord ;
- le golfe du Méxique méxicain ;
- le delta du Congo ;
- le bassin sibérien de Bazhenov.

5.1 Le logiciel "select"

Les statistiques appliquées sont de plus en plus consommatrices de calculs intensifs. De nombreux champs d'étude ont vu leur intérêt se développer considérablement au cours des vingt dernières années grâce à l'extraordinaire explosion des capacités de calcul des ordinateurs modernes. L'un des exemples les plus évidents en la matière est l'engouement unanime du monde des statisticiens pour les techniques de ré-échantillonnage telles le bootstrap, consommatrice importante de capacités de calcul et de stockage mémoire.

Les statistiques appliquées doivent cependant rester, pour l'analyste, l'utilisateur final, dans le domaine du concrêt et de l'accessible au plus grand

nombre. Ainsi, nous sommes partisans de l'idée que l'important recours à l'outil informatique dans les applications des statistiques modernes ne doit en rien faire croire que leur utilisation est fermée au non-spécialiste des statistiques et/ou de l'informatique scientifique. C'est pourquoi, au cours de notre travail de thèse, nous avons investi un temps important à la création d'un logiciel convivial, très facile d'utilisation tant dans la manipulation de ses résultats que de ses sorties graphiques. Ce logiciel est destiné :

- d'une part au statisticien désireux d'effectuer un travail de simulation des données du modèle afin, par exemple, de tester l'efficacité du protocole d'estimation ;
- et d'autre part à l'utilisateur final souhaitant utiliser le protocole d'estimation sur des données réelles.

5.1.1 Motivation de la création d'un logiciel

Étant donnée la complexité de la mise en œuvre de l'estimation des paramètres de notre modèle, l'utilisation d'outils informatiques de calcul scientifique est indispensable. Par ailleurs, les logiciels actuels offrent des possibilités de création d'Interfaces Homme / Machine (interfaces graphiques) tout-à-fait passionnantes et quasiment illimitées.

Définition 5.1.1. *On appelle Interface Homme / Machine (ou IHM par la suite) une application logicielle permettant à l'utilisateur de mettre en œuvre un ou plusieurs protocoles informatiques, de façon interactive, sans avoir recours à la modification d'un code (ou programme) source sous-jacent.*

Dans le domaine du calcul scientifique, le but de la création d'une IHM est donc de permettre au plus grand nombre d'utilisateurs possible de faire fonctionner une application informatique, au besoin en en modifiant un certain nombre de paramètres, sans avoir jamais à se soucier de la programmation des algorithmes sous-jacents. De plus, l'existence d'une IHM facilite la mise en application et l'automatisation de calculs éventuellement très lourds.

Dans notre domaine particulier, il est très clair que les utilisateurs finaux de notre protocole d'estimation des réserves d'hydrocarbures sont, de façon privilégiée, des économistes ou des géologues et n'ont donc pas de raison d'être des statisticiens ou informaticiens chevronnés. Comme on l'a vu en introduction de cette thèse, l'énoncé du problème auquel nous nous sommes intéressés est particulièrement simple, il doit donc en être de même de la forme des résultats que nous fournissons avec notre approche. Le traitement statistique et informatique pour passer de ce problème à la réponse que nous fournissons est quant-à-lui complexe mais doit être transparent pour l'utilisateur (modulo temps de calculs, s'entend).

Nous avons donc choisi de développer un logiciel totalement interfacé sous Matlab® permettant à l'utilisateur de charger depuis un fichier textuel la

liste des données tailles des champs, choisir sa méthode d'estimation puis de visualiser l'ensemble des résultats de façon claire, précise et concise. Une option permet aussi de sauvegarder ces résultats sous forme textuelle donc universellement exportable vers tout type de logiciel en vue de pratiquer d'autres analyses complémentaires.

Enfin, et il s'agit sûrement de l'argument le plus important, les analyses concernant les évaluations de réserves sont sensées évoluer au cours du temps. Il est alors indispensable que nous puissions laisser à nos partenaires industriels un outil qui puisse être utilisé de manière autonome une fois le travail de thèse achevé.

Si les calculs informatiques doivent être transparents pour l'utilisateur, nous devons évidemment les détailler dans cette thèse, ce qui fait l'objet de la partie qui suit.

5.1.2 Description des codes sources mis en œuvre

Nous détaillons et commentons ici le plan et certains des codes source du logiciel "select" que nous avons crée.

Fonctionnalités

Le logiciel **select** que nous avons crée sous environnement Matlab® 5.3[1] vise à mettre en pratique l'estimation d'une densité par sélection de modèles exponentiels de polynômes par morceaux, conformément aux chapitres 3 et 4.

Le logiciel permet de pratiquer :
- des simulations de données par échantillonnage de la loi parente et sous-échantillonnage (deux options de sous-échantillonnage sont proposés) ;
- des simulations à partir de données réelles par sous-échantillonnage ;
- l'estimation de la densité sur un modèle exponentiel de polynômes par morceaux spécifié par l'utilisateur sur données réelles ou simulées par maximum de vraisemblance selon 4 options différentes ;
- l'estimation de la densité par sélection de modèles exponentiels de polynômes par morceaux sur données réelles ou simulées suivant 16 options différentes ;

La manipulation des données ou du simulateur est totalement interfacée. Les résultats peuvent être visualisés graphiquement ou sauvegardés dans un fichier *ad hoc* pour éventuellement être retravaillées par la suite dans tableur ou un logiciel de statistiques quelconque.

[1] Il est à noter que la compatibilité de Matlab® est ascendante, c'est-à-dire que les versions postérieures à la version 5.3 sous laquelle nous avons développé **select** (et notamment l'actuelle version 6) pourront être utilisées pour le faire fonctionner.

Générateur de données

Rappelons ici que les échantillons parent $\mathbb{X} = \{X_1, ..., X_N\}$ et observé $\mathbb{Y} = \{Y_1, ..., Y_n\}$ sont les échantillons des LogTailles des gisements d'hydrocarbures.

Loi parente

Un premier outil extrêmement simple permet de générer un échantillon $\mathbb{X} = \{X_1, ..., X_N\}$ d'une de loi parente de Lévy-Pareto de paramètres α et N spécifiés par l'utilisateur. Dans le logiciel, la statistique d'ordre de l'échantillon $\mathbb{X}$ est directement fournie à l'utilisateur.

Le code source associé à ce simulateur se trouve en 6.2.

Sous-échantillonnage (1)

Supposons-nous donné l'échantillon $\mathbb{X} = \{X_1, ..., X_N\}$. On se donne une famille de poids $\pi = \{\pi_1, ..., \pi_N\}$ et l'on souhaite tirer successivement proportionnellement à la famille π un échantillon $\mathbb{Y} = \{Y_1, ..., Y_n\}$ selon la procédure de la proposée dans la section 2.3.1. Suivant la forme de la fonction ω^1, l'échantillon tiré (à supposer que $\mathbb{X}$ est exponentiel) peut soit être exponentiel de polynômes par morceaux de degré 1 (auquel cas celle-ci est constante par morceaux), soit "non-exponentiel-polynomial".

Le principe est le suivant :
- on considère le segment $[\,\pi_1\,;\pi_1 + \cdots + \pi_N]$ formé de la concaténation de segments de longueurs $(\pi_i)_{1\leq i\leq N}$ ordonnés ;
- on tire un nombre au hasard selon une loi uniforme sur ce segment ;
- on sélectionne l'individu X_i correspondant à la plage de π_i à laquelle appartient le nombre tiré ;
- on réitère le processus en prenant pour nouvelle liste de poids la précédente de laquelle on a retiré le poids associé à l'individu sélectionné ;
- la procédure précédente est menée n fois.

Le code source associé à cet algorithme se trouve en 6.3.1.

Sous-échantillonnage (2)

Toujours sous l'hypothèse que $\mathbb{X}$ est exponentiel, lorsque ce ne sont pas les probabilités d'inclusion au premier tirage qui sont connues, mais directement celles du $n^{\text{ème}}$ tirage et que celles-ci sont supposées constantes par morceaux alors l'échantillon que l'on tire est de densité exponentielle de polynômes par morceaux. Le mode pratique de tirage est bien plus simple que dans le cas qui précède :
- sélectionner les N_I éléments d'une classe de taille sur laquelle la probabilité d'inclusion ω_I est constante ;
- tirer dans cette classe $[N_I\omega_I]$ ou $[N_I\omega_I] + 1$ éléments uniformément ;

– le regroupement de tous les tirages de toutes les classes fournit le sous-échantillon $\{Y_1, ..., Y_n\}$ voulu.

Le code source associé à cet algorithme se trouve en 6.3.2.

Importation de données

Le logiciel Matlab® possède des commandes permettant de charger des matrices de données directement en mémoire. Il est donc possible d'importer, pourvu que le format informatique soit correct (textuel), des données relatives à la loi parente ou directement des données sous-échantillonnées, comme les données réelles.

Outil de sélection de modèles

Nous allons détailler les algorithmes relatifs à l'estimation de la densité $f_{\alpha,\omega}$ par maximum de vraisemblance dans un modèle. Nous verrons ensuite le protocole de sélection de modèles par pénalisation de la vraisemblance, et notamment la méthode de la pente décrite dans la section 4.4.4.

En premier lieu, notons que nous ne serons pas exactement dans le cadre des hypothèses du Théorème 4.4.1. En effet, la densité $f_{\alpha,\omega}$ que nous souhaitons estimer n'est pas minorée par un nombre strictement positif, pas plus qu'elle n'est à support compact.

Il est toujours possible de travailler conditionnellement aux données et l'estimateur par maximum de vraisemblance ne charge alors pas les intervalles vides de données, comme on l'a vu au chapitre 3. Il en résulte que cet estimateur ne sera toujours défini que sur l'intervalle $[\min \mathbb{Y} ; \max \mathbb{Y}]$. Nous supposerons donc que, conditionnellement aux données, notre protocole d'estimation est pratiqué sur cet intervalle compact.

Par ailleurs, il nous est impossible de garantir que la vraie densité à estimer ne s'annule pas. cependant, comme nous l'avons remarqué dans la section 4.4.3, il est très probable qu'en pratique cette hypothèse ne soit pas nécessaire.

Enfin, nous baserons les bornes des intervalles de constance de la fonction ω représentant les probabilités d'inclusion au $n^{\text{ème}}$ tirage sur les données plutôt que de construire des partitions régulières dont les bornes se situent entre les données et peuvent alors paraître (paradoxalement) arbitraires. En effet, dans l'estimation non-paramétrique d'une fonction par maximum de vraisemblance, l'estimateur obtenu est toujours constant entre deux observations (intuitivement, un modèle non basé sur les données est donc moins vraisemblable). Il n'y a donc pas de raison pour nous de contraindre ce penchant naturel, même s'il nous fait sortir une fois de plus du cadre strict du Théorème 4.4.1. En effet, baser les partitions sur les données signifie que l'on rend ces partitions aléatoires et nous n'avons pas tenu compte d'un tel aléa dans le Théorème.

Pour rentrer dans le cadre d'une densité sur $[0\,;1]$, on effectue, conditionnellement aux données $\mathbb{Y} = \{Y_1, ..., Y_n\}$, la transformation suivante :

$$f_{\alpha,\omega}(x) \longleftarrow (\max \mathbb{Y}) \times f_{\alpha,\omega}((\max \mathbb{Y}) \times x),$$

où le membre de gauche est bien une fonction définie sur $[0\,;1]$. C'est cette dernière fonction qui va être estimée dans notre procédure.

En pratique, lorsque nous aurons à définir une partition de l'intervalle d'étude, nous construirons alors la partition régulière (ou sur-partition de cette dernière dans le cas des partitions irrégulières) et considérerons comme partition effective celle dont les bornes sont "calées" sur les données les plus proches (voir en 6.4.1 pour le code associé).

Maximisation de la vraisemblance dans un modèle

Nous souhaitons ici mettre en oeuvre uniquement la procédure d'estimation par maximum de vraisemblance sur un modèle (*i.e.* à partition fixée) avec contrainte de monotonie en ω.

Une partition basée sur les données étant fixée, nous avons vu dans le chapitre 3 que l'estimation de α pouvait se faire par dichotomie sur la fonction φ'_* définie dans la preuve de la Proposition 3.2.8. À chaque valeur étape de α, la valeur de ω correspondante est déterminée par la régression isotonique donnée par la Proposition 3.2.5. Ces estimations permettent aussi de calculer la vraisemblance du modèle, qui va nous être utile dans l'évaluation de la pente dans la calibration de la constante de la fonction de pénalité à prendre en compte ensuite pour le protocole de sélection de modèles.

Les codes sources relatifs à cette estimation (méthode de régression isotonique et dichotomie sur φ'_*) se trouvent en 6.4.2 et 6.4.3.

Sélection de modèles

Pour terminer le protocole d'estimation, il suffit ensuite de pénaliser la vraisemblance calculée sur l'ensemble des modèles par la procédure ci-dessus. La constante multiplicative appliquée à la fonction générique de pénalité – qui dépend de la méthode de partition choisie – est calculée par la méthode de pente (cf. section 4.4.4).

De nombreuses options de partition de l'intervalle d'étude sont proposées :
- partitions en intervalles de longueur régulière ;
- partitions en intervalles de longueur irrégulières basés (sur-partitions) sur une partition fine régulière ;
- partitions en intervalles de mesure empirique (ou fréquence) régulière ;
- partitions en intervalles de mesure empirique irrégulières basés (sur-partitions) sur une partition fine dont les intervalles sont de mesure empirique régulière.

Il est aussi possible de choisir de fixer la valeur du paramètre α si de l'information *a priori* existe par ailleurs, ou de la laisser libre à l'estimation.

Enfin, il est possible de fixer la partition sur laquelle on veut voir se baser l'estimation de ω, contrainte ou non, α étant connu ou non. On n'a alors qu'une estimation par maximum de vraisemblance sur un modèle spécifique.

Un exemple générique des codes source de sélection de modèle, et de détermination de la pente que ces protocoles d'estimation nécessitent, peut être trouvé en 6.5.

Sorties texte et graphique

L'estimation de ω et éventuellement de α étant disponible, le logiciel permet de sauvegarder un fichier ASCII (au format ".txt") contenant :
- la valeur estimée de α ;
- des bornes des plages de constance de ω sélectionnées par l'algorithme ;
- des valeurs estimées de ω sur ces plages ;
- des valeurs estimées du nombre total de champs sur chaque plage au moyen des estimateurs de type Horvitz-Thompson (cf. section 3.3.1) ;
- des valeurs estimées du total des réserves sur chaque plage au moyen des estimateurs de type Horvitz-Thompson (cf. section 3.3.3).

Une sortie graphique conviviale reprenant chacun de ces éléments est de plus fournie. Il est aussi possible, lorsque les données sont issues d'une simulation de la loi parente, de représenter sur le graphe de l'estimation de ω les vraies valeurs des probabilités d'inclusion calculées sur le modèle sélectionné par l'algorithme ou l'utilisateur.

Nous ne décrivons pas ici les codes permettant ces sorties informatiques car ils n'ont évidemment aucun intérêt algorithmique. Il faut cependant remarquer que le code source de l'interfaçage du logiciel **select** lui-même, et qui fait appel à l'ensemble des divers codes fournis en annexe, est incomparablement plus volumineux que ces derniers et représente l'essentiel du temps de développement informatique. Le mode d'emploi et le mode d'interfaçage du logiciel figurent en annexe 2.

Grâce à cet outil, nous sommes en mesure d'effectuer des simulations de la performance de notre méthode.

5.2 Simulations

Nous allons distinguer deux cas de simulation de l'échantillon des observés proposés par le logiciel **select**. Dans chacun d'entre eux, la population parente suit une loi de Lévy-Pareto de paramètre α mais :

- dans le premier cas, la fonction de pondération est une fonction en escalier. La densité des LogTailles résultante appartient alors un modèle exponentiel de polynômes par morceaux de degré 1 ;
- dans le second, la fonction de pondération est une fonction puissance. La densité des LogTailles résultante n'appartient alors pas aux modèles dans lesquels nous pratiquons les estimation.

Remarque : il est à noter que tous les protocoles d'estimation détectent quasiment sans aucun défaut un tirage aléatoire simple, ou assimilé d'une loi de Lévy-Pareto. Par "simple ou assimilé", nous entendons soit
- un échantillonnage i.i.d. d'une loi de Lévy-Pareto ;
- un sous-échantillonnage non biaisé au sein d'une population parente elle même échantillonnage d'une loi de Lévy-Pareto.

Dans ce dernier cas, l'estimation de α est très bonne, mais les hypothèses (Hnc) et (Hc) ne sont évidemment pas respectées puisque $\max \omega = p \neq 1$ où p est le taux de sondage. Il en résulte N et R sont sous-estimés d'un facteur multiplicatif $1/p$.

Ces deux cas ne présentant pas d'intérêt, nous ne présentons pas de simulations les concernant.

Revenons à nos deux premiers cas. Dans le premier d'entre eux qui concerne l'estimation d'une densité appartenant à un modèle exponentiel de polynômes par morceaux, les 8 modes d'estimation suivants sur le couple (α, ω) sont simulés :
- NC REG : sans contrainte de monotonie en ω sur partitions dont les intervalles sont de longueur régulière ;
- C REG : *sous* contrainte de monotonie en ω sur partitions dont les intervalles sont de longueur régulière ;
- NC IRREG : sans contrainte de monotonie en ω sur partitions qui sont sur-partitions (voir définition 3.2.7) d'une partition fine dont les intervalles sont de longueur régulière ;
- C IRREG : *sous* contrainte de monotonie en ω sur partitions qui sont sur-partitions d'une partition fine dont les intervalles sont de longueur régulière ;
- NC FREC : sans contrainte de monotonie en ω sur partitions dont les intervalles sont de fréquence d'observation régulière ;
- C FREC : *sous* contrainte de monotonie en ω sur partitions dont les intervalles sont de fréquence d'observation régulière ;
- NC IRFREQ : sans contrainte de monotonie en ω sur partitions qui sont sur-partitions d'une partition fine dont les intervalles sont de fréquence d'observation régulière ;
- C IRFREQ : *sous* contrainte de monotonie en ω sur partitions qui sont sur-partitions d'une partition fine dont les intervalles sont de fréquence d'observation régulière.

Nous n'avons pas simulé les modes d'estimation de ω à α connu, car ils ne sont d'aucun intérêt pratique dans notre cas et sont à peu de choses près équivalents aux modèles d'histogrammes qui sont simulés dans Castellan [20].

Dans le second cas où la densité à estimer n'appartient pas à un modèle exponentiel de polynômes par morceaux, seuls les 4 protocoles d'estimation sous contraintes sont simulés.

5.2.1 Simulation au moyen d'une densité exponentielle de polynômes par morceaux

Nous avons testé chaque protocole d'estimation sur trois densités différentes, l'objectif étant de voir si la valeur du paramètre α ou le nombre de morceaux intervenant dans la partition m ont une influence sur les performances des estimations.

Notons que dans ce cas, le nombre d'observations n est aléatoire et dépend de α et ω (voir section 2.3.2).

A chaque densité testée est associée une figure représentant un exemple d'échantillonnage/sous-échantillonnage correspondant aux caractéristiques de la simulation.

Les résultats détaillés des estimations se trouvent en annexe 8.1. Dans chaque case de chaque tableau de résultat on trouve la moyenne robuste du paramètre étudié (calculée sur le ventre représentant 95 % de la distribution), l'écart-type robuste ainsi qu'en dessous, les fractiles d'ordre 5 %, 50 % et 95 % de sa distribution sur les simulations effectuées. Les estimations des paramètres dont nous donnons les caractéristiques des distributions sont :
– le paramètre α de la loi de Lévy-Pareto sous-jacente ;
– l'erreur relative commise sur le nombre de champs *restant à découvrir* ;
– l'erreur relative commise sur le montant des réserves *restant à découvrir*.

Première densité : habitat très concentré

Caractéristiques de $f_{\alpha,\omega}$ (voir figure 5.1) :
- $\alpha = 0{,}6$;
- effectif de la population parente : $N = 2000$;
- partition de $\mathbb{R}^+$ en les 4 morceaux suivants :

I	$[0\ ;20[$	$[20\ ;200[$	$[200\ ;2000[$	$[2000\ ;+\infty[$
ω	0,1	0,3	0,6	1

- moyenne de n : 311,11 (écart-type : 6,51).

Les résultats se trouvent dans le tableau 5.1.

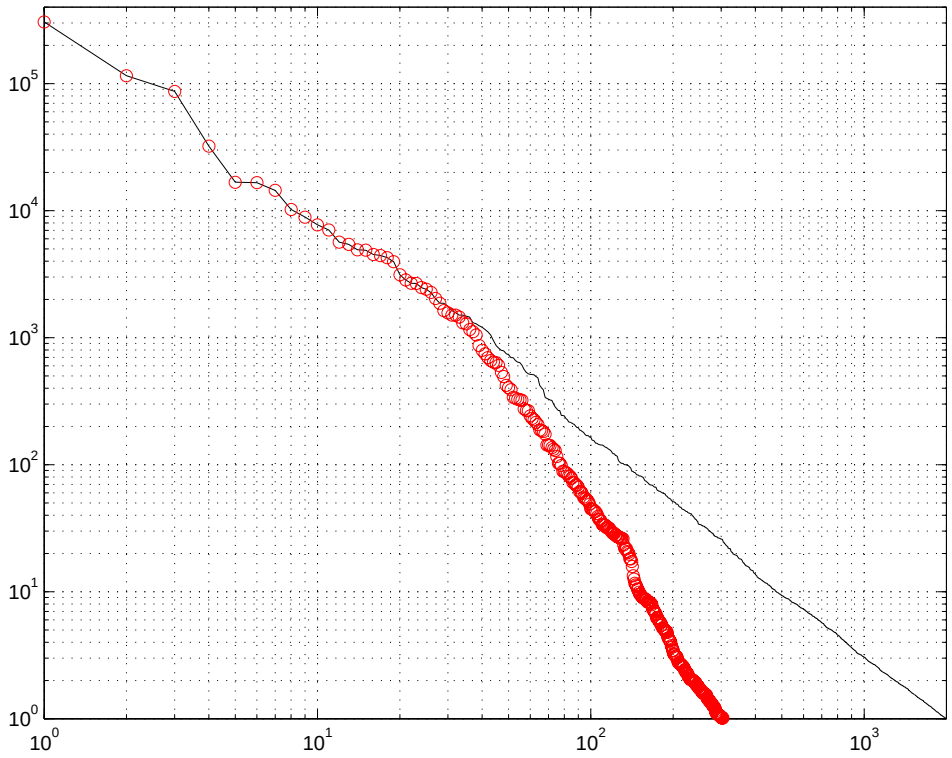

FIG. 5.1 – Diagramme LogLog d'une densité d'habitat très concentré et sous-échantillon associé.

Seconde densité : habitat concentré

Caractéristiques de $f_{\alpha,\omega}$:
– $\alpha = 0{,}8$;
– effectif de la population parente : $N = 2000$;
– partition de $\mathbb{R}^+$ en les 5 morceaux suivants :

I	$[0\ ;10[$	$[10\ ;50[$	$[50\ ;100[$	$[100\ ;500[$	$[500\ ;+\infty[$
ω	0,1	0,3	0,5	0,8	1

– moyenne de n : 311,55 (écart-type : 7,01).

Les résultats se trouvent dans le tableau 5.2.

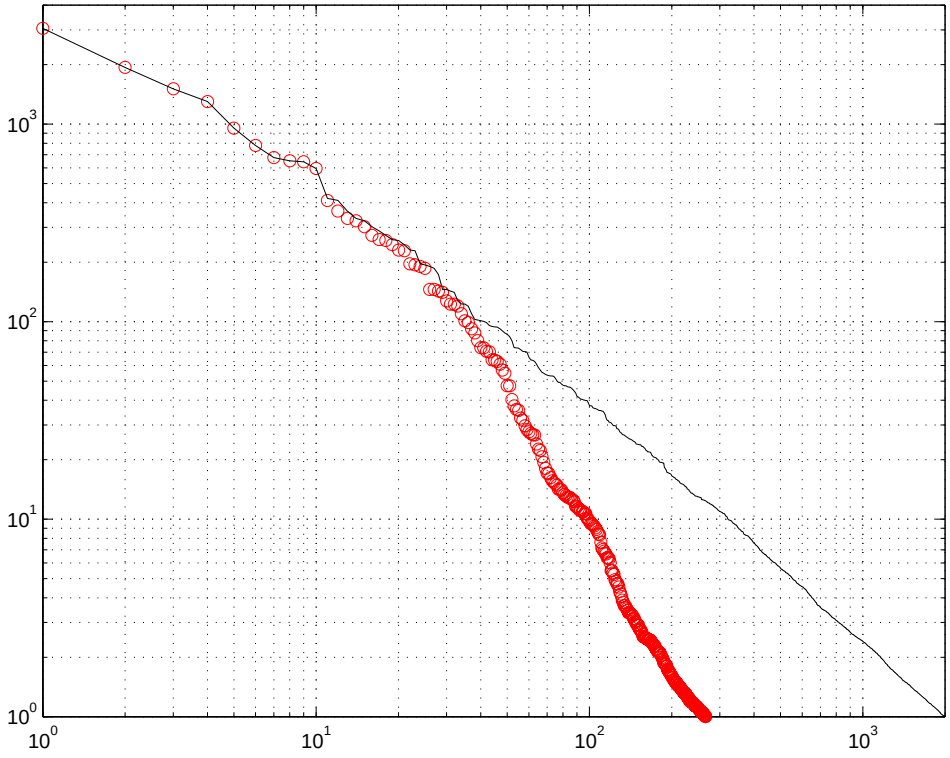

FIG. 5.2 – Diagramme LogLog d'une densité d'habitat concentré et sous-échantillon associé.

	$\hat{\alpha}$	$\frac{(\widehat{N}-n)-(N-n)}{N-n}$	$\frac{(\widehat{R}-r)-(R-r)}{R-r}$
NC REG	0,65 (0,12)	2,79 (5,63)	2,56 (3,53)
200 sim.	0,48 / 0,63 / 0,96	$-0,54$ / 0,41 / 14,11	$-0,80$ / 1,22 / 11,13
C REG	0,67 (0,12)	4,69 (11,65)	3,22 (5,33)
200 sim.	0,52 / 0,63 / 1,02	$-0,47$ / 0,85 / 19,28	$-0,56$ / 1,49 / 13,59
NC IRREG	0,64 (0,06)	1,10 (1,41)	1,86 (1,86)
100 sim.	0,52 / 0,64 / 0,80	$-0,39$ / 0,85 / 3,25	$-0,20$ / 1,26 / 6,13
C IRREG	0,65 (0,07)	1,37 (1,61)	1,86 (1,62)
100 sim.	0,56 / 0,65 / 0,79	$-0,28$ / 0,85 / 5,07	$-0,07$ / 1,33 / 5,33
NC FREQ	0,62 (0,09)	0,99 (1,96)	0,70 (1,27)
200 sim.	0,46 / 0,62 / 0,83	$-0,65$ / 0,38 / 5,00	$-0,76$ / 0,34 / 3,16
C FREQ	0,67 (0,07)	1,57 (1,89)	1,02 (1,23)
200 sim.	0,55 / 0,66 / 0,84	$-0,31$ / 0,88 / 5,49	$-0,39$ / 0,65 / 3,52
NC IRFREQ	0,63 (0,07)	0,80 (1,24)	0,94 (1,04)
100 sim.	0,51 / 0,63 / 0,80	$-0,45$ / 0,48 / 3,13	$-0,41$ / 0,69 / 3,06
C IRFREQ	0,60 (0,06)	0,31 (0,43)	0,55 (0,47)
60 sim.	0,43 / 0,31 / 0,70	$-0,50$ / 0,30 / 1,02	$-0,68$ / 0,48 / 1,20

TAB. 5.1 – Résultats des simulations pour la densité d'habitat très concentré.

Troisième densité : habitat dispersé

Caractéristiques de $f_{\alpha,\omega}$:
– $\alpha = 1{,}2$;
– effectif de la population parente : $N = 2000$;
– partition de $\mathbb{R}^+$ en les 4 morceaux suivants :

I	$[0\ ;5[$	$[5\ ;10[$	$[10\ ;50[$	$[50\ ;+\infty[$
ω	0,1	0,3	0,6	1

– moyenne de n : 310,96 (écart-type : 6,66).

Les résultats se trouvent dans le tableau 5.3.

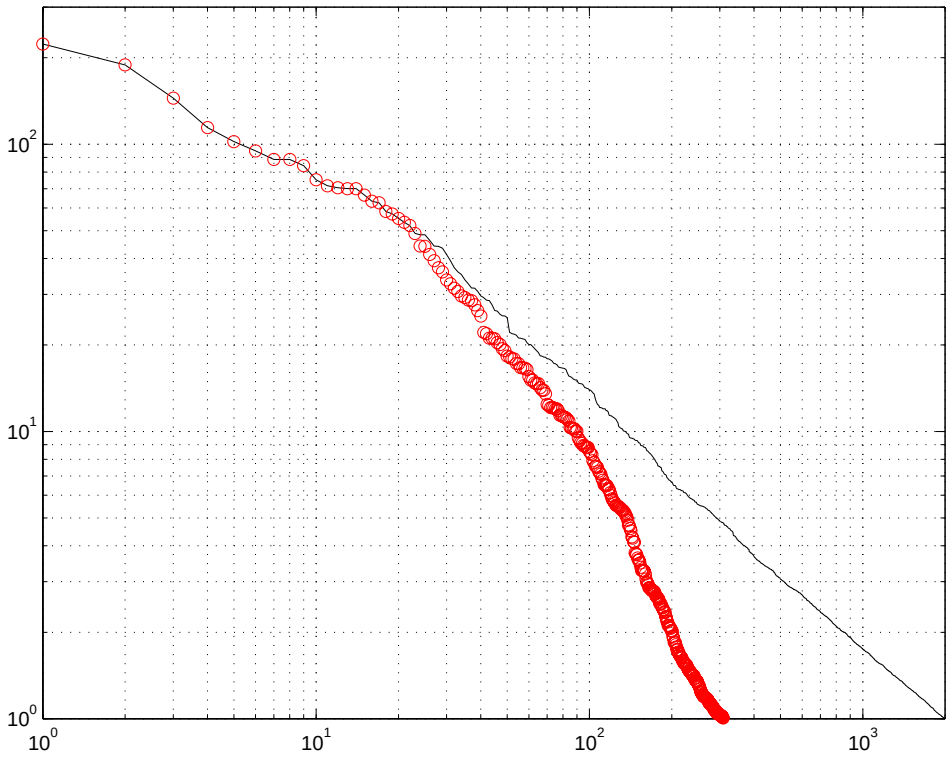

FIG. 5.3 – Diagramme LogLog d'une densité d'habitat dispersé et sous-échantillon associé.

	$\hat{\alpha}$	$\frac{(\widehat{N}-n)-(N-n)}{N-n}$	$\frac{(\widehat{R}-r)-(R-r)}{R-r}$
NC REG	0,83 (0,20)	4,10 (14,10)	3,19 (9,37)
200 sim.	0,56 / 0,79 / 1,32	$-0,82$ / 0,12 / 19,33	$-0,86$ / 0,47 / 14,59
C REG	0,88 (0,17)	3,24 (6,48)	2,85 (4,82)
200 sim.	0,61 / 0,87 / 1,27	$-0,71$ / 0,92 / 16,48	$-0,69$ / 1,14 / 12,45
NC IRREG	0,85 (0,10)	1,24 (1,77)	1,41 (1,65)
100 sim.	0,70 / 0,84 / 1,10	$-0,27$ / 0,54 / 5,35	$-0,30$ / 0,86 / 4,72
C IRREG	0,87 (0,09)	1,44 (1,63)	1,78 (1,61)
100 sim.	0,72 / 0,86 / 1,07	$-0,19$ / 0,86 / 5,50	0,00 / 1,17 / 4,79
NC FREQ	0,80 (0,10)	0,20 (0,94)	0,12 (0,74)
200 sim.	0,63 / 0,79 / 1,03	$-0,68$ / $-0,11$ / 2,45	$-0,71$ / $-0,08$ / 1,80
C FREQ	0,87 (0,09)	0,82 (1,18)	0,62 (0,94)
200 sim.	0,73 / 0,87 / 1,08	$-0,37$ / 0,43 / 3,41	$-0,46$ / 0,31 / 2,54
NC IRFREQ	0,80 (0,09)	0,27 (0,64)	0,34 (0,64)
100 sim.	0,64 / 0,81 / 0,96	$-0,52$ / 0,18 / 1,34	$-0,48$ / 0,23 / 1,42
C IRFREQ	0,81 (0,10)	0,48 (1,34)	0,37 (1,19)
60 sim.	0,65 / 0,82 / 1,14	$-0,46$ / 0,19 / 2,01	$-0,51$ / 0,31 / 2,61

TAB. 5.2 – Résultats des simulations pour la densité d'habitat concentré.

Commentaires

En tout premier lieu, notons que si le nombre de simulations diffère d'une technique d'estimation à l'autre, c'est bien évidemment en raison du temps de calcul nécessaire à leur execution. À titre d'exemple, les temps de calcul UC sur PC équipé d'un processeur intel celeron^{TM} cadencé à 800 MHz vont de 15 à plus de 75 heures pour les simulations irrégulières sous contraintes pour 100 simulations, avec un nombre maximal de 14 morceaux pour la partition la plus fine.

Concernant $\hat{\alpha}$, nous notons que quelle que soit la valeur de α, les écart-types des distributions sont de l'ordre de 20 % à 25 % sur les méthodes NC REG et C REG, là où elles sont plutôt de l'ordre de 10 à 12 % sur toutes les autres méthodes. Nécessairement, cette volatilité se traduit sur les estimations du nombre de champs restant et des réserves restants à découvrir dont les distributions sont largement plus dispersées dans ces deux cas. Si cela est moins évident sur le cas $\alpha = 1,2$, cela est dû au fait que la distribution de départ est elle-même beaucoup moins dispersive. Par ailleurs, la valeur α semble être légèrement surestimée, peut-être encore plus fortement du fait que α est élevé. L'estimation des paramètres du modèle semble donc "forcer" un peu trop le biais observé dans le tirage. Il faut remarquer que cette tendance est plus marquée lorsque les estimations sont réalisées sous contraintes de monotonie en ω. Cette dérive est explicable. En effet, à modèle fixé, l'estimation sous contrainte de α est généralement supérieure ou égale à son estimation non contrainte. Il est donc raisonnable que cette tendance se retrouve en sélection de modèles.

Cette surestimation de α implique nécessairement une surestimation du nombre de champs et des réserves restant à découvrir, ce qui apparaît dans les caractéristiques des distributions de ces deux statistiques. La dispersion beaucoup plus faible autour de α pour toutes les simulations autres que NC REG et C REG se retrouve sur les statistiques d'intérêt, où moyenne et médiane sont toujours strictement positives. Notons que les valeurs faibles des écarts-types des distributions relatives à $\alpha = 0,6$ et au protocole C IRFREQ sont douteuses et compte tenu du nombre de simulations relativement faible. On peut penser qu'un effet de tirage a biaisé ces simulations.

Dernière remarque, les écarts-types de toutes les statistiques d'intérêt sont systématiquement plus faibles lorsque l'on travaille sur les protocoles basés sur les blocs statistiquement équivalents. Sachant qu'il est toujours plus avantageux de travailler avec des estimateurs de faible variance, même s'ils sont légèrement biaisés[2], peut-être ces estimateurs auront-ils notre préférence. Notons que les performances des estimateurs des protocoles NC IRREG et C IRREG ne sont pas éloignés de ces derniers au vu de ces statistiques. Si

[2]Le biais est en effet souvent lui aussi estimable et l'on peut donc trouver des solutions pour le corriger.

	$\hat{\alpha}$	$\frac{(\widehat{N}-n)-(N-n)}{N-n}$	$\frac{(\widehat{R}-r)-(R-r)}{R-r}$
NC REG	1,20 (0,25)	1,26 (2,98)	1,18 (2,69)
200 sim.	0,84 / 1,14 / 1,68	$-0,78$ / $-0,14$ / 8,88	$-0,80$ / 0,00 / 8,25
C REG	1,19 (0,24)	1,20 (3,83)	1,09 (3,19)
200 sim.	0,90 / 1,13 / 1,82	$-0,75$ / $-0,21$ / 8,27	$-0,75$ / $-0,15$ / 7,25
NC IRREG	1,32 (0,13)	1,34 (1,36)	1,36 (1,28)
100 sim.	1,08 / 1,31 / 1,54	$-0,12$ / 0,93 / 3,85	$-0,12$ / 1,06 / 4,05
C IRREG	1,31 (0,14)	1,19 (1,44)	1,18 (1,27)
100 sim.	1,04 / 1,30 / 1,63	$-0,20$ / 0,76 / 3,81	$-0,16$ / 0,84 / 3,68
NC FREQ	1,27 (0,18)	0,77 (1,33)	0,71 (1,17)
200 sim.	0,98 / 1,26 / 1,64	$-0,62$ / 0,31 / 3,89	$-0,62$ / 0,37 / 3,29
C FREQ	1,33 (0,14)	1,07 (1,35)	0,97 (1,15)
200 sim.	1,08 / 1,32 / 1,64	$-0,34$ / 0,69 / 4,33	$-0,37$ / 0,65 / 3,55
NC IRFREQ	1,28 (0,14)	0,71 (1,09)	0,71 (0,98)
100 sim.	1,04 / 1,28 / 1,60	$-0,33$ / 0,48 / 2,88	$-0,34$ / 0,52 / 2,76
C IRFREQ	1,29 (0,15)	0,85 (1,24)	0,84 (1,13)
60 sim.	0,96 / 1,29 / 1,60	$-0,21$ / 0,42 / 3,40	$-0,25$ / 0,41 / 3,37

TAB. 5.3 – Résultats des simulations pour la densité d'habitat dispersé.

le nombre de morceaux des partitions irrégulières que nous utilisons était plus grand, il est certain que l'ensemble de ces estimateurs auraient des performances similaires. Les estimateurs basés sur les partitions en fréquences d'observations régulières sont probablement plus informatifs sur la structure des données. D'un point de vue théorique, il serait aussi très intéressant de pouvoir disposer d'un résultat du type du Théorème 4.4.1 sur les blocs statistiquement équivalents. Cela reste une piste de recherche à l'heure actuelle.

5.2.2 Simulation au moyen d'une densité exponentielle non polynômiale par morceaux

Dans cette section, nous nous intéressons aux simulations mettant en jeu une population parente de loi de Lévy-Pareto de paramètre α et une sous-échantillon observé de cette population issue d'un tirage successif biaisé par effet taille comme dans la section section 2.3.1. Les probabilités d'inclusion au premier tirage π^1 sont directement proportionnelles à la taille, comme par exemple dans Bickel *et al.* [13].

Pour toutes les simulations, l'effectif N de la population parente est de $N = 2000$ et l'effectif de l'échantillon observé est de $n = 300$.

Pour cette simulation, nous avons testé uniquement les protocoles d'estimation sous contrainte de monotonie en ω sur trois densités différentes. L'objectif étant ici de voir si la valeur du paramètre α, lorsque la fonction de pondération ω est inconnue, a une influence sur les performances des estimations.

Comme dans la section précédente, à chaque densité testée est associée une figure représentant un exemple d'échantillonnage/sous-échantillonnage correspondant aux caractéristiques de la simulation.

Les résultats détaillés des estimations se trouvent en annexe 8.2. Dans chaque case de chaque tableau de résultat on trouve la moyenne robuste du paramètre étudié (calculée sur le ventre représentant 95 % de la distribution), l'écart-type robuste ainsi qu'en dessous, les fractiles d'ordre 5 %, 50 % et 95 % de sa distribution sur les simulations effectuées.

Première densité : habitat très concentré

Caractéristiques de $f_{\alpha,\omega}$:
- $\alpha = 0{,}6$;
- effectif de la population parente : $N = 2000$;
- effectif observé : $n = 300$;
- ω est inconnue mais les probabilités d'inclusion au premier tirage π^1 sont directement proportionnelles à la taille.

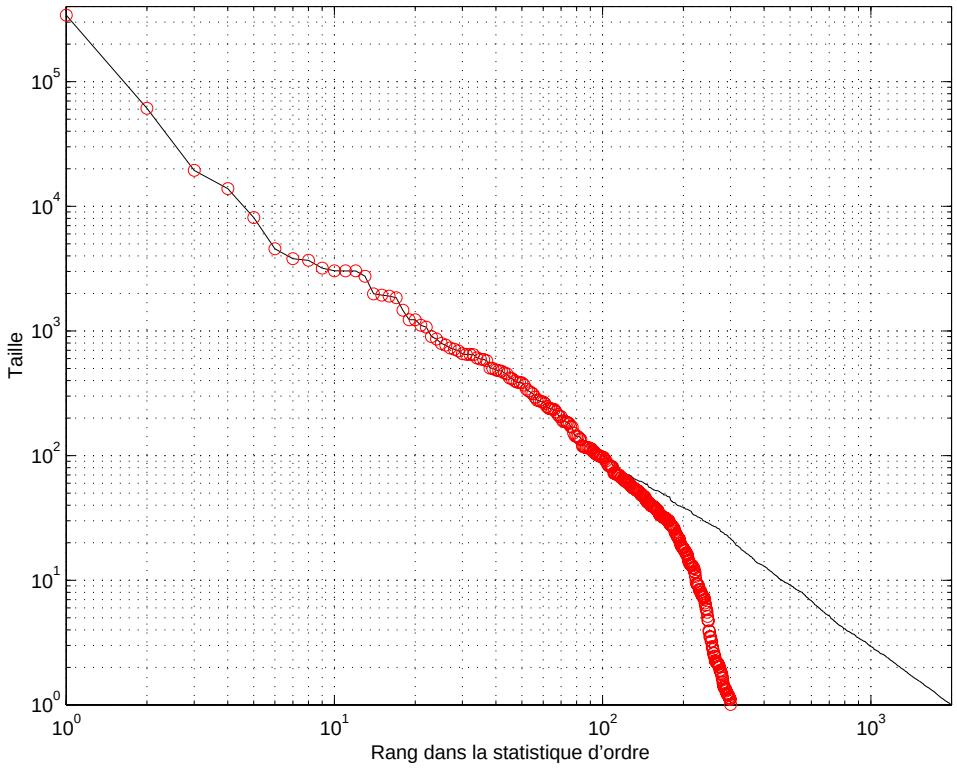

FIG. 5.4 – Diagramme LogLog d'une densité d'habitat très concentré et sous-échantillon par tirage successif proportionnel à la taille associé.

	$\hat{\alpha}$	$\frac{(\widehat{N}-n)-(N-n)}{N-n}$	$\frac{(\widehat{R}-r)-(R-r)}{R-r}$
C REG	0,56 (0,08)	0,06 (1,10)	1,16 (2,93)
200 sim.	0,44 / 0,55 / 0,77	$-0,67$ / $-0,32$ / 1,96	$-0,51$ / $-0,02$ / 7,88
C IRREG	0,58 (0,05)	$-0,11$ (0,35)	0,19 (0,66)
100 sim.	0,50 / 0,58 / 0,68	$-0,53$ / $-0,16$ / 0,50	$-0,42$ / 0,00 / 1,71
C FREQ	0,59 (0,07)	0,05 (0,72)	0,48 (1,16)
200 sim.	0,46 / 0,58 / 0,77	$-0,59$ / $-0,19$ / 1,82	$-0,52$ / 0,03 / 3,21
C IRFREQ	0,60 (0,06)	0,00 (0,43)	0,29 (0,75)
60 sim.	0,50 / 0,60 / 0,70	$-0,49$ / $-0,09$ / 0,79	$-0,48$ / 0,01 / 1,92

TAB. 5.4 – Résultats des simulations pour la densité d'habitat très concentré, ω inconnue.

Seconde densité : habitat concentré

Caractéristiques de $f_{\alpha,\omega}$:
- $\alpha = 0{,}8$;
- effectif de la population parente : $N = 2000$;
- effectif observé : $n = 300$;
- ω est inconnue mais les probabilités d'inclusion au premier tirage π^1 sont directement proportionnelles à la taille, comme dans Bickel *et al.* [13].

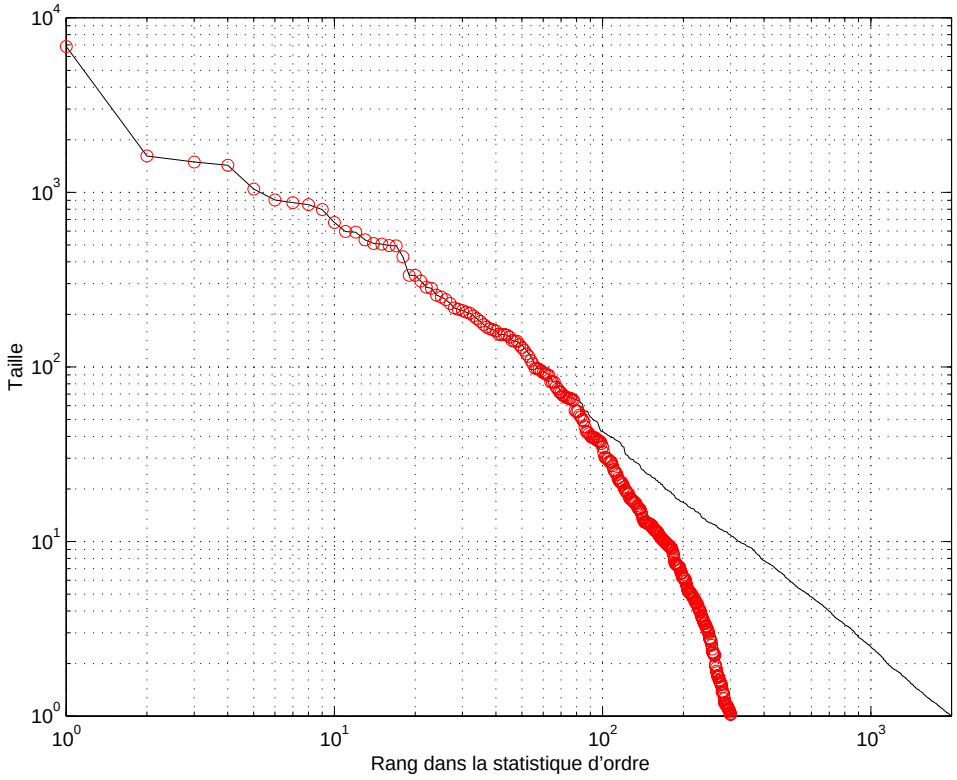

FIG. 5.5 – Diagramme LogLog d'une densité d'habitat concentré et sous-échantillon par tirage successif proportionnel à la taille associé.

	$\hat{\alpha}$	$\frac{(\widehat{N}-n)-(N-n)}{N-n}$	$\frac{(\widehat{R}-r)-(R-r)}{R-r}$
C REG	0,74 (0,12)	0,10 (1,37)	0,39 (1,73)
200 sim.	0,56 / 0,71 / 1,06	$-0,72$ / $-0,38$ / 3,66	$-0,71$ / $-0,24$ / 4,64
C IRREG	0,79 (0,07)	0,04 (0,47)	0,23 (0,64)
100 sim.	0,67 / 0,79 / 0,94	$-0,49$ / $-0,04$ / 1,03	$-0,54$ / 0,06 / 1,52
C FREQ	0,79 (0,10)	0,09 (0,73)	0,20 (0,79)
200 sim.	0,62 / 0,78 / 1,02	$-0,62$ / $-0,12$ / 1,63	$-0,65$ / $-0,03$ / 1,84
C IRFREQ	0,79 (0,08)	0,04 (0,65)	0,12 (0,67)
100 sim.	0,67 / 0,78 / 1,03	$-0,46$ / $-0,12$ / 1,22	$-0,40$ / $-0,07$ / 1,61

TAB. 5.5 – Résultats des simulations pour la densité d'habitat concentré, ω inconnue.

Troisième densité : habitat dispersé

Caractéristiques de $f_{\alpha,\omega}$:
- $\alpha = 1{,}2$;
- effectif de la population parente : $N = 2000$;
- effectif observé : $n = 300$;
- ω est inconnue mais les probabilités d'inclusion au premier tirage π^1 sont directement proportionnelles à la taille, comme dans Bickel *et al.* [13].

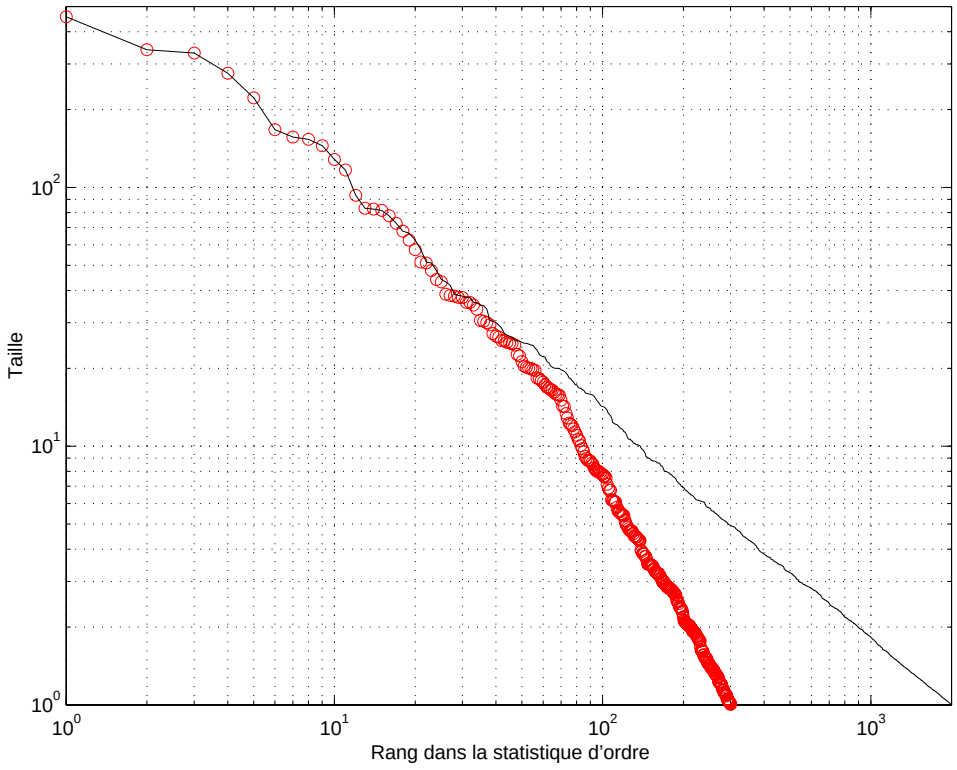

FIG. 5.6 – Diagramme LogLog d'une densité d'habitat dispersé et sous-échantillon par tirage successif proportionnel à la taille associé.

	$\hat{\alpha}$	$\frac{(\widehat{N}-n)-(N-n)}{N-n}$	$\frac{(\widehat{R}-r)-(R-r)}{R-r}$
C REG	1,11 (0,20)	0,10 (1,39)	0,15 (1,37)
200 sim.	0,81 / 1,09 / 1,54	$-0,84$ / $-0,41$ / 2,76	$-0,84$ / $-0,35$ / 3,33
C IRREG	1,17 (0,14)	0,15 (0,70)	0,23 (0,72)
100 sim.	0,92 / 1,16 / 1,44	$-0,63$ / $-0,08$ / 1,77	$-0,57$ / $-0,05$ / 1,61
C FREQ	1,04 (0,18)	$-0,39$ (0,57)	$-0,39$ (0,56)
200 sim.	0,76 / 1,03 / 1,43	$-0,90$ / $-0,58$ / 0,81	$-0,93$ / $-0,58$ / 0,85
C IRFREQ	1,21 (0,11)	0,23 (0,65)	0,26 (0,62)
60 sim.	0,99 / 1,22 / 1,44	$-0,43$ / 0,08 / 1,66	$-0,50$ / 0,11 / 1,45

TAB. 5.6 – Résultats des simulations pour la densité d'habitat dispersé, ω inconnue.

Commentaires

Nous notons dans un premier temps que les estimations de α sur toutes les simulations semblent cette fois-ci soit légèrement biaisées à gauche pour les méthodes régulières C REG et C FREQ, soit centrées pour les méthodes irrégulières C IRREG et C IRFREQ. Les ordres de grandeur des écart-types des estimations de α sont un peu plus faibles que pour les simulations précédentes, c'est-à-dire de l'ordre de 15 % sur les méthodes régulières et 10 % sur les méthodes irrégulières. Compte tenu du meilleur centrage dont elles font état dans tous les cas, nous aurons donc tendance à penser que les méthodes irrégulières ont des qualités d'approximation bien meilleures que les méthodes régulières.

Par ailleurs, les protocoles basés sur les blocs statistiquement équivalents semblent de nouveau être plus précis que ceux basés sur les partitions homogènes en longueurs.

Il est aussi intéressant de constater que la méthode sélectionne généralement des modèles de dimension relativement faible, de l'ordre de 6 en moyenne pour un nombre maximal de morceaux autorisé de 20 pour les protocoles réguliers et 14 pour les protocoles irréguliers. Il aurait pu paraître naturel de sélectionner le plus grand nombre possible de morceaux pour gagner en précision dans le représentation du biais dans le tirage, mais le risque d'estimation

aurait alors alors été dégradé et un nombre "raisonnable" de morceaux pour la partition est sélectionné par la méthode. Ce résultat est à rapprocher du commentaire de la section 4.3.2.

Le meilleur centrage sur α implique évidemment une bien meilleure qualité d'estimation du nombre de champs et des réserves restant à découvrir. L'étendue des distributions constatées de ces statistiques est bien plus faible que dans le cas de la section 5.2.1, et les résultats de ces simulations sont étonnamment meilleurs. Ils montrent en particulier que nos estimations fournissent des résultats dignes de confiance.

La lecture de ces résultats de simulation peut paraître curieuse. En effet, les estimations semblent meilleures lorsque la densité estimée n'appartient pas aux modèles dans lesquels on l'estime. Il est possible de justifier ce phénomène de manière théorique en rappelant que l'estimation par sélection de modèles vise à faire un compromis entre biais d'estimation par le modèle et variance d'estimation à l'intérieur d'un modèle. Lorsque la densité à estimer appartient à un modèle, le terme de biais de l'oracle (au sens de la définition 4.3.1) doit être nul, et l'estimateur a donc fort à faire pour se rapprocher de l'oracle. Lorsque celle-ci n'appartient pas au modèle, le terme de biais de l'oracle est non nul est il est donc moins difficile à l'estimateur de s'approcher de celui-ci.

Du point de vue de l'interprétation pratique de ce phénomène, nous pensons que dans le cas de la section 5.2.1, une mauvaise estimation du nombre de morceaux ou de la localisation des bornes de la partition a un effet bien plus grave que dans le cas où la densité à estimer n'appartient pas au modèle car un effet de lissage s'opère. À titre d'analogie, il faut se rendre compte qu'estimer par un histogramme une densité constante par morceaux en ne localisant pas bien les bornes ou le nombre de morceaux (même avec des erreurs faibles) rend au final une densité de forme souvent très éloignée de l'originale alors qu'estimer de cette façon une densité régulière aura tendance à rendre un résultat ressemblant, ce avec une grande latitude sur les bornes ou le nombre de morceaux. Notre problème est très proche de cette considération, comme le montre la remarque (1) de la Proposition 3.2.3.

5.2.3 Quelques pas vers des intervalles de confiance

Nous ne donnons dans cette section que des pistes d'investigations. Une étude complète et détaillée basée notamment sur des simulations n'étant pas encore disponible pour venir étayer avec sûreté les arguments ci-dessous.

Étant donnée la complexité de nos estimateurs, il est hors de notre portée que de tenter d'évaluer de quelque manière que ce soit leurs lois asymptotiques, *a fortiori* non-asymptotiques en vue de construire des intervalles de confiance. Comment faire alors pour construire des intervalles de confiance

plausibles ? Nous avons tenté d'élaborer plusieurs méthodes dont nous décrivons les principes dans cette section.

Lorsqu'aucune théorie asymptotique n'est secourable, il est possible de faire appel à des simulations intensives de type Monte-Carlo dont le but est de générer un grand nombre d'échantillons de la loi dont que l'on suppose régir le phénomène auquel on s'intéresse. L'étude empirique de la statistique d'intérêt sur ces nombreux échantillons nous renseigne alors sur son comportement. Cette approche est cependant problématique à mettre en œuvre dans notre cas, car la loi de notre échantillon est non seulement inconnue, mais aussi non paramétrique... La simuler directement est donc délicat. Il est nécessaire de repasser par la loi de Lévy-Pareto parente (inconnue...) pour re-simuler ensuite l'échantillon observé grâce aux probabilités d'inclusion estimées.

Une autre approche possible peut être imaginée du côté du Bootstrap. Suivant le principe de base de cette méthode, nous devrions ré-échantillonner avec remise uniformément à l'intérieur de l'échantillon des observés et faire agir notre procédure sur chaque sous-échantillon généré. Il intuitif que ceci a peu de chances de fonctionner avec notre échantillon dans la mesure où, au départ, celui-ci est tiré d'une population parente de façon hautement non-uniforme. Tirer uniformément dans cet échantillon nous ferait alors perdre énormément d'information quant à sa structure d'échantillonnage, principalement en ce qui concerne les objets de tailles les plus importantes. Une technique de bootstrap stratifié (voir Bertail et Combris [12]) est alors certainement plus intéressante dans la mesure où celle-ci respecte le mode d'échantillonnage des données.

Il était aussi possible d'opter pour une technique mixant les deux méthodes. Celle-ci viserait à combiner un aléa sur la génération de la population parente par l'approche Monte-Carlo et un aléa sur le sous-échantillonnages biaisé via l'approche bootstrap.

Enfin, les simulations de la section précédente nous donnent aussi une idée relativement précise du biais et de la variabilité de nos estimateurs suivant le protocole adopté. Il est donc aussi envisageable d'adapter ces résultats pour tenter de construire des intervalles de confiance en fonction des estimations pratiquées.

Nous décrivons dans la suite les trois premiers points de vue. En pratique, nous aurons tendance à préférer soit une méthode de Monte-Carlo simple, qui donne des résultats tout-à-fait comparables à ceux de l'approche mixte, tout en ayant une complexité algorithmique moindre, soit la méthode basée sur les simulations intensives de la section 5.2.

L'approche de type Monte-Carlo

Pour simuler la loi parente tout en restant au plus proche des données, nous créons un échantillon d'une loi de Lévy-Pareto que nous appelons "sur-échantillon" et dont nous définissons de suite les caractéristiques :

Définition 5.2.1. *Soit $\mathbb{Y}$ un échantillon observé. Soient m une partition de $[\min \mathbb{Y}\,;\, \max \mathbb{Y}]$ et $\omega = \sum_{I \in m} \omega_I \mathbb{1}_I$ une fonction en escalier adaptée à m à valeurs dans $[0\,;\,1]$. Soit de plus $\alpha > 0$.*

On dit que $\mathbb{X}$ est un (m, ω, α)-sur-échantillon de $\mathbb{Y}$ si :
- *$\mathbb{Y} \subset \mathbb{X}$;*
- *$\mathbb{X}$ est un échantillon d'une loi de Lévy-Pareto de paramètre α ;*
- *pour tout I de m : $|\mathbb{X} \cap I| = \left[\dfrac{|\mathbb{Y} \cap I|}{\omega_I}\right]$ ou $|\mathbb{X} \cap I| = \left[\dfrac{|\mathbb{Y} \cap I|}{\omega_I}\right] + 1$. L'application $[\cdot]$ désignant classiquement la partie entière.*

Lorsque $\mathbb{Y}$ est un échantillon observé sur lequel on fait agir notre protocole d'estimation et s'il n'y a aucune ambiguïté, on appellera simplement "sur-échantillon" un $(\tilde{m}, \tilde{\omega}, \tilde{\alpha})$-sur-échantillon[3] de $\mathbb{Y}$, que l'on notera alors $\tilde{\mathbb{X}}$.

Remarque : $|\mathbb{Y} \cap I|$ et $|\mathbb{X} \cap I|$ jouent ici les rôles de n_I et N_I de la formule (3.9). On accepte de définir N_I à un individu près car il n'y a bien sûr aucune chance pour que $|\mathbb{Y} \cap I|/\omega_I$ soit un entier. En pratique, nous prendrons pour N_I la valeur arrondie à l'entier le plus proche.

Un sur-échantillon contient donc les données de départ et représente une simulation plausible de la population parente de l'échantillon observé.

Une approche de type Monte-Carlo consiste :
- à simuler un grand nombre C de sur-échantillons $(\tilde{\mathbb{X}}_c)_{1 \leq c \leq C}$ associés à $\mathbb{Y}$;
- puis à construire $(\mathbb{Y}_c)_{1 \leq c \leq C}$ tel que pour tout c, $\mathbb{Y}_c$ est un $\tilde{\omega}$-sous-échantillon de $\tilde{\mathbb{X}}$;
- et enfin faire agir le protocole d'estimation sur chaque $\mathbb{Y}_c$, $c = 1, ..., C$ pour récupérer une collection d'estimations $(\tilde{m}_c, \tilde{\omega}_c, \tilde{\alpha}_c)_{1 \leq c \leq C}$.

Dans l'absolu, les intervalles de confiance sur chacun des paramètres sont alors obtenus comme les fractiles d'un niveau donné des séries de simulations effectuées. Il est alors immédiat que se pose un problème sérieux quant aux intervalles de confiance pour les bornes de la partition. En effet, rien ne garantit a priori que le nombre de morceaux $(|\tilde{m}_c|)_{1 \leq c \leq C}$ des partitions estimées par simulation soit constant et égal à $|\tilde{m}|$. Les fractiles des séries associées aux bornes des intervalles de la partition ainsi qu'aux valeurs estimées des probabilités d'inclusion n'ont alors plus de sens.

[3] Ici, $\tilde{m}$, $\tilde{\omega}$ et $\tilde{\alpha}$ sont bien sûr associés à l'estimation $\tilde{f}_{\alpha,\omega} = f_{\tilde{\alpha},\tilde{\omega}}$ de $f_{\alpha,\omega}$ donnée par le Théorème 4.4.1.

On peut alors de suite envisager trois méthodes sensiblement analogues à celle décrite ci-dessus mais dont le but est de stabiliser les estimations $(\tilde{m}_c, \tilde{\omega}_c, \tilde{\alpha}_c)_{1\leq c\leq C}$. Par ordre de "rigidité" croissante imposée aux estimations sur les simulations,

- la première méthode consiste à pratiquer l'ensemble des estimations de $(\tilde{m}_c, \tilde{\omega}_c, \tilde{\alpha}_c)_{1\leq c\leq C}$ en utilisant le même protocole de sélection de modèles que pour $\mathbb{Y}$.
- la seconde méthode consiste à pratiquer l'ensemble des estimations des couples $(\tilde{\omega}_c, \tilde{\alpha}_c)_{1\leq c\leq C}$ à partition fixée de l'intervalle $[\min \mathbb{Y}\,;\, \max \mathbb{Y}]$ égale à $\tilde{m}$;
- la dernière enfin consisterait à ne pratiquer que les estimation des $(\tilde{\omega}_c)_{1\leq c\leq C}$ à partition fixée égale à $\tilde{m}$ et paramètre de la loi exponentielle sous-jacente fixé égal à $\hat{\alpha}$.

Le problème évident de la première méthode réside dans le fait que chaque partition sélectionnée à chaque étape de la procédure de Monte-Carlo risque d'être différente. Nous risquons ainsi de récupérer au final des résultats qui n'ont pas grand sens... La seconde méthode est quant à elle plus proche du processus générateur de données des simulations $(\tilde{\mathbb{X}}_c)_{1\leq c\leq C}$ et promet donc d'être plus stable (mais peut-être moins réaliste). La dernière méthode est évidemment la plus proche (trop proche ?) du processus générateur de données, et sera donc aussi la plus stable. C'est aussi la moins coûteuse en temps de calcul.

Les trois méthodes sont très faciles à implémenter. Les deux dernières retiendront notre attention car ce sont celles qui, pour nous, ont le plus de sens. En effet, pour les estimations par classe de taille du nombre total de champs et du montant total des réserves, seules l'estimation de ω nous importe réellement. La valeur de α n'est pour nous qu'une information auxiliaire.

L'approche Bootstrap

Celle-ci repose sur le fait que la loi empirique de l'échantillon observé $\mathbb{Y}$ de taille n est proche de la loi théorique de cet échantillon via par la Théorème de Glivenko-Cantelli. Heuristiquement, une statistique d'intérêt T étant donnée, regarder le comportement de T sur tous les sous-échantillons (avec remise, et généralement de même effectif) possibles au sein de l'échantillon de départ renseigne donc sur le comportement de T, et en particulier sur sa variabilité. On peut alors en déduire des intervalles de confiance "bootstrap" construits à partir des fractiles observés sur la distribution de T sur la loi empirique de l'échantillon observé. Bien entendu, il est très rapidement impossible d'exhiber tous les sous-échantillons de $\mathbb{Y}$ avec remise qui sont au nombre de n^n. L'idée première de la procédure de Bootstrap consiste donc à n'en tirer un grand nombre B au hasard. La distribution empirique de T sur ces B "réplications" approche alors (de nouveau via Glivenko-Cantelli) la

vraie distribution de T et les fractiles de cette distribution (dite distribution de la statistique T bootstrapée) fournissent alors les intervalles de confiance souhaités.

Très simple dans l'esprit, cette méthode ne peut absolument pas s'appliquer dans notre cas. En effet, pour avoir une chance de fonctionner, il est essentiel que le mode de ré-échantillonnage qui crée le sous-échantillon bootstrap soit proche du mode d'échantillonnage de l'échantillon de départ (voir Bertail et Combris [12]). Dans l'introduction ci-dessus, les observations ont été supposées issues d'un tirage aléatoire simple dans une loi de probabilité quelconque (en pratique, par exemple par image réciproque par la fonction de répartition de la loi en question d'un échantillon uniforme sur $[0\,;1]$). Dans notre cas, les observations, les données sont issues du ré-échantillonnage par tirage aléatoire fortement biaisé au sein d'un premier échantillon qui est lui-même un tirage aléatoire d'une certaine loi de probabilité.

Pour pouvoir utiliser une méthode de type bootstrap, nos sous-échantillons successifs doivent être obtenus par une méthode qui imite le protocole d'échantillonnage initial, qui nous est évidemment inconnu.

Comme nous supposons que sur chaque intervalle I de la partition $\tilde{m}$ retenue par la méthode de sélection de modèles, le tirage est un sondage aléatoire simple dont le taux de sondage est ω_I estimé par $\tilde{\omega}_I$, on peut effectuer un ré-échantillonnage avec remise simple à l'intérieur de chaque intervalle. C'est le principe du bootstrap stratifié, qui s'interprète comme un bootstrap pondéré particulier (voir Barbe et Bertail [9]).

Il est aussi possible d'envisager des techniques de bootstrap particulières adaptées aux lois stables, comme celle de Lévy-Pareto lorsque $\alpha < 2$. Des conditions nécessaires existent pour que le bootstrap d'une moyenne existe dans ce contexte (voir notamment Athreya [7] dont les résultats ont été affinés et étendus ensuite par Giné et Zinn [33]). Cette méthode est décrite en pratique dans Lepez et Mandonnet [53] et consiste à effectuer un boostrap classique sur un échantillon privé de ses plus grandes observations, c'est-à-dire celles qui ont le comportement le plus déviant. Nous interprétons aujourd'hui un tel protocole comme un bootstrap pondéré moins fin que celui que nous pourrions mettre en place grâce aux estimations que nous avons des probabilités d'inclusion par classe de taille.

Une approche mixte

Il est aussi possible de combiner les deux techniques décrites précédemment pour tenter de reproduire le "double" aléa qui constitue nos échantillons de données :

– aléa dans le tirage de la population parente, qui est reproduit par l'approche Monte-Carlo ;

– aléa dans le tirage du sous-échantillon des observations, qui est plutôt figuré par l'approche bootstrap.

L'idée d'un tel protocole consisterait donc à générer à chaque étape un sur-échantillon dans lequel on tirerait une sous-échantillon biaisé par effet taille auquel serait appliqué un bootstrap stratifié. Cette étape serait ensuite reproduite un grand nombre de fois.

On imagine aisément la lourdeur de traitement informatique d'une telle méthode qui paraît pourtant séduisante, sans parler des problèmes de calage des nombres de réplications et de simulation Monte-Carlo à mettre en place. Le gain par rapport à une méthode de Monte-Carlo simple semble assez clairement marginal dans notre contexte et nous en resterons donc là.

Venons-en, pour terminer, aux applications sur données réelles.

5.3 Estimation du potentiel de réserves d'un système pétrolier

Pour l'ensemble des applications qui suivent, les cartes et les descriptions géologiques fournies sont adaptées de l'USGS [80] et les données sont issues des bases Pétroconsultants 1998 [69]. Le découpage en systèmes pétroliers est conforme à celui de l'USGS.

Nous détaillons particulièrement le premier exemple, celui du Viking Graben de mer du Nord car il a été notre "terrain d'éssai" privilégié tout au long de la thèse. Le Viking Graben est en effet une région mature très bien connue des géologues de TotalFinaElf et de l'IFP sur laquelle de nombreuses données fiables sont disponibles.

Pour les autres régions, nous ne détaillons pas autant les résultats et commentaires. Ces exemples ont pour but de montrer les résultats du protocole d'estimation dans des régions du globe très différentes.

Nous rappelons l'hypothèse essentielle selon laquelle le plus gros champ du système pétrolier considéré a été trouvé de façon sûre, c'est à dire que sa probabilité d'inclusion au moment de l'étude est égale à 1.

5.3.1 Le Viking Graben de mer du Nord

Présentation

Un graben est une formation géologique située le long d'un axe de faille issue d'un étirement perpendiculaire de la roche, comme cela est représenté sur la figure 5.7. Si couches de roche mère et couvertures imperméables sont superposées, la formation d'un graben peut conduire à la création de pièges structuraux (cf. 1.1.1). Il s'agit d'un mode de piégeage classique dont le nord de la mer du Nord possède de nombreux représentants.

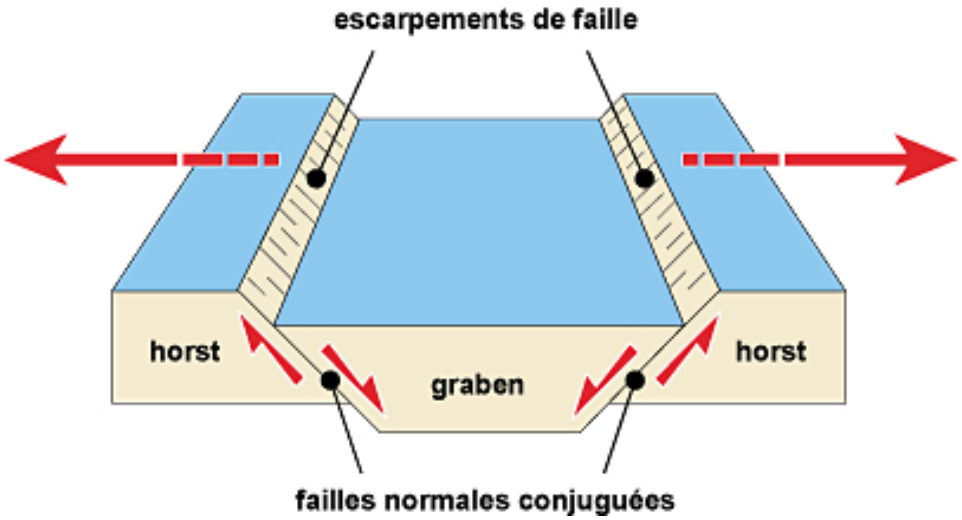

FIG. 5.7 – Formation d'un graben (source : photothèque www.cnrs.fr).

Le Viking Graben est le graben le plus au nord de mer du Nord. Nous considérons, suivant l'USGS [80], qu'il forme un système pétrolier à part entière et respecte donc l'homogénéité géologique des paramètres :
– roche mère ;
– modes de migration ;
– type de piégeage[4].

L'ensemble des champs pétroliers et gaziers de ce système proviennent de dépôts sédimentaires (essentiellement planctoniques) de la fin du Jurassique et du début du Crétacé, adjacents au Viking Graben. Après migration latérale ou verticale au sein de roches perméables ou de fractures, le kérogène s'est trouvé piégé dans des roches du Trias ou des sables du début du Jurassique d'excellente porosité, comme ceux du groupe Brent[5]. Les pièges rencontrés sont stratigraphiques ou structuraux et sont dus aux phénomènes de rifts induits par la formation du graben.

La figure 5.8 montre l'ensemble des réservoirs d'hydrocarbures qui composent le Viking Graben. On peut en particulier voir qu'ils sont dans l'ensemble situés sur le territoire Norvégien.

Au moyen des données relatives à ce système pétrolier, nous pratiquons maintenant l'estimation des réserves restant à y découvrir.

[4]Il faut noter que certains spécialistes pensent que pour garantir cette homogénéité, il serait nécessaire de descendre à l'échelle du *play* – ou horizon géologique – *i.e.* imposer de plus une cohérence de date de dépôt des sédiments. Le problème du nombre d'observations minimal se pose alors.

[5]Nom d'un champ qui a aussi donné son nom au pétrole brut caractéristique de mer du Nord.

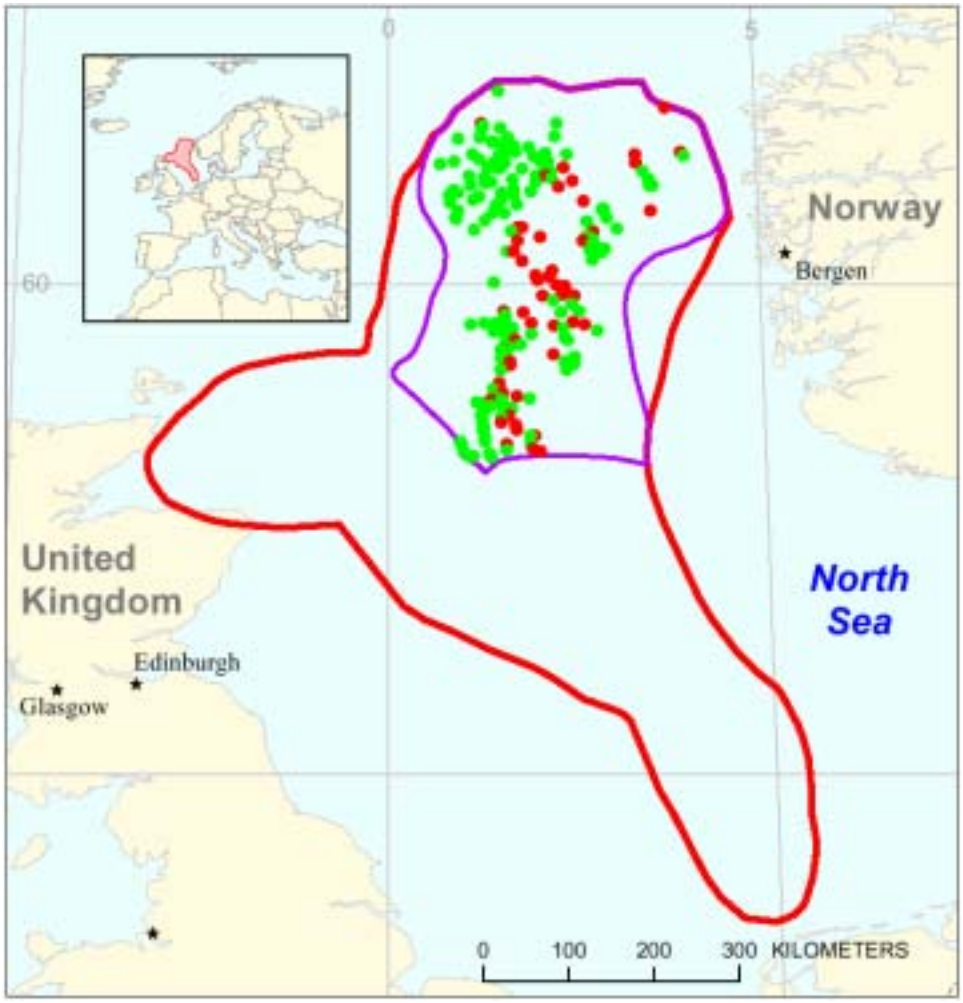

FIG. 5.8 – Carte du Viking Graben de mer du Nord. En rouge sont les champs gaziers, en vert les champs de pétrole.

Estimation

Nous considérons un échantillon de $n = 252$ champs (huile et gaz confondus) de taille supérieure à 1 Mbep. L'ensemble de ces champs représente un montant de réserves prouvées $R = 51$ Gbep. Le diagramme LogLog de cet échantillon se trouve en rouge sur la figure 5.10.

La figure 5.9 résume l'ensemble des résultats de la modélisation et de l'estimation décrite aux chapitres 2, 3 et 4 pour les données précédentes. Il a été choisi une procédure d'estimation :
- par sélection de modèles ;
- sur partitions irrégulières sur blocs statistiques de fréquence d'observation égales ;
- sans contrainte de monotonie ;
- sans spécification préalable de α.

Notons que ce protocole d'estimation est celui qui est choisi par l'algorithme lorsque tous les modèles sont mis en compétition. Il se trouve que, sans l'avoir spécifié dans les contraintes, la fonction $\hat{\omega}$ est croissante, ce qui est conforme à la "philosophie" des chapitres 1 et 2.

Commentaire

Le protocole d'estimation nous montre que le viking Graben de mer du Nord possède un habitat concentré (cf. section 2.1.2) avec $\hat{\alpha} \simeq 0,75$ donc $\hat{h} \simeq 1,33$. Le montant total des réserves estimées est de $\hat{R} = 65,5$ Gbep pour un nombre total de champs $\hat{N} = 2432$.

En détaillant ces résultats, on voit que les champs de taille supérieure à environ 150 Mbep auraient tous été trouvés. Par ailleurs, il resterait 14,5 Gbep de réserves à découvrir, dont plus de la moitié se trouverait disséminée dans pas moins de 2000 champs de taille inférieure à 13 Mbep, c'est-à-dire des champs trop petits pour être la cible de campagnes d'exploration et dont l'exploitation n'a de chance d'être rentable uniquement s'ils se trouvent au voisinage direct de structures préexistantes.

Les 7 autres Gbep restant à découvrir se trouveraient dans environ 170 champs de taille comprise entre 13 et 143 Mbep, soit une moyenne de taille d'un peu plus de 42 Mbep[6]. Il ne faut cependant pas perdre de vue que sur l'intervalle de taille [13 ; 143] la distribution sous-jacente est supposée être exponentielle. C'est à dire que les petits objets (plus proches de 13 que de 143 Mbep) y sont prépondérants. Compte tenu de ce caractère exponentiel, on peut prévoir que seuls une soixantaine devraient dépasser les 42 Mbep et une quinzaine la valeur de 100 Mbep.

[6] Le seuil actuel de taille pour le développement rentable de champs en mer du Nord est d'environ 30 mbep.

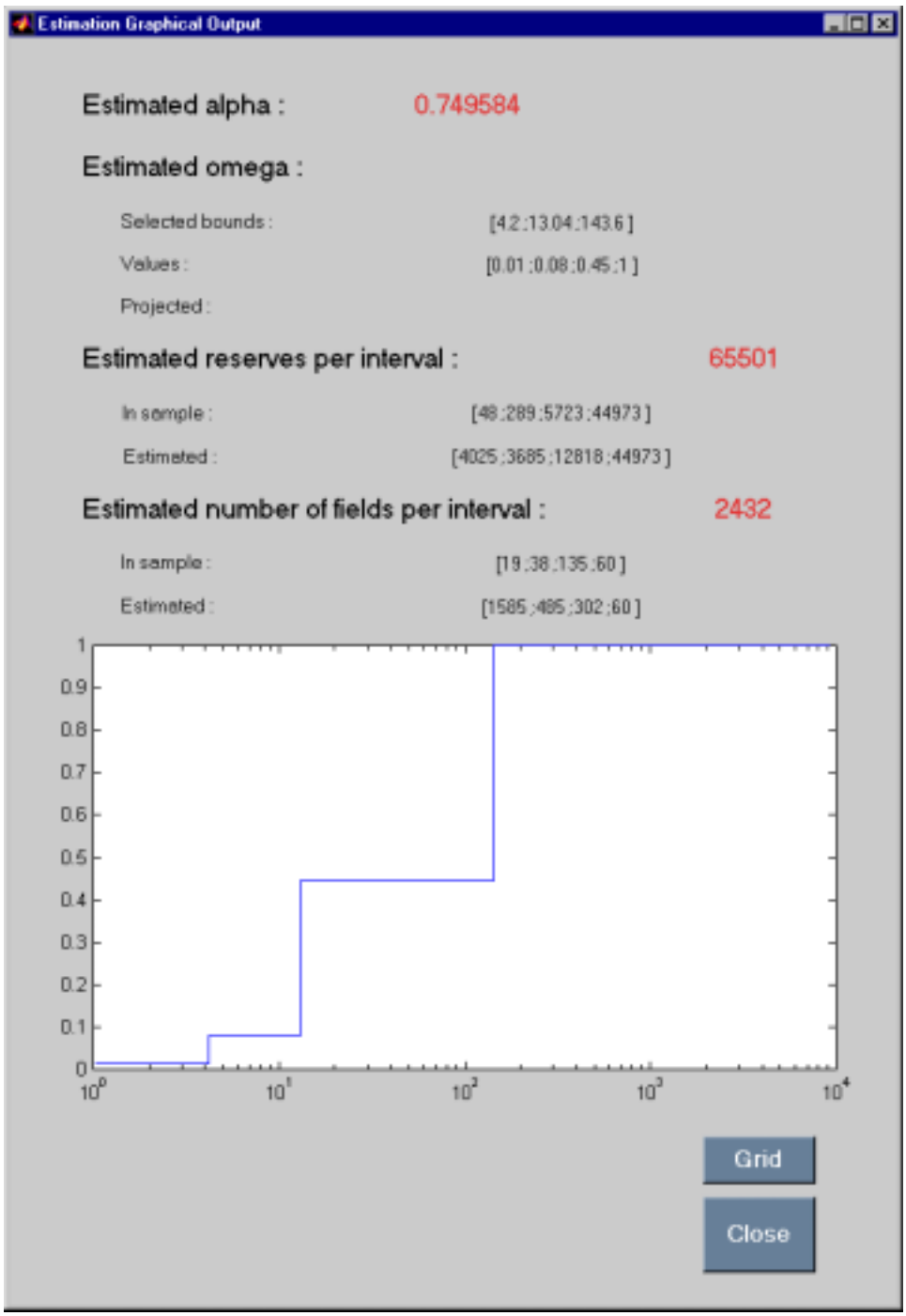

FIG. 5.9 – Sortie graphique du logiciel **select** pour le Viking Graben de mer du Nord (données 1998).

Pour ces derniers, notons que la discussion avec les experts géologues de terrain montre qu'il est plus probable que ces champs soient plutôt des champs gaziers que pétroliers et donc des cibles d'un intérêt stratégique discutable à très court terme.

Terminons cette étude du Viking Graben de mer du Nord par une simulation plausible au vu de nos estimations de ce que pourrait être la distribution ultime des réserves dans cette région (courbe en noir sur la figure 5.10).

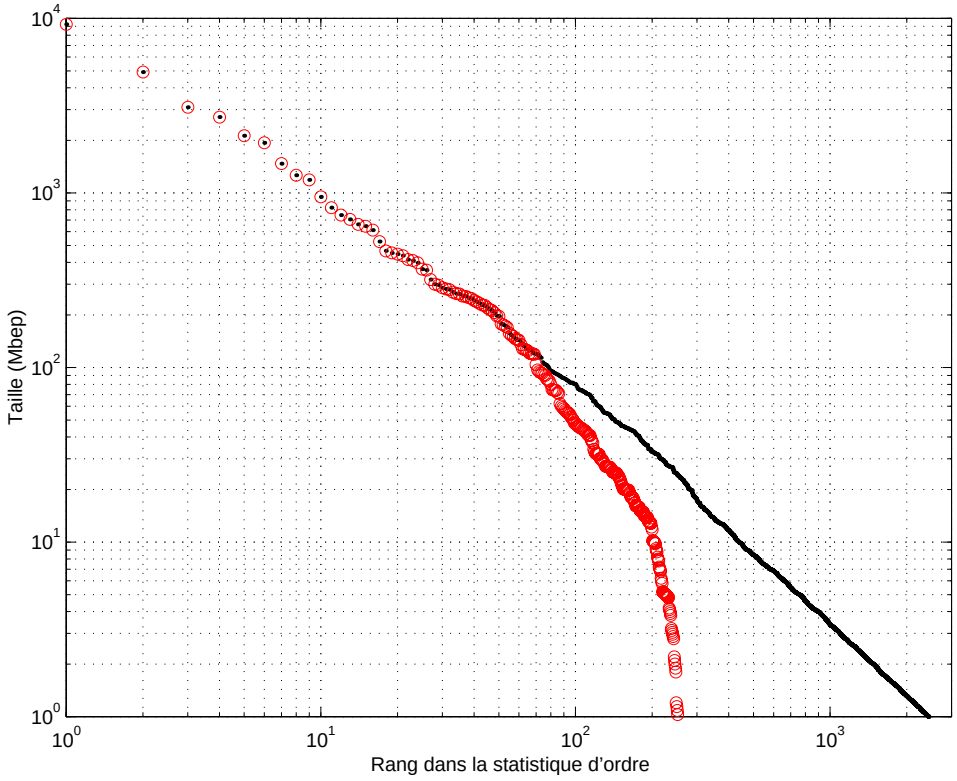

FIG. 5.10 – Diagramme LogLog du Viking Graben de mer du nord (données 1998 en rouge) et sur-échantillon de la loi parente estimée (en noir).

5.3.2 l'offshore du delta du Congo

Présentation

La très grande majorité des champs du delta du fleuve Congo sont des champs offshore qui sont actuellement exploités dans des zones présentant jusqu'à 2000 m de profondeur d'eau (voir figure 5.11). Certaines formations turbiditiques ont été mises en évidence jusqu'à des profondeurs d'eau de 4000 m et restent encore aujourd'hui évidemment inexploitables. Il est considéré (USGS

[80]) que cette région apportera une contribution majeure à l'offre pétrolière des 30 prochaines années.

FIG. 5.11 – Carte de l'offshore du Congo. En rouge sont les champs gaziers, en vert les champs de pétrole.

Les hydrocarbures sont généralement piégés dans des réservoirs de roche sableuse et proviennent d'une roche mère très riche en éléments organiques provenant de dépôts lacustres. De nombreux pièges sont constitué d'un dôme de sel recouvrant la roche réservoir. Ce dôme étant lui-même issu de la fracturation d'une couche de sel produite par les effets de rift de la marge atlantique.

Estimation

Nous considérons un échantillon de $n = 220$ champs (huile et gaz confondus) de taille supérieure à 1 Mbep. L'ensemble de ces champs représente un

montant de réserves prouvées $R = 18,7$ Gbep. Le diagramme LogLog de cet échantillon se trouve en rouge sur la figure 5.13.

La figure 5.12 résume l'ensemble des résultats de la modélisation et de l'estimation décrite aux chapitres 2, 3 et 4 pour les données précédentes. Il a été choisi une procédure d'estimation :
- par sélection de modèles ;
- sur partitions régulières sur intervalles de log-longueur régulière ;
- sans contrainte de monotonie ;
- sans spécification préalable de α.

Cette méthode a été choisie car son résultat est celui sélectionné lorsque l'on met en compétition toutes les méthodes régulières, au sens du critère pénalisé. La même compétition intégrant les méthodes irrégulières sort une estimation manifestement aberrante ($\hat{\alpha}$ proche de 2 et une estimation de réserves associée gigantesque...). Comme pour le Viking Graben, l'estimation de ω est croissante sans que cela ait été spécifié dans les hypothèses.

Commentaire

Il est intéressant de constater, outre les résultats aberrants de certaines des méthodes irrégulières, que les protocoles sélectionnent tous des modèles pour lesquels $\hat{\alpha}$ est compris entre 0,92 et 1,19. Pour la méthode retenue, la valeur $\hat{\alpha} = 1,00$ conduit à un habitat dispersé $\hat{h} = 1/\hat{\alpha} = 1,00$. On retrouve alors ici une caractéristique de dispersion des zones de delta évoquée dans la section 2.1.2.

Les réserves estimées sont $\widehat{R} = 46$ Gbep, soit un montant restant à découvrir de $\widehat{R} - R = 27,4$ Gbep. Comme pour le cas de la mer du nord, on peut supposer que seuls les champs d'une taille supérieure à 20 Mb peuvent constituer des objectifs économiques. A nombre d'environ 200, ils représenteraient environ 10 Gbep. Au vu de ces estimations, il est clair que les objets les plus intéressant sont les 13 champs restant à découvrir dans la zone de 90 à 383 Mbep qui représenteraient environ 2,2 Gbep.

La forte croissance de la courbe représentative de la fonction $\hat{\omega}$ montre que peu d'objectifs de taille importante ont été manqués et que l'exploration a été efficace en se concentrant sur les plus gros objets. Une étude de rétroprévision (cf. 2.4) a tendance à confirmer cette observation.

Enfin, la figure 5.13 montre la diagramme LogLog des données ainsi qu'une simulation possible d'un échantillon de la loi parente d'effectif 5792 et de paramètre $\alpha = 1,00$.

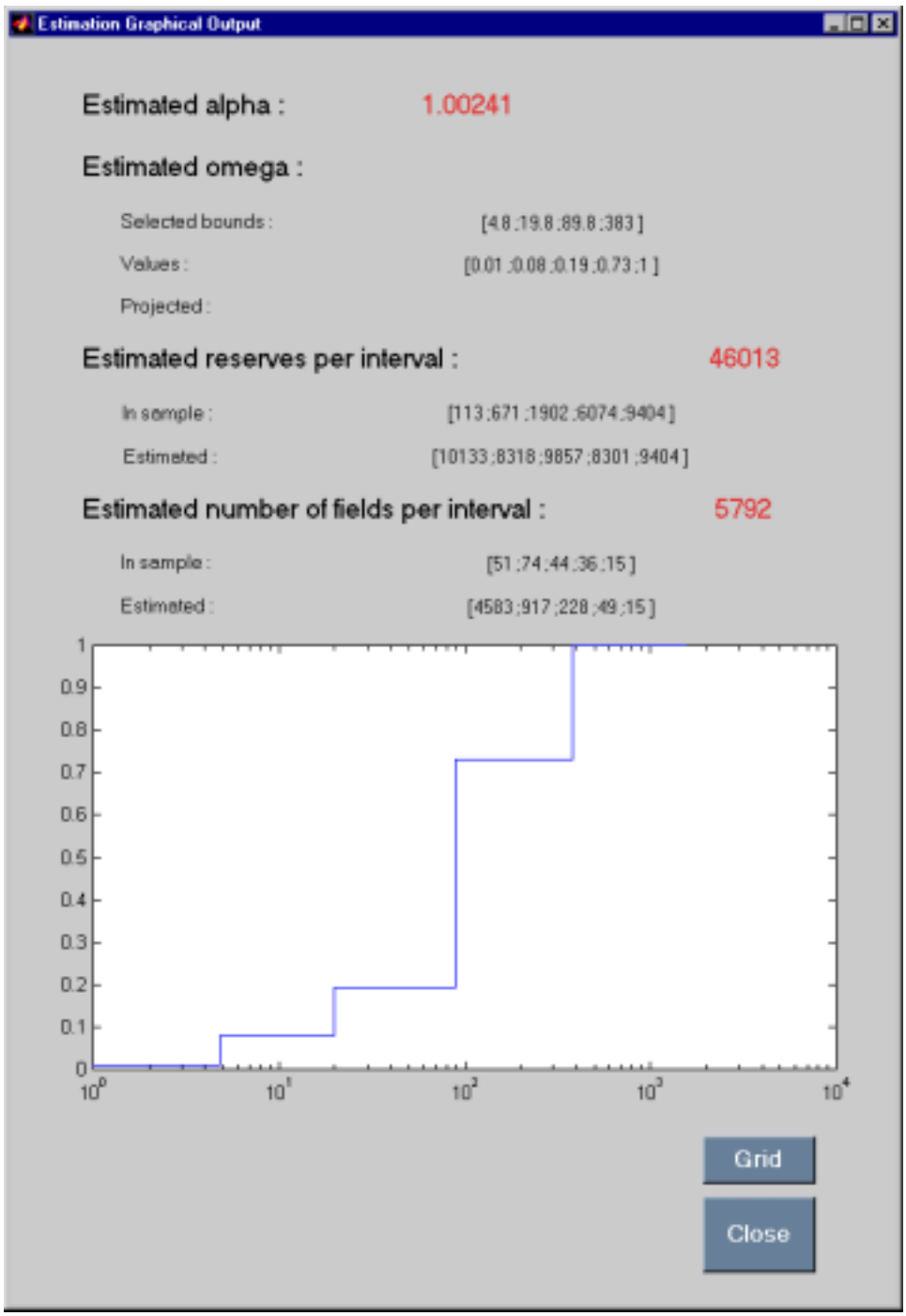

FIG. 5.12 – Sortie graphique du logiciel **select** pour l'offshore du Congo (données 1998).

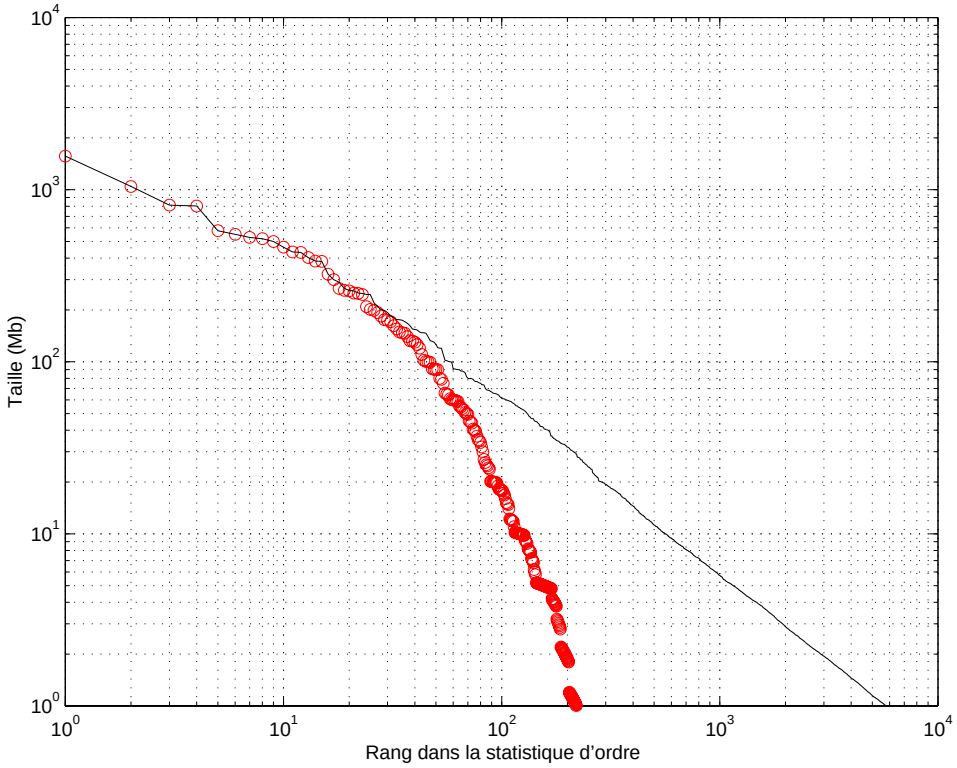

FIG. 5.13 – Diagramme LogLog de l'offshore du Congo (données 1998 en rouge) et sur-échantillon de la loi parente estimée (en noir).

5.3.3 Le système Tamabra du golfe du Mexique

Présentation

Le système pétrolier de l'offshore du golfe du Mexique mexicain représenté sur la figure 5.14 résulte de 100 millions d'années de stabilité d'activité tectonique et de dépôt sédimentaire entre le Jurassique supérieur et le Paléocène. Cette combinaison a fourni d'excellentes conditions de formation d'une roche mère sédimentaire très riche et de carbonates (roches réservoirs) au cours du Crétacé et du Paléocène. La sédimentation Tertiaire qui a suivi a engendré le mouvement des couches de sel qui ont crée les pièges et permis la maturation de la roche mère sous-jacente.

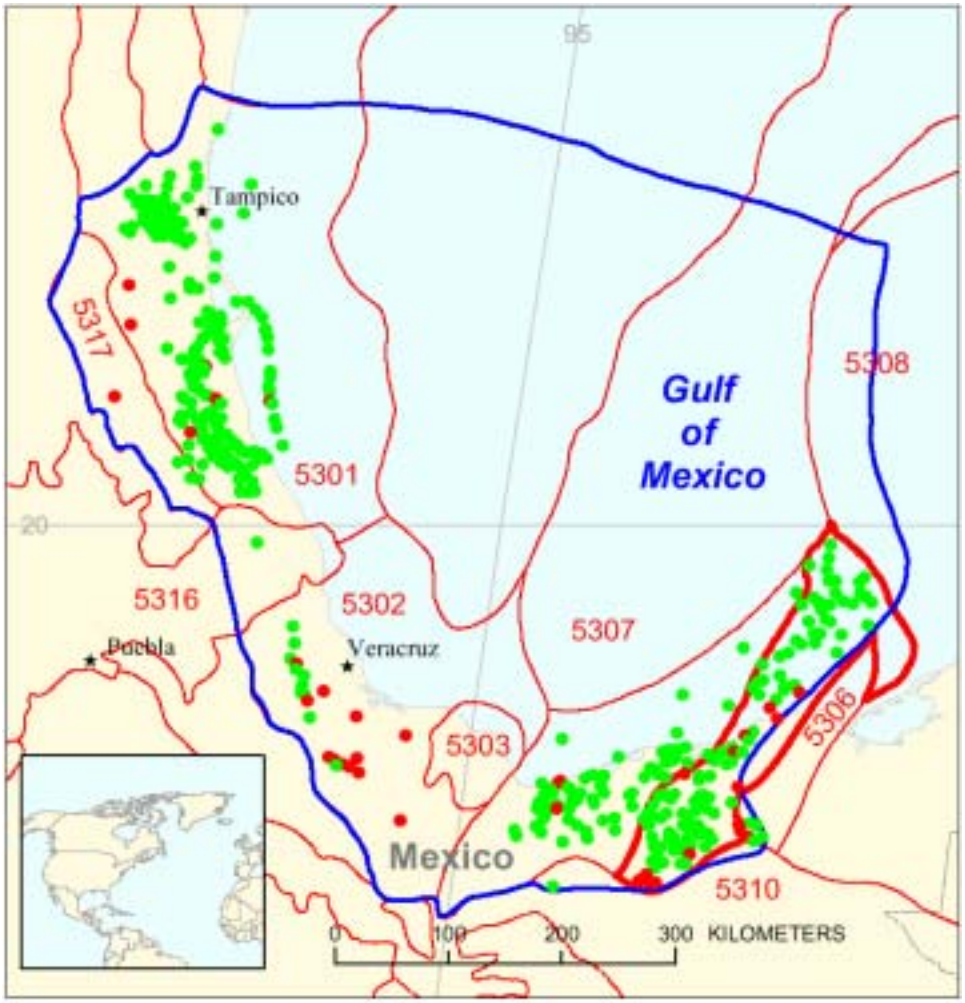

FIG. 5.14 – Carte du système pétrolier de Tamabra (golfe du Mexique mexicain). En rouge sont les champs gaziers, en vert les champs de pétrole.

Estimation

Nous considérons un échantillon de $n = 141$ champs (huile et gaz confondus) de taille supérieure à 1 Mbep. L'ensemble de ces champs représente un montant de réserves prouvées $R = 40,4$ Gbep. Le diagramme LogLog de cet échantillon se trouve en rouge sur la figure 5.16.

La figure 5.15 résume l'ensemble des résultats de la modélisation et de l'estimation décrite aux chapitres 2, 3 et 4 pour les données précédentes. Il a été choisi une procédure d'estimation :
- par sélection de modèles ;
- sur partitions irrégulières sur blocs statistiques de fréquence d'observation régulière ;
- avec contrainte de monotonie ;
- sans spécification préalable de α.

Sur ce jeu de données, deux protocoles rendent une solution sur un modèle de dimension 1, ce qui signifierait qu'il ne resterait plus rien à découvrir, ce qui est évidemment aberrant. Les autres protocoles rendent des valeurs de $\hat{\alpha}$ proches de 0.7 et celui qui tend à réaliser le meilleur score au sens du critère pénalisé est celui dont nous donnons les résultats.

Commentaire

L'habitat de ce système est concentré : $\hat{\alpha} = 0,68$ et $\hat{h} = 1/\hat{\alpha} = 1,46$. La protocole a sélectionné un modèle à 11 dimensions, un nombre proche du nombre maximal autorisé qui est de 14.

Il ressort des résultats de l'estimation qu'il resterait environ 6,4 Gbep à découvrir sur environ 860 champs. A noter que sur les champs restant à découvrir :
- 1 champ aurait une taille de 450 Mb ;
- 3 champs auraient une taille comprise entre 100 et 200 Mbep ;
- 9 champs auraient une taille comprise entre 45 et 100 Mbep ;
- 12 champs auraient une taille comprise entre 30 et 45 Mbep ;
- 21 champs auraient une taille comprise entre 20 et 30 Mbep.

les autres champs étant considérés comme trop petit pour constituer des objectifs économiques.

Le grand nombre de morceaux dans la partition fournit ici une impression de précision quant aux résultats. Il faut pourtant se souvenir que le nombre d'individus présents dans chaque classe est relativement faible, de l'ordre de 12, et qu'en conséquence les estimations des probabilités d'inclusion ont probablement une variabilité importante.

Sur la figure 5.16 nous représentons le diagramme LogLog des données du système pétrolier ainsi qu'une simulation d'une possible population parente.

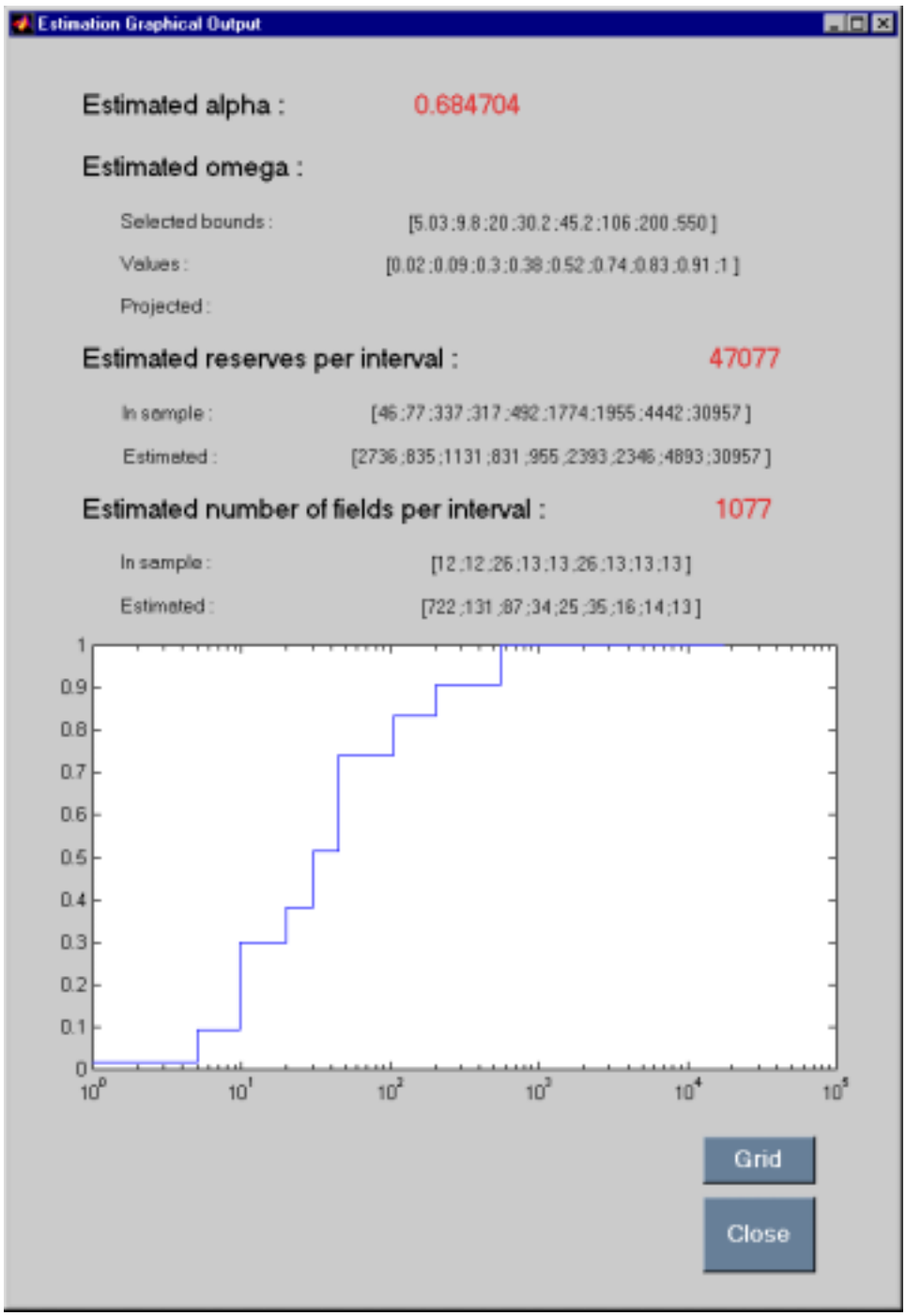

FIG. 5.15 – Sortie graphique du logiciel **select** pour le système pétrolier de Tamabra (données 1998).

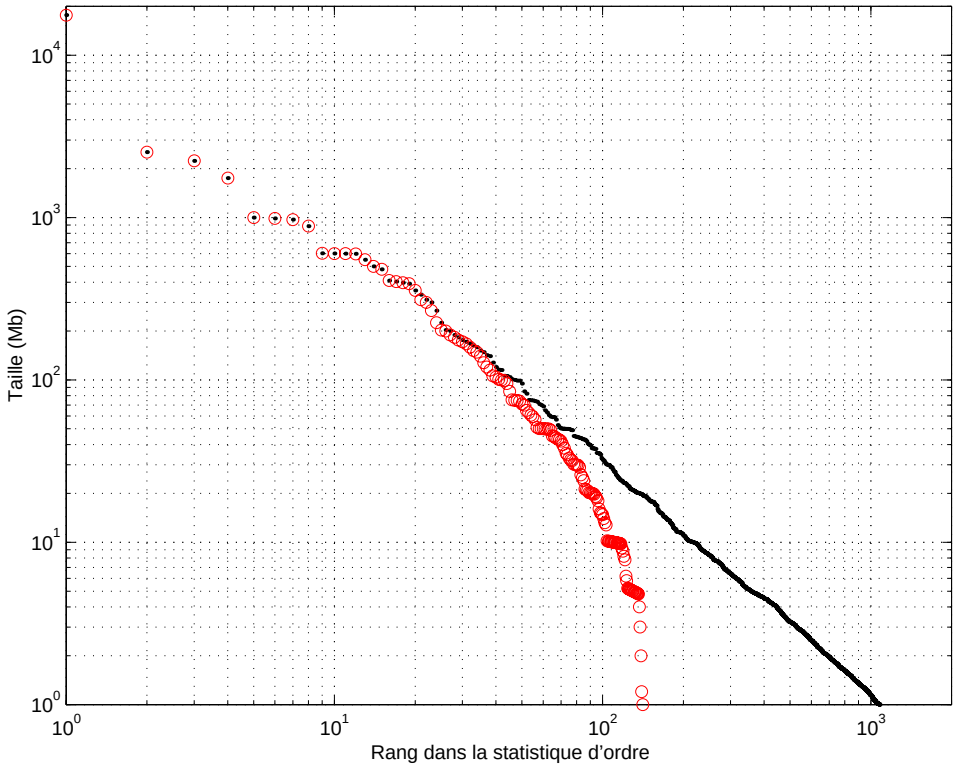

FIG. 5.16 – Diagramme LogLog du système pétrolier de Tamabra (données 1998 en rouge) et et sur-échantillon de la loi parente estimée (en noir).

5.3.4 Un exemple volontairement non optimal : le système de Bazhenov en Russie

Le dernier exemple que nous traitons est volontairement non optimal, dans le sens où la méthode que nous présentons n'est pas la plus performante au sens du critère pénalisé sur plusieurs protocoles. Nous souhaitons montrer ici un exemple d'estimation par sélection de modèle sur données réelles dont le résultat de l'estimation de ω n'est pas une fonction monotone.

Présentation

La roche mère de ce système pétrolier Sibérien (voir figure 5.17) est essentiellement constituée de schistes continentaux datés du Jurassique moyen et, pour quelques représentants, du Trias. Les roches réservoirs sont des sables de silice du Jurassique et du Crétacé. On y trouve aussi quelques carbonates. Les pièges par dôme de sel sont assez rares dans cette région. Les épais schistes du crétacé et du Trias qui couvrent les réservoirs fournissent des pièges suffisamment efficaces pour préserver les hydrocarbures.

Estimation

Nous considérons un échantillon de $n = 542$ champs (huile et gaz confondus) de taille supérieure à 1 Mbep. L'ensemble de ces champs représente un montant de réserves prouvées $R = 120,6$ Gbep. Le diagramme LogLog de cet échantillon se trouve en rouge sur la figure 5.20.

La figure 5.18 résume l'ensemble des résultats de la modélisation et de l'estimation décrite aux chapitres 2, 3 et 4 pour les données précédentes. Pour cet exemple particulier, il a été choisi une procédure d'estimation :
- par sélection de modèles ;
- sur partitions régulières sur blocs statistiques de fréquence d'observation égales ;
- sans contrainte de monotonie ;
- sans spécification préalable de α.

Il faut noter que le meilleur modèle au sens du critère pénalisé est celui qui correspond à une partition régulière en fréquence avec contrainte de monotonie, donné sur la figure 5.19. La dimension du modèle est alors de 13 et l'estimation générale bien plus optimiste que ce qui est présenté ici.

Commentaire

Le premier commentaire à fournir est celui de l'allure globalement croissante de l'estimation de ω. Celle-ci montre (tout comme le cas du Viking Graben et du Congo) que l'hypothèse de monotonie que nous avons faite tout au long de la thèse semble une fois de plus justifiée.

FIG. 5.17 – Carte du système pétrolier de Bazhenov (Russie). En rouge sont les champs gaziers, en vert les champs de pétrole.

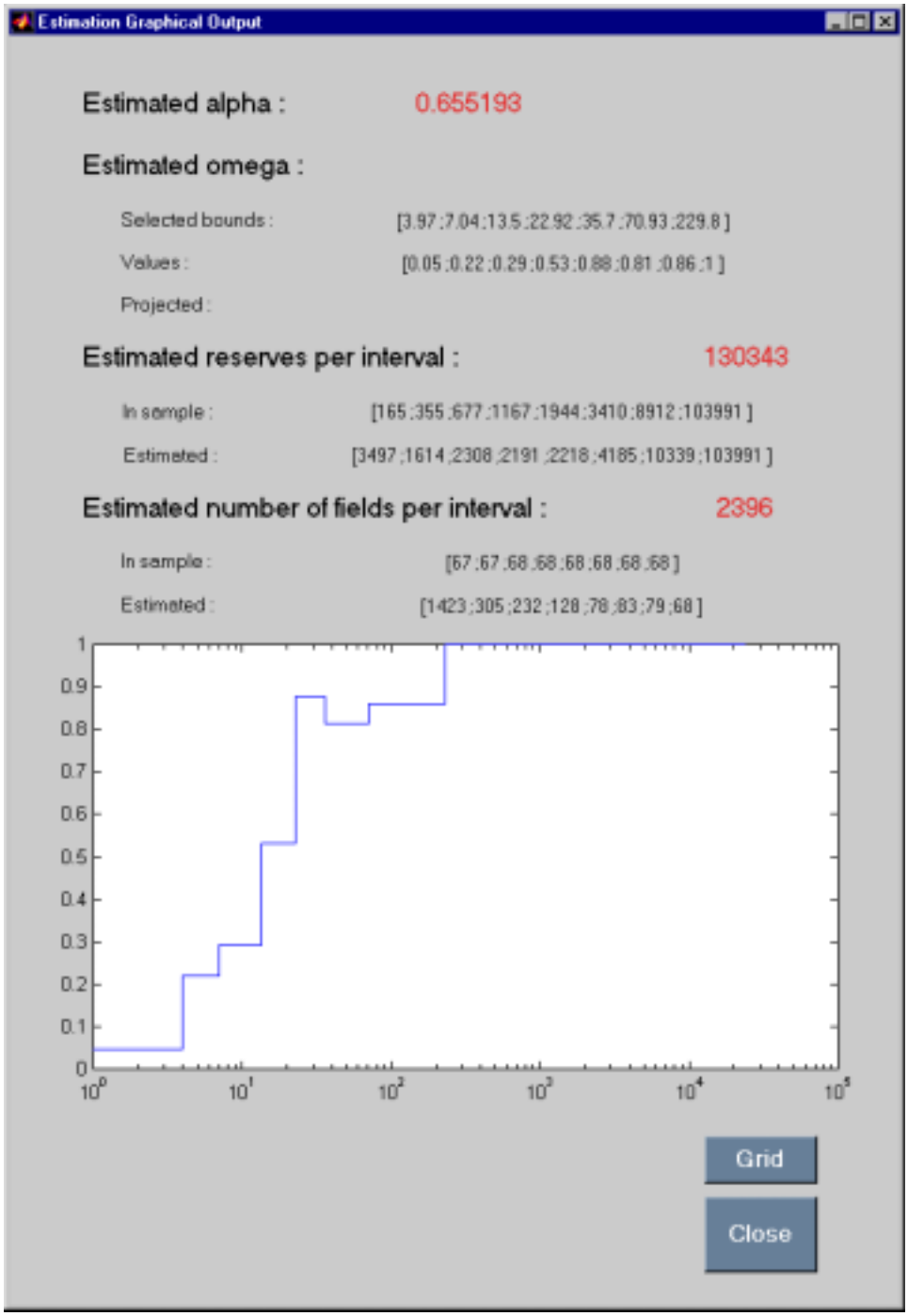

FIG. 5.18 – Sortie graphique du logiciel **select** pour le système pétrolier de Bazhenov (non "optimale", données 1998).

Nous ne commentons pas ici les résultats détaillés des estimations car ils ne sont pas les résultats optimaux au sens du critère pénalisé. Nous voyons cependant que l'habitat est concentré sur cette zone, ce qui est confirmé par l'estimation sous contrainte de la sortie graphique 5.19.

Nous notons sur les deux sorties graphiques des figures 5.18 et 5.19 que la variabilité sur $\hat{\alpha}$ (environ 20 % d'un cas à l'autre) a un impact considérable sur les estimations de réserves et du nombre total de champs. Il faut cependant remarquer que l'essentiel de la différence constatée se situe sur les classes de très petite taille, c'est à dire moins de 20 Mbep. En effet, sur cette classe on trouve un écart de réserves restant à découvrir d'environ 11 Gbep, qui explique plus de la moitié de l'écart constaté entre les deux estimations de réserves et 90 % de l'écart sur le nombre total de champs. Une erreur d'estimation sur α a donc un impact bien plus important sur les estimations relatives aux petits objets qu'aux objets de taille importante.

La figure 5.20 montre une simulation de loi parente dans le cas non optimal. Si cette estimation était proche de la réalité, nous voyons alors très clairement sur ce graphe que nous serions très proches de l'exhaustion.

5.3.5 Conclusion générale sur les estimations

Retenons ici que sans un expert de terrain, le statisticien ne peut rien.

En amont du traitement statistique, même si nous avons vu que les estimations nous confortent parfois dans le choix des hypothèses, notamment celle de la monotonie des probabilités d'inclusion, il nous est impossible de tirer des conclusions pertinentes si les échantillons de départ ne sont pas correctement spécifiés.

Une fois les résultats statistiques obtenus, Le dernier exemple du bassin de Bazhenov nous montre que l'avis de l'expert géologue et l'information *a priori* qu'il peut apporter sont essentielles pour valider un modèle. Les différences d'estimation peuvent en effet être considérables d'un protocole à l'autre sans que l'un soit spécialement plus plausible que l'autre pour le seul statisticien.

Mentionnons aussi pour conclure qu'il est impossible de valider la méthode en dehors des simulations et de rétroprévisions non contradictoires. En effet, aucun bassin sédimentaire du globe ne peut aujourd'hui être considéré comme totalement mature, c'est-à-dire considéré comme un bassin où *tous* les gisements d'hydrocarbures auraient été découverts. Les rétroprévisions peuvent aider à la validation d'un modèle en donnant des informations sur sa stabilité au cours du temps.

Nous terminons par un exemple qui illustre à ce dernier point. La figure 5.21 représente les résultats de l'estimation de la densité des observations du Viking Graben de mer du Nord avec les données de 1985 (170 champs

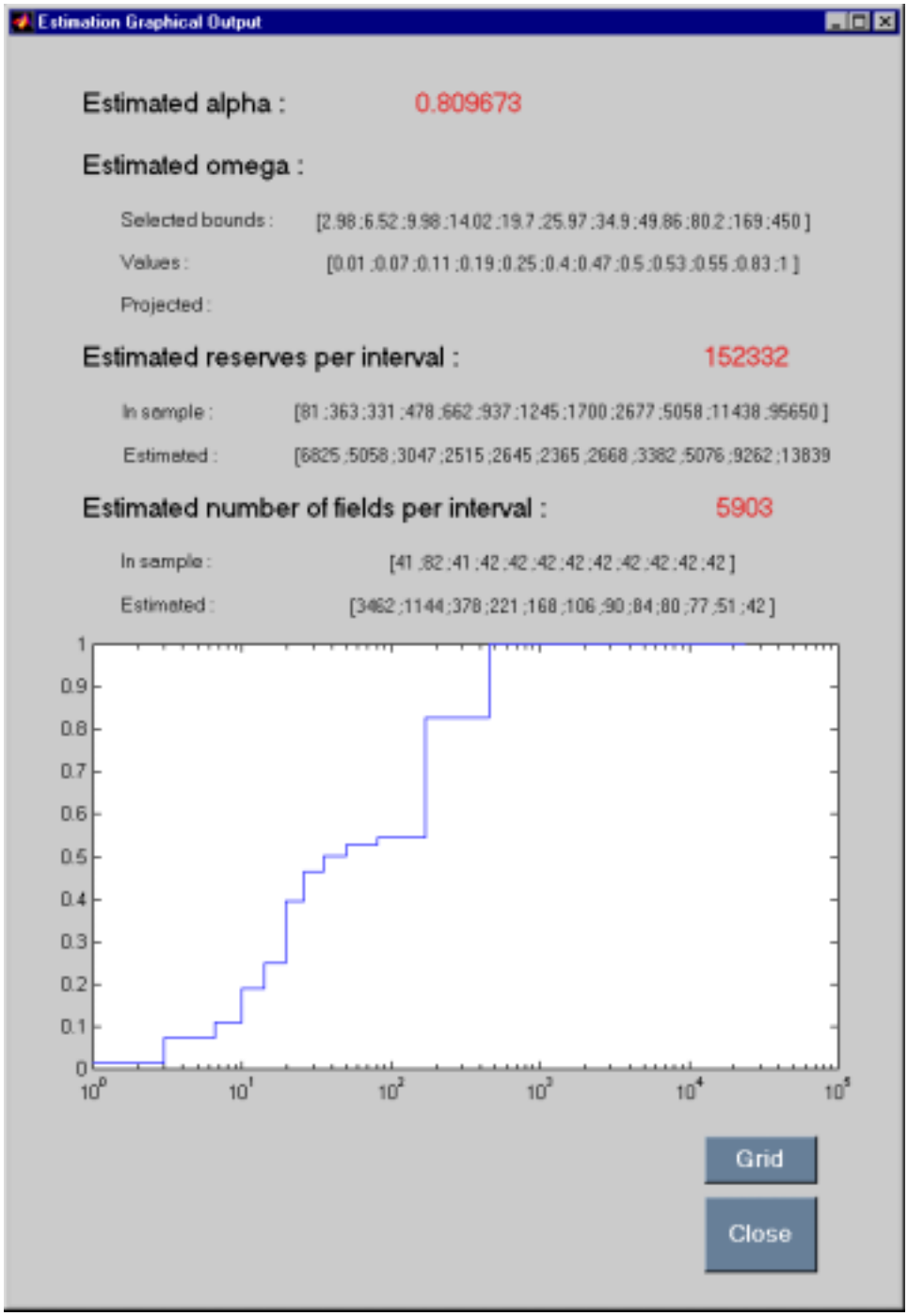

FIG. 5.19 – Sortie graphique du logiciel **select** pour le système pétrolier de Bazhenov ("optimale" données 1998).

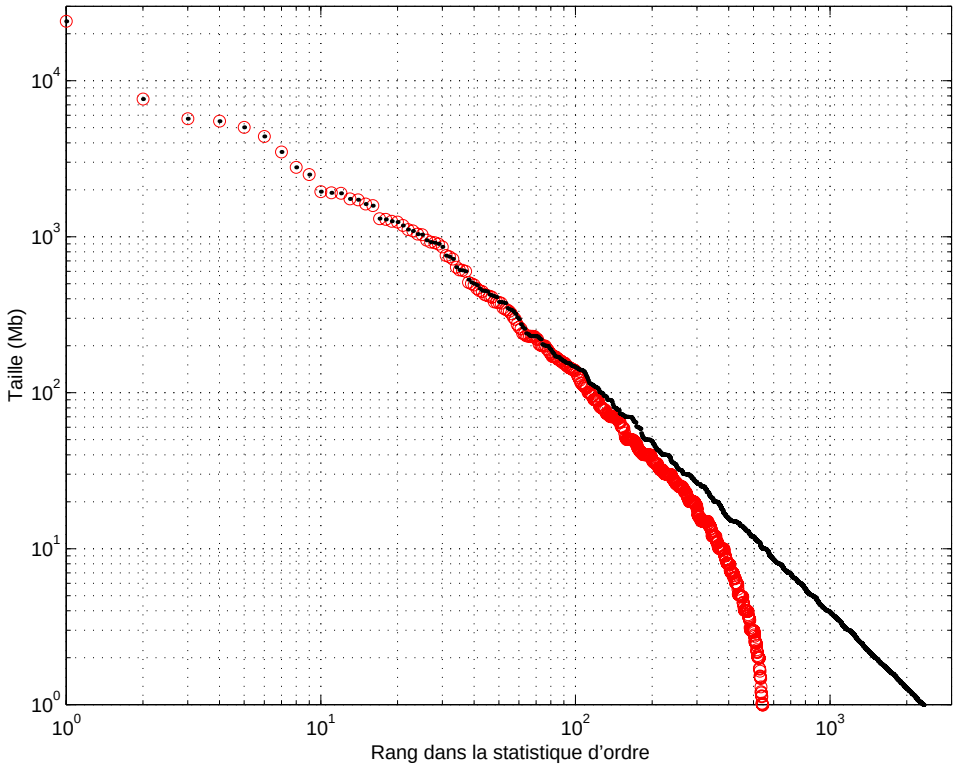

FIG. 5.20 – Diagramme LogLog du système pétrolier de Bazhenov (données 1998 en rouge) et sur-échantillon de la loi parente estimée (en noir).

contre 252 en 1998, voir la section 5.3.1). Le modèle présenté est celui qui minimise le critère pénalisé sur tous les modèles possibles. On peut voir que l'estimation de α sur ces données est très proche de celle de 1998 (environ 0,75 pour les deux) et que le modèle de 1985 prévoit que l'on découvre encore un champ parmi les plus grandes observations de l'époque (c'est-à-dire les champs de taille supérieure à 410 Mbep), ce qui a exactement été vérifié depuis.

Pourtant, si nous avions pratiqué nos estimations en 1985, nous aurions probablement imposé la contrainte de monotonie qui aurait alors masqué ce dernier grand champ restant à découvrir...

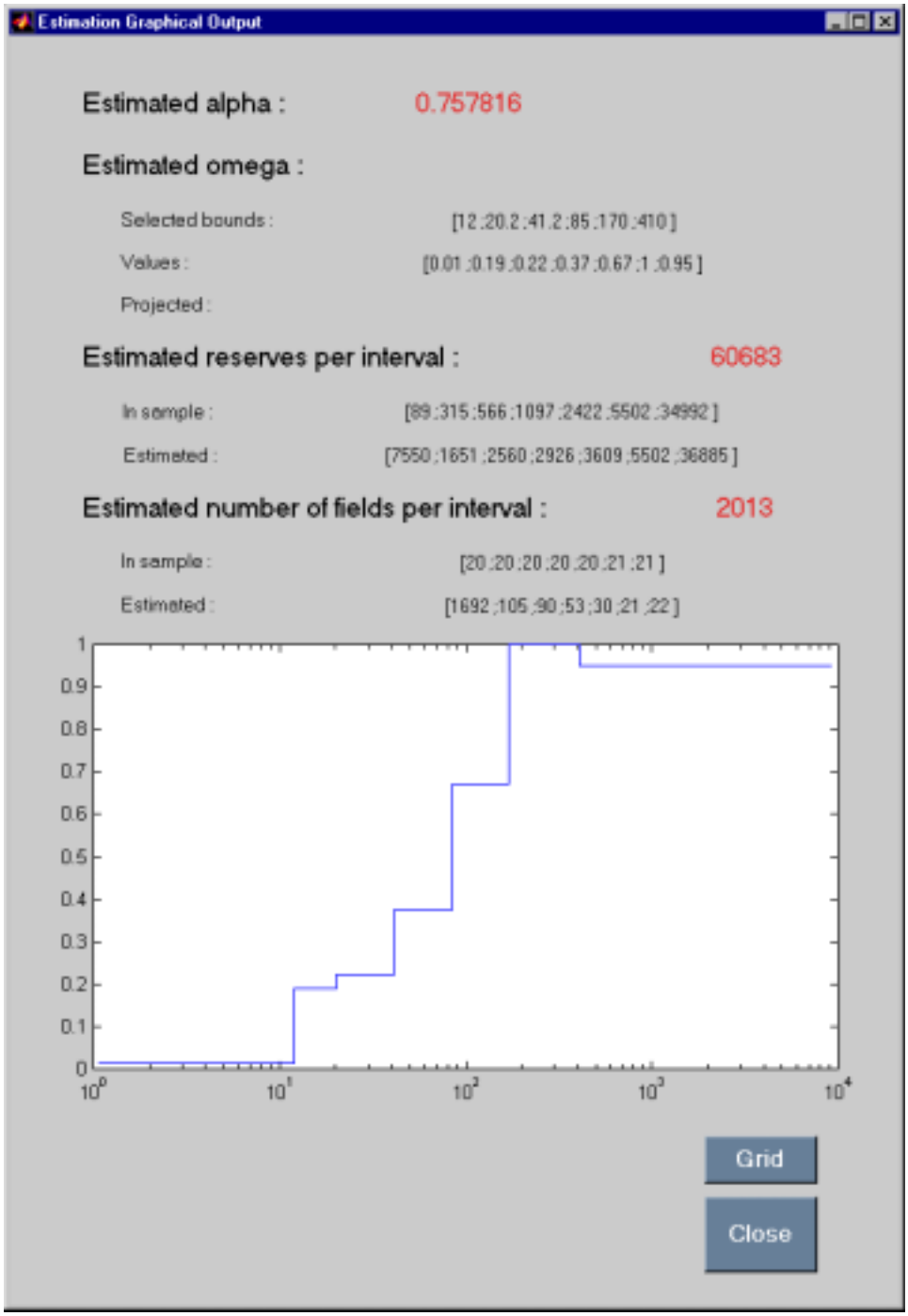

FIG. 5.21 – Sortie graphique du logiciel **select** pour le Viking Graben de mer du Nord (données 1985).

Chapitre 6

Annexe 2 : quelques codes source

L'ensemble des codes reproduits ci-dessous sont rédigés en langage Matlab® version 5.3. Excepté le premier d'entre-eux, et bien qu'ils ne représentent qu'un très petit volume de l'ensemble des codes nécessaires au fonctionnement du logiciel **select**, ils sont au coeur de celui-ci et illustrent le travail d'estimation correspondant à la description théorique des chapitres 3 et 4 en termes d'algorithmique.

Lorsque ces codes font appel à des échantillons de données, il s'agit partout d'échantillons de logtailles de champs

6.1 Calcul de probabilités d'inclusion

Le programme `probincl` qui suit permet de calculer les probabilités d'inclusion au $k^{\text{ème}}$ tirage sans remise successif dans un échantillon de taille N associé à des probabilités d'inclusion ω au premier tirage. Ce programme est inspiré de la formule de récurrence donnée par la Proposition 2.3.1.

```
function y = probincl(omega,i,long)

%
% PROBINCL calcule la probabilité d'inclusion d'une unité
% d'indice 'i' dans un échantillon de taille 'p' d'un
% omega-ordonné-PPSWOR (Proportional to Size Sampling
% WithOut Replacement).
%

global p n oi
```

```
p = long;
n = length(omega);
oi = omega(i);

y = oi/sum(omega);

y = y*parsum(omega([1:i-1 i+1:n]));

%
% Sous-Fonction récursive PARSUM
%

function t = parsum(o)

global p n oi

m = length(o);

t = 1;

if m > n-p
   U = sum(o)+oi;
   for k = 1:m
      t = t + o(k)/(U-o(k))*parsum(o([1:k-1 k+1:m]));
   end
end
```

6.2 Simulation de la loi parente

Le code `pialpha` suivant permet de créer un ou plusieurs échantillons d'une loi de Lévy-Pareto au moyen de sa fonction de queue de répartition (voir section 2.1).

```
function A = pialpha(alpha,n,m)

%
% PIALPHA cree un ou plusieurs echantillons de la loi de Pareto
% Elle prend comme argument l'exposant de Pareto 'alpha',
% la taille des echantillons 'n' et leur nombre 'm'.
% Par defaut, si 'm' n'est pas specifie alors la fonction
```

```
% ne cree qu'un echantillon. Dans le cas contraire, les
% echantillons sont donnes sous forme de colonnes d'une matrice (n,m).
%

if nargin == 2, m = 1; end

A = rand(n,m).^(-1/alpha);
```

6.3 Simulation du sous-échantillonnage

6.3.1 Tirage successif

Le code `wsample`[1] suivant permet de créer un ou plusieurs sous-échantillons de taille donnée d'une liste de valeurs par tirage successif biaisé par effet taille selon une pondération elle aussi donnée.

```
function V = wsample(X,W,m)

%
% WSAMPLE tire sans remise dans proportionnellement
% à une mesure de taille. Elle prend pour argument
% le vecteur base, le vecteur des poids (de meme
% taille imperativement), ainsi que la taille du
% sous echantillons voulu.
% Elle renvoie la matrice des sous échantillons
% generes.
%

n = length(X);
X = reshape(X,n,1);
W = reshape(W,n,1);

for k = 1:m

   a = min(W);
   l = 0;
   j = a + rand*(sum(W)-a);

   while j >= a
```

[1] La commande `sortd` qui apparaît dans ce code est un utilitaire de tri de vecteur par valeur décroissante.

```
      l = l+1;
      j = j - W(l);
   end

   if k == 1
      V(1,1) = X(l);
   else
      V(1:k,k) = sortd([V(1:k-1,k-1);X(l)]);
   end

   if l == n
      X = [X(1:n-1)];
      W = [W(1:n-1)];
   else
      X = [X(1:l-1);X(l+1:n)];
      W = [W(1:l-1);W(l+1:n)];
   end

   n = n-1;

end
```

6.3.2 Tirage global

Le code `wempirsample`[2] suivant permet de créer un sous-échantillon d'une liste de valeurs par tirage direct dans des classes de tailles données, au moyen de probabilités d'inclusion elles aussi données. La taille du sous-échantillon est alors une variable aléatoire, comme dans la remarque de la définition 2.3.2.

```
function Y = wempirsample(X,omega,bornes)

%
% WEMPIRSAMPLE tire la proportion omega sans remise
% d'individus dans 'X' entre des bornes internes en [[.
%
% Y = wempirsample(X,bornes,omega). 'X' doit être trié
% décroissant, 'bornes' trié croissant et 'omega' correspondre
% à l'ordonnancement de bornes.
```

[2] La commande `sortcle` qui apparaît dans ce code est un utilitaire qui permet de trier les lignes d'une matrice en utilisant comme clé une colonne spécifiée de cette matrice.

```
%

n = length(X);
p = length(bornes);

eps = min(abs(X(1:n-1)-X(2:n)))/10;

bornes = [X(n) ; bornes ; X(1)+2*eps];

k = 1;

for i = p+1:-1:1
   Z = X(find((X>=(bornes(i)-eps)) & (X<(bornes(i+1)-eps))));
   Z = sortcle([Z rand(length(Z),1)],2);
   q = round(omega(i)*length(Z));
   Y(k:k+q-1,1) = sortd(Z(1:q,1));
   k = k+q;
end
```

6.4 Estimation dans un modèle

6.4.1 Partition de l'intervalle d'étude

Le code `closeto` qui suit permet notamment de construire des partitions de l'intervalle d'étude basées sur les données à partir de partitions régulières. Il peut aussi utilisé pour construire des sur-échantillons des données en vue de l'établissement d'intervalles de confiance (voir section 5.2.3).

```
function Y = closeto(X,bornes)

%
% CLOSETO permet de sortir le vecteur issu d'un vecteur
% "bornes" le plus proche mais basé sur les statistiques
% d'ordre d'un vecteur 'X'.
%
% Y = closeto(X,bornes), où 'bornes' est le vecteur des
% bornes internes en [[. Attention, CLOSETO garde le même
% ordonnancement que celui de 'bornes'.
%

bornes = bornes(find((bornes >= min(X)) & (bornes < max(X))));
```

```
Y(1)   = min(X(find(X>=bornes(1))));
k      = 1;

for i = 2:length(bornes)
   T = min(X(find(X>=bornes(i))));
   if T ~= Y(k)
      k    = k + 1;
      Y(k) = T;
   end
end
```

6.4.2 Régression isotonique

Calculer la régression isotonique d'une fonction constante par morceaux associée à une famille de poids nécessite en premier lieu de disposer d'un outil qui permet de construire la minorante convexe d'un nuage de points. C'est le rôle du code `convexhull`[3].

```
function Z = convexhull(X)

%
% CONVEXHULL rend la minorante convexe d'une famille de
% points ordonnée en abscisse.
%
% CONVEXHULL prend pour argument la matrice des coordonnées,
% 1ère colonne absisces, 2ème colonne ordonnées et 3eme
% colonne poids des n points, où n est le nombre de
% lignes. Elle rend une matrice du même type contenant
% les coordonnées des points de la minorante convexe.
%

X = sortcle(X,1);
[Y,Z] = epure(X(:,1));
Y(:,2) = X(Z,2);

Z = Y(1,:);
n = length(Y);
i = 1;
```

[3] Ce code utilise un utilitaire appelé `epure` qui élimine les doublons d'un vecteur de données.

```
j = 2;

while i < n
   B = (Y(i+1:n,2)-Y(i,2))./(Y(i+1:n,1)-Y(i,1));
   i = i + max(find(B==min(B)));
   Z(j,:) = Y(i,:);
   j = j + 1;
end
```

La régression isotonique d'une fonction constante par morceaux associée à une famille de poids définis pour chaque morceaux est la dérivée de la minorante convexe de la fonction qui à chaque saut de la fonction de départ associe la somme cumulée des hauteurs des sauts pondérée par les poids (cf. chapitre 3 ou [72] pour les détails). Le code `isotonic` qui suit met en œuvre cette idée.

```
function Z = isotonic(X);

%
% ISOTONIC renvoie la regression isotonique d'un
% ensemble de points.
%
% Elle prend et renvoie le même genre d'arguments
% que CONVEXHULL
%

[n,p] = size(X);

if p == 2
   X(:,3)=ones(n,1);
end

n = n(1); Z = cumsum(X(:,2).*X(:,3)); H = cumsum(X(:,3));

Z = convexhull([H Z]); n = length(Z);

Z(2:n,2) = (Z(2:n,2)-Z(1:n-1,2))./(Z(2:n,1)-Z(1:n-1,1));

Z(1,2) = X(1,2) ;

for i = 1:n
```

```
   Z(i,1) = X(min(find(H==Z(i,1))),1);
end
```

6.4.3 Estimation sous contraintes

La fonction `vraisemblanceC`[4] qui suit rend la solution du programme de maximisation de la vraisemblance en α et ω dans un modèle fixé, sous contrainte de monotonie. Le code suit la preuve de la proposition 3.2.8 en distinguant le cas où le modèle est de dimension 1 des dimensions plus grandes. Ce code est divisé en deux parties. La première a pour rôle d'effectuer la dichotomie en α sur la fonction φ'_* qui intervient dans la preuve de la Proposition 3.2.8 ; la seconde évalue la fonction φ'_*, tout les paramètres étant fixés, au moyen de la régression isotonique des $(\omega_I)_{I\in m}$.

À noter qu'il existe quatre versions de ce code suivant le type d'estimation que l'utilisateur souhaite mettre en œuvre :
– la version non contrainte en ω à α connu ;
– la version contrainte en ω à α connu ;
– la version non contrainte en ω à α inconnu ;
– la version contrainte en ω à α inconnue (présentée ci-dessous).

```
function [alpha,omega, isobornes, loglik] = ...
                                  vraisemblanceC(X,bornes)

n      = length(X);
xbar   = mean(X);
eps    = 10^-5;

if isempty(bornes)

   Pn     = 1;
   bornes = [0 ; inf];
   m      = length(bornes);

%
% Calibrage du point de départ de la dichotomie en alpha
%
```

[4] Cette fonction utilise un utilitaire `saucissonne` qui permet de savoir combien d'éléments d'un vecteur se trouvent entre des bornes spécifiées à l'avance.

```
   alpha = 1;
   while phiprim(alpha,bornes,m,Pn,xbar) > 0
      alpha = alpha + 1;
   end

%
% Dichotomie en alpha
%

   alph = alpha - 1;
   while alpha-alph > eps
      alp = (alpha+alph)/2;
      if phiprim(alp,bornes,m,Pn,xbar) > 0
         alph  = alp;
      else
         alpha = alp;
      end
   end

%
% Renvoi d'alpha, omega, log-vraisemblance
%
   alpha  = (alpha+alph)/2;
   omega  = 1;
   loglik = alpha*xbar - log(alpha);
   isobornes = [];

else

   Pn     = (saucissonne(X,bornes))'/n;
   bornes = [0 ; bornes ; inf];
   m      = length(bornes);

%
% Calibrage du point de départ de la dichotomie en alpha
%

   alpha = 1;
   while phiprim(alpha,bornes,m,Pn,xbar) > 0
      alpha = alpha + 1;
   end

%
% Dichotomie en alpha
```

```
%

   alph = alpha - 1;
   while alpha-alph > eps
      alp = (alpha+alph)/2;
      if phiprim(alp,bornes,m,Pn,xbar) > 0
         alph  = alp;
      else
         alpha = alp;
      end
   end

%
% Renvoi d'alpha, omega, log-vraisemblance
%

   alpha  = (alpha+alph)/2;
   ex        = exp(-alpha*bornes);
   ex        = ex(1:m-1)-ex(2:m);
   omega     = Pn./ex;
   T         = isotonic([bornes [0;omega] [0;ex]]);
   omega     = T(2:length(T),2);
   omega     = omega/max(omega);

   if isempty(T(2:length(T)-1,1))
      loglik = alpha*xbar - log(alpha);
      isobornes = [];
   else
      tex    = exp(-alpha*T(1:length(T)-1,1)) - ...
                        exp(-alpha*T(2:length(T),1));
      loglik = alpha*xbar - log(alpha) - sum(log(omega).* ...
                (saucissonne(X,T(2:length(T)-1,1)))'/n) ...
                + log(sum(omega.*tex));
      isobornes = T(2:length(T)-1,1);
   end
end

%
% Fonction intermédiaire PHIPRIM
%

function y = phiprim(alpha,bornes,m,Pn,xbar)

ex     = exp(-alpha*bornes);
```

```
ex     = ex(1:m-1)-ex(2:m);

omega  = Pn./ex;
T      = isotonic([bornes [0;omega] [0;ex]]);
dex    = [T(1:length(T(:,1))-1,1);0].*exp(-alpha*T(:,1));

k      = length(dex);

y      = 1/alpha - xbar + sum(T(2:k,2).*(dex(1:k-1)-dex(2:k)));
```

6.5 Sélection de modèles

Le code source de la fonction `modelselectCirreg`[5] qui suit donne les boucles nécessaires à l'évaluation des critères pénalisés sur l'ensemble des modèles basés sur une partition régulière fixée.

Cette fonction prend pour argument les données ainsi que la constante de calibration devant la fonction de pénalité (voir la section 4.4.4). Pour des raisons de temps de calcul, un maximum de 13 morceaux est admit pour la partition fine régulière sous-jacente aux partitions régulières visitées.

À noter que le protocole de sélection de modèles ci-dessous (ainsi que le code nécessaire à la calibration de la constante par la méthode de la pente `penteCirreg` – voir la section 4.4.4 – aussi donné ci-après) existe en 16 versions différentes suivant que les modèles choisis sont :
– avec sou sans contraintes en ω ;
– supposés à α connu ou inconnu ;
– basés sur des partitions homogènes en longueur ou en fréquence ;
– réguliers ou irréguliers.

```
function [alpha, omega, isobornes, crit] = ...
   modelselectCirreg(X,theta)

%
% MODELSELECTCIRREG est un protocole complet de selection
% de modeles sur un vecteur colonne ordonné décroissant
% pareto w-sous-échantillonne, sur un ensemble de partitions
% irregulieres en longueurs. Omega-contraint croissant.
```

[5] Ce code utilise un utilitaire `concat` nécessaire à la création des sur-partitions basées sur une partition régulière initiale.

```
%
% 'X' est le vecteur des données ordonné décroissant,
% 'theta' est une constante de calibration de la pénalité.
%
% 'alpha' est l'estimation finale de alpha
% 'omega' est l'estimation finale de omega
% 'isobornes' est le vecteur des sauts de omega
% 'crit' est la valeur de critère pénalisé.
%

n = length(X);
p = min(floor(sqrt(n)),12);
k = ceil(n/(log(n))^2);

[alpha, omega, isobornes, loglik] = vraisemblanceC(X,[]);

crit = loglik + theta*1/n*(log(n) + 2.5); isobornes = [];

% Sélection de modèles

for i = 2:p
   born = (closeto(X,X(n) + (X(1)-X(n))*(1:(i-1))/i))';
   born = born(find(born<X(1) & born>X(n)));
   born = concat(X,born,k);
   l = length(born);
   for j = 1:l
      A = nchoosek(born,j);
      for kk = 1:nchoosek(l,j)
         borne = A(kk,:)';
         b = length(borne)+1;
         [alph, omeg, isoborne, logli] = ...
                        vraisemblanceC(X,borne);
         cri = logli + theta*b/n*(log(n/(l+1)) + 2.5);
         if cri < crit
            crit  = cri;
            alpha = alph;
            omega = omeg;
            isobornes = isoborne;
         end
      end
   end
end
```

Code de calibration de la pente `penteCirreg` dans le cas d'une estimation sous contrainte de monotonie en ω, à α inconnu.

```
function theta = penteCirreg(X);

%
% PENTECIRREG donne la valeur de la pente efficace minimale
% sur l'enveloppe convexe du nuage de points (\pen,\gamma)
% pour le protocole de sélection de modèles sur partitions
% irrégulières basées sur une partition fine régulière.
%

n = length(X);
p = min(floor(sqrt(n)),12);
k = ceil(n/(log(n))^2);

yy(1:p) = inf;

xx(1) = 1/n*(log(n) + 2.5);

[alpha, omega, bornes, loglik] = vraisemblanceNC(X,[]);

yy(1) = loglik;

% Sélection de modèles

for i = 2:p
   born = (closeto(X,X(n) + (X(1)-X(n))*(1:(i-1))/i))';
   born = born(find(born<X(1) & born>X(n)));
   born = concat(X,born,k);
   l = length(born);
   if l >= 1
       for j = 1:l
          A = nchoosek(born,j);
          for kk = 1:nchoosek(l,j)
            borne = A(kk,:)';
            b = length(borne) + 1;
            [alph, omeg, borne, loglik] = ...
                        vraisemblanceC(X,borne);
            if (loglik < yy(b) & alph >= alpha)
               xx(b) = b/n*(log(n/(l+1)) + 2.5);
               yy(b) = loglik;
             end
```

```
          end
       end
   end
end

xx = xx(find(xx~=0)); yy = yy(find(yy~=inf));

A = convexhull([xx' yy']); l = length(A(:,1));

t = min(find(A(2:l,2)-A(1:l-1,2) >= 0));

if ~isempty(t)
   A = A(1:t,:);
end

l = length(A(:,1));

for i = 1:l
   D(i) = find(xx==A(i,1));
end

DD = D(2:end) - D(1:length(D)-1);

i = min(find(DD == max(DD)));

theta = (A(i+1,2)-A(i,2))/(A(i+1,1)-A(i,1));
```

Chapitre 7

Annexe 2 : Interfaçage et mode d'emploi de select

L'ouverture du logiciel **select** sous Matlab® fait apparaître la fenêtre représentée sur la figure 7.1.

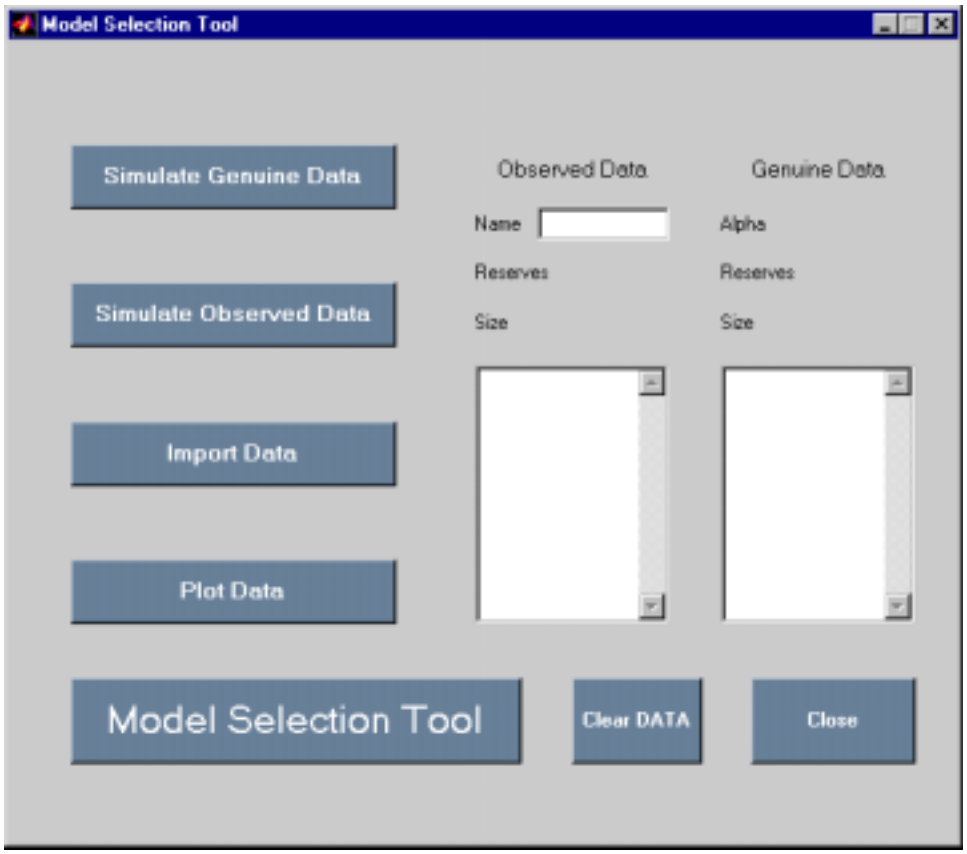

FIG. 7.1 – Fenêtre d'entrée du logiciel "select".

7.1 Simuler les données de la population parente

Dans un premier temps et dans l'objectif de pratiquer des simulations, cette fenêtre interactive permet en cliquant sur le bouton "*Simulate Genuine Data*" de simuler un échantillon d'une loi de Lévy-Pareto (avec un minimum fixé à l'équivalent de 1 Mbep). Apparaît alors la fenêtre de la figure 7.2, qui permet à l'utilisateur de rentrer le paramètre α de la loi de Lévy-Pareto (0,7 dans l'exemple) ainsi que l'effectif N (2000 ici) de l'échantillon. Cet échantillon représente une réalisation de la population parente $\mathbb{X} = \{X_1, ..., X_N\}$ de notre modèle. Si l'utilisateur souhaite utiliser l'outil directement sur des données réelles, se rendre directement au commentaire de la figure 7.7.

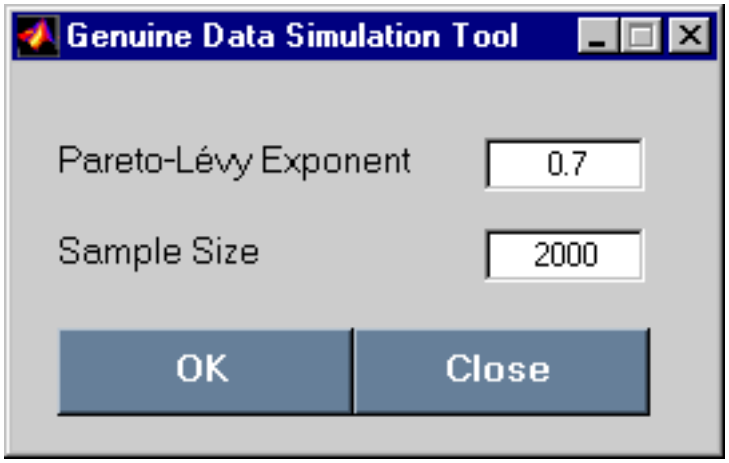

FIG. 7.2 – Fenêtre de simulation de la population parente.

Lorsque les entrées des paramètres sont effectuées, un clic sur le bouton "OK" actualise automatiquement la fenêtre **select** de départ, comme le montre la figure 7.3. L'effectif, le cumul ainsi que la valeur du paramètre de la loi sont indiqués au dessus de la boîte de visualisation de la population parente.

7.2 Simuler les données de l'échantillon observé

Il est ensuite nécessaire de simuler le tirage de l'échantillon observé $\mathbb{Y} = \{Y_1, ..., Y_n\}$ grâce au bouton "*Simulmate Observed Data*" de la fenêtre d'entrée. Deux options sont possibles suivant que l'on souhaite définir des probabilités d'inclusion aux premiers tirages "à la Rosén" (se reporter en 2.3.1 ou [74] et [75]) et un taux de sondage ; ou bien que l'on souhaite directement spécifier des probabilités d'inclusion dans l'échantillon observé de classes de taille données dans la population parente. Le taux de sondage dépend alors directement de ces s probabilités d'inclusion et des bornes fixées par l'utilisa-

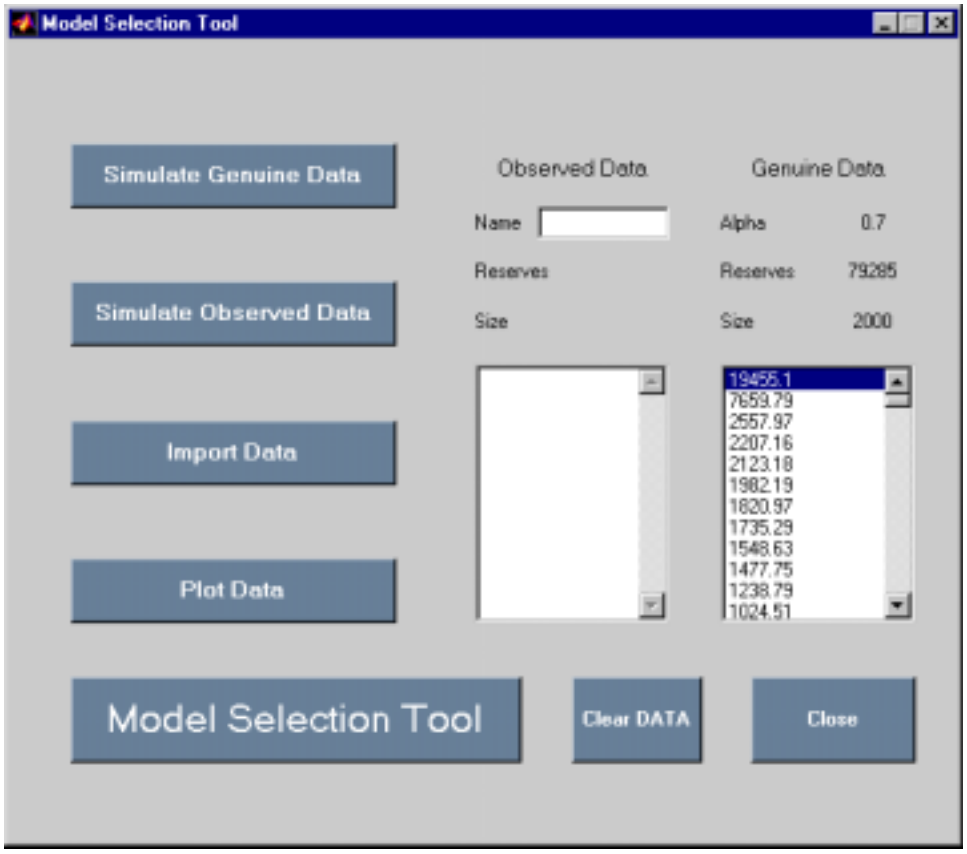

FIG. 7.3 – Fenêtre d'entrée actualisée de la présence de la population parente.

teur. La figure 7.4 montre comment sont introduits les deux choix possibles dans la fenêtre de simulation.

- La fenêtre qui figure à gauche permet de fixer les probabilités d'inclusion au premier tirage (*First Order Probability Inclusions* – FOPI) au moyen d'une fonction puissance dont l'utilisateur rentre la valeur (0,4 dans l'exemple). Plus celle-ci est faible et plus le tirage est uniforme ; plus celle-ci est élevée et plus les objets de taille importante sont favorisés. Il est aussi nécessaire de d'entrer la taux de sondage (0,15 ici, soit 300 individus). Le tirage est effectué de manière successive et conforme à la méthode décrite en 2.3.1.
- Dans la fenêtre de droite sont spécifiées les bornes des intervalles des plages de constance de ω ainsi que les valeurs des probabilités d'inclusion désirées (*Last Order Probability Inclusions* – LOPI). Par exemple, si une classe de la population parente contient 100 individus et que la probabilité d'inclusion fixée par l'utilisateur est de 0.2, la contribution de ladite classe à l'échantillon des observations sera de 20 individus tirés selon une loi uniforme au sein de cette classe. Dans notre exemple, l'échantillon simulé l'est par édition des LOPI avec pour bornes, en Mbep : 5 10 50 100 500 et 1000 et les probabilités d'inclusion associées sont 0,05 0,1 0,2 0,4 0,6 0,8 et 1 pour les plus grandes tailles. Cette méthode autorise à spécifier des probabilités d'inclusion non nécessairement croissantes en fonction de la taille.

Il est à noter que les graphes tracés en bas de la fenêtre de simulation sont ceux des FOPI et LOPI suivant les cas, et représentent donc les probabilités d'inclusion respectives au premier et dernier tirage successif.

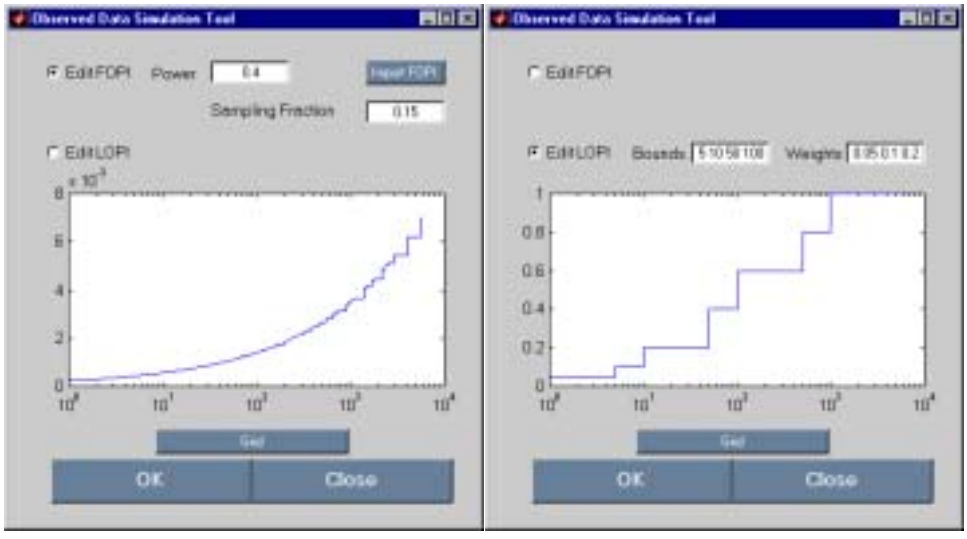

FIG. 7.4 – Variantes de la fenêtre de simulation de l'échantillon observé.

La fenêtre d'entrée s'actualise de l'échantillon des observations une fois le bouton "OK" pressé. L'effectif, le cumul associé ainsi qu'un nom pour cet échantillon (à des fins de sauvegarde ultérieure des données) sont automatiquement affichés, comme le montre la figure 7.5. L'effectif de l'échantillon observé est de 217 individus pour un montant de réserves cumuleés de 62285 contre 79285 Mbep dans la population parente.

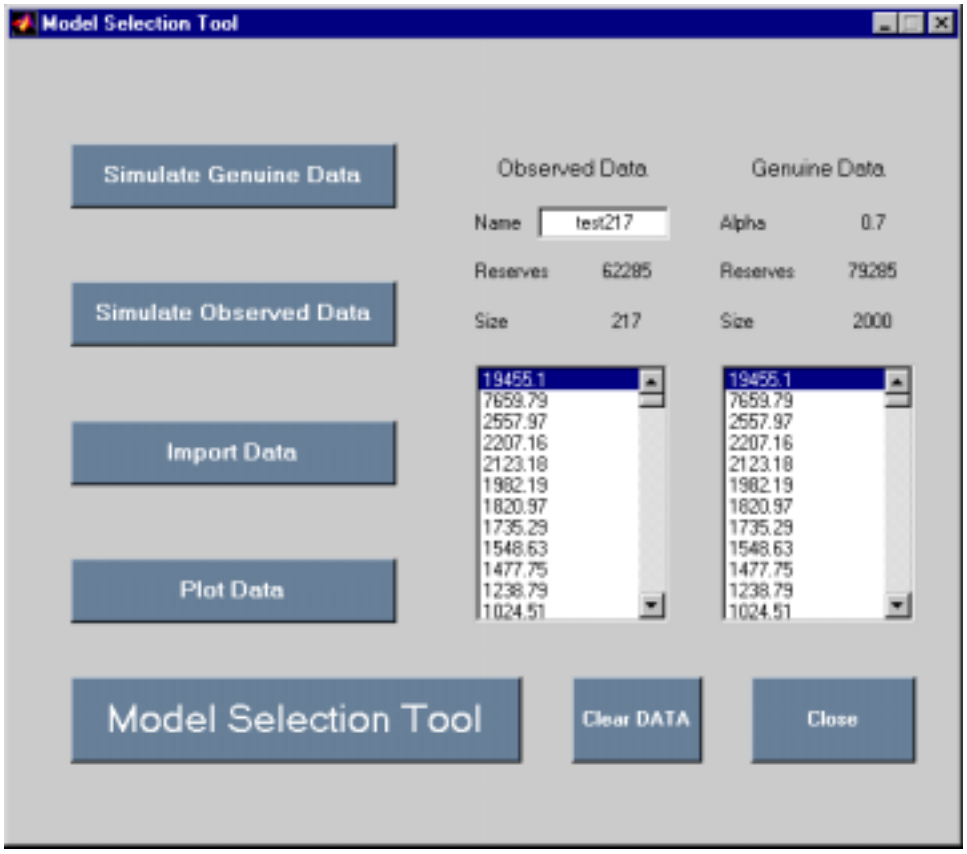

FIG. 7.5 – Fenêtre d'entrée actualisée de la présence de l'échantillon des observés.

Remarque : lorsque l'on choisit de simuler des donnés observées en spécifiant les probabilités d'inclusion au premier tirage grâce à une fonction puissance, la vraie densité à estimer $f_{\alpha,\omega}$ n'appartient pas à l'espace des densités exponentielles de polynômes par morceaux. À l'inverse, si ce sont les probabilités d'inclusion au dernier tirage qui sont spécifiées alors la densité à estimer appartient bien à cet espace.

7.3 Importer les données

Il est aussi possible d'importer directement soit des données de population parente $\mathbb{X}$ (à des fins de simulation par exemple), soit des données observées $\mathbb{Y}$ (notamment pour effectuer tests sur données réelles) grâce au bouton "*Import Data*" de la fenêtre d'entrée. S'affiche alors la fenêtre de choix 7.6. Le type de données étant choisi, il est essentiel que le fichier que l'utilisateur va sélectionner ensuite présente des données ASCII (format texte ".txt") en colonne[1].

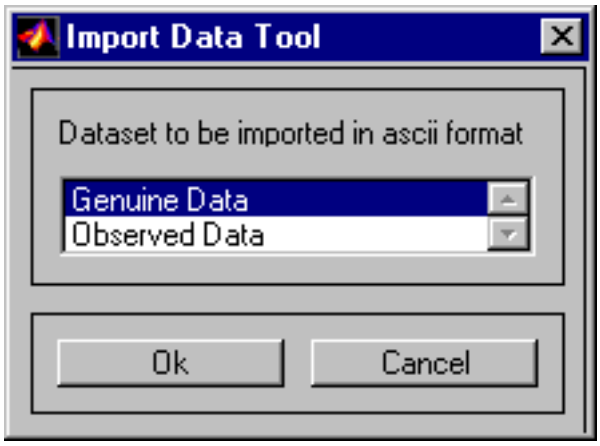

FIG. 7.6 – Fenêtre d'importation de données.

7.4 Tracer les diagrammes LogLog des données

Les données (population parente et/ou échantillon observé) étant prêtes à exploiter dans la fenêtre d'entrée de la figure 7.5, celles-ci peuvent être tracées dans un diagramme LogLog (voir 1.2.5) au moyen du bouton "*Plot Data*". Une fenêtre graphique classique Matlab® apparaît alors avec les données population $\mathbb{X}$ (si elles existent) en bleu et les données échantillon $\mathbb{Y}$ (si elles existent) en rouge, comme le montre la figure 7.7.

Les données étant au complet, nous sommes en mesure de lancer un ou plusieurs protocoles d'estimation sur les données.

7.5 Choix du type de partition

Pour faire apparaître la fenêtre d'options d'estimation, cliquer sur le bouton "*Model Selection Tool*" de la fenêtre d'entrée.

[1] Ce type de fichier de données peut-être produit par tous les logiciels du marché : traitement de texte, tableurs, calcul scientifique...

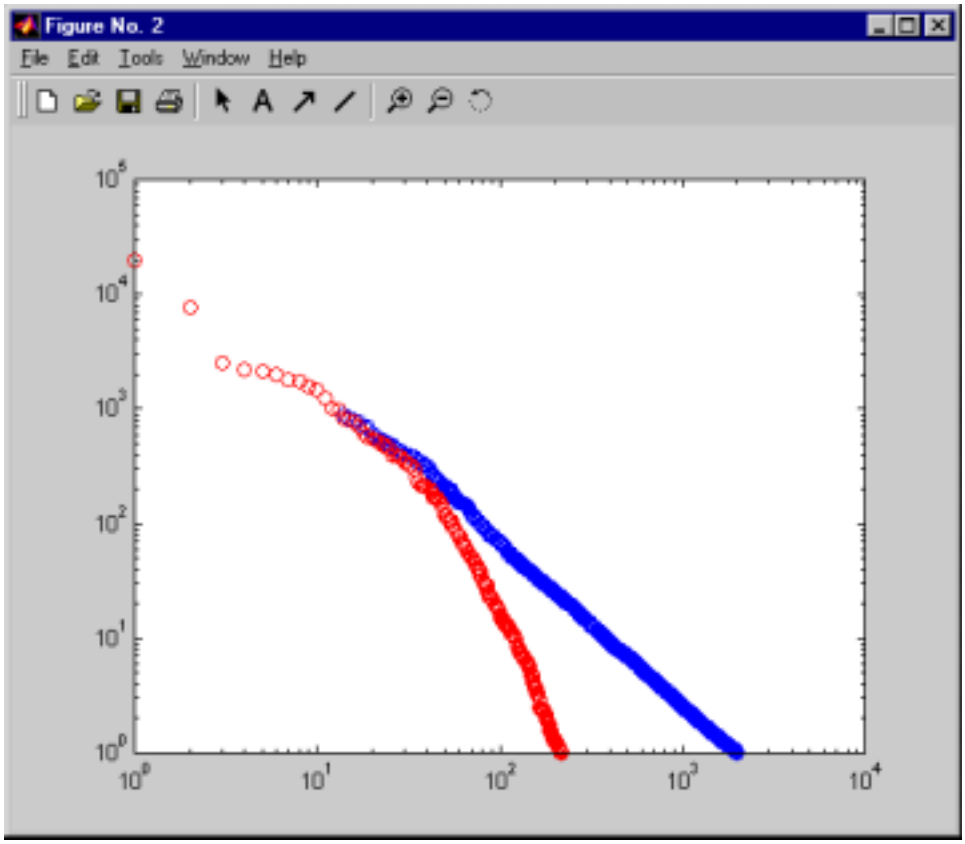

FIG. 7.7 – Diagramme LogLog des données population et échantillon.

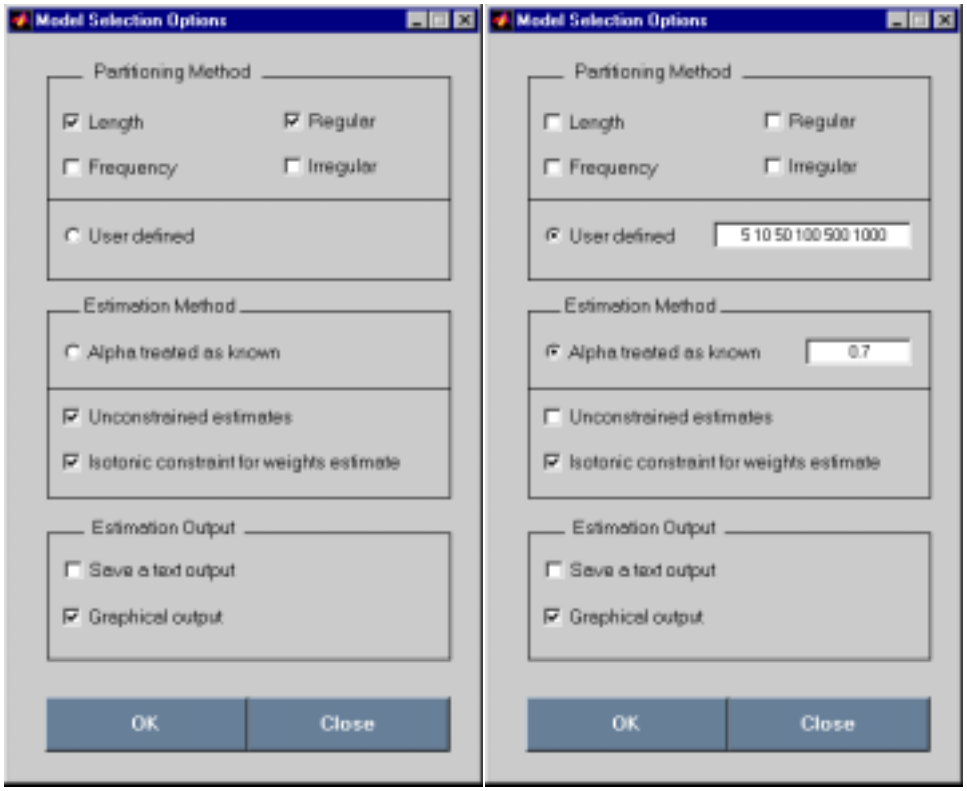

FIG. 7.8 – Variantes de la fenêtre de choix de méthode d'estimation.

Cette fenêtre d'options fait permet à l'utilisateur de cocher plusieurs *checkboxes*. Le premier groupe de ces *checkboxes* concerne la méthode de partition de l'intervalle des données (groupe "*Partitioning Method*"). Il est possible de spécifier une partition en
- Log-longueur d'intervalles constants (*checkbox* "*Length*") ;
- effectifs constants, ou blocs statistiques équivalents selon la terminologie *ad hoc* (*checkbox* "*Frequency*").

Il est ensuite nécessaire de préciser si l'algorithme doit aller visiter
- uniquement les partitions régulières (*checkbox* "*Regular*") ;
- toutes les partitions irrégulières pouvant être construites sur une grille régulière (*checkbox* "*Irregular*").

L'utilisateur peut aussi spécifier les bornes qu'il souhaite pour l'estimation (*checkbox* "*User Defined*"). On sort alors évidemment du cadre de la sélection de modèles puisque par cette manœuvre le modèle dans lequel sont estimés les paramètres est entièrement fixé.

Remarque (1) : les bornes des plages de constance de la fonction de pondération $\widehat{\omega}$ estimée sont calées sur les données. Elles peuvent donc varier de la définition de l'utilisateur ou bien ne pas définir des plages de Log-longueur parfaitement régulières dans le cas où cette option aurait été retenue. Du point de vue algorithmique, les bornes sont choisies dans la collection des données observées comme étant les données les plus proches de celles spécifiées par l'utilisateur ou bien les plus proches d'une partition parfaitement régulière.

Remarque (2) : lorsque l'on ne se trouve pas en situation *User Defined*, il est possible de cocher plusieurs méthodes en même temps. Toutes les partitions correspondant à toutes les méthodes possibles vont alors être testées. Celle qui est sélectionnée est celle qui minimise le critère du maximum de vraisemblance pénalisé parmi tous les modèles. On peut ainsi éventuellement mettre en compétition plusieurs méthodes de partition.

7.6 Choix de la méthode d'estimation

De nouveau, on peut envisager plusieurs options possibles pour les estimations des paramètres des modèles visités (groupe "*Estimation Method*").

Il est possible de travailler à α fixé par pour l'utilisateur[2] (*checkbox* "*Alpha treated as known*"), ou bien de laisser libre cette valeur qui sera alors estimée par l'algorithme.

Remarque : notons que si α est fixé, les valeurs des $(\omega_I)_{I \in \hat{m}}$, où $\hat{m}$ est la partition finale choisie, ne dépendent plus que du nombre d'observations dans chaque intervalle. En effet, l'estimateur du maximum de vraisemblance

[2] rappelons que α est l'inverse de l'habitat du système.

de ω correspond alors à l'histogramme des Log-données contre la mesure exponentielle de paramètre α comme on l'a vu dans la remarque (1) de la proposition 3.2.3.

Enfin, on peut soit laisser les estimations des $(\omega_I)_{I\in\hat{m}}$ libres de contraintes (*checkbox* "*Unconstrained Estimates*"), ou bien les contraindre à la monotonie (*checkbox* "*Isotonic constraint for weights estimate* "). On peut encore mettre ces deux méthodes en compétition l'une contre l'autre en cochant les deux *checkboxes*.

Enfin, il est possible de demander une sortie graphique "*Graphical Output*" et/ou une sauvegarde des résultats "*Save a Text Output*" au format ASCII à l'intérieur du dossier *work* de Matlab® dans un fichier dont un nom par défaut est proposé. Nous détaillons les résultats de la sortie graphique dans la section qui suit.

Deux exemples de protocoles possibles d'estimation sont présentés sur la figure 7.8.

7.7 Sortie graphique des résultats

Les résultats du protocole de sélection de modèles ou de la simple estimation des paramètres du modèle spécifié par l'utilisateur sont représentés sur la figure 7.9.

On trouve sur cette fenêtre :
- la valeur estimée, ou spécifiée de α ;
- les bornes $(\hat{b}_1, ..., \hat{b}_{|\hat{m}|-1})$ des plages de constance de $\widehat{\omega}$ spécifiées par l'utilisateur ou sélectionnées par l'algorithme ;
- les valeurs estimées des probabilités d'inclusion entre les bornes mentionnées au dessus (les premières et dernières valeurs de $\widehat{\omega}$ étant définies respectivement pour les intervalles $[\min \mathbb{Y}; \hat{b}_1]$ et $[\hat{b}_{|\hat{m}|-1}; \max \mathbb{Y}]$;
- dans le cas où les il existe une population parente simulée, les vraies valeurs des probabilités d'inclusion entre les bornes $(\hat{b}_1, ..., \hat{b}_{|\hat{m}|-1})$ sont données à titre de comparaison avec les valeurs estimées[3] (cette option – bouton "*Display projected omegas*" n'est pas activée sur la figure 7.9) ;
- la valeur des réserves totales estimées et le détail, intervalle par intervalle, des réserves de l'échantillon des observations et des réserves estimées. Une estimation des réserves restant à découvrir est donc la différence entre les deux ;
- le nombre total de champs estimé et le détail, classe de taille par classe de taille, du nombre de champs dans l'échantillon des observations et du

[3]Ces "vraies valeurs" correspondent aux coefficients de la projection Kullback de la vraie densité devant les vecteurs $(\mathbf{I}_I)_{I\in\hat{m}}$, c'est-à-dire la fonction $\bar{\omega}$, conformément aux notations du chapitre 4.

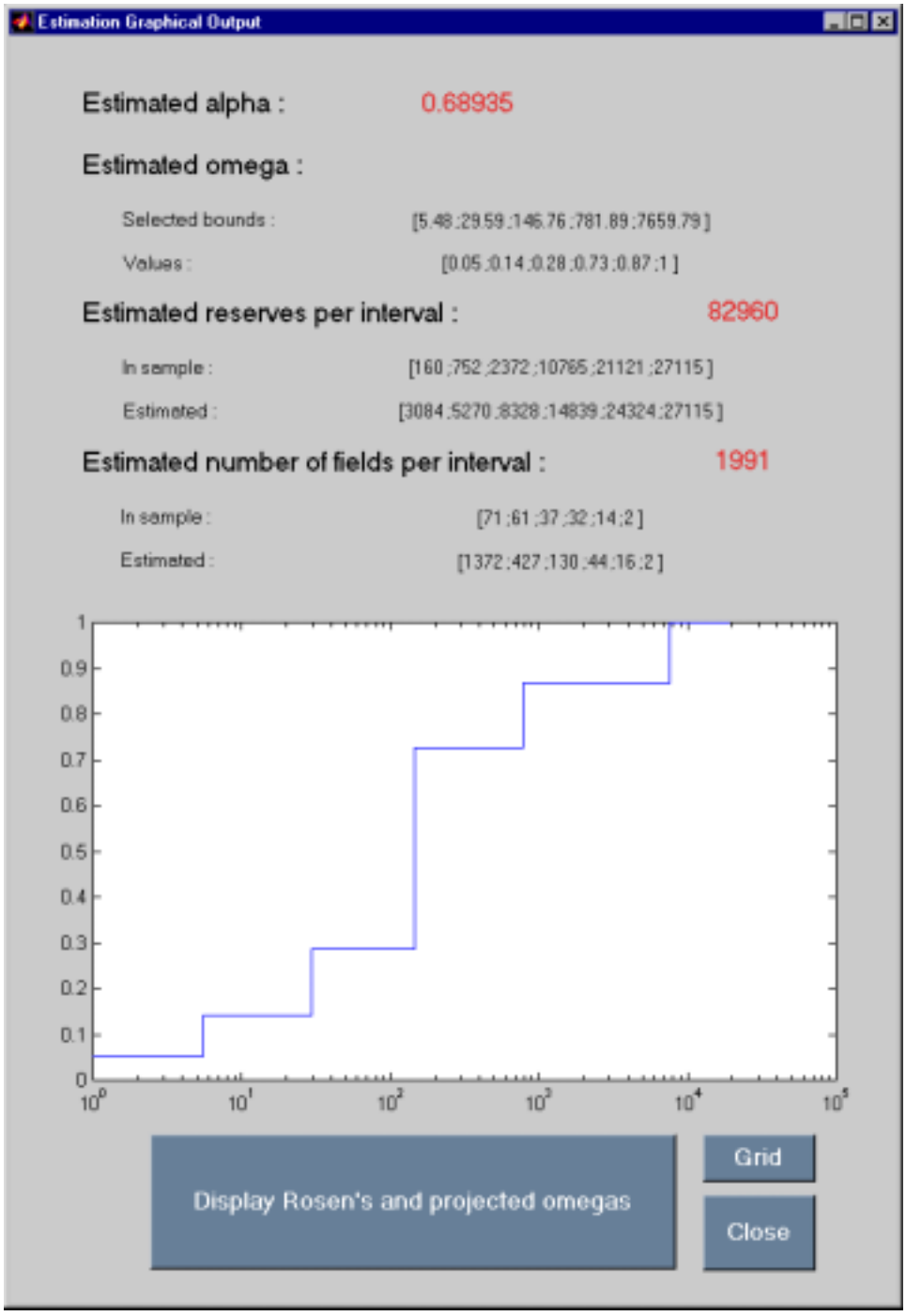

FIG. 7.9 – Sortie graphique du protocole d'estimation.

nombre de champs estimé. Comme pour les réserves, le nombre de champs restant à découvrir s'estime par la différence.

– le graphe de la fonction $\widehat{\omega}$ dans un diagramme semi-Log. Si les données de la population parente sont simulées, un bouton permet d'afficher aussi la projection Kullback $\bar{\omega}$ représentant les vraies probabilités d'inclusion de chaque classe de taille.

Chapitre 8

Annexe 3 : résultats détaillés des simulations

Dans cette annexe se trouvent certains résultats détaillés des simulations de la section 5.2. Pour chaque simulation ci-dessous on retrouve un histogramme[1] :
- de la valeur estimée du paramètre α ;
- du nombre de morceaux de la partition sélectionnée ;
- de l'erreur relative d'estimation du nombre de champs restant à découvrir ;
- de l'erreur relative du montant de réserves restant à découvrir.

Chaque histogramme est assorti du calcul de la médiane, de la moyenne et de l'écart-type empiriques de la distribution observée.

8.1 Densité exponentielle polynômiale par morceaux

8.1.1 Habitat très concentré

Rappel : les caractéristiques de la densité $f_{\alpha,\omega}$ à estimer sont les suivantes :
- $\alpha = 0{,}6$;
- effectif de la population parente : $N = 2000$;
- partition de $\mathbb{R}^+$ en 4 morceaux avec les caractéristiques suivantes :

I	$[0\ ;20[$	$[20\ ;200[$	$[200\ ;2000[$	$[2000\ ;+\infty[$
ω	0,1	0,3	0,6	1

[1] Histogramme non construit par sélection de modèles !

Estimation sans contrainte,
partitions en intervalles de longueur régulière

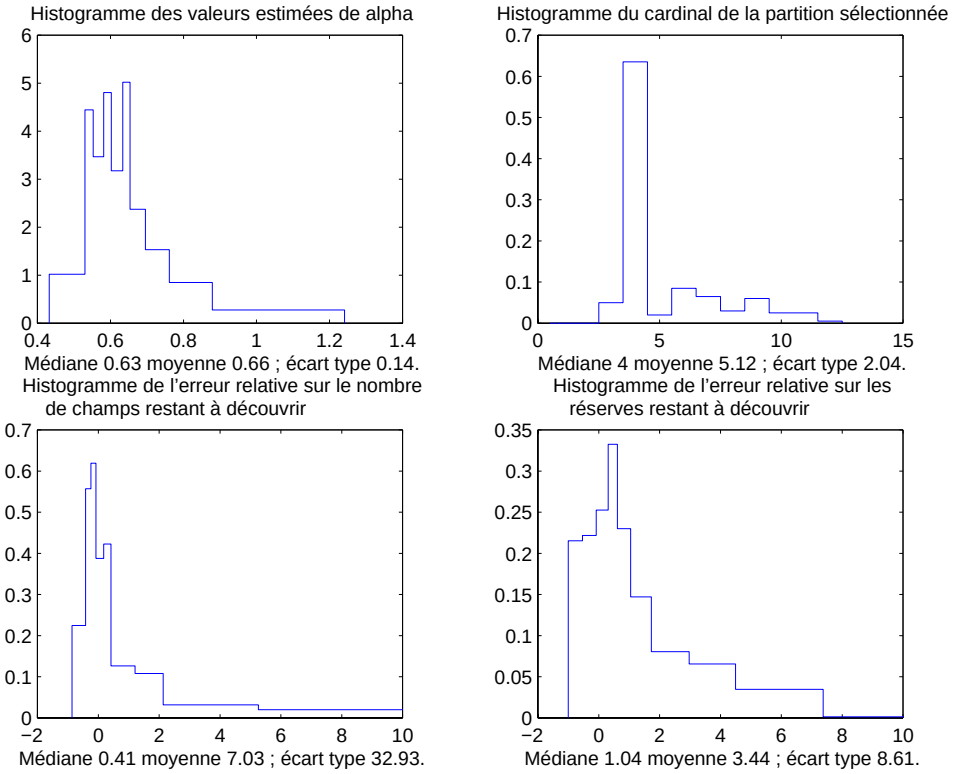

FIG. 8.1 – Estimation d'une densité d'habitat très concentré, sans contrainte de monotonie en ω sur partitions en intervalles de longueur régulière. Protocole de 200 simulations.

Estimation sous contrainte,
partitions en intervalles de longueur régulière

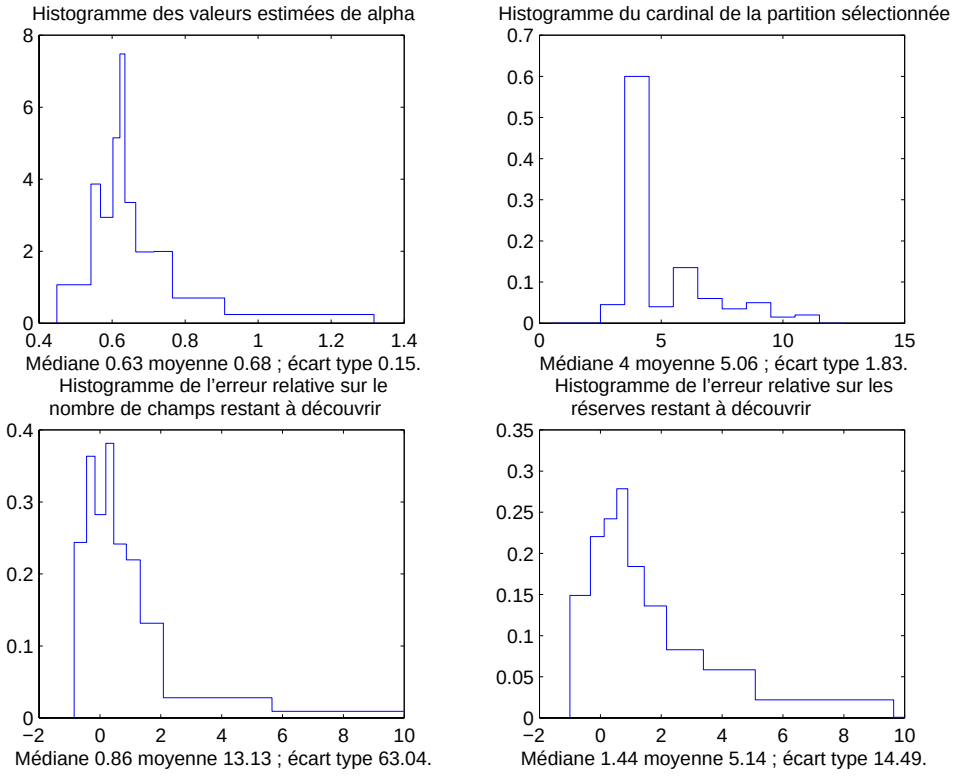

FIG. 8.2 – Estimation d'une densité d'habitat très concentré, sous contrainte de monotonie en ω sur partitions en intervalles de longueur régulière. Protocole de 200 simulations.

Estimation sans contrainte,
partitions en intervalles de longueur irrégulière

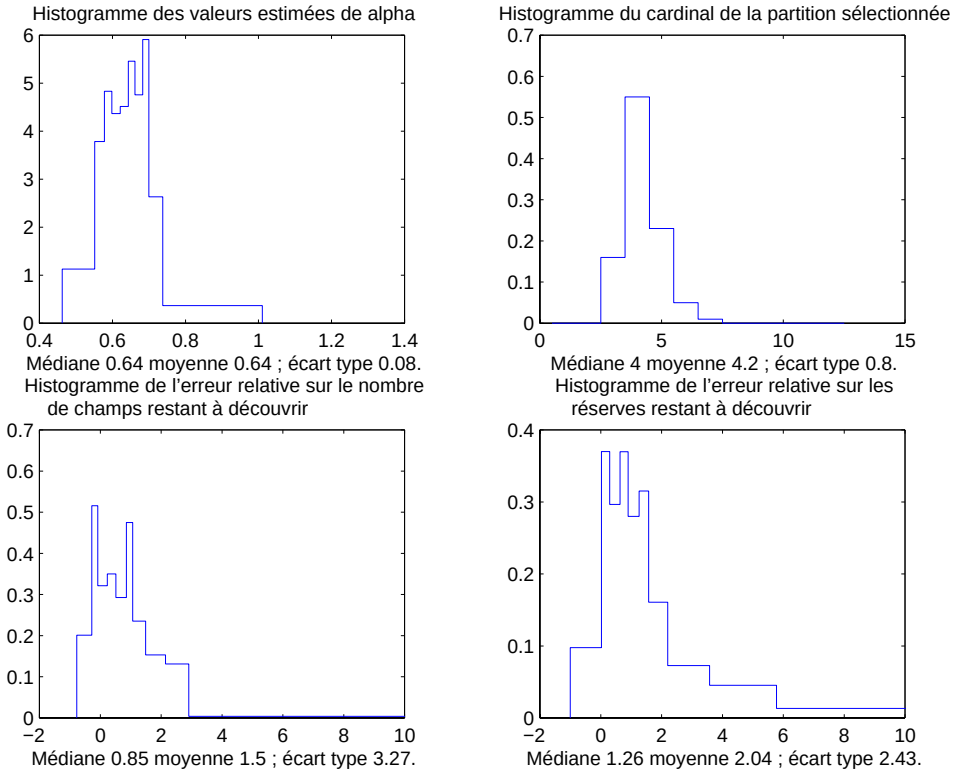

FIG. 8.3 – Estimation d'une densité d'habitat très concentré, sans contrainte de monotonie en ω sur partitions en intervalles de longueur irrégulière. Protocole de 100 simulations.

Estimation sous contrainte,
partitions en intervalles de longueur irrégulière

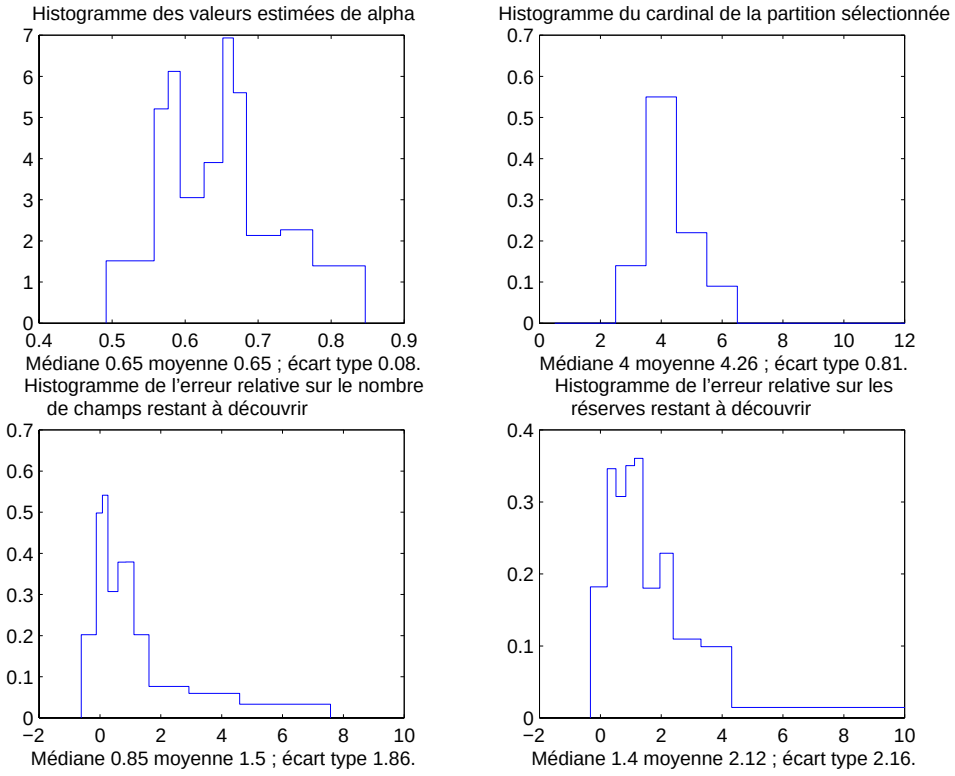

FIG. 8.4 – Estimation d'une densité d'habitat très concentré, sous contrainte de monotonie en ω sur partitions en intervalles de longueur irrégulière. Protocole de 100 simulations.

Estimation sans contrainte,
partitions en intervalles de fréquence régulière

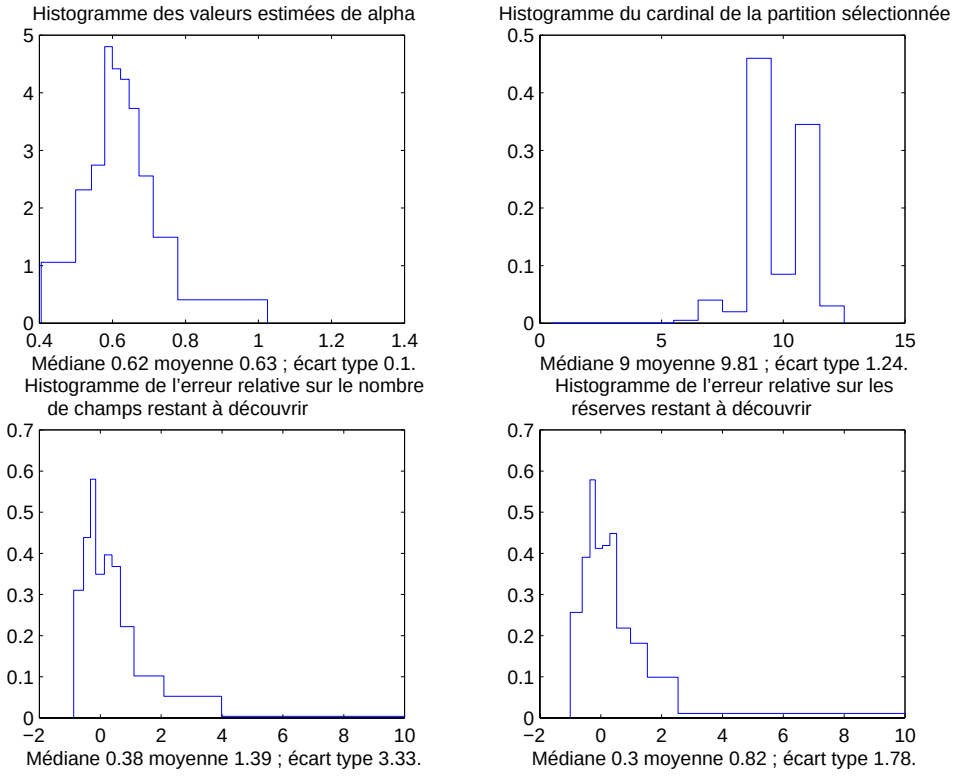

FIG. 8.5 – Estimation d'une densité d'habitat très concentré, sans contrainte de monotonie en ω sur partitions en intervalles de fréquence régulière. Protocole de 200 simulations.

Estimation sous contrainte,
partitions en intervalles de fréquence régulière

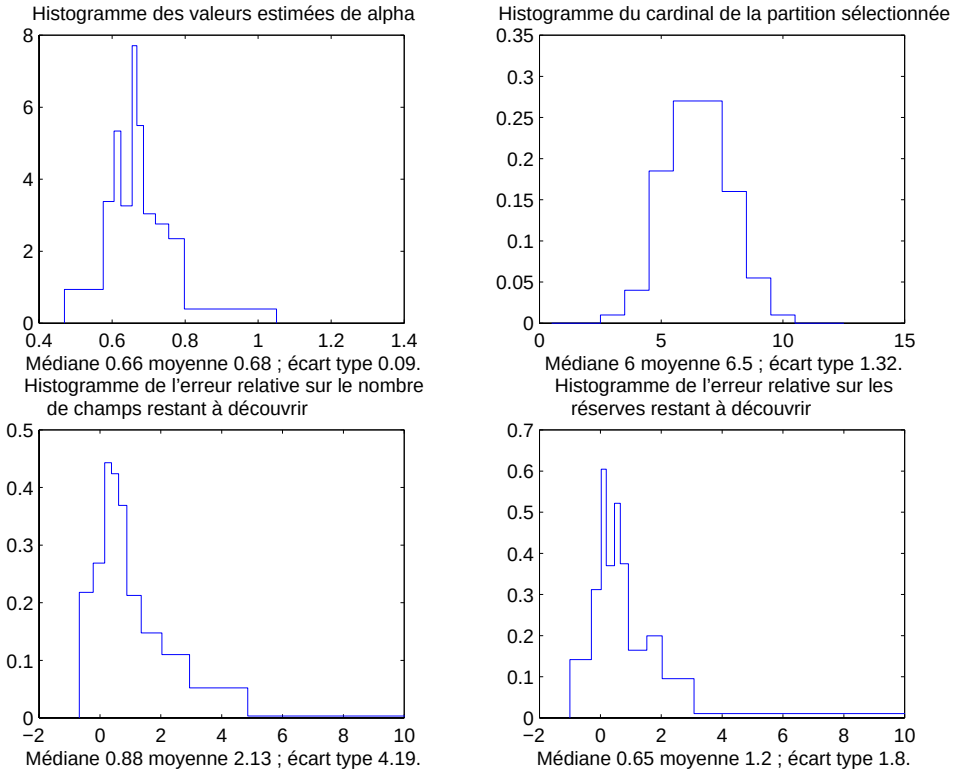

FIG. 8.6 – Estimation d'une densité d'habitat très concentré, sous contrainte de monotonie en ω sur partitions en intervalles de fréquence régulière. Protocole de 200 simulations.

Estimation sans contrainte,
partitions en intervalles de fréquences irrégulières

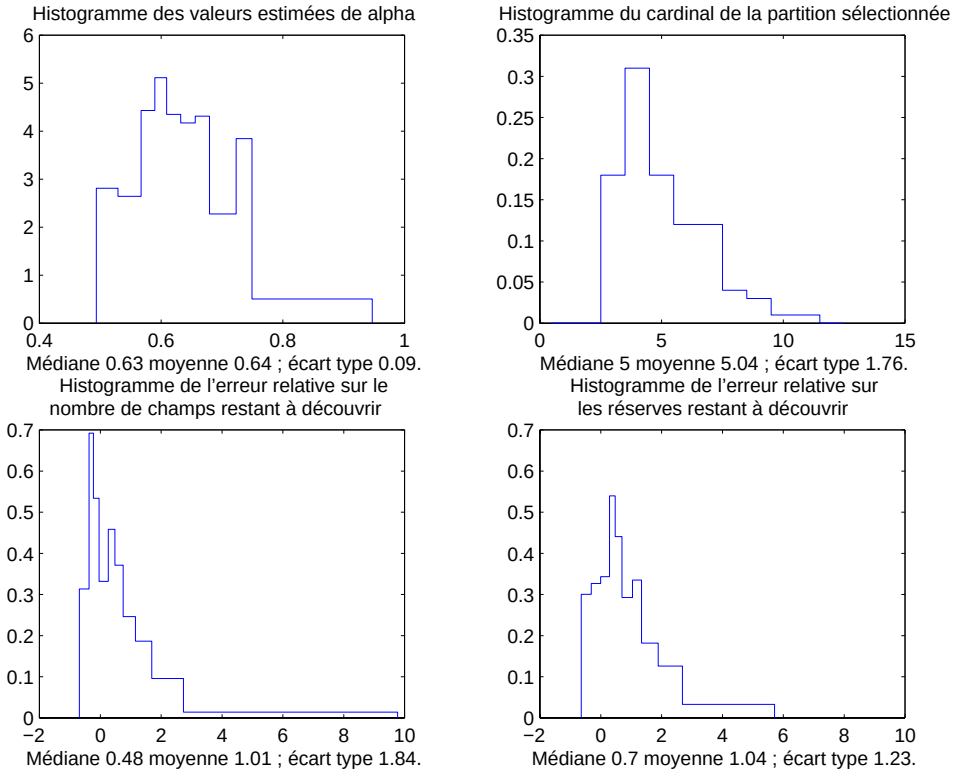

FIG. 8.7 – Estimation d'une densité d'habitat très concentré, sans contrainte de monotonie en ω sur partitions en intervalles de fréquences irrégulières. Protocole de 100 simulations.

Estimation sous contrainte,
partitions en intervalles de fréquences irrégulières

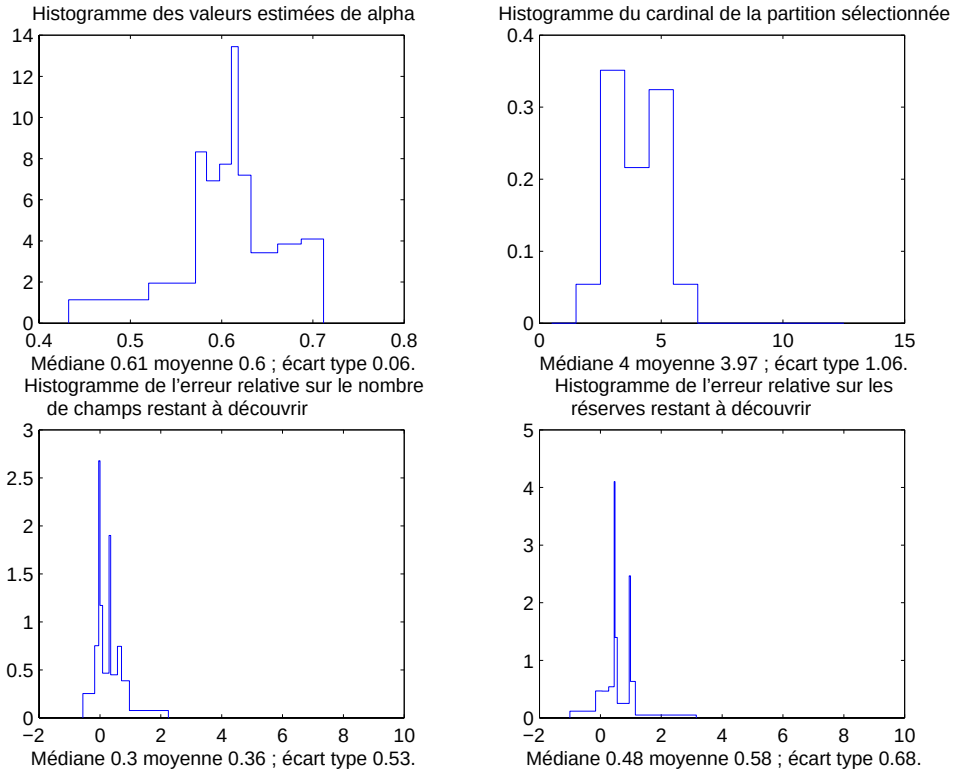

FIG. 8.8 – Estimation d'une densité d'habitat très concentré, sous contrainte de monotonie en ω sur partitions en intervalles de fréquences irrégulières. Protocole de 60 simulations.

8.1.2 Habitat concentré

Rappel : les caractéristiques de la densité $f_{\alpha,\omega}$ à estimer sont les suivantes :
- $\alpha = 0{,}8$;
- effectif de la population parente : $N = 2000$;
- partition de $\mathbb{R}^+$ en 5 morceaux avec les caractéristiques suivantes :

I	$[0\ ;10[$	$[10\ ;50[$	$[50\ ;100[$	$[100\ ;500[$	$[500\ ;+\infty[$
ω	0,1	0,3	0,5	0,8	1

Estimation sans contrainte,
partitions en intervalles de longueur régulière

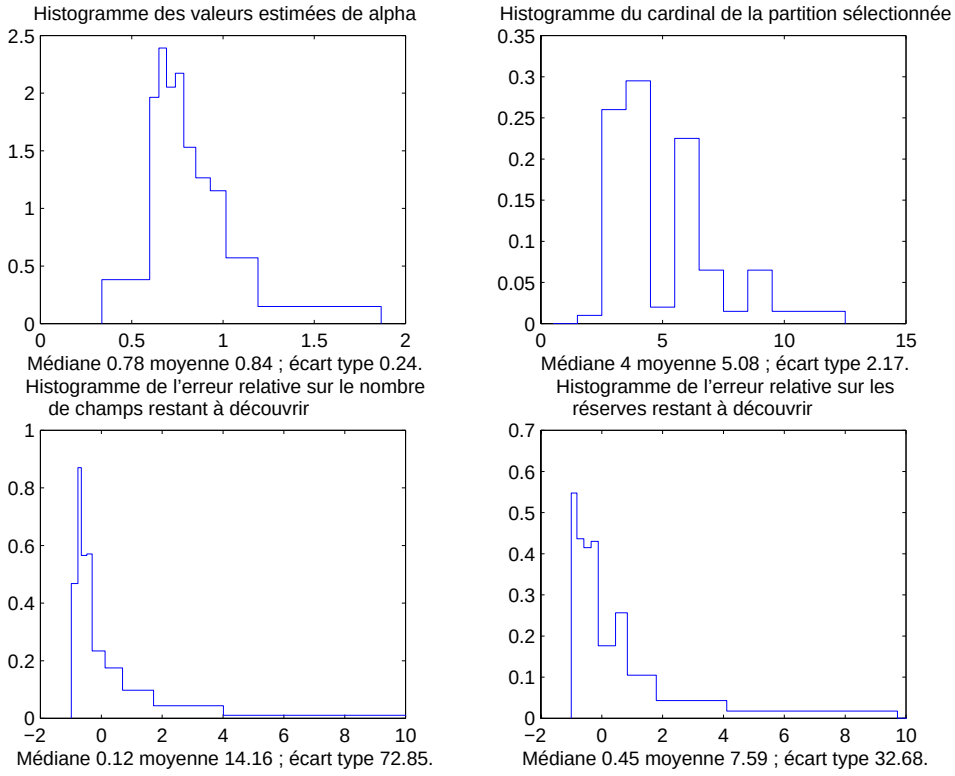

FIG. 8.9 – Estimation d'une densité d'habitat concentré, sans contrainte de monotonie en ω sur partitions en intervalles de longueur régulière. Protocole de 200 simulations.

Estimation sous contrainte,
partitions en intervalles de longueur régulière

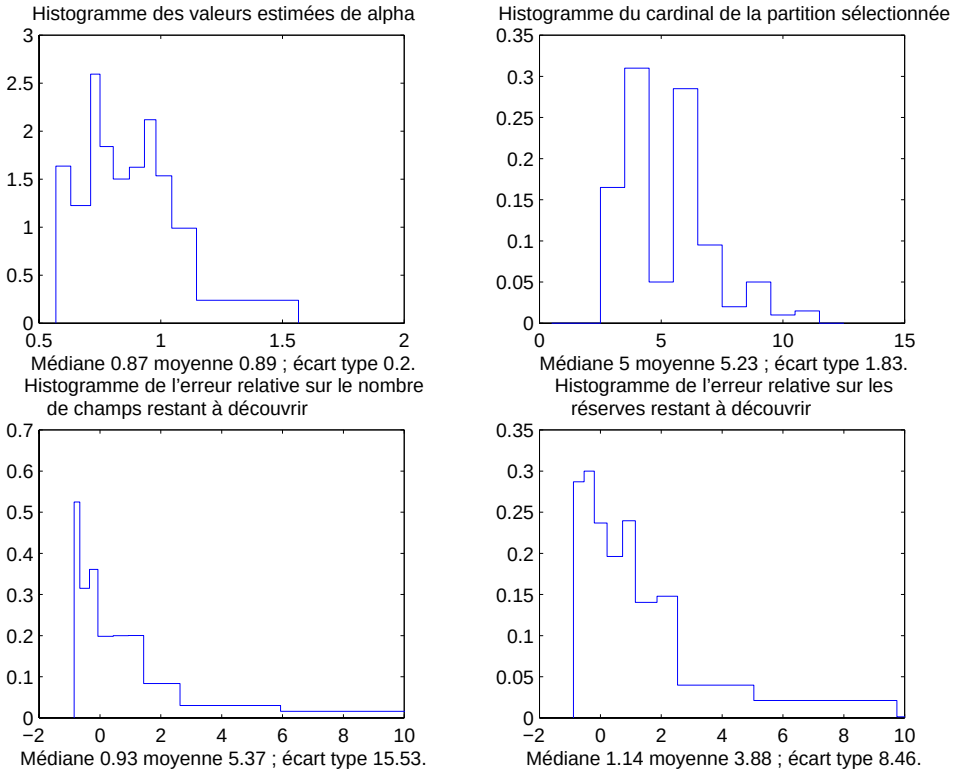

FIG. 8.10 – Estimation d'une densité d'habitat concentré, sous contrainte de monotonie en ω sur partitions en intervalles de longueur régulière. Protocole de 200 simulations.

Estimation sans contrainte,
partitions en intervalles de longueur irrégulière

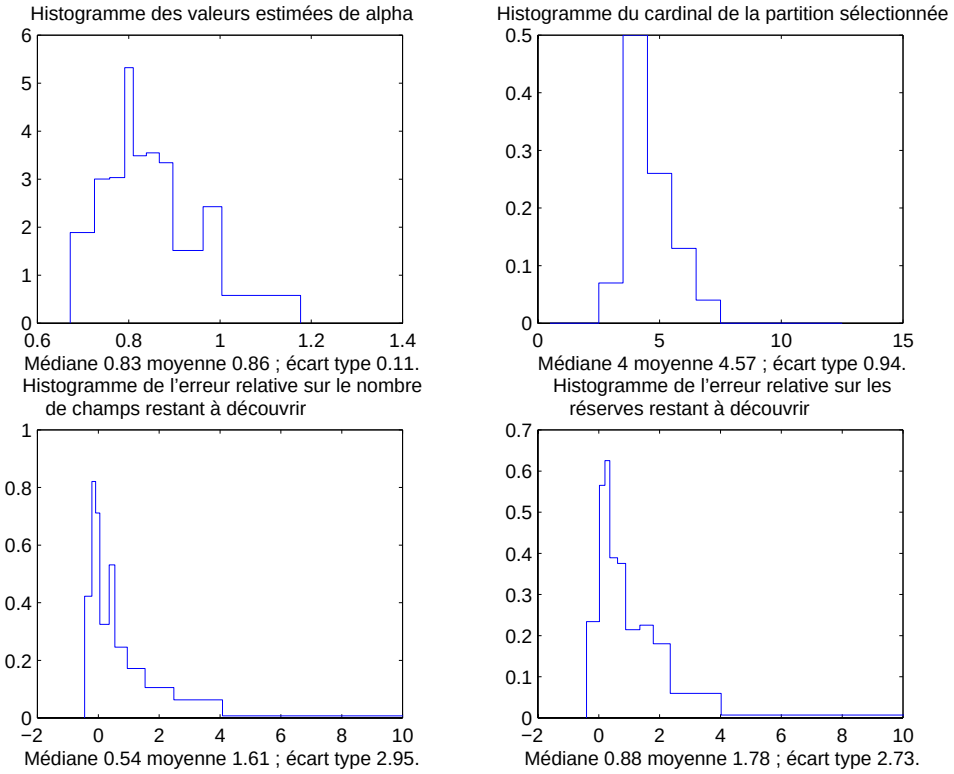

FIG. 8.11 – Estimation d'une densité d'habitat concentré, sans contrainte de monotonie en ω sur partitions en intervalles de longueur irrégulière. Protocole de 100 simulations.

Estimation sous contrainte,
partitions en intervalles de longueur irrégulière

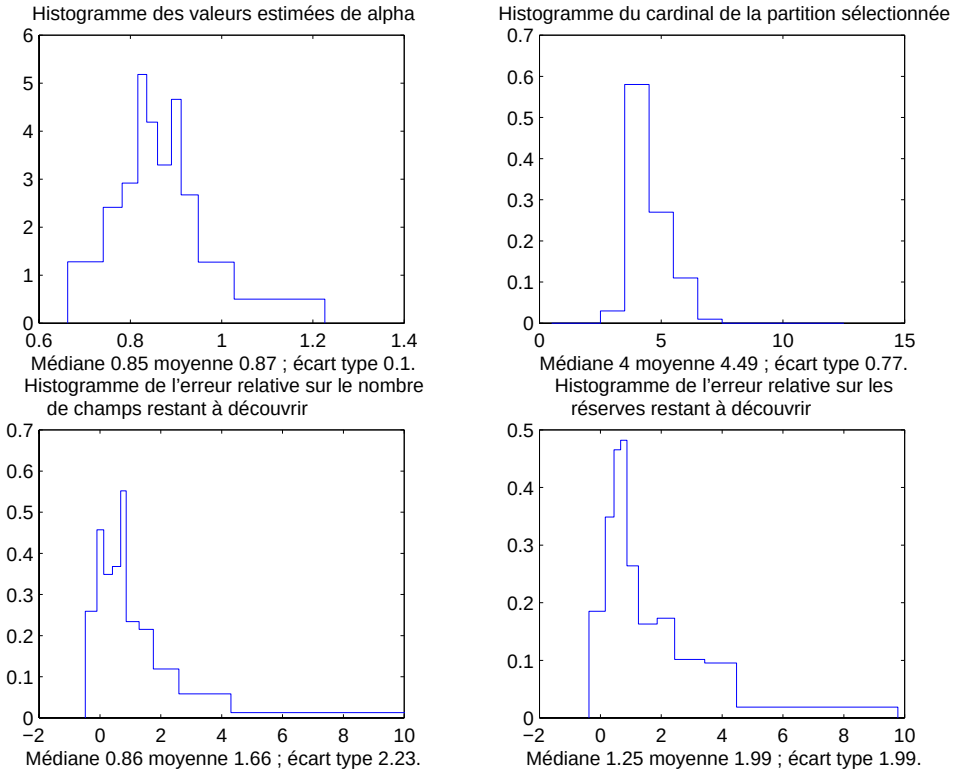

FIG. 8.12 – Estimation d'une densité d'habitat concentré, sous contrainte de monotonie en ω sur partitions en intervalles de longueur irrégulière. Protocole de 100 simulations.

Estimation sans contrainte,
partitions en intervalles de fréquence régulière

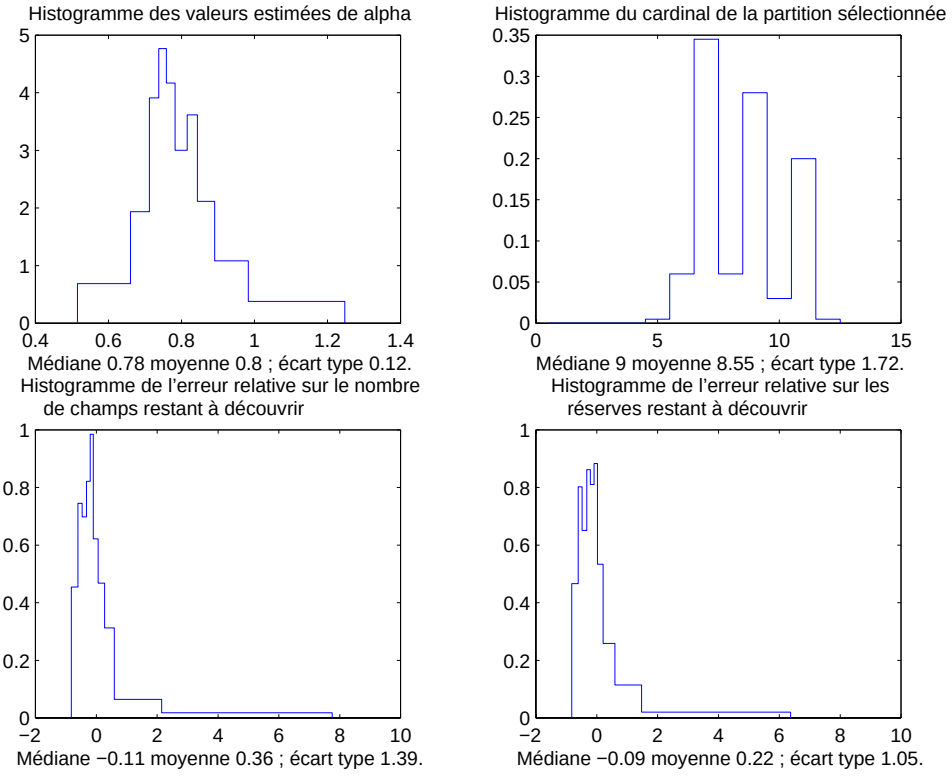

FIG. 8.13 – Estimation d'une densité d'habitat concentré, sans contrainte de monotonie en ω sur partitions en intervalles de fréquence irrégulière. Protocole de 200 simulations.

Estimation sous contrainte,
partitions en intervalles de fréquence régulière

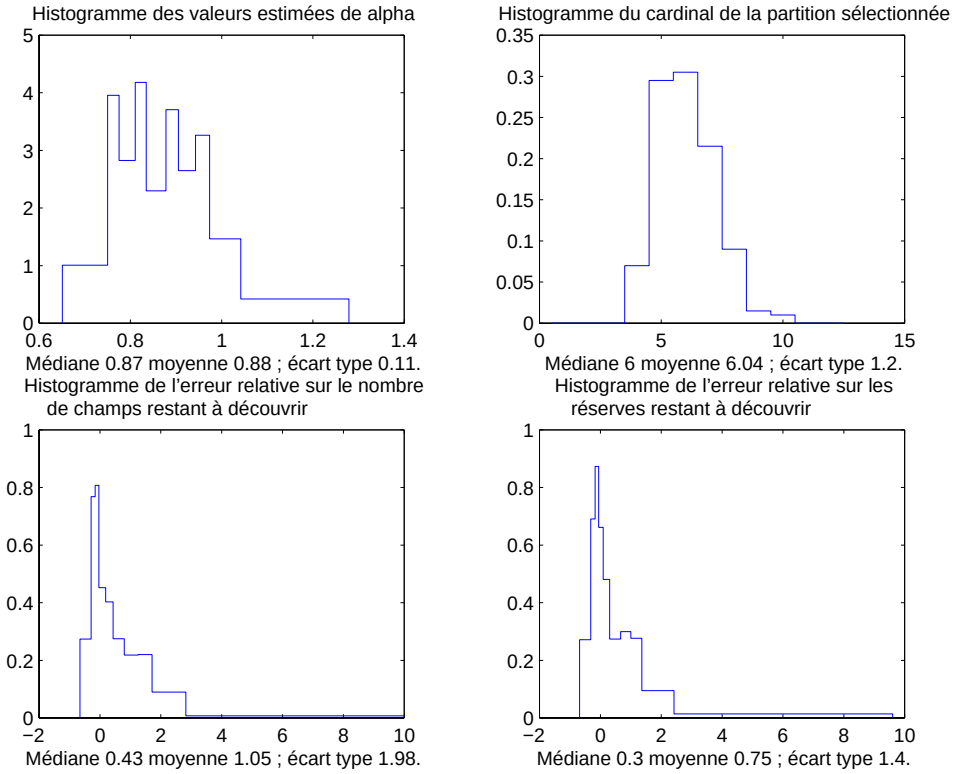

FIG. 8.14 – Estimation d'une densité d'habitat concentré, sous contrainte de monotonie en ω sur partitions en intervalles de fréquence régulière. Protocole de 200 simulations.

Estimation sans contrainte,
partitions en intervalles de fréquences irrégulières

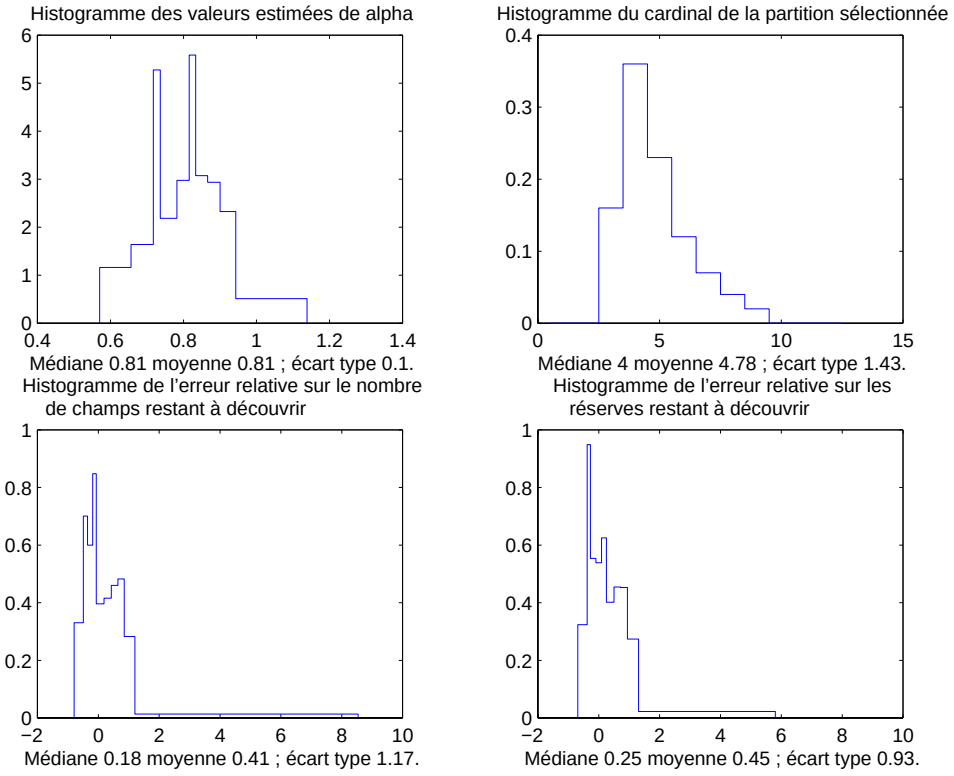

FIG. 8.15 – Estimation d'une densité d'habitat concentré, sans contrainte de monotonie en ω sur partitions en intervalles de fréquences irrégulières. Protocole de 100 simulations.

Estimation sous contrainte,
partitions en intervalles de fréquences irrégulières

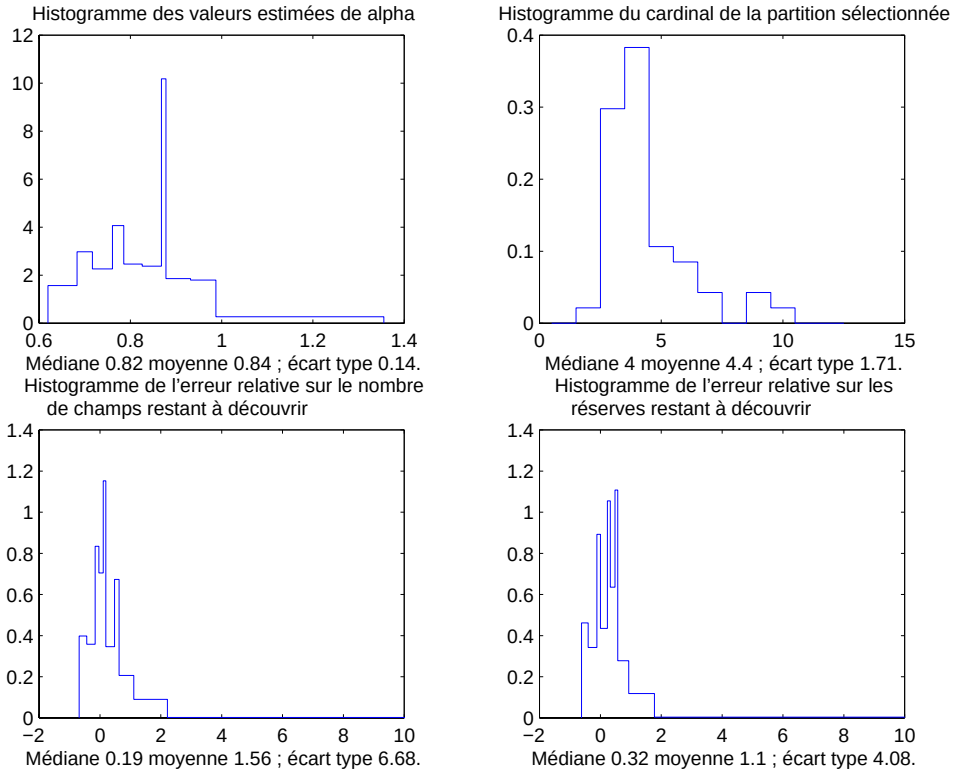

FIG. 8.16 – Estimation d'une densité d'habitat concentré, sous contrainte de monotonie en ω sur partitions en intervalles de fréquences irrégulières. Protocole de 60 simulations.

8.1.3 Habitat dispersé

Rappel : les caractéristiques de la densité $f_{\alpha,\omega}$ à estimer sont les suivantes :
- $\alpha = 1{,}2$;
- effectif de la population parente : $N = 2000$;
- partition de $\mathbb{R}^+$ en 4 morceaux avec les caractéristiques suivantes :

I	$[0\ ;5[$	$[5\ ;10[$	$[10\ ;50[$	$[50\ ;+\infty[$
ω	0,1	0,3	0,6	1

Estimation sans contrainte,
partitions en intervalles de longueur régulière

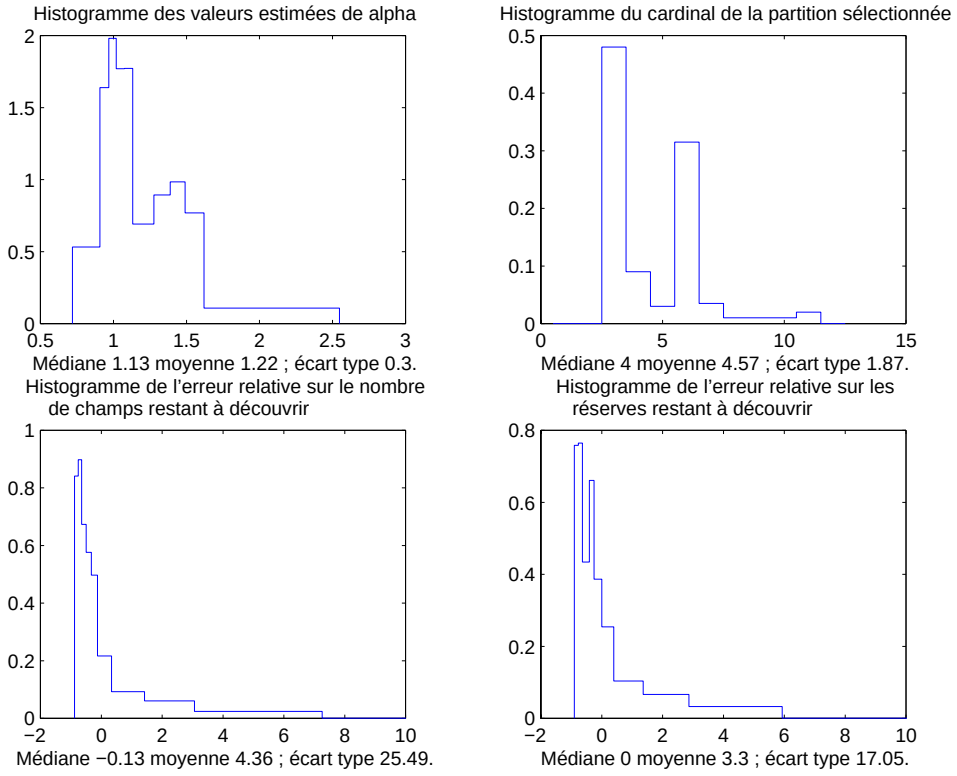

FIG. 8.17 – Estimation d'une densité d'habitat dispersé, sans contrainte de monotonie en ω sur partitions en intervalles de longueur régulière. Protocole de 200 simulations.

Estimation sous contrainte,
partitions en intervalles de longueur régulière

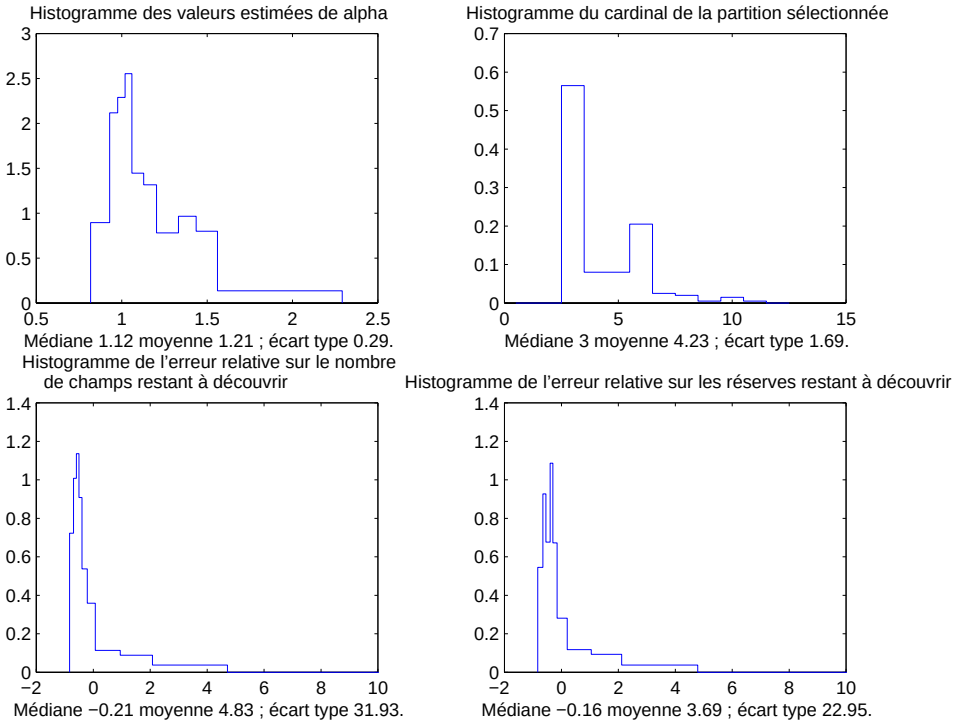

FIG. 8.18 – Estimation d'une densité d'habitat dispersé, sous contrainte de monotonie en ω sur partitions en intervalles de longueur régulière. Protocole de 200 simulations.

Estimation sans contrainte,
partitions en intervalles de longueur irrégulière

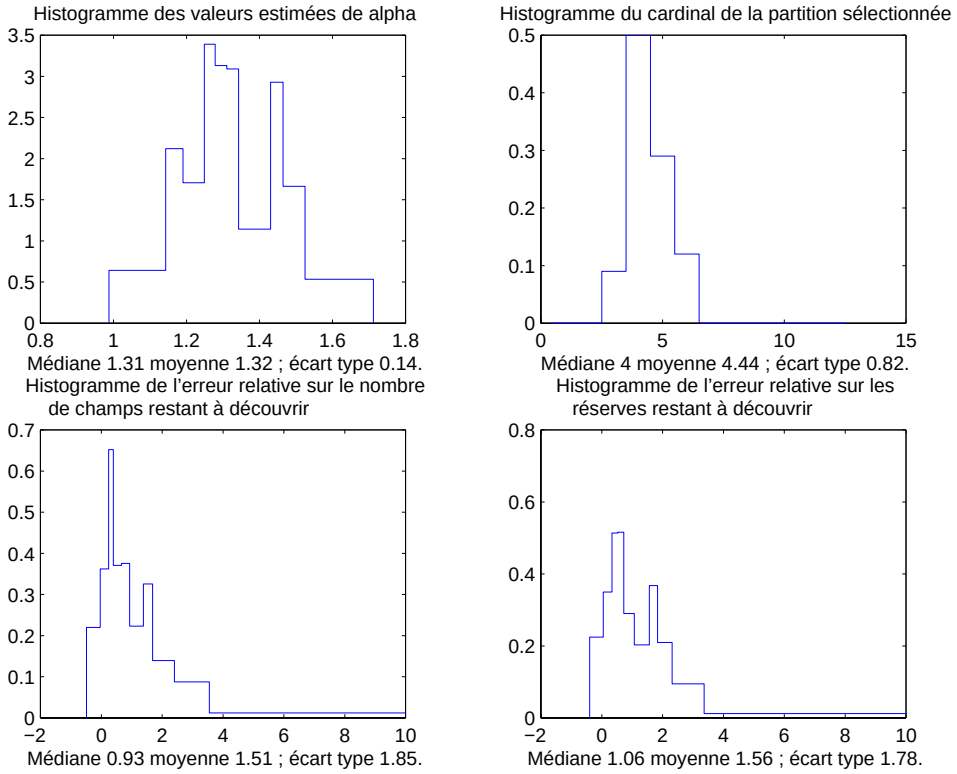

FIG. 8.19 – Estimation d'une densité d'habitat dispersé, sans contrainte de monotonie en ω sur partitions en intervalles de longueur irrégulière. Protocole de 100 simulations.

Estimation sous contrainte,
partitions en intervalles de longueur irrégulière

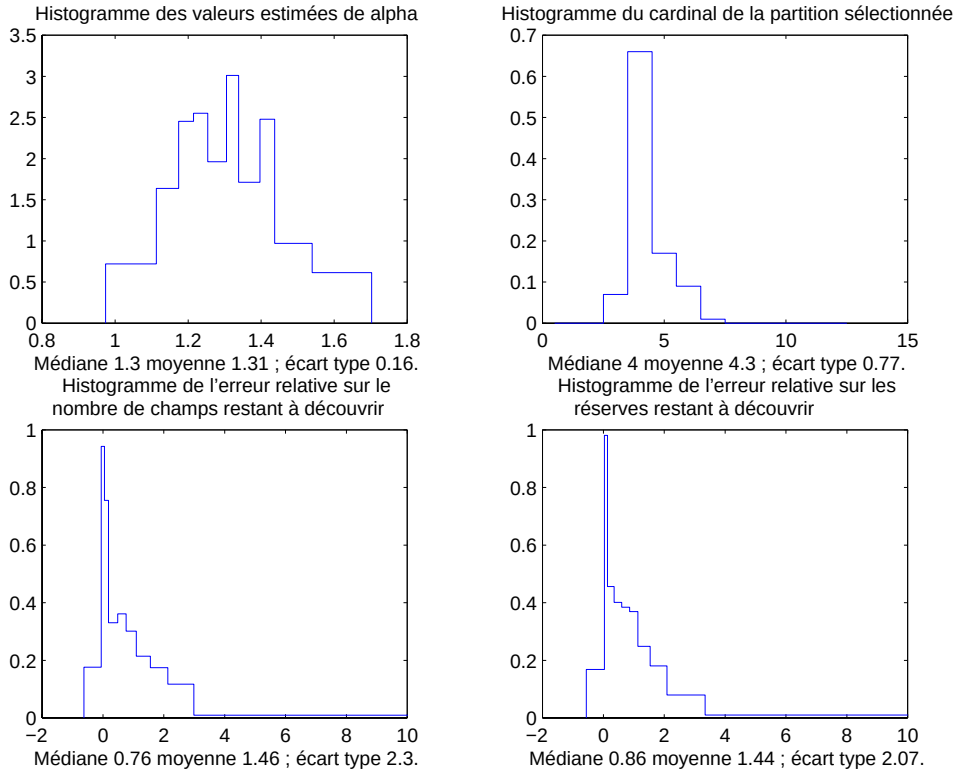

FIG. 8.20 – Estimation d'une densité d'habitat dispersé, sous contrainte de monotonie en ω sur partitions en intervalles de longueur irrégulière. Protocole de 100 simulations.

Estimation sans contrainte,
partitions en intervalles de fréquence régulière

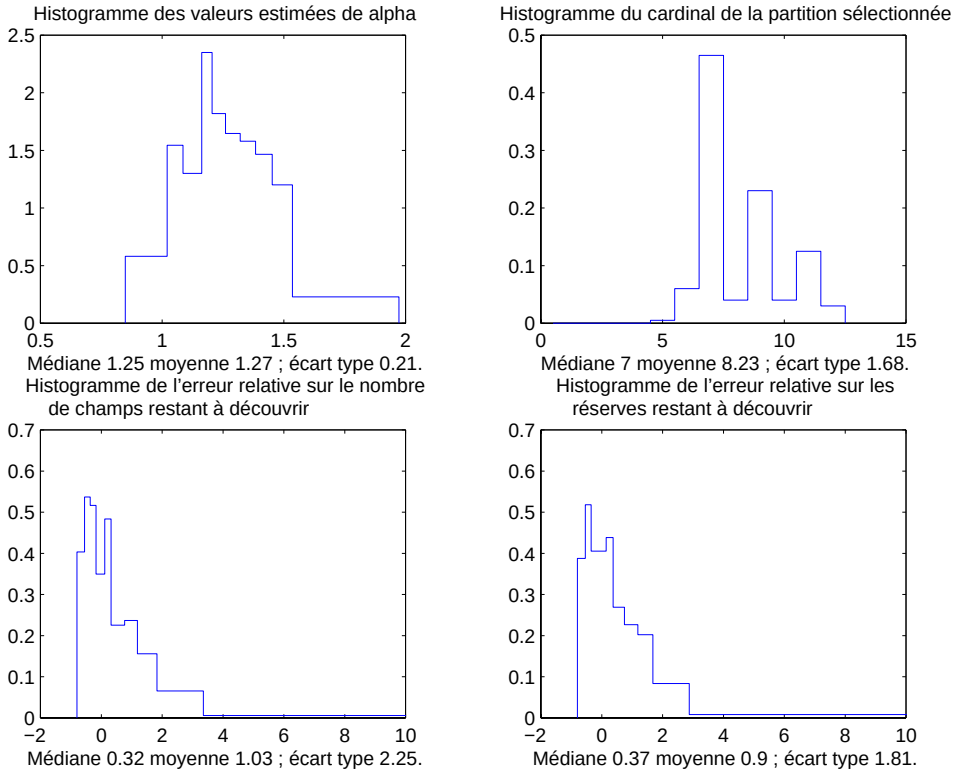

FIG. 8.21 – Estimation d'une densité d'habitat dispersé, sans contrainte de monotonie en ω sur partitions en intervalles de fréquence régulière. Protocole de 200 simulations.

Estimation sous contrainte,
partitions en intervalles de fréquence régulière

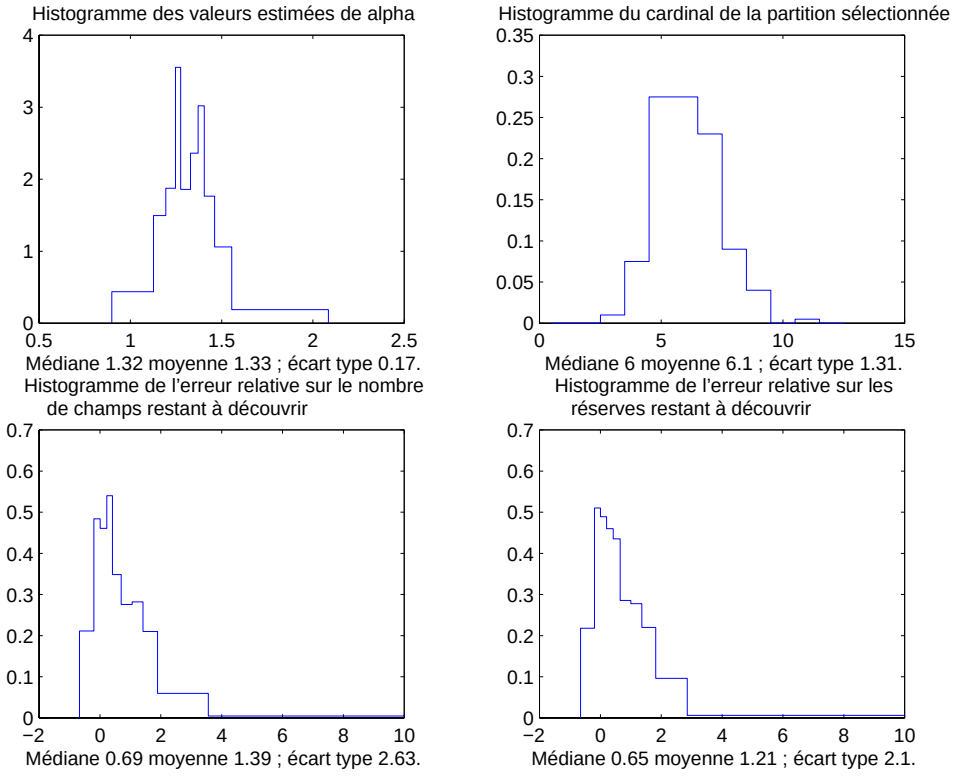

FIG. 8.22 – Estimation d'une densité d'habitat dispersé, sous contrainte de monotonie en ω sur partitions en intervalles de fréquence régulière. Protocole de 200 simulations.

Estimation sans contrainte,
partitions en intervalles de fréquences irrégulières

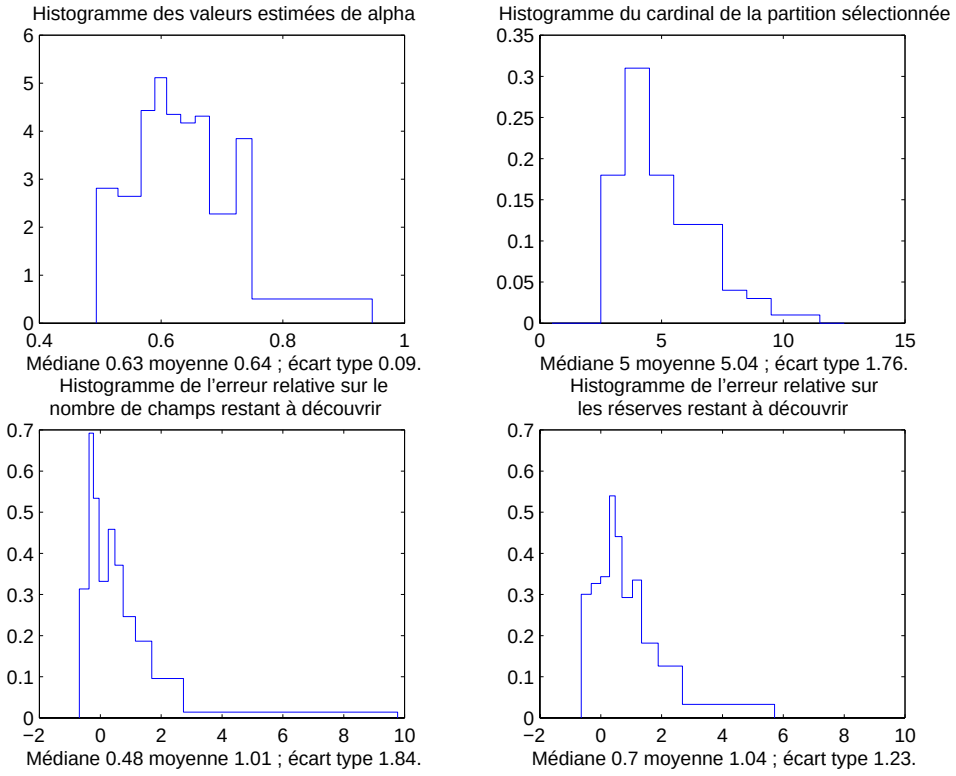

FIG. 8.23 – Estimation d'une densité d'habitat dispersé, sans contrainte de monotonie en ω sur partitions en intervalles de fréquences irrégulières. Protocole de 100 simulations.

Estimation sous contrainte,
partitions en intervalles de fréquences irrégulières

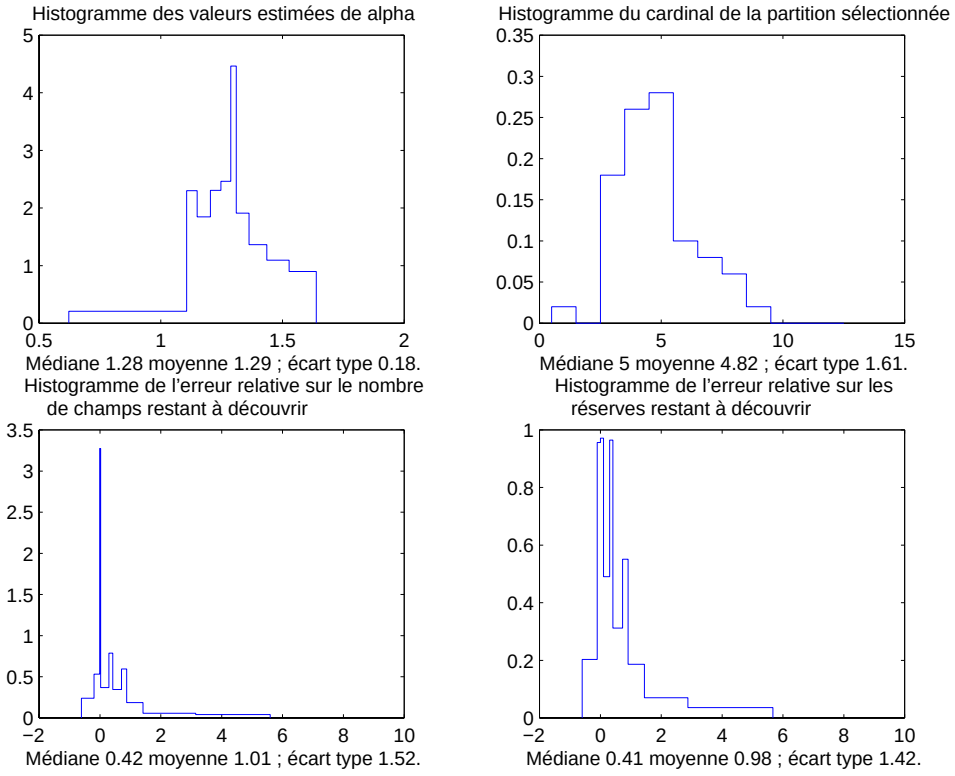

FIG. 8.24 – Estimation d'une densité d'habitat dispersé, sous contrainte de monotonie en ω sur partitions en intervalles de fréquences irrégulières. Protocole de 60 simulations.

8.2 Densité exponentielle non polynômiale par morceaux

8.2.1 Habitat très concentré

Rappel : les caractéristiques de la densité $f_{\alpha,\omega}$ à estimer sont les suivantes :
- $\alpha = 0{,}6$;
- effectif de la population parente : $N = 2000$;
- effectif de la population observée : $n = 300$;
- tirage successif biaisé par effet taille directement proportionnel à la taille.

Estimation sous contrainte,
partitions en intervalles de longueur régulière

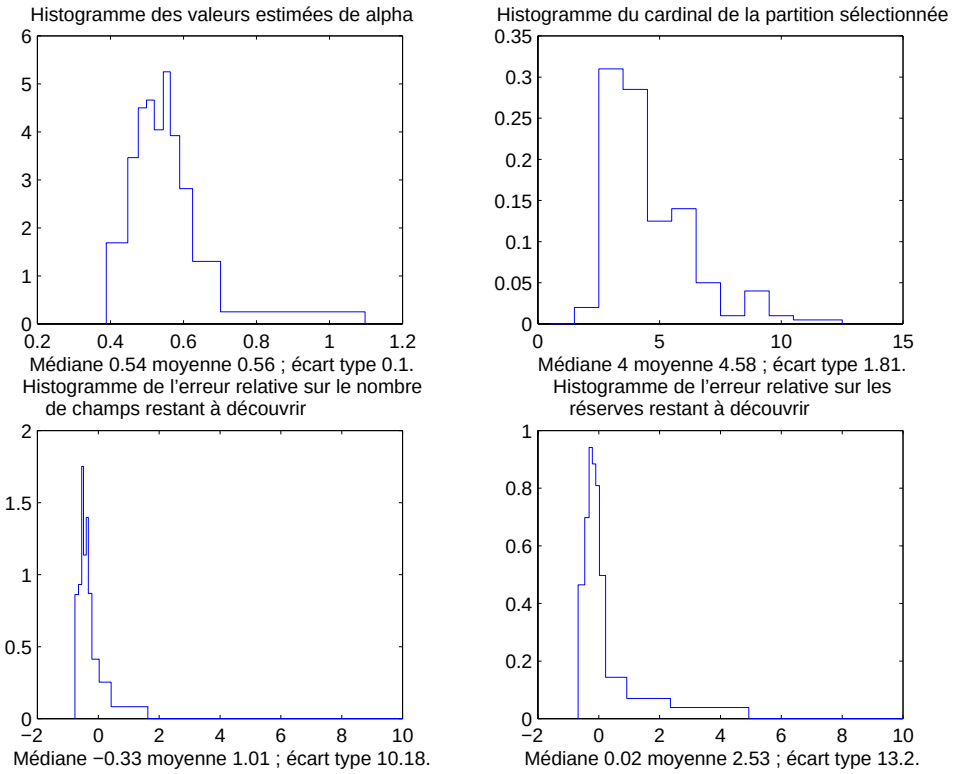

FIG. 8.25 – Estimation d'une densité d'habitat très concentré, sous contrainte de monotonie en ω sur partitions en intervalles de longueur régulière. Protocole de 200 simulations.

Estimation sous contrainte,
partitions en intervalles de longueur irrégulière

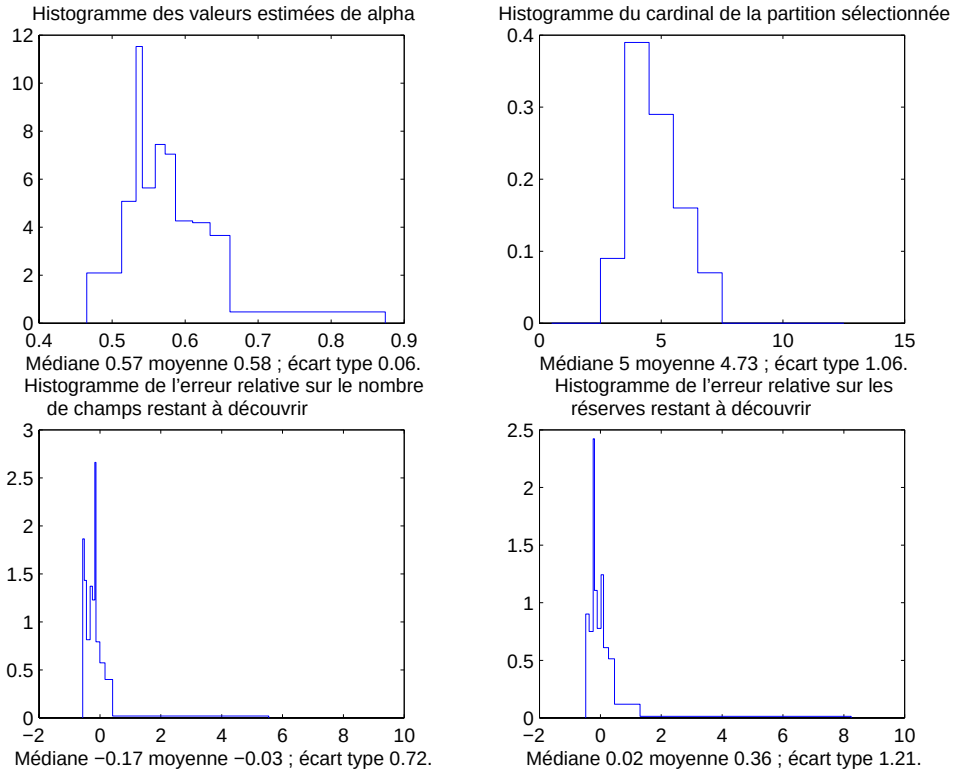

FIG. 8.26 – Estimation d'une densité d'habitat très concentré, sous contrainte de monotonie en ω sur partitions en intervalles de longueur irrégulière. Protocole de 100 simulations.

Estimation sous contrainte,
partitions en intervalles de fréquence régulière

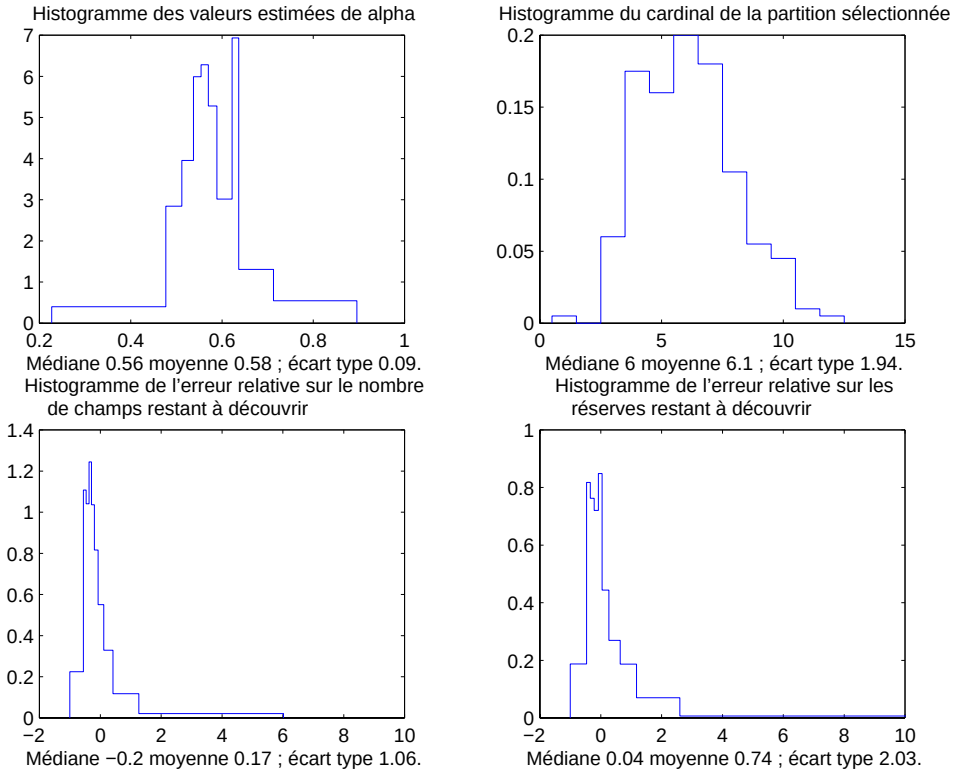

FIG. 8.27 – Estimation d'une densité d'habitat très concentré, sous contrainte de monotonie en ω sur partitions en intervalles de fréquence régulière. Protocole de 200 simulations.

Estimation sous contrainte,
partitions en intervalles de fréquences irrégulières

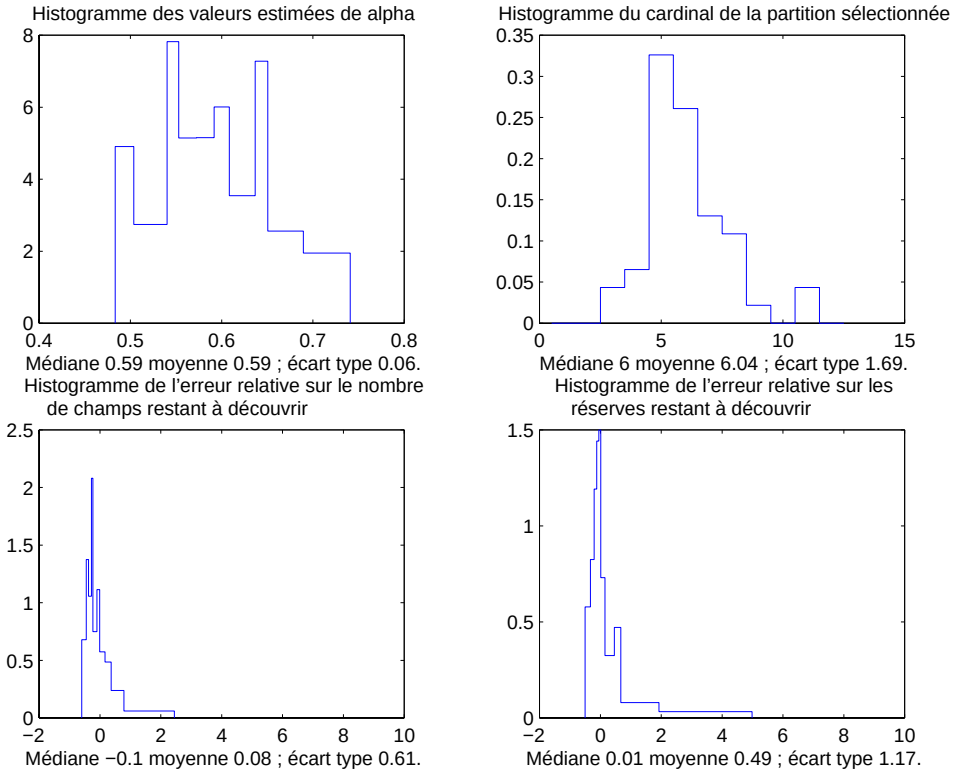

FIG. 8.28 – Estimation d'une densité d'habitat très concentré, sous contrainte de monotonie en ω sur partitions en intervalles de fréquences irrégulières. Protocole de 60 simulations.

8.2.2 Habitat concentré

Rappel : les caractéristiques de la densité $f_{\alpha,\omega}$ à estimer sont les suivantes :
- $\alpha = 0{,}8$;
- effectif de la population parente : $N = 2000$;
- effectif de la population observée : $n = 300$;
- tirage successif biaisé par effet taille directement proportionnel à la taille.

Estimation sous contrainte,
partitions en intervalles de longueur régulière

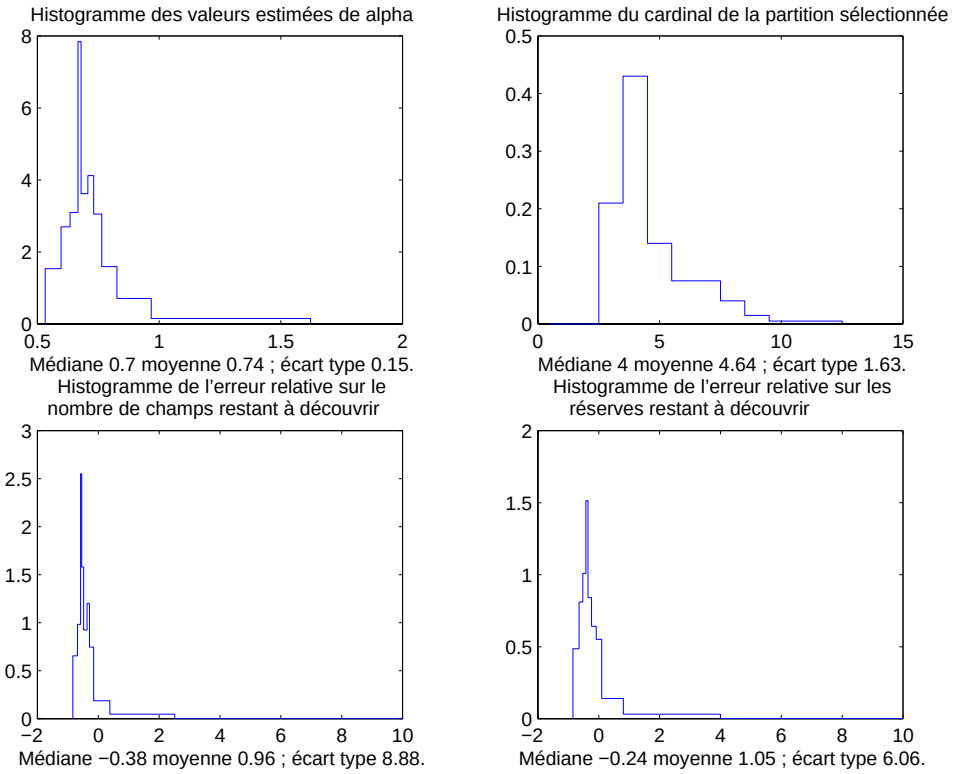

FIG. 8.29 – Estimation d'une densité d'habitat concentré, sous contrainte de monotonie en ω sur partitions en intervalles de longueur régulière. Protocole de 200 simulations.

Estimation sous contrainte,
partitions en intervalles de longueur irrégulière

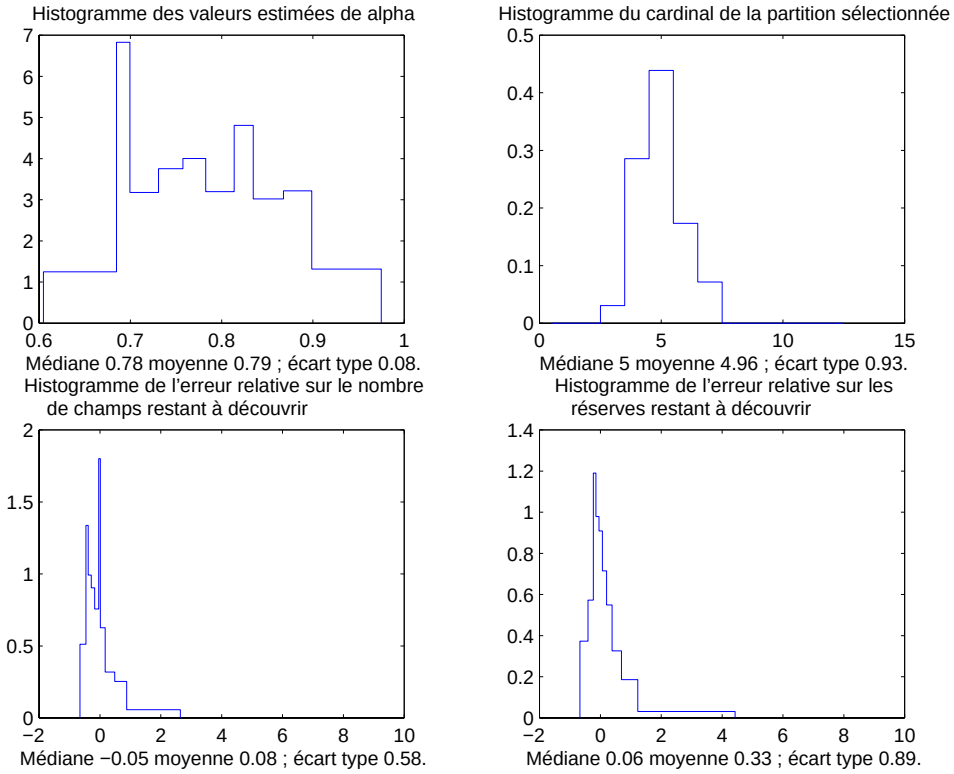

FIG. 8.30 – Estimation d'une densité d'habitat concentré, sous contrainte de monotonie en ω sur partitions en intervalles de longueur irrégulière. Protocole de 100 simulations.

Estimation sous contrainte,
partitions en intervalles de fréquence régulière

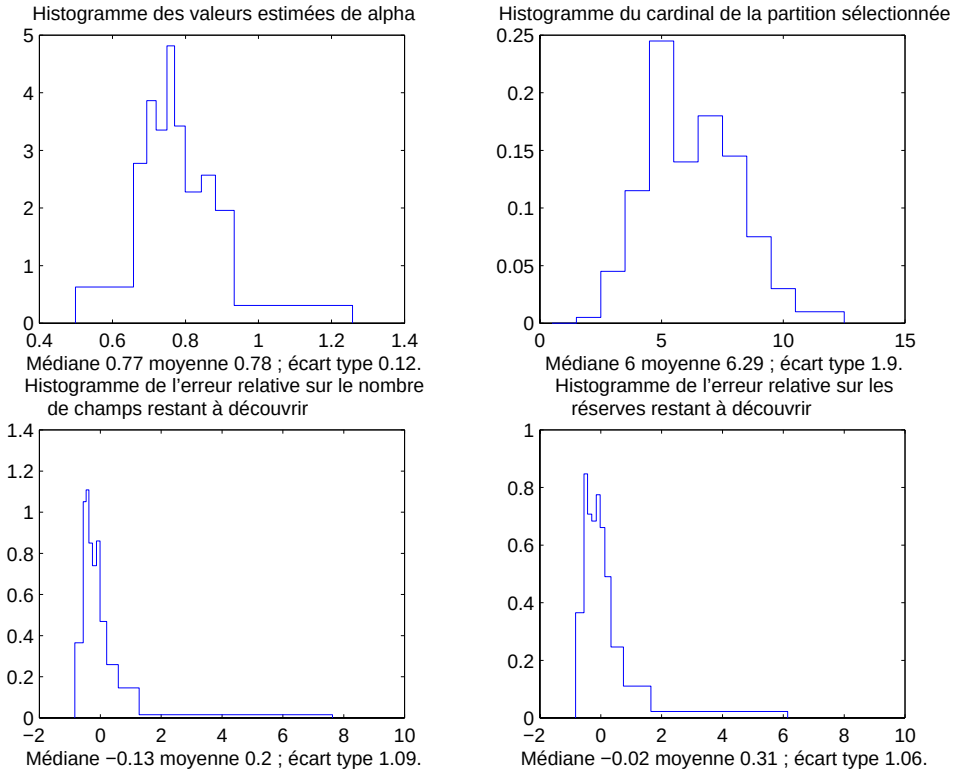

FIG. 8.31 – Estimation d'une densité d'habitat concentré, sous contrainte de monotonie en ω sur partitions en intervalles de fréquence régulière. Protocole de 200 simulations.

Estimation sous contrainte,
partitions en intervalles de fréquences irrégulières

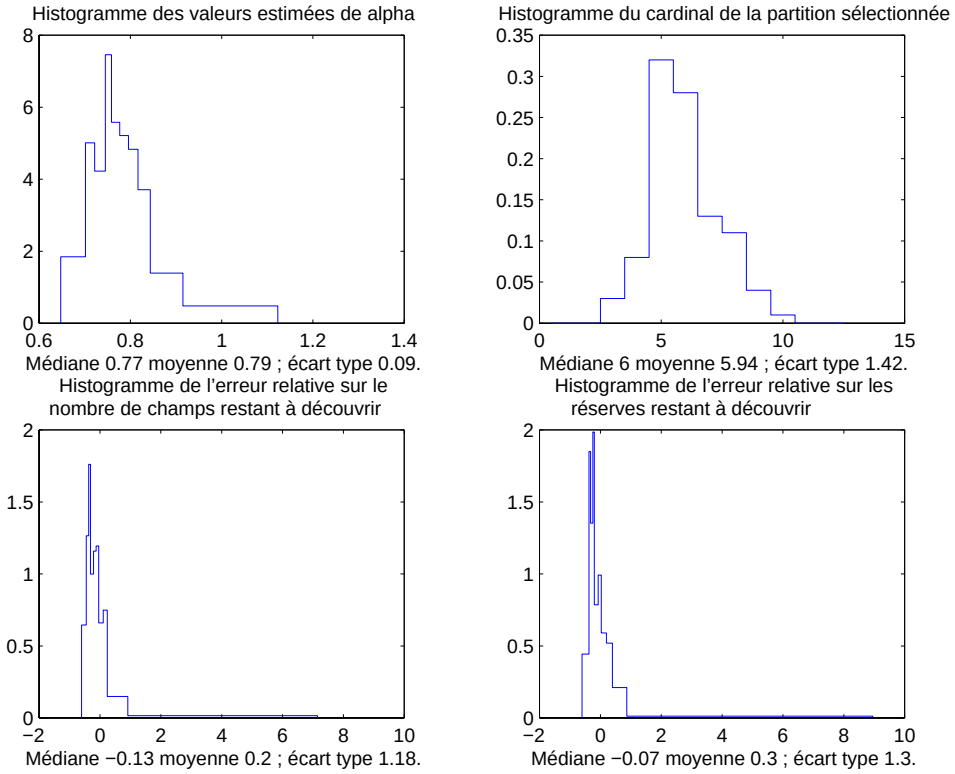

FIG. 8.32 – Estimation d'une densité d'habitat concentré, sous contrainte de monotonie en ω sur partitions en intervalles de fréquences irrégulières. Protocole de 100 simulations.

8.2.3 Habitat dispersé

Rappel : les caractéristiques de la densité $f_{\alpha,\omega}$ à estimer sont les suivantes :
- $\alpha = 1{,}2$;
- effectif de la population parente : $N = 2000$;
- effectif de la population observée : $n = 300$;
- tirage successif biaisé par effet taille directement proportionnel à la taille.

Estimation sous contrainte,
partitions en intervalles de longueur régulière

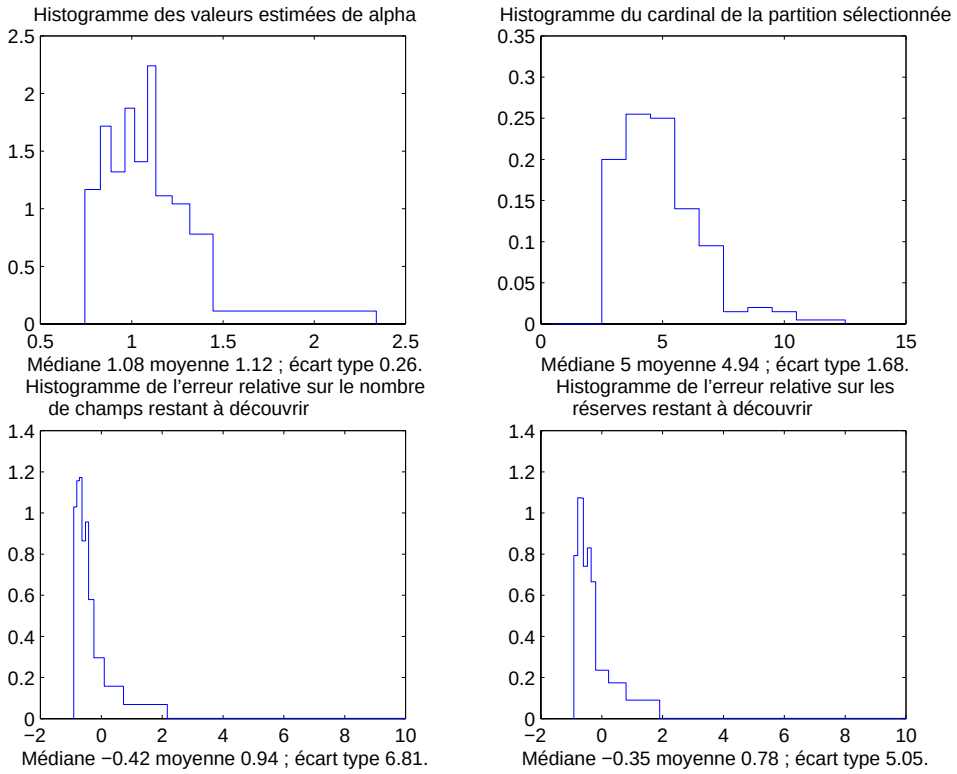

FIG. 8.33 – Estimation d'une densité d'habitat dispersé, sous contrainte de monotonie en ω sur partitions en intervalles de longueur régulière. Protocole de 200 simulations.

Estimation sous contrainte,
partitions en intervalles de longueur irrégulière

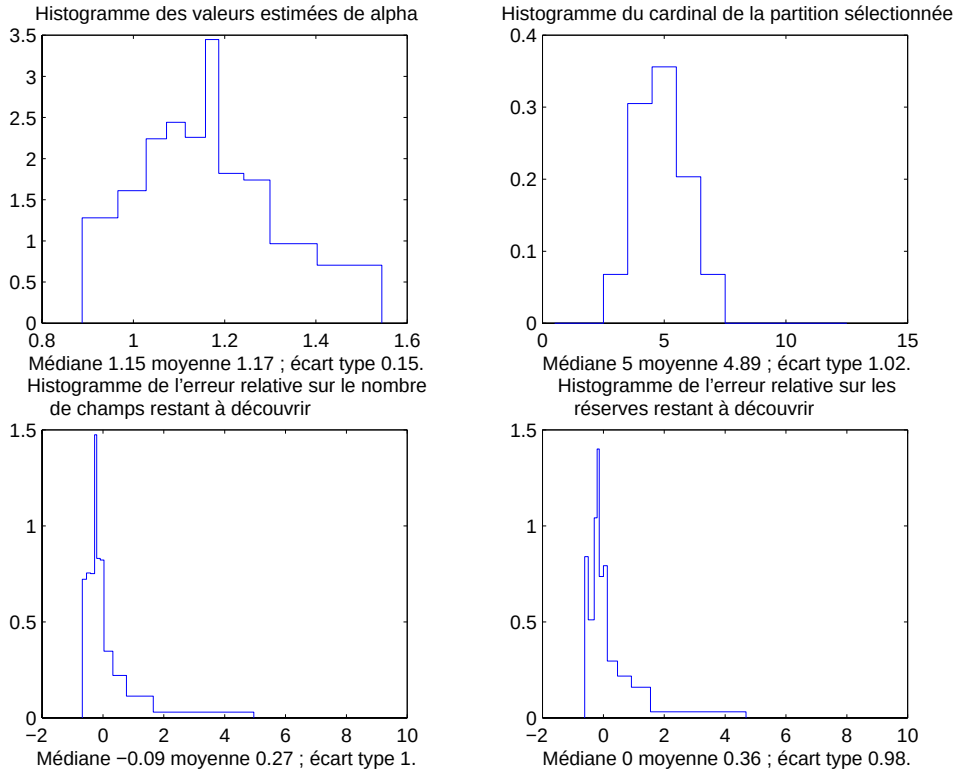

FIG. 8.34 – Estimation d'une densité d'habitat dispersé, sous contrainte de monotonie en ω sur partitions en intervalles de longueur irrégulière. Protocole de 100 simulations.

Estimation sous contrainte,
partitions en intervalles de fréquence régulière

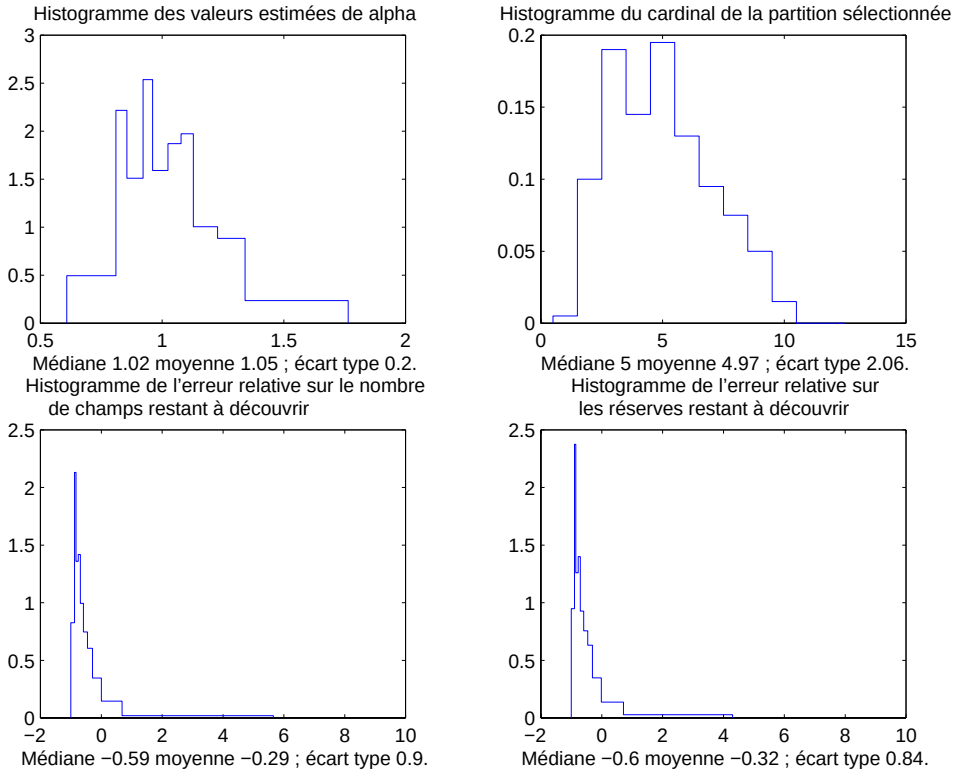

FIG. 8.35 – Estimation d'une densité d'habitat dispersé, sous contrainte de monotonie en ω sur partitions en intervalles de fréquence régulière. Protocole de 200 simulations.

Estimation sous contrainte,
partitions en intervalles de fréquences irrégulières

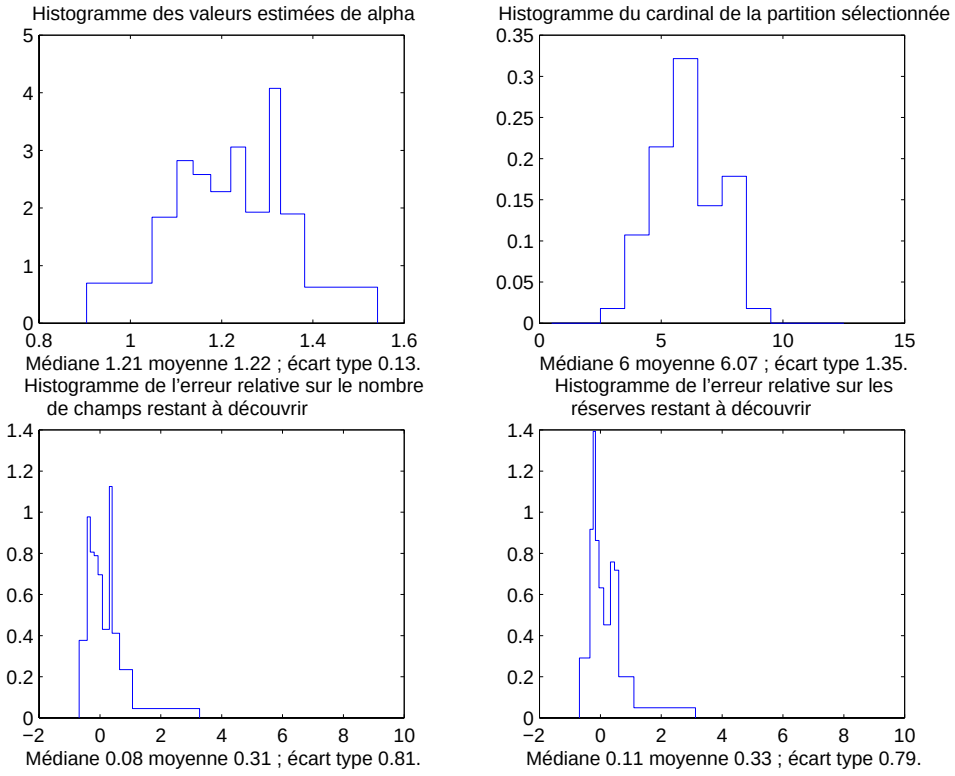

FIG. 8.36 – Estimation d'une densité d'habitat dispersé, sous contrainte de monotonie en ω sur partitions en intervalles de fréquences irrégulières. Protocole de 60 simulations.

Bibliographie

[1] WPC 1997, *Beijing world petroleum congress 1997*, SPE Press releases, 1997.

[2] H Akaïke, *Information theory and an extension of the maximum likelyhood principle. in p. petrov and f. csaki (eds.)*, Proceedings 2nd international symposium on information theory **Akademia Kiado, Budapest** (1973), 267–281.

[3] Nathalie Alazard, Jean Laherrère, and Alain Perrodon, *Réserves et ressources de pétrole et de gaz des pays méditerranéens*, Revue de l'énergie **441** (1992).

[4] G. Andreatta and G. M. Kaufman, *Estimation of finite population properties when sampling is without replacement and proportional to magnitude*, Journal of the american statistical association **81 (395)** (1986), 657–666.

[5] A. Araujo and E. Giné, *The central limit theorem for real and banach valued random variables*, John Wiley and Sons Inc., New York, 1980.

[6] Patrice Ardilly, *Les techniques de sondage*, Technip, Paris, 1994.

[7] K. B. Athreya, *Bootstrap of the mean in the infinite variance case*, The annals of statistics **15** (1987), 724–631.

[8] Denis Babusiaux, *Décision d'investissement et calcul économique dans l'entreprise*, Economica & Technip, Paris, 1990.

[9] Philippe Barbe and Patrice Bertail, *The weighted bootstrap*, Springer-Verlag, New York, 1995.

[10] Andrew Barron and C. Sheu, *Approximation of density functions by sequences of exponential families*, The Annals of Statistics **19** (1991), 1054–1347.

[11] Pierre Bauquis, *Un point de vue sur les besoins et les approvisionnements en énergie à l'horizon 2050.*, Revue de l'Énergie **509** (1999).

[12] Patrice Bertail and Pierre Combris, *Bootstrap généralisé d'un sondage*, Annales d'économie et de statistique **46** (1997), 49–83.

[13] Peter J. Bickel, Vijayan N. Nair, and Paul C. C. Wang, *Nonparametric inference under biased sampling from a finite population*, The Annals of Statistics **20 (2)** (1992), 853–878.

[14] Lucien Birgé, *Approximation dans les espaces métrique et théorie de l'estimation*, Z. Wahrscheinlichkeitstheorie Verw. Geb. **65** (1983), 181–237.

[15] Jean-Noël Boulard, Revue de l'énergie (1999).

[16] Olivier Bousquet, *A bennett concentration inequality and its application to suprema of empirical processes*, Comptes rendus de l'académie des sciences **Série I, 334** (2002), 495–500.

[17] D. A. Brobst and Pratt W. P., *United states mineral resources*, US Government Printing Office (1973), 1–8.

[18] Colin Campbell and Jean Laherrère, *The end of cheap oil*, Scientific American **March** (1998), 80–85.

[19] Colin Campbell, Jean Laherrère, and Alain Perrodon, *The world's non-conventional oil and gas*, Petroleum economist **March** (1998).

[20] Gwenaelle Castellan, *Sélection d'histogrammes ou de modèles exponnentiels de polynômes par morceaux à l'aide d'un critère de type akaike*, Thèse de doctorat, Université de Paris-Sud, 2000.

[21] ———, *Density estimation via exponential model selection*, Annales de l'IHP **à paraître** (2002).

[22] CEG IFP (Vincent Lepez co auteur), *La recherche et la production du pétrole et du gaz*, Technip, Paris, 2001.

[23] Clifford B. Cordy, *An extension horvitz-thompson theorem point sampling from continuous universe*, Statistics and probability letters **18** (1993), 353–362.

[24] B. R. Crain, *Estimation of distributions using orthogonal expansions*, The annals of statistics **2** (1974), 453–463.

[25] Robert A. Crovelli and Christopher C. Barton, *Fractals and the pareto distribution applied to petroleum accumulation-size distributions*, USGS Open-File Report **91-18** (1993).

[26] I. Csiszár, *I-divergence geometry of probability distributions and minimization problems*, The Annals of Probability **3** (1975), 146–158.

[27] Pierre Delfiner, *Estimation du potentiel de découverte dans un bassin pétrolier*, Note interne TotalFinaElf (2000).

[28] Gérard Demaison and Bradley J. Huizinga, *Genetic classification of petroleum systems*, The American Association of Petroleum Geologists Bulletin **75 (10)** (1991), 1626–1643.

[29] Wietse Dol, Ton Steerneman, and Tom Wansbeek, *Matrix algebra and sampling theory : the case of the horvitz-thompson estimator*, Linear algebra and its applications **237/238** (1996), 225–238.

[30] William Feller, *An introduction to probability theory and its applications. vol. ii.*, John Wiley and Sons Inc., New York, 1971.

[31] Kevin F. Forbes and Ernets M. Zampelli, *Technology and the exploration success rate in the u.s. offshore*, The Energy Journal **21-1** (2000), 109–120.

[32] Uriel Frisch and Didier Sornette, *Extreme deviations and applications*, Journal of Physics **I France 7** (1997), 1155–1171.

[33] Evarist Giné and Joel Zinn, *Necessary conditions for the bootstrap of the mean*, The annals of statistics **17 (2)** (1989), 684–691.

[34] I. J. Good and R. A. Gaskins, *Nonparametric roughness penalties for probability densities*, Biometrika **58** (1971), 155–277.

[35] Louis Gordon, *Estimation for large successive samples with unknown inclusion probabilities*, Advances in applied mathematics **14** (1993), 89–122.

[36] Christian Gourieroux, *Mesures d'inégalités, de pauvreté, de concentration*, ensae, Malakoff, 1993.

[37] Antonion Aznar Grasa, *Econometric model selection : a new approach*, Kluwer, 1989.

[38] Muhammad Hanif and K. R. W. Brewer, *Sampling with unequal probabilities without replacement : a review*, International statistical review **48** (1980), 317–335.

[39] Lars Holst, *Some limit theorems with applications in sampling theory*, The Annals of Statistics **1 (4)** (1973), 644–658.

[40] D. G. Horvitz and D. J. Thompson, *A generalization of sampling without replacement from a finite universe*, Journal of the american statistical association **47 (260)** (1952), 663–685.

[41] J. C. Houghton, *Use of the truncated shifted pareto distribution in assessing size distribution of oil and gas fields*, Mathematical geology **20 - 8** (1988), 907–937.

[42] G. M. Kaufman, *Statistical decision and related techniques in oil and gas exploration : Englewood cliffs*, N. J., Prentice-Hall, Paris, 1963.

[43] H. D. Klemme, *Field size distribution related to basin characteristics*, Oil and gas journal **December** (1983), 168–176.

[44] Alexey E. Kontorovich, Viktor I. Dyomin, and Valery R. Livshits, *Size distribution and dynamics of oil and gas fielddiscoveries in petroleum basins*, The American Association of Petroleum Geologists Bulletin **85 (9)** (2001), 1609–1922.

[45] Jean Laherrère, *Comment estimer le potentiel résiduel d'un bassin pétrolier ? lognormal ou fractal ?*, Manuscrit non publié (1991).

[46] ———, *Nouvelle approche des réserves ultimes. application aux réserves de gaz des états-unis*, Pétrole et Techniques **397** (1994).

[47] ———, *Distributions de type "fractal parabolique" dans la nature*, Comptes rendus de l'académie des sciences de Paris **322 (II a)** (1996), 535–541.

[48] ———, *Évolution des réserves mondiales d'hydrocarbures*, Pétrole et techniques **416** (1999), 61–79.

[49] Jean Laherrère and Didier Sornette, *Stretched exponential distributions in nature and economy : "fat tails" with characteristic scales*, European physical journal **B (2)** (1998), 525–539.

[50] Émilie Lebarbier, *Model selection for the detection of multiple change points in the mean of a random process via a new heuristic method for estimating the penalty function*, Technical report - Université de Paris-Sud (2002).

[51] P. J. Lee and P. C. C. Wang, *Probabilistic formulation of a method for the evaluation of petroleum resources*, Journal of mathematical geology **15** (1983), 163–181.

[52] Vincent Lepez, *Modelling the field size distribution of a petroleum system, lognormal or fractal ? a unifying aproach*, Soumis (2001).

[53] Vincent Lepez and Gentiane Mandonnet, *Problèmes de robustesse dans l'estimation des réserves ultimes de pétrole conventionnel*, Cahiers du CEG - IFP **39** (1999).

[54] Maretheux Louis, *L'épuisement des mines de pétrole aux États-unis*, La technique moderne **XI - 9** (1919), 416.

[55] C. Mallows, *Some comments on cp*, Technometrics **15** (1973), 661–675.

[56] Benoit Mandelbrodt, *Fractales hasard et finances* $3^{ème}$ *éd.*, Flammarion, Paris, 2001.

[57] Pascal Massart, *About the constants in talagrand's concentration inequalities for empirical processes*, The annals of probability **to appear** (1998).

[58] ———, *Some applications of concentration inequalities to statistics*, Annales de la faculté des sciences de Toulouse **IX (2)** (2000), 245–303.

[59] ———, *Heuristique de pente*, Machins d'Orsay **to appear** (2001).

[60] Jean Masseron, *L'économie des hydrocarbures (4. ed.)*, Technip, Paris, 1991.

[61] V. E. McKelvey, *Mineral resource estimates and public policy*, American Scientist **60 (1)** (1972), 32–40.

[62] J. Meisner and F. Demirmen, *The creaming method : a bayesian procedure to forecast future oil and gas discoveries in mature exploration provinces*, Journal of the royal statistical society **144 (A)** (1981), 1–31.

[63] V. Pareto, *Cours d'économie politique*, Rouge et Cie, Lausanne and Paris, 1897.

[64] Alain Perrodon, *Essai de classification des bassins sédimentaires*, Sciences de la terre **16** (1972), 197–227.

[65] ———, *Géodynamique pétrolière : genèse et répartition des gisements d'hydrocarbures*, Masson, Elf-Aquitaine, Paris, 1980.

[66] ———, *Sedimentary basin geodynamics and "petroleum systems"*, Bulletin des centres de recherches en exploration-production d'Elf-Aquitaine **7** (1983), 645–676.

[67] ———, *Vers les réserves ultimes d'hydrocarbures conventionnels*, Bulletin des centres de recherches en exploration-production d'Elf-Aquitaine **15** (1991).

[68] ———, *Quel pétrole demain ?*, Technip, Paris, 1999.

[69] Petroconsultants, *Field data update*, Cd rom, Petroconsultants, 1998.

[70] D. Pollard, *Convergence of stochastic processes*, Springer Verlag, New York, 1984.

[71] Emmanuel Rio, *Une inégalité de bennett pour les maxima de processus empiriques*, Annales de l'IHP **To appear** (2002).

[72] Tim Robertson, F. T. Wright, and R. L. Dykstra, *Order restricted statistical inference*, John Wiley and sons Inc., New York, 1988.

[73] R. Rockafellar, *Convex analysis*, Princeton University Press, Princeton, 1970.

[74] Bengt Rosén, *Asymptotic theory for successive sampling with varying probabilities without replacement, i*, The Annals of Mathematical Statistics **43 (2)** (1972), 373–397.

[75] ———, *Asymptotic theory for successive sampling with varying probabilities without replacement, ii*, The Annals of Mathematical Statistics **43 (3)** (1972), 748–776.

[76] Gilbert Saporta, *Probabilités, statistiques et analyse des données*, Technip, Paris, 1990.

[77] Jiayang Sun and Michael Woodroofe, *Semi-parametric estimates under biased sampling*, Statistica sinica **7** (1997), 545–575.

[78] Michel Talagrand, *New concentration inequalities in product spaces*, Invent. Math. **126(3)** (1996), 505–563.

[79] Philippe Tassi, *Statistique non-paramétrique et robustesse*, Economica, Paris, 1987.

[80] USGS, *World petroleum assesment 2000*, Cd rom, United States Geological Survey, 2000.

[81] Jean Vignes, *Algorithmes numériques – analyse et mise en œuvre 2*, Technip, Paris, 1980.

[82] John J. Wiorkowski, *Estimating volumes of remaining fossil fuel ressources : a critical review*, Journal of the american statistical association **76 (375)** (1981), 534–559.