Ecole nationale supérieure **du pétrole et des moteurs**

Problèmes de robustesse dans l'estimation des réserves ultimes de pétrole conventionnel

Vincent LEPEZ
Gentiane MANDONNET

Centre économie et gestion

INSTITUT FRANÇAIS DU PÉTROLE

Centre Économie et Gestion

Problèmes de robustesse dans l'estimation des réserves ultimes de pétrole conventionnel

Vincent LEPEZ
Gentiane MANDONNET

mars 1999

Cahiers du CEG - n° 32

ÉCOLE DU PÉTROLE ET DES MOTEURS
Centre Économie et Gestion
228-232, avenue Napoléon Bonaparte
92852 RUEIL-MALMAISON CEDEX
télécopieur : 01 47 52 70 66 - téléphone : 01 47 52 64 10

La collection "Cahiers du CEG" est un recueil de présentations de travaux réalisés au Centre Économie et Gestion de l'École du Pétrole et des Moteurs, Institut Français du Pétrole, travaux de recherche ou notes de synthèse. Elle a été mise en place pour permettre la diffusion de ces travaux, parfois sous une forme encore provisoire, afin de susciter des échanges de points de vue sur les sujets abordés.

Les opinions émises dans les textes publiés dans cette collection doivent être considérées comme propres à leurs auteurs et ne reflètent pas nécessairement le point de vue de l'École du Pétrole et des Moteurs ou de l'IFP.

Pour toute information complémentaire, prière de contacter :

Nathalie ALBA-SAUNAL *tél. :* **01 47 52 64 10**

The "Cahiers du CEG" is a collection of articles carried out at the Center for Economics and Management of the IFP School, Institut Français du Pétrole. It is designed to promote an exchange of ideas on the topics covered.

The opinions expressed are the sole responsability of the author(s) and do not necessarily reflect the views of the IFP School or IFP.

For any additional information, please contact :

Nathalie ALBA-SAUNAL *tél. :* **+ 33 1 47 52 64 10**

Institut Français du Pétrole

PROBLÈMES DE ROBUSTESSE DANS L'ESTIMATION DES RÉSERVES ULTIMES DE PÉTROLE CONVENTIONNEL

par

Vincent LEPEZ et Gentiane MANDONNET

Cet article présente les conclusions du stage de fin de DEA que nous avons effectué au sein de l'Institut Français du Pétrole. Ce travail a été réalisé sous la direction de M. Pascal MASSART, M. Yanick HEURTEAUX, respectivement Professeur et Maître de Conférence de l'Université Paris XI–Orsay, d'une part, et Mme Nathalie ALAZARD - TOUX, Ingénieur à la direction Stratégie - Économie - Plan de l'IFP, d'autre part.

RÉSUMÉ – Depuis le début des années 90, de nombreux travaux sont effectués pour fournir une approche probabiliste de l'estimation des réserves ultimes récupérables de pétrole conventionnel. Sur la base de ces études, plusieurs formules d'estimation ont été proposées sans que l'on ait, jusqu'alors, donné d'intervalles de confiance attenants.

C'est dans ce cadre probabiliste que notre travail s'inscrit. Nous cherchons à mettre en place un modèle statistique propre de la distribution de la taille des champs au sein d'un système pétrolier. Ce modèle nous permet de construire les outils nécessaires à l'évaluation de la qualité des formules d'estimation qui existent. Nous mettons alors en évidence leur manque de fiabilité, en fournissant notamment des intervalles de confiance dans lesquelles elles s'inscrivent. Nous proposons ensuite des solutions pour tenter d'améliorer la robustesse de ce modèle et les estimations de ses paramètres.

Les données avec lesquelles nous effectuons les applications numériques sont les chiffres 1997 des réserves dites prouvées sur le système pétrolier défini par l'offshore de la mer du Nord (les 3 grabens de la mer du Nord) — Source IFP.

INTRODUCTION

Cet article a pour objet de présenter une démarche statistique rigoureuse de l'estimation des réserves ultimes de pétrole conventionnel au niveau d'un système pétrolier[1]. Nous y étudions, en particulier, la fiabilité de quelques estimateurs qui ont été proposés au cours de ces dernières années[2].

Bien que placées naturellement dans un schéma où l'avenir est incertain, les prévisions des géologues et des économistes de l'énergie se résument généralement à un chiffre brut. Or, travailler en avenir incertain nécessite que les prévisions soient fournies sous la forme d'un intervalle de confiance, dans lequel on est sûr, à 50 %, 75 % ou 95 %, que le résultat final se trouvera.

Le travail du statisticien est de donner des intervalles de confiance objectifs, c'est-à-dire basés sur la rigueur d'une théorie mathématique et non sur la seule intuition de l'expert.

Depuis le début des années 90, quelques modèles probabilistes de la distribution de la taille des champs de pétrole ont vu le jour (lognormal, fractal-linéaire, etc.). On a donc commencé à avancer l'hypothèse qu'au sein d'un système pétrolier, il existe une loi de probabilité selon laquelle ces tailles sont réparties. En déduisant cette loi de l'observation des champs déjà découverts, on peut extrapoler et ainsi espérer prévoir le volume de ce qui reste à découvrir.

La **première partie** de ce document présente le modèle fractal-linéaire, introduit par Jean Laherrère en 1991. Nous montrons en détail que l'on peut l'interpréter comme étant un modèle statistique paramétrique, dont la loi sous-jacente est une loi de Pareto.

Nous nous intéressons, dans la **deuxième partie**, à l'étude de la mèthode d'estimation des réserves de Jean Laherrère, basée sur le modèle fractal-linéaire. Nous montrons que les estimateurs qu'il propose sont soumis à une extrême variabilité due au comportement erratique des plus gros champs.

Dans La **troisième partie** de cet article, nous donnons, dans le cas de l'offshore de la mer du Nord, les intervalles de confiance gigantesques auxquels conduisent les formules étudiées précédemment.

Après avoir proposé quelques voies d'amélioration de la qualité de l'estimation, nous consacrons une **quatrième partie** aux raffinements possibles du modèle. Nous montrons qu'une approche dite fractale-parabolique, bien que plus complexe, n'est pas mieux adaptée aux données. Nous concluons enfin cet article en proposant un plan d'échantillonnage prometteur pour le modèle fractal-linéaire.

---

[1] Zone géographique qui présente une structure géologique homogène.

[2] Les recherches n'étant pas encore suffisamment avancées, nous avertissons le lecteur que nous ne fournissons aucune prévision. Par suite, nous ne rentrons en aucune façon dans le débat qui existe, en matière de réserves, entre optimistes et pessimistes.

1 Le Modèle Fractal-Linéaire, ou Modèle Pareto

1.1 La notion d'Habitat

On peut classer les champs d'un système pétrolier donné par ordre décroissant de taille de leurs réserves prouvées. Un champ est donc déterminé de manière unique par son rang dans ce classement.

On trace alors un diagramme portant en abscisse le logarithme en base 10 du rang de ces champs et en ordonnée le logarithme en base 10 de leur taille. Nous qualifions un tel diagramme de *diagramme log-log*. La figure 1 est le diagramme log-log dans lequel se répartissent les champs de la mer du Nord.

On définit la notion d'*habitat* d'un système pétrolier par rapport à la pente de la droite de régression que l'on peut tracer sur les premiers points de son diagramme log-log. Soit k l'habitat, égal à la valeur absolue de cette pente. L'habitat est :

$$\begin{cases} \textit{dispersé} & \text{si} \quad k < 1 \\ \textit{normal} & \text{si} \quad k = 1 \\ \textit{concentré} & \text{si} \quad k > 1 \end{cases}$$

L'interprétation de l'habitat est la suivante : **plus la valeur absolue k de la pente est forte et moins il existe de gros champs** (c'est le cas que le plus souvent rencontré). Ces derniers *concentrent* l'essentiel de la masse des réserves. A l'inverse, **si k est faible alors il existe de nombreux champs de grande taille**, sur lesquels la masse des réserves sera *dispersée*. Le terme habitat *normal* n'est que conventionnel car, en pratique, le k mesuré n'est jamais strictement égal à 1.

Le nombre de points à prendre en compte pour tracer la droite de régression varie selon les auteurs. Alain Perrodon définit l'habitat sur les 10 premiers, tandis que Jean Laherrère trace la droite de régression sur l'ensemble du diagramme et la contraint à passer par le premier ou second plus gros champ. Nous supposerons que la droite passe par le plus gros champ. Cela nous permet de construire une loi de probabilité selon laquelle la taille des champs serait distribuée..

1.2 Le Modèle Fractal-Linéaire et la loi de Pareto

Par définition, la droite de régression est celle qui approche au mieux un nuage de points, au sens des moindres carrés. Dès lors, en première approximation, on peut modéliser la distribution des tailles des champs par une droite dans un diagramme log-log. Ce modèle est appelé *fractal-linéaire* (Laherrère — 1991). **Dans cet article nous ne faisons pas le lien entre fractals en tant qu'objets mathématiques et la forme de la distribution de la taille des champs d'un système pétrolier. En effet, pour peu qu'il existe, ce lien n'est pas du tout clairement établi.** Il semble, au jour d'aujourd'hui, que l'utilisation du terme "fractal-linéaire" relève plus d'un effet de mode que d'une vérité mathématique. Aussi préférons-nous la désignation

FIGURE 1

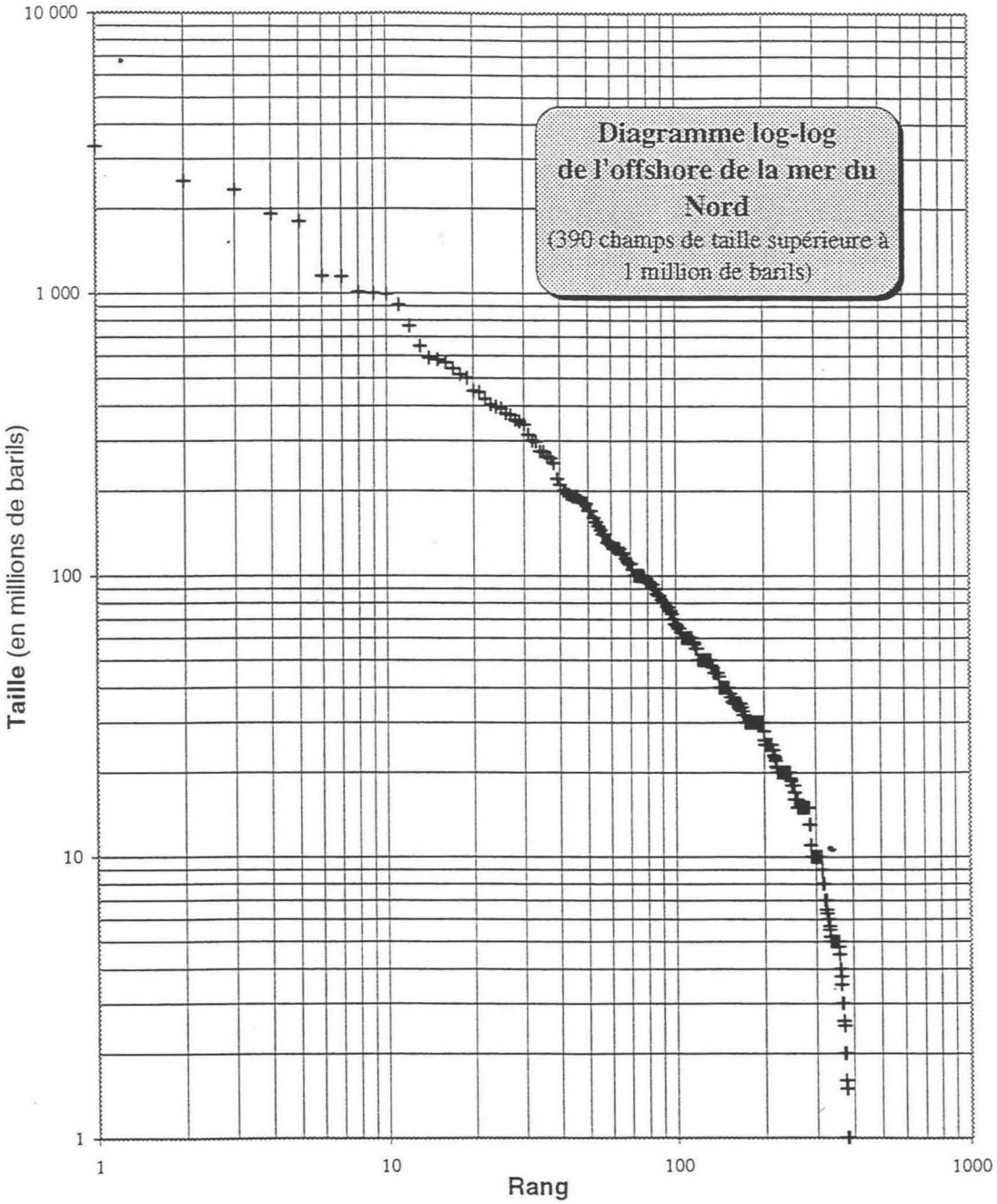


Diagramme log-log
de l'offshore de la mer du Nord
(390 champs de taille supérieure à 1 million de barils)
Taille (en millions de barils)
10 000
1 000
100
10
1
1
10
100
1000
Rang

"modèle Pareto" plutôt que "modèle fractal-linéaire", mais utiliserons l'une ou l'autre indifféremment.

A partir de maintenant, nous nous plaçons au niveau d'un système pétrolier. Soit N, inconnu en pratique, le nombre total de champs du système. **On ne considère que les champs d'une taille supérieure à une certaine unité correspondant à un minimum de rentabilité économique** : 25 millions de barils par exemple.

Soient $X_{(1)}, \ldots, X_{(N)}$ les tailles des champs du système ordonnées en décroissant. On fait l'hypothèse que le nuage de points de coordonnées $(\log i\,;\, \log X_{(i)})_{i\in\mathbb{N}_N}$ est linéaire de pente $-k$ dans un diagramme log-log. Ceci s'écrit :

$$\log X_{(i)} = \log X_{(1)} - k \log i$$

ou encore

$$X_{(i)} = X_{(1)} \times i^{-k}$$

Nous allons voir que la loi de Pareto intervient de façon naturelle dans cette modélisation.

En pratique, seuls $n \leq N$ champs sont découverts et constituent une sous-famille $X_1, \ldots, X_n$ issue de l'échantillon $X_1, \ldots, X_N$. Dans notre modèle, ce dernier échantillon est i.i.d. d'une certaine loi de probabilité que nous cherchons à déterminer au moyen de sa fonction de queue de répartition Q.

La loi empirique de l'échantillon $X_1, \ldots, X_N$ est donnée par sa fonction de queue de répartition empirique Q_N :

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

La quantité $Q_N(x)$ mesure la fréquence des observations supérieures au réel x. En particulier, on a $Q_N(X_{(i)}) = i/N$, donc :

$$X_{(i)} = X_{(1)} \times (N Q_N(X_{(i)}))^{-k}$$

d'où

$$Q_N(X_{(i)}) = \left(\frac{X_{(i)}}{X_{(N)}} \right)^{-1/k}$$

Comme le minimum de rentabilité économique $X_{(N)}$ est égal à une unité, il vient :

$$\forall i \in \mathbb{N}_N \qquad Q_N(X_{(i)}) = X_{(i)}^{-1/k}$$

Le théorème de Glivencko–Cantelli nous assure maintenant la convergence uniforme de la suite de fonctions $(Q_i)_{i\in\mathbb{N}}$ vers Q sur $\mathbb{R}$. Ainsi, asymptotiquement, on a $Q_N \simeq Q$. Donc, pour tout i dans $\mathbb{N}_N$ on a $Q(X_{(i)}) \simeq X_{(i)}^{-1/k}$. Comme $X_{(i)}$ peut

prendre n'importe quelle valeur dans $[1\,;+\infty[$, on peut conclure que $X_1, \ldots, X_N$ est un échantillon i.i.d. de la v.a. X dont la loi est donnée par la formule :

$$Q(x) = \mathbb{P}(X \geq x) = \begin{cases} 1 & \text{si} \quad x \in]-\infty\,;1[\\ x^{-1/k} & \text{si} \quad x \in [1\,;+\infty[\end{cases}$$

Cette loi de probabilité est connue sous le nom de *loi de Pareto*. Dans la suite nous poserons $\alpha = 1/k$. Nous parlerons alors de loi de Pareto d'exposant α. Nous noterons $\Pi(\alpha)$ cette loi et Q_α sa fonction de queue de répartition.

En résumé, notre modèle est le suivant :

> *Un système pétrolier est constitué de N champs de taille supérieure à une unité, correspondant à un minimum de rentabilité économique. Les tailles $X_1, \ldots, X_N$ de ces champs sont considérées comme des v.a. i.i.d. de loi de Pareto d'exposant α. L'ensemble des découvertes représente un tirage $X_1, \ldots, X_n$ sans remise, issu de la famille $X_1, \ldots, X_N$.*
>
> *Les paramètres inconnus du modèle sont N et α, nous devons en donner des estimations*[3]*. Pour ce faire nous disposons des n observations :*
>
> $$X_1 = x_1, \ldots, X_n = x_n$$

La loi de tirage sans remise des $X_1, \ldots, X_n$ parmi les $X_1, \ldots, X_N$ complique fortement le travail d'estimation, car les tirages successifs ne sont plus indépendants. Nous verrons aussi qu'il est nécessaire de tenir compte, dans cette loi, de la tendance qui existe à trouver d'abord les gros champs.

1.3 Propriétés d'intégrabilité de la loi de Pareto

Celles-ci sont liés à la notion d'habitat et conduisent à deux régimes différents.

En dérivant la fonction de queue de répartition Q_α de la loi de Pareto d'exposant α, on obtient sa densité π_α :

$$\pi_\alpha(x) = -Q'_\alpha(x) = \alpha x^{-\alpha-1} \times \mathbb{I}_{[1\,;+\infty[}(x)$$

L'expression précédente montre que X de loi $\Pi(\alpha)$ **est intégrable si et seulement si l'habitat est dispersé** ($\alpha > 1$) :

$$\alpha > 1 \Rightarrow \mathbb{E}(X) = \int_{\mathrm{R}} x\pi_\alpha(x)\,dx = \int_1^{+\infty} \alpha x^{-\alpha}\,dx = \frac{\alpha}{\alpha - 1}$$
$$\alpha \leq 1 \Rightarrow \mathbb{E}(X) = +\infty$$

Le cas concentré ($\alpha < 1$), le plus couramment observé sur des données réelles, est particulièrement délicat. En effet, pour $\alpha < 1$, la loi $\Pi(\alpha)$ ne suit pas la Loi des

[3] Une discussion sur l'estimation de α est menée en Annexe.

Grands Nombres. Nous verrons, dans la seconde partie, que cela aura une grande incidence sur la variabilité des estimations du cumul des réserves.

Dans la partie suivante, Nous étudions les estimateurs du cumul des réserves de Jean Laherrère (1991). En particulier, nous en donnons une interprétation probabiliste à l'aide du modèle Pareto. Grâce à celle-ci, nous montrons quelles sont les limites de cette méthode en donnant des intervalles de confiance auxquels ces formules conduisent.

2 La méthode de Laherrère et ses limites

Nous allons nous appuyer sur les formules que propose Jean Laherrère (1991).

Notre modélisation nécessite de disposer d'estimateurs du nombre total N de champs d'un système pétrolier, avant de proposer des estimateurs du cumul des réserves. L'article de référence ne fait pas mention de ce problème. Nous verrons ensuite que pour le régime $\alpha < 1$ existent des fluctuations quantifiables, mais incontrôlables, qui rendent l'estimateur du cumul proposé inefficace en pratique. Nous verrons enfin, dans les deux cas, que les formules de Jean Laherrère conduisent implicitement à un estimateur de N qui s'avère être très dispersif.

On supposera dorénavant connus α et $X_{(1)}$ (le plus gros champ).

2.1 Estimation du cumul des réserves dans le cas intégrable

L'estimateur de $S_N = X_1 + \cdots + X_N$ que propose Jean Laherrère dans ce cas est :

$$\widehat{S_N} = \frac{\alpha}{\alpha - 1} X_{(1)}^{\alpha}$$

Cette formule vient des approximations que nous avons déjà rencontrées en 1 :

$$X_{(i)} \simeq X_{(1)} \times i^{-1/\alpha}$$

donc

$$\sum_{i=1}^{N} X_i = \sum_{i=1}^{N} X_{(i)} \simeq X_{(1)} \int_1^N \frac{dx}{x^{1/\alpha}}$$

ainsi

$$\sum_{i=1}^{N} X_i \simeq X_{(1)} \times \frac{\alpha}{\alpha - 1} N^{1-1/\alpha}$$

or

$$Q_\alpha(X_{(1)}) = X_{(1)}^{-\alpha} \simeq \frac{1}{N}$$

donc

$$\sum_{i=1}^{N} X_i \simeq \frac{\alpha}{\alpha - 1} X_{(1)} \times X_{(1)}^{\alpha - 1} = \frac{\alpha}{\alpha - 1} X_{(1)}^{\alpha}$$

On peut aussi noter que l'on a $\widehat{S_N} \simeq N \times \mathbb{E}(X_1)$. Ceci n'est autre qu'une formulation de la Loi des Grands Nombres.

Mise à part la sommation des approximations $X_{(i)} \simeq X_{(1)} \times i^{-1/\alpha}$ (qui sont des approximations déterministes de quantités aléatoires), la formule repose essentiellement sur l'équation $Q_\alpha(X_{(1)}) = X_{(1)}^{-\alpha} \simeq 1/N$. Celle-ci nous montre que N est implicitement estimé par la quantité $\widehat{N_1} = 1/Q_\alpha(X_{(1)})$.

Nous retrouverons cette observation dans l'étude de l'estimateur que Jean Laherrère propose dans le cas non intégrable.

2.2 Estimation du cumul des réserves dans le cas non intégrable

Jean Laherrère fournit un estimateur basé sur l'observation suivante :

$$\sum_{i=1}^{N} X_{(i)} \simeq X_{(1)} \sum_{i=1}^{N} \frac{1}{i^{\frac{1}{\alpha}}}$$

Puis, comme $\alpha < 1$, la série précédente converge lorsque N tend vers $+\infty$ vers $X_{(1)} \sum_{i=1}^{+\infty} \frac{1}{i^{\frac{1}{\alpha}}} = X_{(1)} \times \zeta(1/\alpha)$, la formule d'estimation suggérée est alors :

$$\widehat{S_N} = X_{(1)} \times \left[1 + \frac{1}{2^{\frac{\alpha+1}{\alpha}}} \left(1 + \frac{4\,\alpha}{1-\alpha}\right)\right]$$

Le terme en facteur de $X_{(1)}$ n'est autre qu'une approximation de la fonction ζ de Riemann évaluée au point $1/\alpha$.

Pour pouvoir interpréter cette formule en termes de probabilités, nous ne disposons plus de la Loi des Grands Nombres. Il existe cependant un résultat (que nous ne montrerons pas, cf Feller — 1971) qui donne le comportement asymptotique de S_N lorsque N tend vers $+\infty$:

Proposition : *Soient $X_1, \ldots, X_N \rightsquigarrow \Pi(\alpha)$ i.i.d. avec $\alpha < 1$, on a :*

$$\frac{S_N}{N^{1/\alpha}} = \frac{X_1 + \cdots + X_N}{N^{1/\alpha}} \xrightarrow[N \to +\infty]{\mathcal{L}oi} Z_\alpha$$

où Z_α est une variable aléatoire qui suit une loi dite α-stable, distribuée sur $\mathbb{R}^+$.

Ainsi, asymptotiquement on a :

$$S_N \overset{\mathcal{L}oi}{=} N^{1/\alpha} \times Z_\alpha \simeq \left(1/Q_\alpha(X_{(1)})\right)^{1/\alpha} \times Z_\alpha = X_{(1)} \times Z_\alpha$$

La formule de Jean Laherrère nous apprend ainsi que :

– la somme partielle est extrapolée par la somme de la série, ce qui peut être la cause de graves erreurs lorsque α est proche de 1

– N est de nouveau estimé implicitement par $\widehat{N_1} = 1/Q_\alpha(X_{(1)})$

– l'effet de la v.a. Z_α, qui apparait naturellement dans l'étude probabiliste du cas non intégrable, est ignoré. Cette dernière est assimilée à la constante $\zeta(1/\alpha)$, ce qui est fortement démenti par les simulations (cf. figures 2 et 3).

La loi α-stable que suit la v.a. Z_α est mise en évidence dans les figure 2 et 3 dans le cas $\alpha = 0,80$ (cf. Annexe) de la mer du Nord. Une simulation de 10 000 impacts de cette loi et l'histogramme associé montrent qu'elle est fortement dispersée avec une moyenne de plus de 80 et un écart type de plus de 3 500. Elle n'est donc pas du tout assimilable à la valeur $\zeta(1/0,80)$, proche de 5.

Le cas $\alpha < 1$ est, à l'image de la mer du Nord, le cas le plus généralement rencontré quand on positionne les champs d'un système pétrolier dans un diagramme log-log. Dans l'état actuel des recherches, on ne sait pas encore réduire l'effet de la grande dispersion de la loi α-stable sur les estimateurs du cumul. C'est l'un des principaux axes de recherche pour le futur.

2.3 Calcul des intervalles de confiance associés

- Étude de l'estimateur $\widehat{N_1} = 1/Q_\alpha(X_{(1)})$ de N

Nous aurons besoin d'un petit lemme technique préparatoire :

Lemme : *soit $Y_1, \ldots, Y_p$ un p-échantillon d'une v.a. réelle Y de fonction de queue de répartition Q, alors $Q(Y_{(1)}), \ldots, Q(Y_{(p)})$ est un p-échantillon ordonné décroissant de la loi Uniforme sur* $[0\,;\,1]$.

Preuve : soit $x \in [0\,;\,1]$ et $Q^{-1}(x) = \inf\{t \,|\, Q(t) < x\}$, on a :

$$\mathbb{P}(Q(Y) \leq x) = \mathbb{P}(Y \geq Q^{-1}(x)) = Q\left(Q^{-1}(x)\right) = x$$

On reconnait la fonction de répartition de la loi Uniforme sur $[0\,;\,1]$, donc $Q(Y_{(1)}), \ldots, Q(Y_{(p)})$ en est un p-échantillon.

La décroissance de la fonction Q permet de conclure sur l'ordonnancement de cet échantillon. $\diamond$

Pour juger de la qualité de l'estimateur $\widehat{N_1}$, nous cherchons à quantifier ses fluctuations autour de N. Nous allons donc déterminer un intervalle de confiance de $\widehat{N_1}$ au seuil δ, c'est-à-dire trouver $0 < a < 1$ et $b > 1$ tels que l'on ait $\mathbb{P}(\widehat{N_1} \leq Na) = \mathbb{P}(\widehat{N_1} \geq Nb) = \delta/2$.

$$\begin{aligned}\mathbb{P}(\widehat{N_1} \leq Na) &= \mathbb{P}\left(\frac{1}{Q_\alpha(X_{(1)})} \leq Na\right) = \mathbb{P}\left(\frac{1}{Q_\alpha(X_i)} \leq Na\,;\, \forall i \in \mathbb{N}_N\right) \\ &= \mathbb{P}\left(Q_\alpha(X_i) \geq \frac{1}{Na}\,;\, \forall i \in \mathbb{N}_N\right) = \mathbb{P}\left(U \geq \frac{1}{Na}\right)^N\end{aligned}$$

où, par le lemme, U désigne une v.a. de loi Uniforme sur $[0\,;\,1]$.

FIGURE 2

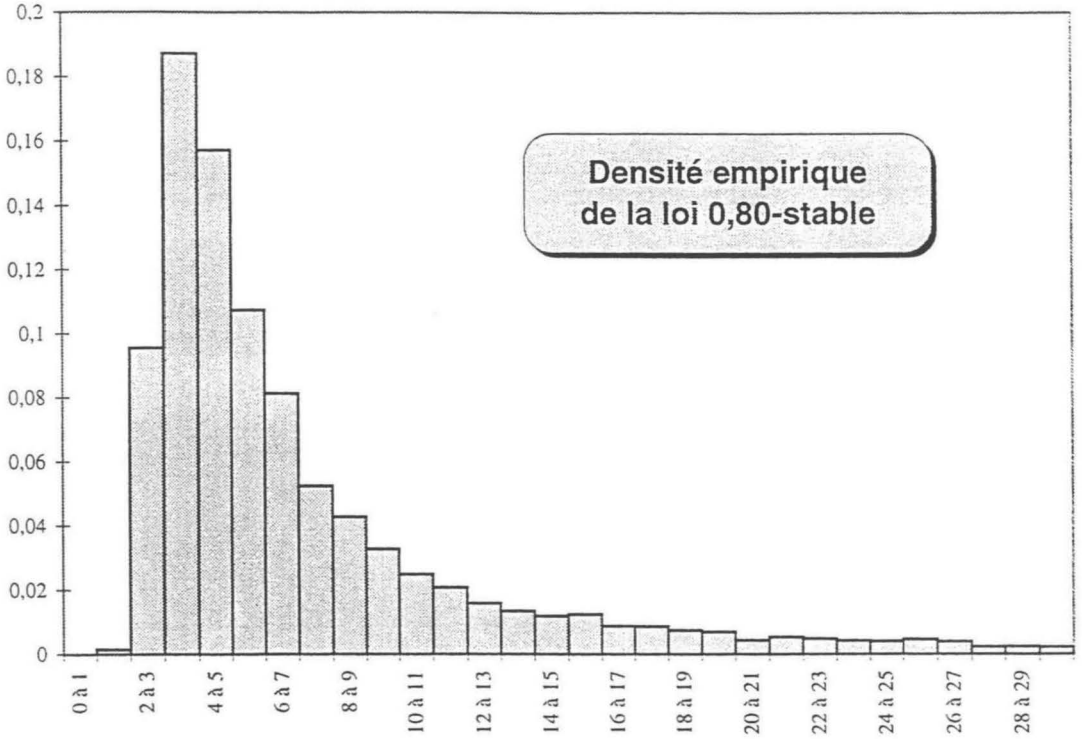
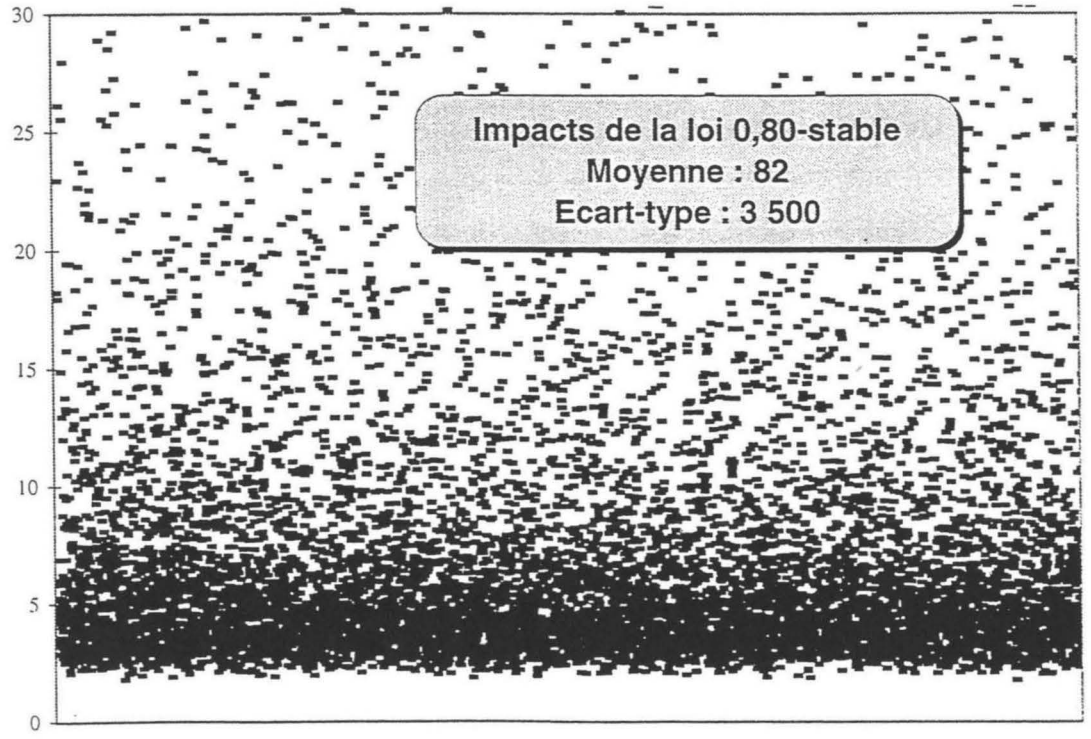


FIGURE 3



Par suite :

$$\mathbb{P}(\widehat{N_1} \leq Na) = \left(1 - \frac{1}{Na}\right)^N \underset{N\to+\infty}{\sim} e^{-1/a}$$

et

$$\mathbb{P}(\widehat{N_1} \geq Nb) = 1 - \left(1 - \frac{1}{Nb}\right)^N \underset{N\to+\infty}{\sim} 1 - e^{-1/b}$$

Ainsi, sous une forme plus explicite

$$\mathbb{P}\left(N \in \left[\frac{\widehat{N_1}}{a}\,;\, b\widehat{N_1}\right]\right) \simeq e^{-1/a} - e^{-b}$$

On peut noter l'écart entre $\mathbb{P}(\widehat{N_1} \geq N) \simeq 0,63$ et $\mathbb{P}(\widehat{N_1} \leq N) \simeq 0,37$. Cette disymétrie montre que $\widehat{N_1}$ tend à surestimer N dans 2/3 des cas.
Voici deux intervalles de confiance pour N aux seuils $\delta = 0.05$ et $\delta = 0.1$:

$$I^N_{90\ \%} = \left[\frac{\widehat{N_1}}{10}\,;\, 5\widehat{N_1}\right] \quad \text{et} \quad I^N_{95\ \%} = \left[\frac{\widehat{N_1}}{20}\,;\, 10\widehat{N_1}\right]$$

Ces deux intervalles montrent à quel point l'estimateur $\widehat{N_1}$ est capable de dévier de la vraie valeur N. La simulation de la figure 4 illustre le comportement éloquent de ce dernier (cf. 3.3 pour les paramètres de la simulation).

De plus, on remarque que l'amplitude par rapport à $\widehat{N_1}$ des intervalles est indépendante de la loi de départ. Cela signifie que, **quel que soit le modèle mis en place, l'estimateur $\widehat{N_1}$ sera toujours aussi déviant par rapport à N.**

Afin de mettre en évidence sur des données réelles ce manque de robustesse des estimateurs de Laherrère, nous effectuons maintenant quelques applications numériques qu cas de l'offshore de la mer du Nord.

3 Application au cas de l'offshore de la mer du Nord

Nous utilisons pour valeur de α la valeur de l'estimateur sans biais $\hat{\alpha}_n^{sb}$ associé à l'estimateur du maximum de vraisemblance décrit en Annexe. Sur les données IFP 1997 de l'offshore de la mer du Nord, on a $\hat{\alpha}_n^{sb} \simeq 0,80$.

3.1 Validation du modèle sur les données

Nous avons effectué un test d'adéquation de Kolmogorov–Smirnov des données de la mer du Nord à une loi de Pareto d'exposant $\alpha = \hat{\alpha}_n^{sb} = 0,80$. La valeur de rejet du test à 90 % est de l'ordre de 1,2 (Lecoutre, Tassi — 1987). Nous obtenons un résultat voisin de 1,0 pour un minimum de rentabilité économique fixé à 25 millions de barils. L'hypothèse d'adéquation n'est donc pas rejettée. **Ce résultat ne signifie pas que la loi que l'on observe est effectivement une loi de Pareto,**

mais que pour les champs de taille supérieure à 25 millions de barils, elle en est proche.

L'objectif étant l'estimation du cumul des réserves, on peut postuler que de légères erreurs, dûes à la modélisation de la distribution de la taille des champs, n'affectent pas notablement la valeur de la somme.

Il faut cependant souligner que la valeur fixée pour le minimum de rentabilité économique peut fortement influer sur le résultat du test. En effet, on peut voir que le diagramme log-log s'incurve fortement au niveau des petits champs (cf. figure 1). Ces derniers ont tendance à sortir de la zone linéaire (cf. figure 9). La raison en est que ces champs sont découverts de façon beaucoup moins systématique que les champs de taille supérieure, car l'effort de recherche se porte beaucoup plus naturellement sur les champs de grande taille. Nous ne considérons donc pas ces "petits" champs dans notre étude. Étants sous-représentés, ils ne correspondent pas au modèle d'échantillonage, dont le tirage est supposé équiprobable. De plus, la somme de leurs réserves prouvées est négligeable devant la somme des réserves prouvées des champs de la zone linéaire. Nous reviendrons sur ce dernier point dans la quatrième partie.

Pour la mer du Nord, sur 390 champs, nous négligeons 179 champs de taille strictement inférieure à 25 millions de barils. Le cumul des réserves des ces champs représente 1,94 milliards de barils soit 4,71 % des 41,29 milliards de barils du total des réserves de la zone linéaire. Si nous ne retirons pas ces petits champs, la valeur du test peut grimper à plus de 8, ce qui nous ferait rejeter l'hypothèse.

Ces résultats tendent à valider l'hypothèse de d'adéquation à la loi de Pareto, au moins pour les champs de taille non négligeable. Acceptant le modèle, nous montrons dans la suite à quel point les déviations dûes aux grandes observations élargissent les intervalles de confiance.

3.2 Les intervalles de confiance bruts pour le cumul des réserves

On peut calculer $\widehat{N_1} = 1/Q_{\hat{\alpha}_n^{sb}}(X_{(1)}) = 51$, alors que le nombre de champs (de plus de 25 milions de barils) découverts en 1997 est de 211 ! Cette valeur est si aberrante, qu'on ne peut fournir aucun intervalle de confiance associé qui soit digne d'intérêt.

L'étude théorique a montré que la formule d'estimation du cumul de Jean Laherrère ne prend pas en compte la loi stable qui apparaît. Nous ne pouvons donc pas l'utiliser ici, ni en donner un intervalle de confiance. En revanche, nous pouvons fournir un intervalle de confiance empirique basé sur l'estimateur le plus proche de celui de Jean Laherrère, qui est conforme à la théorie probabiliste :

$$\widehat{S_N} = \widehat{N_1}^{1/\hat{\alpha}_n^{sb}} \times Z_{\hat{\alpha}_n^{sb}} \simeq \left(1/Q_{\hat{\alpha}_n^{sb}}(X_{(1)})\right)^{1/\hat{\alpha}_n^{sb}} \times Z_{\hat{\alpha}_n^{sb}} = X_{(1)} \times Z_{\hat{\alpha}_n^{sb}}$$

Par simulation, on obtient $[2,71\,;\,44,68]$ comme intervalle de confiance empirique à 90 % de $Z_{\hat{\alpha}_n^{sb}}$. Un intervalle "cohérent" de N est [5, 255]. Pour un minimum de

rentabilité économique fixé à 25 millions de barils, un intervalle de confiance de S_N à $0,9 \times 0,9 = 80$ % est :

$$I_{80\,\%}^{S_N} = [0,5\,;\,1\,100] \quad \text{(en milliards de barils)}$$

En conclusion, l'estimateur conduit à un intervalle d'une amplitude si grande qu'il ne fournit, aucune information exploitable.

On aurait pu imposer aux intervalles de confiance de N et S_N d'avoir leurs bornes de gauche égales aux quantités connues aujourd'hui. Cela aurait eu pour effet de rejeter à l'infini la borne de droite, tout en baissant le seuil de confiance à un taux ridiculement faible. Ceci ne fait que renforcer la très haute incertitude qu'engendrent ces estimateurs ; incertitude liée, ici, à la très grande variabilité de la quantité $X_{(1)}$.

3.3 Quelques pas vers une meilleure fiabilité

Nos efforts d'amélioration de la qualité d'estimation portent essentiellement sur l'estimation de N. On a vu précédemment que l'équation $Q_\alpha(X_{(k)}) \simeq k/N$ n'était exploitée que pour $k = 1$. Cette équation conduit naturellement à considérer la famille d'estimateurs de N suivante :

$$\widehat{N_k} = \frac{k}{Q_\alpha(X_{(k)})} \quad \text{pour } 1 \leq k \leq N$$

Nous allons voir que l'étude théorique de ces estimateurs montre que plus k augmente et meilleure est l'estimation.

Cette amélioration a cependant un prix. Le k maximal que nous considérerons pour calculer la valeur de notre estimateur doit être tel que l'observation $x_{(k)}$ corresponde effectivement au $k^{\text{ième}}$ plus gros champ de la véritable distribution $X_{(1)}, \ldots, X_{(N)}$. Ceci impose donc que l'on soit sûr que les k plus gros champs de la distribution sont découverts. Nous nous en remettons donc aux experts géologues pour nous garantir la valeur maximale k_{max} de k telle que l'on ait la suite d'égalités $x_{(1)} = X_{(1)}, \ldots, x_{(k_{max})} = X_{(k_{max})}$.

Evaluons maintenant l'avantage de prendre k_{max} aussi grand que possible. Voici un Lemme préparatoire (cf. Saporta — 1990) :

Lemme : *Soit X une v.a. de fonction de queue de répartition Q, on a :*

$$Q(X_{(k)}) \overset{Loi}{=} \frac{S_k}{S_{N+1}}$$

où $S_p = \sum_{j=1}^{p} \varepsilon_j$ avec $\varepsilon_1, \ldots, \varepsilon_{N+1} \rightsquigarrow \mathcal{E}(1)$ i.i.d.

Par ce lemme on a $\widehat{N_k} \overset{Loi}{=} \dfrac{k}{S_k} \times S_{N+1}$. Or d'après la loi des grands nombres :

$$\frac{S_k}{k} \xrightarrow[k \to +\infty]{p.s.} \mathbb{E}(\varepsilon_1) = 1$$

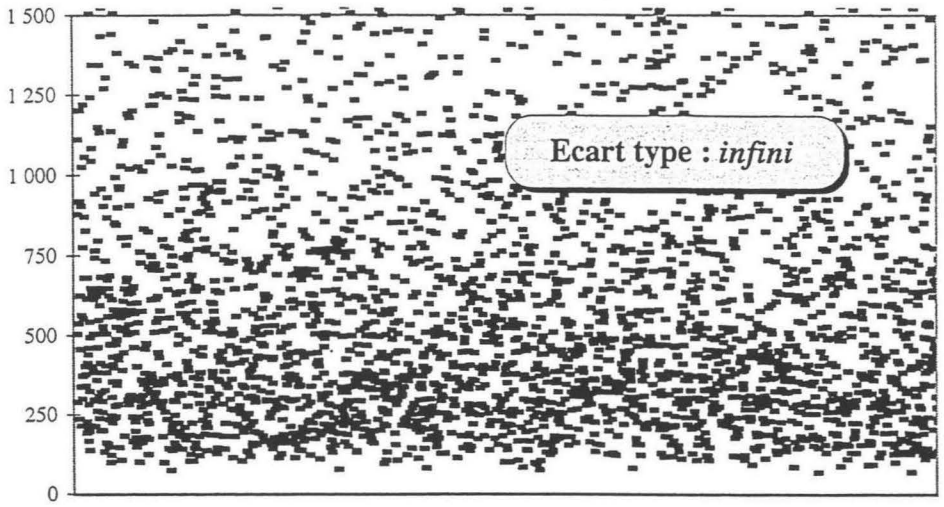


FIGURE 4, N = 500 :

Impacts de l'estimateur $1/Q(X_{(1)})$

Intervalle de confiance empirique à 90 %

[166 ; 10 644]

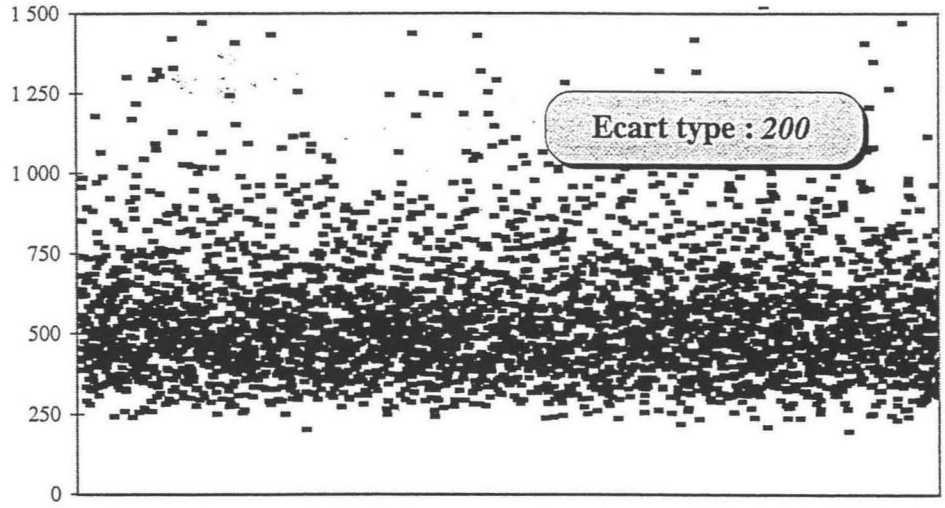


FIGURE 5, N = 500 :

Impacts de l'estimateur $10/Q(X_{(10)})$

Intervalle de confiance empirique à 90 %

[319 ; 917]

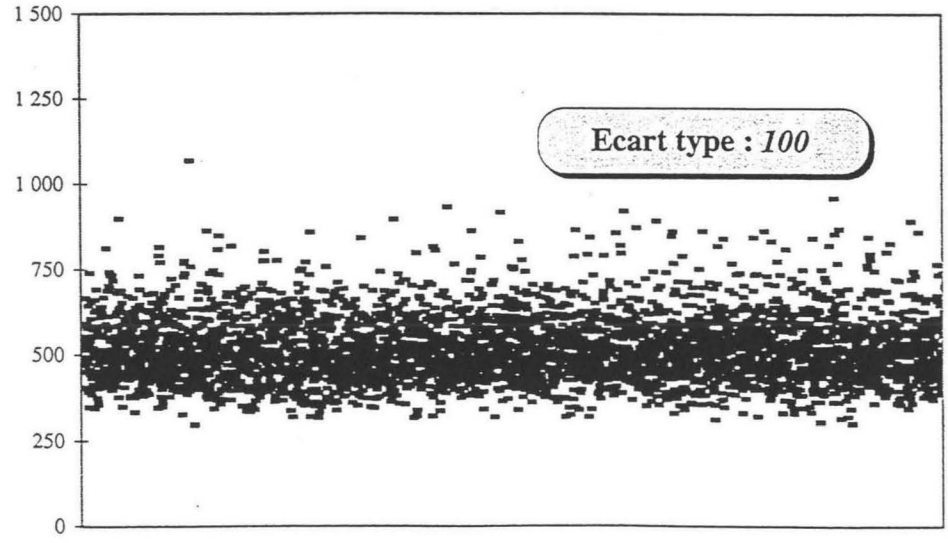


FIGURE 6, N = 500 :

Impacts de l'estimateur $30/Q(X_{(30)})$

Intervalle de confiance empirique à 90 %

[384 ; 686]

donc

$$\frac{\widehat{N_k}}{N} = \left(\frac{k}{S_k}\right) \times \left(\frac{S_{N+1}}{N}\right) \xrightarrow[k,N\to+\infty]{p.s.} 1$$

D'autre part, lorsque k est assez grand, on peut appliquer le théorème central limite à la suite $(\varepsilon_i)_{i\in\mathbb{N}^*}$:

$$\frac{S_p - p}{\sqrt{p}} \xrightarrow[p\to+\infty]{\mathcal{L}oi} \mathcal{N}(0;1)$$

Ainsi, choisissant $(Z, Z') \rightsquigarrow \mathcal{N}(0;1)$ i.i.d., on a :

$$\frac{\widehat{N_k}}{N} \xrightarrow[k,N\to+\infty]{\mathcal{L}oi} \frac{1 + \frac{Z}{\sqrt{N+1}}}{1 + \frac{Z'}{\sqrt{k}}}$$

La suite $\left(\frac{\widehat{N_k}}{N}\right)_{k\in\mathbb{N}_N}$ se comporte donc asymptotiquement comme $1 + \frac{Z}{\sqrt{N}} - \frac{Z'}{\sqrt{k}}$.
L'amplitude de l'intervalle de confiance de $\widehat{N_k}$ autour de N décroit donc en $1/\sqrt{k}$ ce, comme pour $\widehat{N_1}$, indépendamment de la loi définie par Q.

Les figures 4, 5 et 6 montrent une simulation de 5 000 impacts des estimateurs $\widehat{N_k}$ pour $k = 1$, 10 et 30 pour des 500-échantillons d'une loi uniforme. Le gain de précision dans l'estimation de $N = 500$ lorsque k augmente est spectaculaire.

Cependant, dans le cas de la mer du Nord, aussi bonne que soit l'estimation de N, l'estimation du cumul des réserves sera soumise aux fluctuations de la loi stable.

En se basant sur l'intervalle de confiance empirique de N à 90 % obtenu pour $k_{max} = 10$ sur la simulation précédente, un intervalle de confiance de N issu de l'estimateur $10/Q_{\hat{\alpha}_n^{sb}(X_{(10)})}$ à 90 % serait [89 ; 309].

On en déduit alors un intervalle de confiance à $0,9 \times 0,9 = 80$ % pour S_N (on utilise l'estimateur : $\widehat{S_N} = \widehat{N_{10}}^{1/\hat{\alpha}_n^{sb}} \times Z_{\hat{\alpha}_n^{sb}}$:

$$I^{S_N}_{80\,\%} = [18\,;\,1\,400] \quad \text{(en milliards de barils)}$$

La grande amplitude de ce dernier intervalle est essentiellement dûe à la contribution de la loi stable : le rapport "borne supérieure / borne inférieure" n'est que d'environ 3 pour l'estimation de N, et passe a plus de 70 pour S_N. Les recherches en cours visent à s'affranchir de cet effet en "écartant" de la distribution les plus gros champs, qui sont les éléments les plus déviants.

- **Remarque**

On s'est basé plus haut sur des simulations successives de $\widehat{N_k}$ pour $k = 1$, $k = 10$ et $k = 30$ de la loi stable $Z_{\hat{\alpha}_n^{sb}}$. On a donné comme bornes des intervalles de confiance

du cumul les produits des bornes obtenues pour les intervalles de confiance des $\widehat{N_k}$ et de $Z_{\hat{\alpha}_n^{sb}}$.

En simulant directement des valeurs des deux estimateurs du cumul (5 000 simulations) on obtient pour intervalles de confiance empiriques en utilisant $\widehat{N_1}$:

$$I^{SN}_{90\%} = [19\,;\,3\,500] \quad \text{et} \quad I^{SN}_{80\%} = [26\,;\,1\,370] \quad \text{(en milliards de barils)}$$

Et en utilisant $\widehat{N_{10}}$:

$$I^{SN}_{90\%} = [29\,;\,640] \quad \text{et} \quad I^{SN}_{80\%} = [25\,;\,325] \quad \text{(en milliards de barils)}$$

Bien que d'amplitude fortement réduite par rapport aux précédents, ces intervalles restent bien trop larges pour fournir une réelle information.

Le modèle Pareto, ou "fractal-linéaire", qui est le point central de cette étude, a particulièrement retenu notre intérêt, car il a été le premier modèle mathématique destiné à l'estimation des réserves de pétrole. D'autres modèles, qui ne sont d'ailleurs que des raffinements, ont depuis vu le jour. Leur but est "d'ajouter de la courbure" dans la régression du nuage de point log-log, afin que le modèle tienne compte de l'effondrement de ce nuage au niveau des petits champs (cf. figure 9).

Nous consacrons la quatrième et dernière partie de cet article au modèle qualifié de *fractal-parabolique* par Jean Laherrère (1994) ainsi qu'à l'interprétation de la courbure des petits champs dans le modèle Pareto.

4 Mieux approximer la distribution

4.1 Le modèle Fractal-Parabolique

L'idée de départ est de modéliser le nuage de point du diagramme log-log par une parabole (Laherrère 1996) plutôt que par une droite, afin d'introduire dans le modèle la courbure que l'on observe au niveau des petits champs.

Plusieurs méthodes peuvent-être envisagées pour définir cette parabole. La première est d'effectuer une régression parabolique sur le nuage de points. La seconde, utilisée par Jean Laherrère et Collin Campbell (1998), consiste à prendre une enveloppe supérieure parabolique du nuage, passant par le plus gros champ. Ce dernier modèle possède trois paramètres : deux coefficients pour la parabole et le nombre total N de champs (ce qui equivaut à se donner un minimum de rentabilité économique). Le fait de prendre une enveloppe supérieure garantit que le cumul des réserves modélisées par la parabole est supérieur au cumul prouvé. **Chez Laherrère et Campbell, les coefficients de la parabole sont déterminés de manière totalement empirique et dépendent donc dans une large mesure de la subjectivité des auteurs.** Sans méthode statistique bien définie, il n'est pas possible de mesurer la fiabilité de ce modèle.

Nous allons maintenant voir que l'approximation du nuage par une parabole, qu'elle soit issue d'une régression, d'un calcul d'enveloppe ou d'une quelconque autre méthode, est plus difficile à modéliser par une loi de probabilité que l'approximation par une droite.

Dans un diagramme log-log avec une modélisation parabolique, l'échantillon ordonné décroissant $X_{(1)}, \ldots, X_{(N)}$ des tailles des champs d'un système pétrolier suit l'équation suivante :

$$\log X_{(k)} = -a \log^2 k - b \log k + \log X_{(1)} \quad \text{pour} \quad 1 \leq k \leq N$$

où a et b sont les coefficients (positifs) de la parabole.

On cherche une fonction de queue de répartition Q compatible avec cette équation. Compte tenu de l'approximation $Q(X_{(k)}) \simeq k/N$ que nous avons vue en 1.2, l'égalité ci-dessus devient :

$$\log X_{(k)} = -a \log^2 Q(X_{(k)}) - (b + 2a \log N) \log Q(X_{(k)}) + \log X_{(N)}$$

Or $X_{(N)} = 1$ donc $\log X_{(N)} = 0$. Ainsi, $\log Q(X_{(k)})$ est solution de l'équation du second degré en x :

$$ax^2 + (b + 2a \log N)x + \log X_{(k)} = 0$$

Le discriminant de cette équation est $\Delta = (b + 2a \log N)^2 - 4a \log X_{(k)}$. Pour que l'équation ci-dessus admette une solution, il faut $\Delta \geq 0$, soit encore

$$\forall k \in \mathbb{N}_N \qquad 4a \log X_{(k)} \leq (b + 2a \log N)^2$$

Comme l'échantillon est ordonné décroissant, il suffit d'avoir la contrainte ci-dessus valide pour $k = 1$, ce qui équivaut à $b^2 \geq 0$, ce qui est evidemment verifié.

L'équation du second degré en $\log Q(X_{(k)})$ admet donc une unique solution décroissante en k. Soit X une v.a. de loi "parabolique", la fonction de queue de répartition Q que l'on obtient est du type :

$$Q(x) = \mathbb{P}(X > x) = \begin{cases} 1 & \text{si } x \in \quad]-\infty\,;\,1[\\ 10^{-\alpha + \sqrt{\alpha^2 - \beta \log x}} & \text{si } x \in \quad [1\,;\,10^{\alpha^2/\beta}] \\ 0 & \text{si } x \in \quad]10^{\alpha^2/\beta}\,;\,+\infty[\end{cases}$$

On peut noter que cette distribution est à support compact. En particulier, elle possède des moments de tous ordres. La loi de Pareto s'obtient comme cas limite de cette loi lorsque α et β tendent vers 0 sous la contrainte que $\beta/2\alpha$ tende vers une constante strictement positive.

L'estimation des paramètres de cette loi est beaucoup plus difficile à mener que pour la loi de Pareto, car les équations de maximum de vraisemblance ne possèdent pas de solution analytique explicite. On peut cependant utiliser une estimation par

FIGURE 7

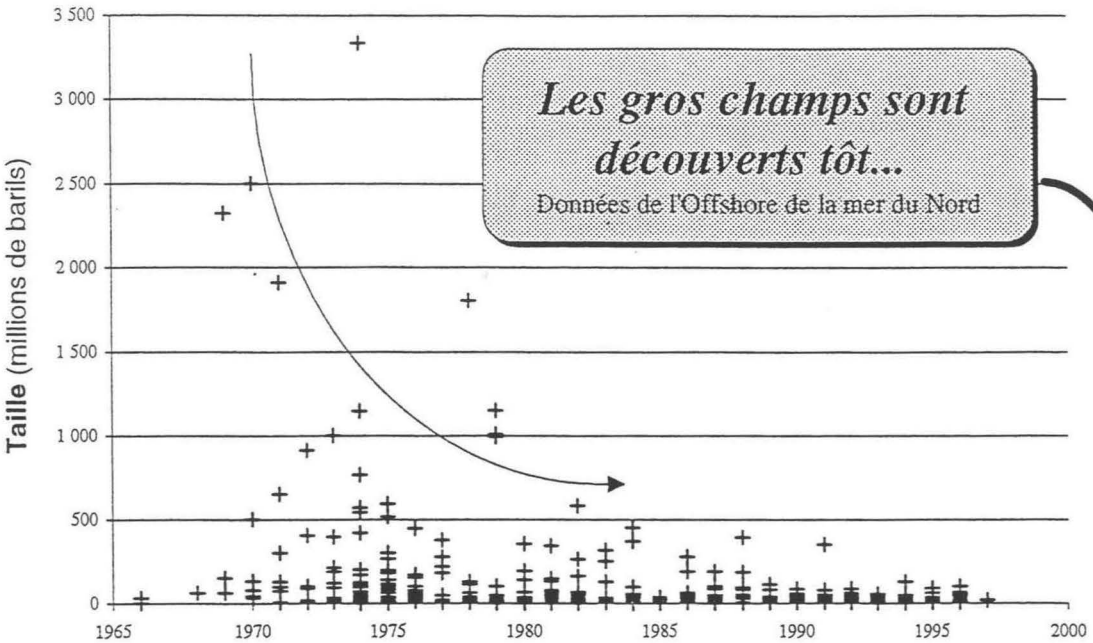
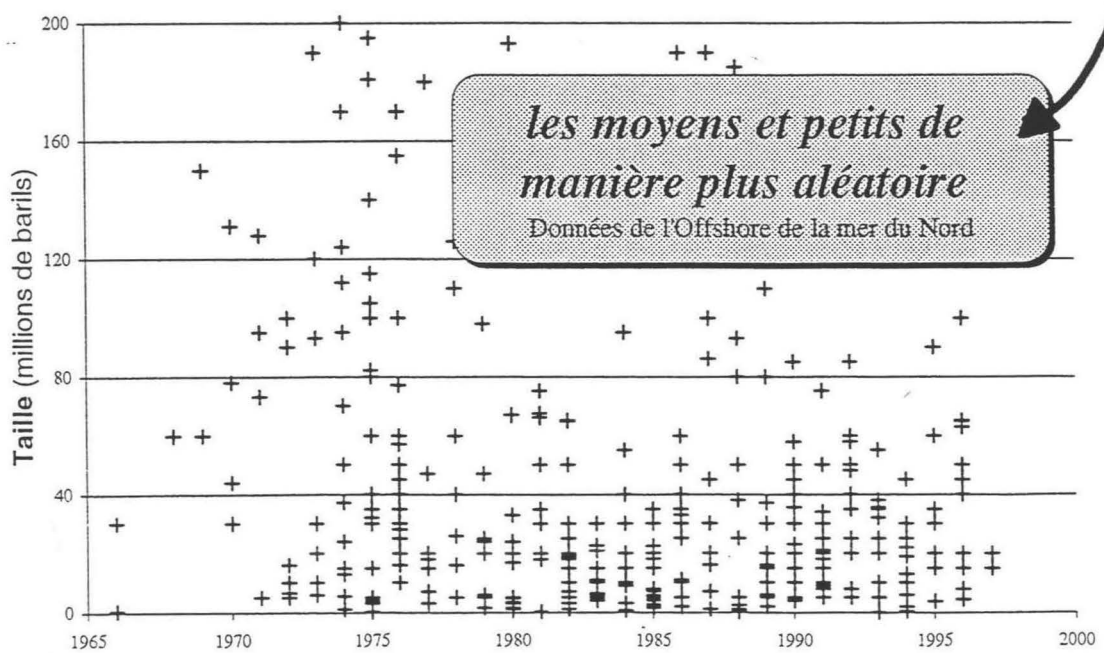


FIGURE 8



Moindres Carrés Ordinaires, mais les tests d'adéquation de Kolmogorov-Smirnov que nous avons effectués sur les données de la mer du Nord avec divers seuils de rentabilité économique, ont tendance à rejeter l'hypothèse d'adéquation.

Cette dernière conclusion et les tests d'adéquation à la loi de Pareto, dont nous avons donné les résultats en 3.1, nous ont invité à poursuivre dans la voie du modèle Pareto. Il nous faut cependant prendre en compte la courbure que l'on observe pour les petits champs. **Pour cela, nous remettons en cause notre modèle d'échantillonage en considérant que celui-ci est biaisé par le fait que l'effort de recherche se porte en priorité sur les gros champs.**

4.2 Le modèle Pareto biaisé

Nous avons vu (cf 3.3) que l'on obtient de meilleurs résultats dans l'estimation de N en utilisant $\widehat{N_k} = k/Q_\alpha(X_{(k)})$. Cela requiert d'utiliser le k qui correspond *effectivement* au $k^{\text{ème}}$ plus gros champ de la distribution naturelle. Nous partons du principe que les n_0 plus gros champs de la distribution naturelle ont été découverts : notre modélisation ne porte donc plus que sur les champs de rangs supérieurs.

L'hypothèse qui soutend cette modélisation est toujours l'hypothèse selon laquelle la taille des champs est distribuée selon une loi de Pareto.

Lorsque l'on étudie les diagrammes log-log sur données réelles, on peut observer presque systématiquement une rupture de pente dans la courbe. Elle se situe souvent entre le $5^{\text{ème}}$ et le $20^{\text{ème}}$ plus gros champ et est suivie d'une zone médiane où la courbe est relativement linéaire. Nous négligeons les champs de petite taille : en effet, ils ne sont que peu exploités car peu rentables économiquement. De fait, ces observations "sortent" de la zone linéaire, car elles sont sous-représentées dans l'échantillon par rapport à la distribution réelle.

On peut alors voir ce modèle d'échantillonage comme un sondage sans remise à probabilités inégales d'inclusion parmi les champs de la distribution réelle :

- les n_0 plus gros champs sont tirés de façon sûre (et seront supposés connus)
- les k champs médians sont tirés avec forte probabilité d'inclusion
- les p petits champs sont tirés avec faible probabilité d'inclusion.

On a bien entendu $n_0 + k + p = n \leq N$.

La figure 7 représente la taille des découvertes par année pour l'offshore de la mer du Nord. On observe nettement la tendance décroissante au cours du temps a partir de la période d'exploitation intensive.

La figure 8 illustre le caractère aléatoire des découvertes de taille moyenne au cours du temps, qui justifie l'hypothèse de tirage à forte probabilité d'inclusion.

La figure 9 est le diagramme offshore des champs de la Mer du Nord sur lequel nous avons isolé les champs moyens. Elle illustre notre hypothèse de plan de sous-échantillonnage.

Les figures 10 et 11 ont pour but de montrer visuellement que ce modèle permet de générer des situations proches de celles que l'on peut observer. La figure 10 est

FIGURE 9

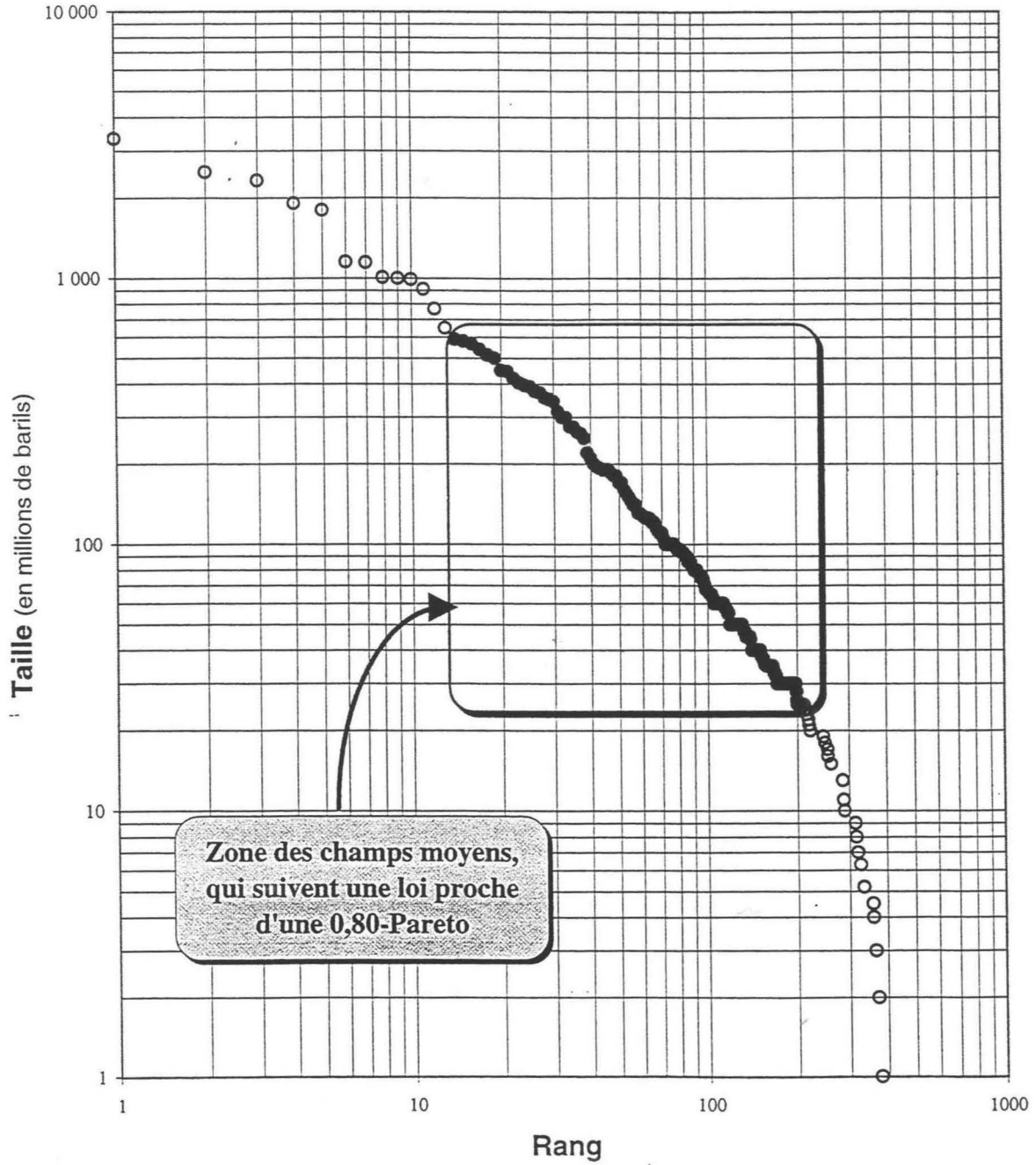
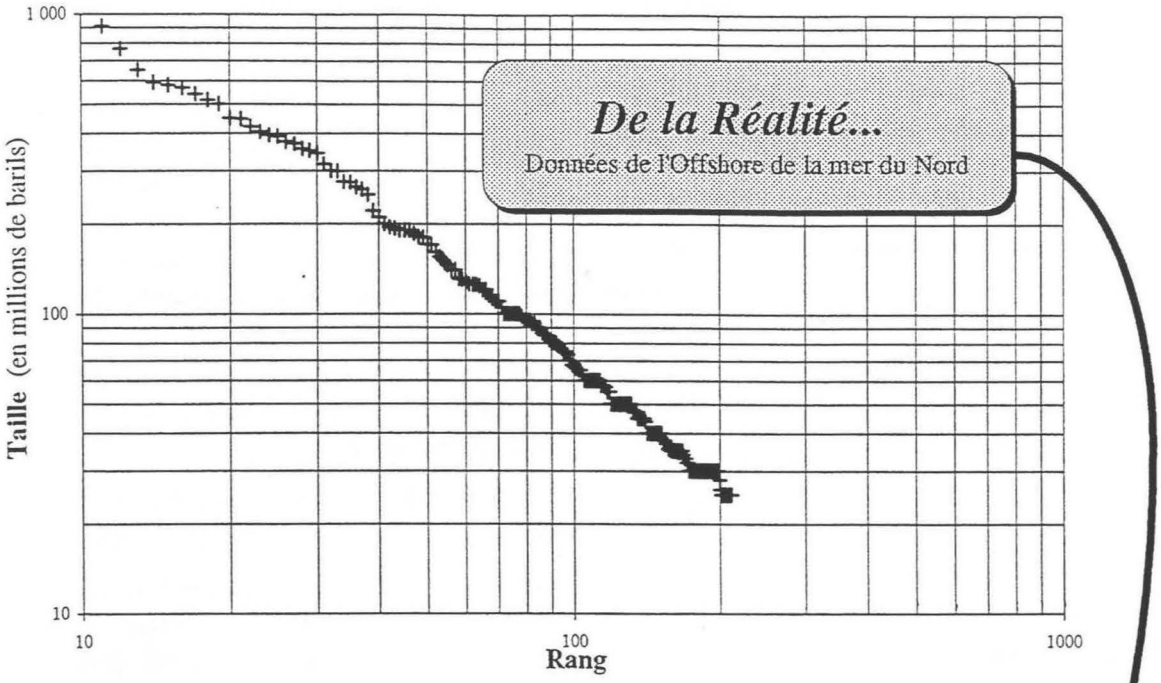


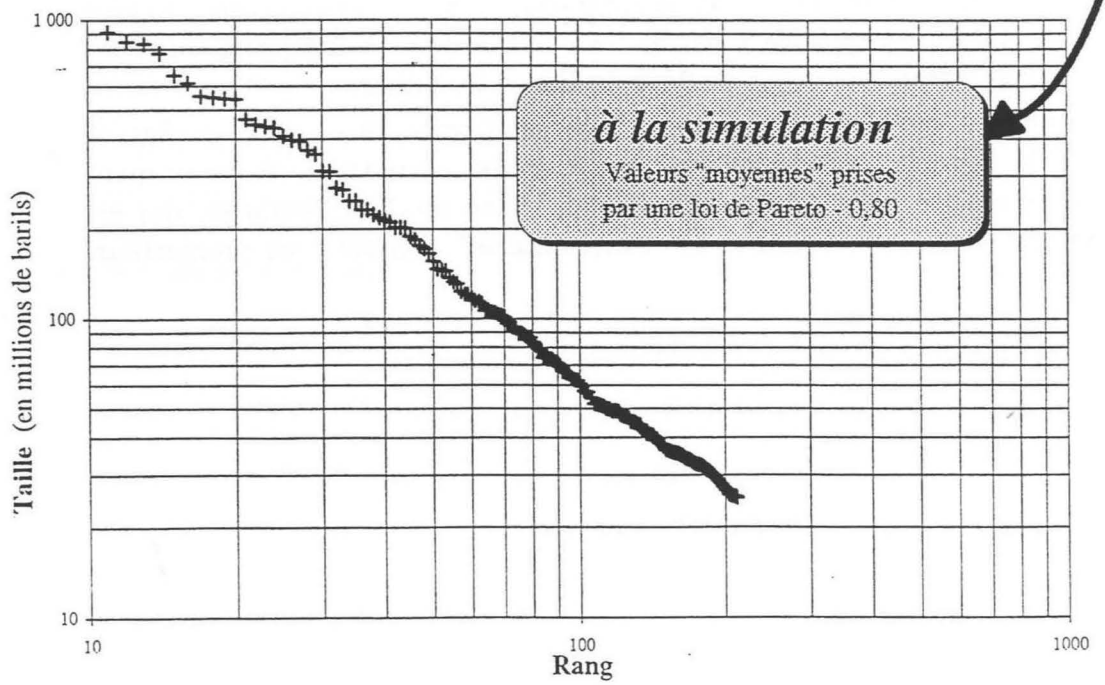
Diagramme log-log de l'offshore de la mer du Nord

FIGURE 10



211 champs de taille moyenne. Les 10 plus gros et 179 plus petits ont été exclus.

FIGURE 11



211-sous-échantillon d'un 300-échantillon d'une loi de pareto d'exposant 0,80. Les 10 plus grandes observations ont été retirées avant sous-échantillonnage.

le diagramme log-log réel de l'offshore de la mer du Nord où sont représentés les seuls champs de la zone linéaire, c'est-à dire, ceux que nous considérons comme étant de taille moyenne.

Les champs de taille moyenne sont l'enjeu véritable de la modélisation car, mis à part les plus gros champs qui ont été trouvés très tôt dans l'histoire de l'exploration et de l'exploitation de la mer du Nord (les 14 plus gros – de taille supérieure à 600 millions de barils – avant 1980), les champs moyens représentent aujourd'hui presque 60 % des réserves prouvées et probablement encore bien plus demain, puisque les champs qui restent à découvrir auront des tailles incluses dans cette plage. Les petits champs quant à eux représentent à ce jour moins de 5 % des réserves prouvées, et il nous semble raisonnable de les négliger ici. Plusieurs raisons motivent ce choix. En premier lieu, on a vu que ces champs sortent de la zone linéaire. Ils sortent donc du cadre de la modélisation Pareto. Ensuite, leur volume est aujourd'hui très inférieur à l'incertitude que l'on peut au mieux espérer sur la modélisation des champs moyens.

La figure 11 donne une illustration du modèle d'échantillonnage des champs moyens à partir d'une loi de Pareto. Les paramètres du modèle sont les suivants :

L'exposant de Pareto choisi est l'estimateur du maximum de vraisemblance calculé sur la distribution de l'offshore mer du Nord — $\hat{\alpha}_n^{sb} \simeq 0,80$. Nous avons effectué un 300-échantillonnage de la loi $\Pi(0,80)$ en considérant un minimum de rentabilité économique situé à 25 millions de barils. Nous avons ensuite pris un 211-sous-échantillon sur le 300-échantillon précédent, de façon à se rapprocher des conditions de la figure 10.

Du point de vue des calculs de cumuls, **le total des réserves prouvées des champs moyens de l'offshore mer du Nord est de 24,1 milliards de barils et le total des valeurs de cette modélisation est de 24,0 milliards de barils.**

Le résultat obtenu est très convaincant : la ressemblance entre les deux figures montre que **le modèle est capable de générer des situations étonnamment proches de la réalité.** Le changement de plan de sous-échantillonnage, qui est à la base du modèle que nous appelons "Pareto biaisé", semble donc prometteur.

CONCLUSION

Au-delà des aspects techniques développés dans cet article, nous avons cherché à montrer quels pouvaient être les apports d'une démarche statistique rigoureuse dans l'estimation des réserves de pétrole. En effet, l'intuition ne saurait être le seul instrument de mesure de la fiabilité d'un modèle. Comme nous l'avons vu, l'outil statistique permet d'évaluer les qualités, les défauts et souvent de préciser les corrections qui doivent être apportées à un modèle.

Le modèle Pareto, défini en 1, présente des difficultés liées à la robustesse de l'estimation de ses paramètres. Mais, malgré les conclusions pessimistes de l'étude détaillée que nous avons menée en 2 et 3, nous souhaitons insister sur le potentiel qu'il recèle (plus certainement dans sa version biaisée, dont nous avons brièvement exposé les fondements en 4). En effet, nous avons été surpris de la pertinence de ce modèle, qui paraît si simple de prime abord. C'est pourquoi nous restons persuadés qu'au prix de quelques raffinements, vers lesquels se portent nos recherches, nous pourrons grandement améliorer la qualité des estimations.

Certains auteurs privilégient aujourd'hui des modèles beaucoup plus complexes, dont le gain reste douteux et non quantifié. Nous avons la certitude qu'il est encore trop tôt pour faire le deuil du modèle Pareto, dont nous sommes encore loin d'avoir exploité toutes les ressources.

Un des principaux axes de recherche pour l'avenir est la mise en œuvre de techniques de rééchantillonnage (de type Bootstrap). Elles consistent à augmenter artificiellement le nombre d'observations dans le but d'améliorer la fiabilité des estimateurs. Cependant, des problèmes théoriques se posent dans l'application de ces méthodes au modèle Pareto, car il ne rentre pas dans leur cadre habituel de validité.

D'autres pistes peuvent encore être explorées, notamment du côté de la statistique non paramétrique. La validation finale du modèle doit aussi tenir compte de l'intervention de nos partenaires géologues et économistes, afin qu'ils contrôlent sa cohérence. Le champ des investigations s'avère donc aussi large que le sujet des réserves est passionnant.

BIBLIOGRAPHIE

• Articles

ALAZARD N., PERRODON A., LAHERRÈRE J.H. [1992] : Réserves et ressources de pétrole et de gaz des pays méditerranéens, *Revue de l'énergie* n° 441 – Août/Septembre.

IVANHOE L.F. [1976] : Oil/gas potential in basins estimated, *Oil and Gas Journal* – 6 Décembre.

LAHERRÈRE J.H. [1991] : Comment estimer le potentiel résiduel d'un bassin pétrolier ? Lognormal ou fractal ?, *manuscrit non publié.*

LAHERRÈRE J.H. [1994] : Nouvelle approche des réserves ultimes – Application aux réserves de Gaz des États-Unis, *Pétrole et Techniques* n° 392 – Décembre.

LAHERRÈRE J.H. [1996] : Distributions de type fractal parabolique dans la nature, *Comptes Rendus de l'Académie des Sciences* – Série 7 – 4 Avril.

LAHERRÈRE J.H., CAMPBELL C. [1998] : La fin du Pétrole bon marché, *Pour la Science* n° 247 – Mai.

PERRODON A. [1991] : Vers les réserves ultimes d'hydrocarbures conventionnels, *Bulletin du Centre de Recherche en Exploration-Production d'Elf-Aquitaine* n° 15 – 4 Décembre

• Ouvrages de référence

FELLER W. [1971] : *An introduction to probability theory and its applications*, volume 2, Wiley.

GOURIEROUX C., MONFORT A. [1989] : *Statistique et modèles économétriques*, 2 volumes, Economica.

LECOUTRE J.P., TASSI P. [1987] : *Statistique non paramétrique et robustesse*, Economica.

SAPORTA G. [1990] : *Probabilités – Analyse des données et Statistique*, Technip.

ANNEXE : Estimation de l'exposant de Pareto

Nous supposerons que nous sommes, dans le cas de l'offshore de la mer du Nord, assez proche de l'exhaustion. Cela signifie que l'on se trouve, en première approximation, dans un cas où les tirages sans remise parmi l'échantillon des champs existant dans la nature le représentent en quasi-totalité. L'échantillon réel étant supposé i.i.d., nous considérons donc que le sous-échantillon des champs découverts de 1997 avec lequel nous travaillons l'est aussi.

Jean Laherrère (Alazard, Perrodon, Laherrère — 1992) estime l'habitat $k = 1/\alpha$ par une régression linéaire de type Moindres Carrés Ordinaires. Nous avons choisi d'utiliser la méthode du Maximum de Vraisemblance parce qu'elle conduit à des estimateurs statistiquement plus puissants et fournit, de plus, des formules plus simples à manipuler.

Si l'on suppose que la distribution de la taille des champs est linéaire en log-log, alors l'estimateur du maximum de vraisemblance représente la pente la plus probable de la droite.

Soit $X_1, \ldots, X_n$ le n-échantillon observé, que l'on supposera i.i.d. de loi $\Pi(\alpha)$. Sa densité de probabilité $\pi_{n,\alpha}$ est donnée par la formule :

$$\pi_{n,\alpha}(x_1, \ldots, x_n) = \alpha^n \prod_{i=1}^{n} x_i^{-\alpha-1} \times \mathbb{I}_{\min\{x_i \,|\, 1 \leq i \leq n\} \geq 0}$$

L'estimateur du maximum de vraisemblance de α est alors la valeur $\hat{\alpha}_n$ telle que $\pi_{n,\alpha}(\hat{\alpha}_n) = \max\{\ln \pi_{n,\alpha}(X_1, \ldots, X_n) \,|\, \alpha \in [0\,;\, +\infty[\,\}$ donc :

$$\hat{\alpha}_n = \mathrm{Argmax}_\alpha \left(n \ln \alpha - \sum_{i=1}^{n} \ln X_i \right) = \frac{n}{\sum_{i=1}^{n} \ln X_i}$$

- Propriétés de l'estimateur $\hat{\alpha}_n$

Proposition : $\mathbb{E}(\hat{\alpha}_n) = \dfrac{n}{n-1}\alpha$, *$\hat{\alpha}_n$ est donc un estimateur asymptotiquement sans biais.*

Nous aurons besoin d'un lemme qui fait le lien entre la loi de Pareto et la loi exponentielle.

Lemme : *si $X \rightsquigarrow \Pi(\alpha)$, alors $\ln X \rightsquigarrow \mathcal{E}(\alpha)$, où $\mathcal{E}(\alpha)$ désigne la loi Exponentielle de paramètre α.*

Preuve du lemme : soit $\varphi : [1\,;\, +\infty[\longrightarrow \mathbb{R}$ mesurable bornée et $X \rightsquigarrow \Pi(\alpha)$, on a :

$$\mathbb{E}\left(\varphi(\ln X)\right) = \int_{\mathrm{R}} \varphi(\ln x) f_\alpha(x)\, dx = \int_1^{+\infty} \alpha x^{-\alpha-1} \varphi(\ln x)\, dx$$

puis le changement de variable ($x = e^t$) donne

$$\mathbb{E}\left(\varphi(\ln X)\right) = \int_0^{+\infty} \alpha e^{-\alpha t} \varphi(t)\, dt$$

La densité de la v.a. $\ln X$ est donc $\mathcal{E}_\alpha(x) = \alpha e^{-\alpha x} \times \mathbb{1}_{[0;+\infty[}(x)$, c'est-à-dire la densité de la loi Exponentielle de paramètre α. $\diamond$

Preuve de la proposition : Par le lemme, la loi de $\sum_{i=1}^n \ln X_i$ est la loi d'une somme de n variables aléatoires i.i.d. de loi $\mathcal{E}(\alpha)$. C'est donc une loi $\mathcal{E}(\alpha)^{\star n} = \Gamma(n, \alpha)$ de densité :

$$\gamma_{n,\alpha}(x) = \frac{\alpha^n}{(n-1)!} x^{n-1} e^{-\alpha x} \times \mathbb{1}_{[0;+\infty[}(x)$$

donc

$$\mathbb{E}(\hat{\alpha}_n) = \int_{\mathrm{R}} \frac{n}{x} \gamma_{n,\alpha}(x)\, dx = \frac{n}{(n-1)!} \int_0^{+\infty} \alpha^n x^{n-2} e^{-\alpha x}\, dx$$

posant ($t = \alpha x$), il vient

$$\mathbb{E}(\hat{\alpha}_n) = \frac{n\alpha}{(n-1)!} \int_0^{+\infty} t^{n-2} e^{-t}\, dt = n\alpha \frac{(n-2)!}{(n-1)!} = \frac{n}{n-1}\alpha$$

mais $\lim\limits_{n\to+\infty} \dfrac{n}{n-1}\alpha = \alpha$ donc $\hat{\alpha}_n$ est asymptotiquement sans biais. $\diamond$

On peut également en conclure que $\hat{\alpha}_n^{sb} = \dfrac{n-1}{n}\hat{\alpha}_n = \dfrac{n-1}{\sum_{i=1}^n \ln X_i}$ estime sans biais le paramètre α.

Proposition : *$\hat{\alpha}_n$ et $\hat{\alpha}_n^{sb}$ sont consistants, c'est-à-dire qu'ils convergent p.s. vers α lorsque n tend vers l'infini.*

Preuve : d'après le lemme, on sait que la famille $\ln X_1, \ldots, \ln X_n$ est i.i.d. de loi $\mathcal{E}(\alpha)$ intégrable. On peut donc lui appliquer la Loi Forte des Grands Nombres :

$$\frac{1}{n}\sum_{i=1}^n \ln X_i \xrightarrow[n\to+\infty]{p.s.} \mathbb{E}(\ln X_1) = \frac{1}{\alpha}$$

donc

$$\hat{\alpha}_n = \frac{n}{\sum_{i=1}^n \ln X_i} \xrightarrow[n\to+\infty]{p.s.} \alpha$$

et

$$\hat{\alpha}_n^{sb} = \frac{n-1}{n} \frac{n}{\sum_{i=1}^n \ln X_i} \xrightarrow[n\to+\infty]{p.s.} \alpha \quad \diamond$$

Ces résultats préconisent l'emploi de $\hat{\alpha}_n$ ou de $\hat{\alpha}_n^{sb}$ pour estimer l'exposant α de Pareto. Dans le paragraphe suivant, notre but est de donner un critère statistique

portant sur l'étude de la variance de ces estimateurs pour déterminer lequel d'entre eux est le plus efficace.

- Efficacité de l'estimation : évaluation de la variance de $\hat{\alpha}_n$ et $\hat{\alpha}_n^{sb}$

La variance minimale que peut posséder un estimateur sans biais de α est donnée par la borne de Fréchet–Darmois–Cramer–Rao (FDCR_n) du modèle :

$$\mathrm{FDCR}_n = -\left(\mathbb{E}\left(\frac{\partial^2 \ln \pi_{n,\alpha}(X_1,\ldots,X_n)}{\partial \alpha^2}\right)\right)^{-1} = \frac{\alpha^2}{n}$$

Proposition : *$\hat{\alpha}_n^{sb}$ est asymptotiquement efficace, c'est-à dire que* $\mathrm{Var}(\hat{\alpha}_n^{sb})$ *est équivalente à* FDCR_n *lorsque n tend vers l'infini.*

Preuve : calculons la variance de $\hat{\alpha}_n^{sb}$, on a

$$\mathbb{E}\left((\hat{\alpha}_n^{sb})^2\right) = \frac{(n-1)^2}{(n-1)!}\,\alpha^n \int_0^{+\infty} x^{n-3} e^{-\alpha x}\,dx$$

posant $(t = \alpha x)$, il vient

$$\mathbb{E}\left((\hat{\alpha}_n^{sb})^2\right) = \frac{n-1}{(n-2)!}\,\alpha^2 \int_0^{+\infty} t^{n-3} e^{-t}\,dt = \frac{(n-1)(n-3)!}{(n-2)!}\,\alpha^2 = \frac{n-1}{n-2}\,\alpha^2$$

ainsi

$$\mathrm{Var}(\hat{\alpha}_n^{sb}) = \mathbb{E}\left((\hat{\alpha}_n^{sb})^2\right) - \mathbb{E}(\hat{\alpha}_n^{sb})^2 = \frac{n-1}{n-2}\,\alpha^2 - \alpha^2 = \frac{\alpha^2}{n-2}$$

puis

$$\lim_{n\to+\infty} \frac{\mathrm{Var}(\hat{\alpha}_n^{sb})}{\mathrm{FDCR}_n} = \lim_{n\to+\infty} \frac{n}{n-2} = 1$$

donc

$$\mathrm{Var}(\hat{\alpha}_n^{sb}) \underset{n\to+\infty}{\sim} \mathrm{FDCR}_n \quad \diamond$$

De plus, $\hat{\alpha}_n = \dfrac{n}{n-1}\,\hat{\alpha}_n^{sb}$, ainsi $\mathrm{Var}(\hat{\alpha}_n) = \left(\dfrac{n}{n-1}\right)^2 \times \mathrm{Var}(\hat{\alpha}_n^{sb}) > \mathrm{Var}(\hat{\alpha}_n^{sb})$. L'estimateur $\hat{\alpha}_n$ est donc plus dispersif que $\hat{\alpha}_n^{sb}$. Nous choisirons donc pour estimateur de α :

$$\hat{\alpha}_n^{sb} = \frac{n-1}{\sum_{i=1}^n \ln X_i}$$

- Détermination d'intervalles de confiance pour $\hat{\alpha}_n^{sb}$

Appliquons le Théorème Central Limite à la famille $\ln X_1, \ldots, \ln X_n$:

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\ln X_i - \frac{1}{\alpha}\right) \xrightarrow[n\to+\infty]{\mathcal{L}oi} \mathcal{N}(0, \mathrm{Var}(\ln X_1))$$

or

$$\mathrm{Var}(\ln X_1) = \frac{1}{\alpha^2}$$

donc

$$\alpha\sqrt{n}\left(\frac{n-1}{n} \times \frac{1}{\hat{\alpha}_n^{sb}} - \frac{1}{\alpha}\right) \xrightarrow[n\to+\infty]{\mathcal{L}oi} \mathcal{N}(0,1)$$

Soit $Z \rightsquigarrow \mathcal{N}(0,1)$, on a :

$$\mathbb{P}(Z \in [-\varepsilon\,;\,\varepsilon]) = \mathbb{P}\left(\alpha \in \left[\frac{n}{n-1}\,\hat{\alpha}_n^{sb} - \frac{\sqrt{n}}{n-1}\,\varepsilon\alpha_n^{sb}\,;\,\frac{n}{n-1}\,\hat{\alpha}_n^{sb} + \frac{\sqrt{n}}{n-1}\,\varepsilon\alpha_n^{sb}\right]\right)$$

Pour un seuil δ ($\delta = 0,1$ par exemple), on choisit ε tel que $\mathbb{P}(Z \in [-\varepsilon\,;\,\varepsilon]) = 1-\delta$. On obtient alors avec les données de la mer du Nord et un minimum de rentabilité économique fixé à 25 millions de barils :

$\hat{\alpha}_n^{sb} = 0,80$ et $\alpha \in [0,72\,;\,0,90]$ avec probabilité 90 %. L'estimateur $\hat{\alpha}_n^{sb}$ présente donc une déviation modeste, de l'ordre de 10 % par rapport à α.