PROBLEMES STATISTIQUES DE LA TRES GRANDE VARIABILITE

MULLON CH.[1], *PICHON G.*[2]

RESUME

L'analyse expérimentale des phénomènes présentant une très grande variabilité semble correspondre à une limitation de l'usage des lois des grands nombres. Nous examinons comment cette constatation expérimentale s'explique naturellement lorsque les phénomènes étudiés suivent des lois de Pareto. Nous indiquons ensuite quelles raisons nous conduisent à envisager que ces lois de Pareto ne sont pas exceptionnelles et correspondent en fait à des situations hiérarchisées rencontrées assez souvent dans l'analyse des phénomènes complexes, notamment à l'occasion de modélisations fractales.

Dans ce sens, cet exposé qui ne contient pas de résultats originaux peut être considéré comme une introduction à la partie statistique de la théorie de Mandelbrot.

INTRODUCTION

Le problème fondamental du transfert d'échelle, entre données quantitatives, réside dans la difficulté d'extrapoler des mesures effectuées sur des parties à une mesure valable sur le tout : on se trouve parfois dans la situation où on ne peut accepter le résultat, manifestement faux, du calcul d'une moyenne ou d'une somme. Si l'explication la plus courante de cette situation est "fonctionnelle" et réside dans une nécessaire prise en compte d'interactions entre les parties, on doit cependant considérer que l'emploi de méthodes issues de la statistique inférentielle peut, en certaines circonstances, ne pas être possible, et se révéler tout à fait inapproprié à des calculs de moyennes et de sommes. En fait ces circonstances sont celles de la très grande variabilité.

[1]ORSTOM-LIA
[2] ORSTOM-LIA

Lorqu'une équipe de recherche répète une même expérience et obtient des résultats chaque fois différents sans pouvoir indiquer les causes de ces différences, elle en déduit une très grande variabilité pour le phénomène étudié. La théorie statistique enseigne que cette variabilité observée peut être inhérente aux phénomènes, mais également provenir du dispositif expérimental employé, notamment dans sa partie statistique; c'est ce point que nous voudrions développer ici.

En effet, la situation de la grande variabilité apparait paradoxale sur un plan statistique : elle semble correspondre à une remise en cause de la loi forte des grands nombres ou du théorème central-limite, qui précisent dans quelle mesure une augmentation du nombre des essais dans une expérience produit une augmentation de la précision du résultat obtenu.

STATISTIQUE DE L'EXPERIMENTATION

Donnons très rapidement quelques rappels sur la théorie statistique de l'expérimentation.

La première hypothèse est que le résultat d'une expérience peut être vu comme une variable aléatoire X ayant une loi de densité $f(x)$. La loi de répartition de X est donnée par :

$$F(x) = P(X < x) = \int_{-\infty}^{x} f(x)dx.$$

La valeur cherchée, le résultat de la mesure, qui est en fait la moyenne de la variable aléatoire X, est donnée par :

$$M(X) = \int_{-\infty}^{\infty} xf(x)dx.$$

La variabilité de cette quantité, la sensibilité de la mesure, est indiquée par son écart-type

$$\sigma(X) = \sqrt{V(X)}$$

où $V(X)$ est la variance de la variable aléatoire X :

$$V(X) = \int_{-\infty}^{\infty} (x-M(X))^2 f(x)dx$$

Pour estimer la quantité $M(X)$, on effectue N essais expérimentaux et on en note les résultats

$$\{ X_i , i=1,N \}.$$

Alors, on estime $M(X)$ par :

$$M_{echant}(X) = \frac{1}{N} \sum_{i=1}^{N} X_i$$

Et, les lois des grands nombres nous indiquent qu'un indicateur de la qualité de cette estimation est son écart-type et qu'il est donné par :

$$\sigma_{estimateur} = \frac{\sigma(X)}{\sqrt{N}}$$

où l'on voit clairement que l'augmentation du nombre d'essais conduit à une augmentation de la qualité de l'estimation. On utilise ainsi le théorème central-limite qui permet de déduire des résultats du type :

Il est probable à 95% que $M(X)$ appartienne à l'intervalle

$$\left[M_{echant}(X) - 1.96 \frac{\sigma(X)}{\sqrt{N}} , M_{echant}(X) + 1.96 \frac{\sigma(X)}{\sqrt{N}} \right].$$

Or une première difficulté vient du fait que dans presque tous les cas, on ne connait pas la valeur exacte de $\sigma(X)$; lorsque l'on veut aller plus loin et avoir une mesure effective de la sensibilité de l'estimation, on estime $V(X)$ et $\sigma(X)$ par:

$$V_{echant}(X) = \frac{1}{N-1} \sum_{i=1}^{N} (X_i - M_{echant})^2$$

$$\sigma_{echant}(X) = \sqrt{V_{echant}(X)}$$

Nanti de ces deux estimations, et de l'assimilation entre σ et $\sigma_{echant}(X)$ on déduit :

Il est probable à 95% que M(X) appartienne à l'intervalle

$$\left[M_{echant}(X)-1.96\frac{\sigma_{echant}(X)}{\sqrt{N}},M_{echant}(X)+1.96\frac{\sigma_{echant}(X)}{\sqrt{N}}\right]$$

LE CAS DE LA TRES GRANDE RARETE

Un premier exemple très simple où les résultats d'une expérimentation peuvent conduire à un diagnostic de très grande variabilité est celui de la grande rareté. Considérons en effet, le cas où la variable aléatoire peut prendre les valeurs 0 et 1 avec les probabilités :

$$P(X=1)=p$$

$$P(X=0)=q=1-p$$

Alors on trouve immédiatement que :

$$E(X) = p$$

$$V(X) = pq = p(1-p)$$

Lorsque l'on effectue une estimation de la moyenne de X, en fait de p, on compte le nombre n de cas où la valeur X_t a été trouvée égale à 1 à l'occasion des N essais; on estime ainsi :

$$E_{echant}(X) = \frac{n}{N},\ V_{echant}(X) = \frac{n}{N}(1-\frac{n}{N}),\ \sigma_{echant}(X) = \sqrt{\frac{n}{N}(1-\frac{n}{N})}$$

Donnons une application directe de ces calculs :

Lorsque l'on veut estimer le nombre de personnes atteintes en France d'une maladie rare (à priori de l'ordre de 1 pour 1000) et que l'on effectue une enquête sur 10.000 personnes parmi lesquelles 16 se révèlent atteintes, on ne peut conclure qu'à un résultat du type : il y a 95% de chances pour que le nombre de malades parmi 50 millions de Français appartienne à l'intervalle *[25.000 , 135.000]*; en première estimation il peut être estimé égal à 80.000, mais un autre sondage dans des conditions identiques aurait pu tout aussi bien conduire à une valeur de 120.000. Si l'on effectue une enquête sur 100.000 personnes parmi lesquelles 160 sont atteintes, cet intervalle devient *[68.000 , 92.000]*.

Ce cas trivial de la grande rareté conduit à des résultats expérimentaux très variables et on comprend bien pour quelles raisons seule l'augmentation du nombre d'essais peut conduire à une plus grande précision. Il n'y a là aucune difficulté d'ordre théorique, seulement

l'exigence de protocoles expérimentaux très lourds dont on cherche parfois, mais en vain, à faire l'économie.

LA LOI DE PARETO

Nous voudrions évoquer une situation beaucoup plus intéressante dans laquelle la variabilité a une cause moins immédiate. Considérons la loi hyperbolique, cas particulier de la loi de Pareto. Sa fonction de densité est donnée par la formule :

$$f(x)=\begin{cases} 0 \text{ si } x<1 \\ \alpha x^{-\alpha-1} \text{ si } x\geq 1 \text{ où } \alpha>0 \end{cases}$$

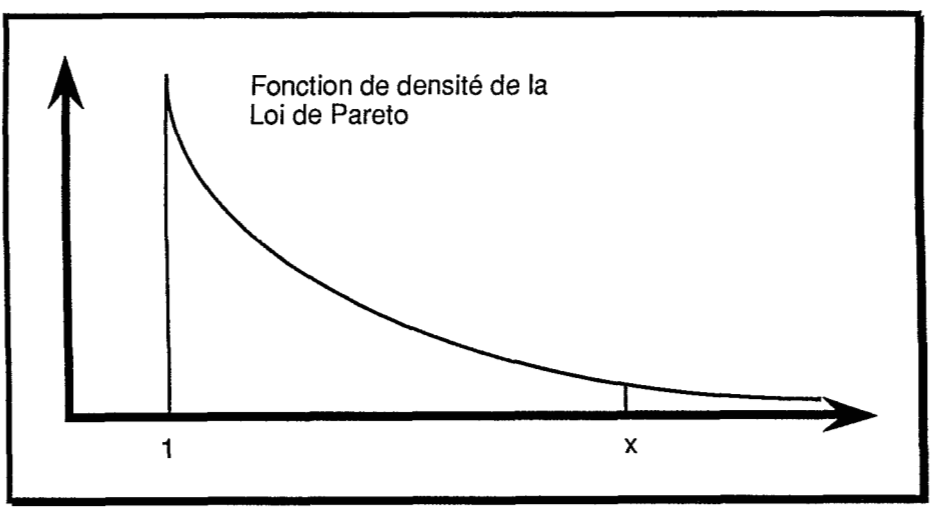

Figure 1 : Distribution de la loi de Pareto

La fonction de répartition F(X) est donnée par :

$$F(x)=P(X<x)=\int_1^x f(x)dx=\int_1^x \alpha x^{-\alpha-1}dx=\left[-x^{-\alpha}\right]_1^x=1-x^{-\alpha}$$

Si $i<\alpha$, le $i^{ième}$ moment de cette loi est donné par :

$$E^i(X)=\int_1^{\infty} x^i f(x)dx=\int_1^{\infty} -\alpha x^{-\alpha-1+i}dx=\left[\frac{\alpha}{i-\alpha}(x^{i-\alpha}\right]_1^{\infty}=\frac{\alpha}{\alpha-i}$$

Mais si $i\geq\alpha$, le $i^{ième}$ moment est infini.

En particulier, la moyenne et la variance de cette loi sont données par :

$$M(X)=\frac{\alpha}{\alpha-1} \text{ si } \alpha>1$$

$$V(X)=\frac{\alpha}{(\alpha-1)^2(\alpha-2)} \text{ si } \alpha>2$$

Et on voit qu'une difficulté surgit lorsque α est compris entre 1 et 2: en effet, dans ce cas, la loi de X a une moyenne, mais a une variance infinie!

Il est alors nécessaire d'être très prudent à l'occasion d'une expérimentation visant à étudier un phénomène qui suit une telle loi. En effet, dans toute expérimentation, on peut calculer une moyenne d'échantillonnage et une variance d'échantillonnage et ensuite appliquer directement les résultats ci-dessus, sans se souvenir de l'assimilation *de s(X)* à $\sigma_{echant}(X)$, qui ici n'est absolument pas justifiable.

SIMULATION DE LA LOI DE PARETO

On trouve sur le tableau suivant, les résultats d'une simulation informatique : on a tiré N nombres selon la loi de Pareto de paramètre α et on en estime la moyenne pour un certain nombre de valeurs de α et de N. On y voit que lorsque $\alpha>2$, la variabilité de l'estimation décroit bien lorsque le nombre d'essais augmente et que l'expérience permet d'aboutir au chiffre exact. Par contre, dés que α se trouve entre 1 et 2, on trouve une variabilité très grande de l'ordre de la moyenne recherchée, et de plus qui ne diminue pas en fonction du nombre d'essais expérimentaux.

α=1.1 N = 100 E = 11		
Essai	M_{echant}	$\frac{\sigma_{echant}}{\sqrt{N}}$
0	3.38639	0.55158
1	7.94646	2.61697
2	5.94094	1.77283
3	4.18416	0.69072
4	4.97818	0.85373
5	7.08925	2.20945
6	5.24802	1.24577
7	5.69620	2.56436
8	4.84120	0.80710
9	3.61352	0.84695

α=1.1, N = 1000, E =11		
Essai	M_{echant}	$\frac{\sigma_{echant}}{\sqrt{N}}$
0	4.74137	0.60386
1	5.81076	0.80448
2	4.08109	0.41123
3	5.85745	1.23124
4	21.99337	13.56232
5	6.65098	2.19758
6	5.03957	0.66299
7	5.26892	0.54118
8	7.64773	1.71144
9	7.05129	1.22852

α=2.62,N=100,=1.61		
Essai	M_{echant}	$\frac{\sigma_{echant}}{\sqrt{N}}$
0	1.57471	0.07844
1	1.52587	0.06719
2	1.69158	0.12495
3	1.75215	0.11953
4	1.60421	0.09688
5	1.67909	0.09756
6	1.45424	0.04941
7	1.59257	0.08640
8	1.71441	0.10508
9	1.53915	0.08069

α=2.62,N=1000,E=1.61		
Essai	M_{echant}	$\frac{\sigma_{echant}}{\sqrt{N}}$
0	1.64515	0.03248
1	1.69928	0.05005
2	1.58737	0.02664
3	1.59020	0.02670
4	1.62183	0.03766
5	1.55981	0.02917
6	1.69408	0.05530
7	1.61385	0.05726
8	1.66936	0.04053
9	1.63071	0.03520

REMARQUE SUR LA FORME DE LA DISTRIBUTION DE LA LOI DE PARETO

La forme de la fonction de densité de la loi de Pareto est classique, avec un mode à l'origine et une décroissance régulière. Elle ne diffère, en fait, de lois plus connues, comme la loi géométrique, que par la forme de sa décroissance à l'infini, relativement lente. L'importance, dans les calculs statistiques, de cette décroissance est d'autant plus grande que le phénomène étudié est quantifié sur une très large gamme d'échelles : on peut ainsi considérer que, s'il y a entre les valeurs extrêmes des distributions observées un rapport inférieur à 100, l'utilisation de la loi géométrique ou de la loi binomiale négative s'avère justifié, notamment par le fait qu'elle conduit à des calculs fiables. Par contre, dès que le rapport entre les valeurs extrêmes est supérieur à 1000, par exemple dans le cas des distributions de revenus, on est conduit à s'intéresser principalement au comportement à l'infini de la distribution, au détriment d'autres caractéristiques plus apparentes telles le mode (l'existence d'un sommet dans la courbe), et on doit donner un rôle prépondérant aux lois de Pareto.

P Levy (1925) a développé complétement la théorie statistique de la loi de Pareto et a montré que de façon asymptotique, les lois normales sont des cas particuliers des lois de Pareto.

APPLICATIONS DE LA LOI DE PARETO

La loi de Pareto est couramment utilisée en économie et linguistique.

En économie, elle est reconnue pour fournir une bonne représentation de la distribution des revenus. On l'écrit sous la forme :

$$Proba(Revenu>x) = x^{\alpha+1} \quad pour\ x>1$$

On a montré (Mandelbrot, 1968) qu'une telle distribution se déduit naturellement, à la fois dans le cas des revenus salariaux et des revenus spéculatifs à partir de l'hypothèse d'une structure hierarchisée : à chaque agent d'un niveau de revenu donné correspond un nombre précis d'agents subordonnés et il existe un rapport constant entre le revenu d'un agent et le revenu de ses subordonnés; le nombre de subordonnés pour un agent et ce rapport sont des variables aléatoires qui, dans l'intervalle d'observation, ne dépendent ni de l'agent, ni de son niveau de revenu. On considère ainsi que l'ensemble des revenus d'une population est structuré par une propriété d'homothetie interne.

En linguistique, la loi de Pareto se déduit naturellement de la célèbre loi de Zipf qui assure que la fréquence d'un mot dans un texte est liée de façon très simple à son rang parmi tous les mots de ce texte; plus précisément, si l'on classe les N mots d'un texte selon leur fréquence dans ce texte, alors la fréquence du $r^{-ième}$ est approximativement égale à :

$$N_r = N\, r^{-d}$$

Là encore, on a donné une explication de cette distribution en terme de structure hierarchisée de l'ensemble des mots d'un texte: on considère qu'à chaque mot du lexique correspond un nombre fixe de mots qui lui sont associés et que le rapport entre la fréquence d'un mot et celle d'un de ses associés est constant et on en déduit simplement la loi de Zipf . Ces deux hypothèses correspondent en fait à un principe d'optimalité de l'utilisation d'un alphabet fini.

On trouvera une intéressante application de cette théorie dans l'exposé de Yves Lemaître (1989) sur la pharmacopée tahitienne.

PHENOMENES DE PERCOLATION

La théorie de la percolation est à la fois très simple et d'un très vaste champ d'application. Présentons-en très rapidement les grandes lignes dans le cas très suggestif des feux de forêts. Pour un exposé plus général on pourra consulter l'article de Deutcher et al (1983).

La situation que l'on veut formaliser se présente sous la forme d'un réseau maillé

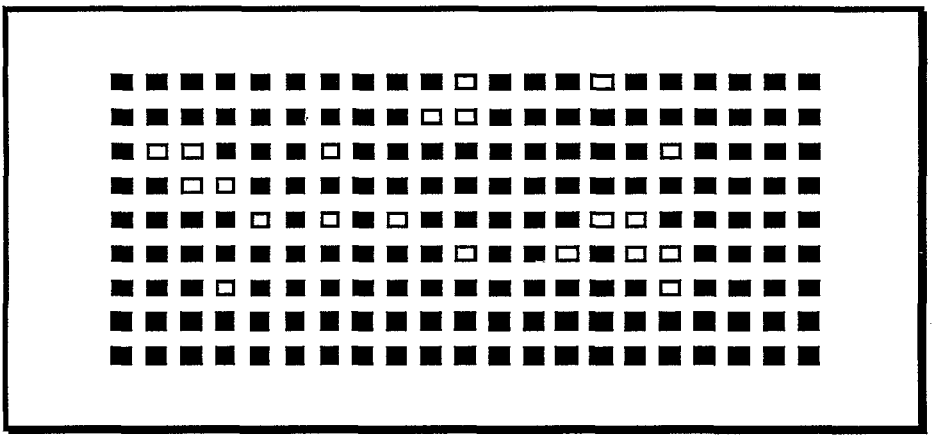

Figure 2 : Réseau de percolation

Chaque point de ce réseau représente un arbre. Parmi les arbres un certain nombre, en blanc sur le schéma, sont inflammables; les autres ne le sont pas. La proportion des arbres inflammables, notée p, est une caractéristique essentielle du réseau; les arbres inflammables sont supposés répartis au hasard dans le réseau. Un arbre est enflammé en bordure de la forêt et le feu se propage de la façon suivante : à chaque pas de temps, tous les voisins inflammables d'un arbre enflammé sont enflammés à leur tour. On conçoit aisément que selon la valeur de la proportion p d'arbres inflammables, le feu se propage de façons très diverses. Les résultats de la théorie de la percolation sont les suivants :

- il existe une valeur de p notée p_0 à partir de laquelle il est à peu près certain qu'un incendie déclenché en bordure de la forêt se propagera jusqu'à la bordure opposée. Cette valeur p_0 est appelée "seuil de percolation". Elle est en fait une caractéristique de la topologie du réseau

- un grand nombre de caractéristiques de la dynamique du système étudié se comportent de façon très particulière autour du seuil de percolation. Ainsi, si nous représentons, la valeur

moyenne de la durée de feux de forêts en fonction du seuil de percolation, nous obtenons une courbe d'équation :

$$D = | p\text{-}p_0 |^{-\alpha}$$

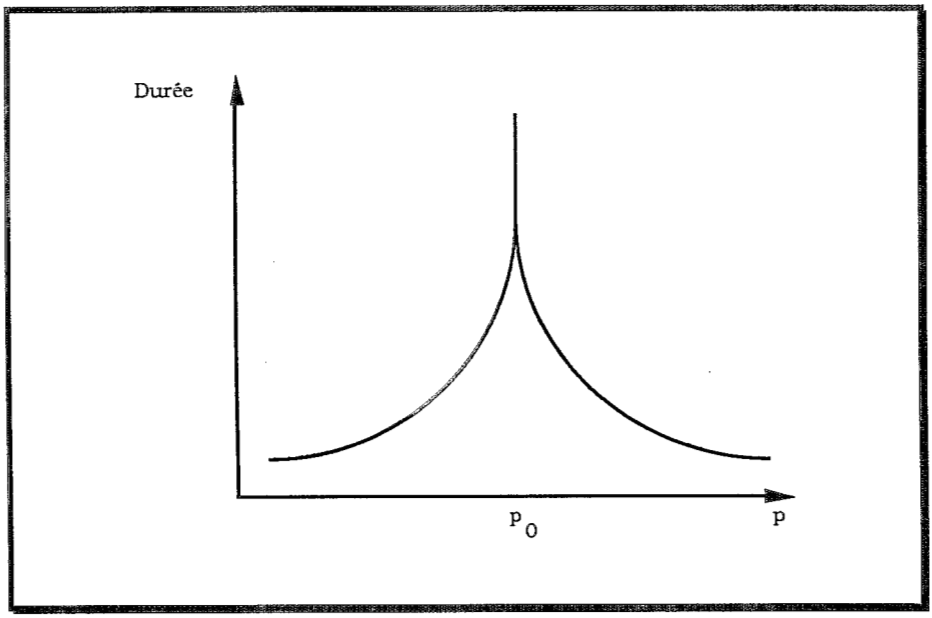

Figure 3: Exposants critiques

α est appelé "exposant critique. C'est également une caractéristique de la topologie du réseau.

Il est alors assez facile de montrer que si l'on est en présence d'un ensemble de forêts réparties autour du seuil de percolation, alors la durée des feux de forêts suivra une loi de Pareto. Et, du calcul des exposants critiques, qui ne dépend que de la topologie du réseau, et en fait principalement de la dimension de l'espace sous-jacent, on peut déduire la valeur de l'exposant de la loi de Pareto et dans de très nombreux cas, le trouver dans la zone [1,2] qui correspond à des lois pour lesquelles on ne peut pas appliquer la loi des grands nombres. Pour la mise en évidence de la relation entre topologie du réseau et valeur des exposants critiques, on pourra consulter l'article de Wilson (1989).

CONCLUSION

Nous avons essayé de montrer dans cet exposé élémentaire que l'emploi trop rapide des outils de l'expérimentation statistique pouvait conduire aux situations paradoxales de la très grande variabilité; il suffit que le phénomène étudié suive une loi dont la

variance est infinie. Nous avons ensuite exhibé une telle loi et essayé de montrer que loin d'être pathologique, elle est assez naturelle et correspond à des situations que l'on peut rencontrer couramment en étudiant un phénomène complexe, possédant une propriété structurelle d'homothétie interne. Enfin en évoquant le cas des phénomènes de percolation, nous avons indiqué comment cette propriété structurelle pouvait provenir de la seule spatialisation du phénomène.

Les applications pratiques de ces remarques élémentaires sont diverses :

- dans le cas d'une enquête socio-économique, visant par exemple à déterminer un revenu moyen, lorsque l'on veut assurer la pertinence et la stabilité du résultat, la solution est bien connue, même si elle n'est pas toujours justifiée théoriquement : il faut constituer deux strates à partir d'un seuil de revenu et procéder à une enquête exhaustive dans la strate supérieure.

- dans les cas où il n'existe pas de base de sondage finie, par exemple dans l'estimation d'une biomasse, il n'y a pas de solution simple évitant la très grande variabilité, alors inhérente au phénomène. Simplement d'après ce que l'on vient de présenter, la recherche des conditions de spatialisation du phénomène aboutissant à la mise en évidence de seuils de percolation et d'exposants critiques permet d'identifier les situations critiques et d'en proposer une approche séparée.

BIBLIOGRAPHIE

AHARONY and Stauffer, 1987, Percolation, The Encyclopedia of Physical Science and Technology, n°10, 226-244, San Diego CA Academic

DEUTCHER G et al, 1983, Percolation Strcutures and Processes, Annals Israel Physical Socitety, n°5

LEMAITRE Yves, 1988, Un modèle cognitif de l'invention dans la pharmacopée tahitienne traditionnelle, in Seminfor 2, La modélisation : aspects pratiques et méthodologie, Paris, Editions de l'ORSTOM

JOHNSON Norman, KOTZ Samuel, 1969, Discrete Distributions, Boston, Houghton Mifflin Company

JOHNSON Norman, KOTZ Samuel, 1969, Continuous Distributions, Boston, Houghton Mifflin Company

LEVY Paul, 1925 , Calcul des probabilités, Paris, Gauthier-Villars

MANDELBROT Benoit, 1968, Les constantes chiffrées du discours, in Le Langage, Paris, Encyclopédie de la Pléiade

MANDELBROT Benoit, 1984, Les objets fractals, Paris, Flammarion

MANDELBROT Benoit, 1960, The Pareto-Levy law and the distribution of income, Inyternational Economic Review, n°1

MANDELBROT Benoit, 1963, The variation of certain speculation price, Journal of Business, University of Chicago, n°36, n°40

WILSON Kenneth, 1989, Les phénomènes physiques et les échelles de longueur, in L'ordre du Chaos, Paris, Pour la Science-Belin