MATHÉMATIQUES ET SCIENCES HUMAINES

M. PETRUSZEWYCZ

Loi de Pareto et processus markovien

Mathématiques et sciences humaines, tome 3 (1963), p. 21-29

<http://www.numdam.org/item?id=MSH_1963__3__21_0>

M. PETRUSZEWYCZ

LOI DE PARETO ET PROCESSUS MARKOVIEN

A la fin du XIXe siècle, Pareto étudiant la distribution des revenus dans divers pays et à diverses époques, en constatait la grande stabilité et l'exprimait en une loi empirique qu'on peut écrire: $\log N(x) = -\alpha \log x$, où $N(x)$ représente le nombre des individus percevant un revenu supérieur ou égal à x, et α, un coefficient d'inégalité. On a remarqué depuis que beaucoup d'autres phénomènes présentaient une semblable loi de répartition: les mots suivant leur fréquence, les villes selon leur population, les entreprises selon le nombre des syndiqués, pour ne citer que ces exemples. Cette diversité pose un problème à ceux qui, depuis la découverte de Pareto, rechercheraient la loi théorique sous-jacente car l'hétérogénéité des domaines évoqués permet de mettre en doute l'existence d'un modèle unique. Aucun des modèles (1) présentés pour expliquer la distribution des revenus n'a, à ce jour, emporté une adhésion générale. Cela se justifie sans doute en partie par le caractère très complexe du phénomène étudié, car il semble que pour des phénomènes beaucoup plus simples on puisse présenter un schéma satisfaisant.

L'étude de certaines activités socio-économiques a permis à J. Durand (2) de signaler une attraction pour les "nombres ronds" qu'il définit, pour certains des exemples cités, comme les multiples et sous-multiples de 10 ou d'une puissance de 10, ou la somme des deux.

L'auteur étudie des relevés de vente d'articles se débitant par lots où se remarquent des concentrations importantes sur certaines valeurs privilégiées. Il présente une analyse statistique de ces séries basée sur la relation rang-fréquence qu'il rapporte à la loi de Zipf tout en rappelant que c'est une forme particulière de la loi de Pareto. J. Durand évoque ensuite un modèle markovien à propos de la succession des chiffres constituant les nombres et donne à preuve les résultats d'une simulation établie pour une série de 313 ventes de fil de fer galvanisé. Cette série servira précisément d'exemple pour une application numérique du modèle, exposé tout d'abord, d'un processus markovien produisant une distribution de type parétien.

*

* *

(1) Un exposé des divers modèles proposés se trouve dans Etudes de Comptabilité Nationale, Publication du Ministère des Finances et des Affaires Economiques, Service des Etudes Economiques et Financières, Avril 1960, Imprimerie Nationale: THIONET - Sur la distribution des revenus et les modèles qui s'y rapportent p. 15.

(2) *L'attraction des nombres ronds et ses conséquences économiques* - Revue Française de Sociologie, 1961, II, 3, 131, 151.
L'attraction des nombres ronds et quelques aspects sociologiques du nombre dans: Comptes rendus séminaire C.M.S.S., 1960-61, fasc. 1.

I - Le modèle cherche à rendre compte d'une série de nombres, entendus par un observateur ou présentés en relevés. Ces nombres expriment les décisions d'achats d'agents qui ont d'autre part en mémoire leur estimation des besoins les motivant. On reviendra sur ces points après avoir présenté les hypothèses.

Chacun des nombres est considéré comme résultant de l'énoncé successif des chiffres y figurant. La deuxième hypothèse veut que les chiffres soient énoncés ou lus de gauche à droite. Par exemple si le premier chiffre est 1, le nombre dont il est le premier chiffre est ou bien compris entre 10 et 19, ou entre 100 et 199, etc... (Fig. 1). Le premier chiffre détermine dans l'ensemble des

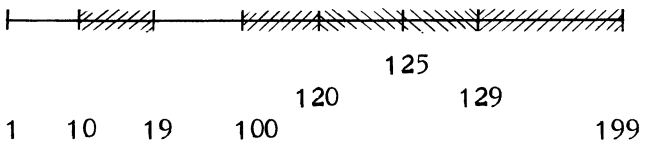


Fig. 1

nombres entiers et "ronds" une famille d'intervalles disjoints dont un seul contiendra le nombre finalement énoncé ou lu. Si ce nombre a plus d'un chiffre, le chiffre suivant détermine à son tour, dans chacun des intervalles un sous-intervalle. L'ensemble des sous-intervalles ainsi sélectionnés en contient un où se trouve nécessairement le nombre définitif. Si le nombre a n chiffres, on l'obtiendra au bout de n opérations de ce genre, c'est-à-dire à la nième subdivision du nième intervalle en partant de la gauche.

Il faut ensuite définir l'alphabet dont disposent les agents qu'on limitera, quant à son étendue, d'une part pour présenter un modèle très simple, mais surtout pour respecter le phénomène même mis en évidence par J. Durand: l'utilisation des seuls chiffres: 1, 2, 5, 0. L'alphabet comporte donc cinq signes; quatre représentent ces chiffres: u, d, c, z, et le cinquième le signe "stop": permet de délimiter l'arbre des nombres possibles qui, sans cela, serait illimité. Cet arbre (Fig. 2) comporte trois branches seulement au départ, zéro ne pouvant jamais être énoncé en premier mais présente ensuite quatre branches au départ de tout carrefour. Il faut enfin introduire l'hypothèse markovienne: à chaque instant l'agent peut choisir un nouveau signe dans l'alphabet et ce choix ne dépend que du dernier chiffre choisi.

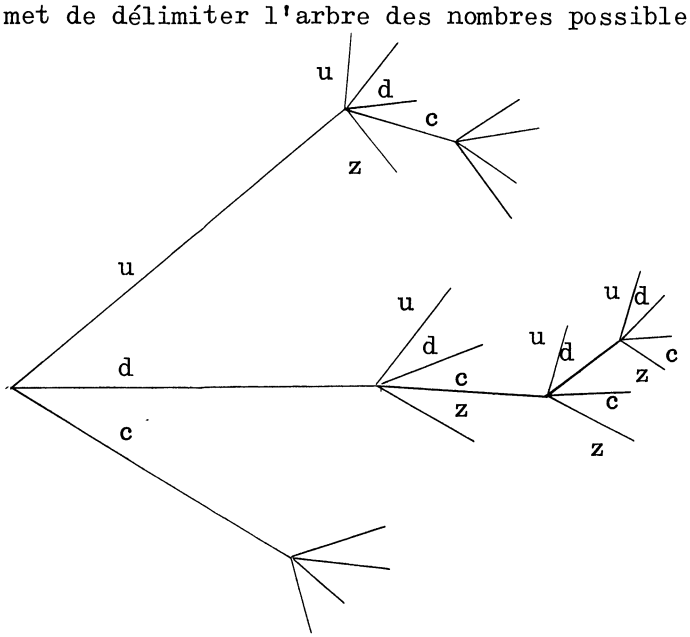


Fig. 2

Quelques remarques au sujet de ces hypothèses. La procédure de sélection que décrit la seconde semble valable, dans sa formalisation, aussi bien pour l'agent que pour l'observateur. Mais il

semblerait raisonnable d'admettre que la réalité vécue s'accompagne d'une connaissance respectivement préalable ou quasi-simultanée de la longueur de la chaîne. Pour l'agent, la prise de conscience du besoin motivant l'achat l'amène à une détermination de l'intervalle qui l'intéresse et parfois même son expérience lui permet de se situer d'emblée dans l'un des sous-intervalles sans le secours d'un instrument de mesure, mètre ou chaîne d'arpenteur. Pour le lecteur, la séparation en tranches de trois chiffres par colonnes comptables, espace ou point est perçue immédiatement - dans une zone "usuelle" -, bien avant la lecture détaillée. La dernière hypothèse semble assez bien rendre compte des approximations successives que réalise l'agent qui passant pour ainsi dire de colonne en colonne comptable n'a besoin de retenir que le chiffre énoncé en dernier s'il a en même temps une idée de la longueur de la chaîne.

Sous ces hypothèses, le diagramme des choix possibles peut être représenté par le schéma markovien habituel. On trouvera fig. 3 le schéma des choix possibles dans la série retenue pour servir d'application numérique. Ce diagramme

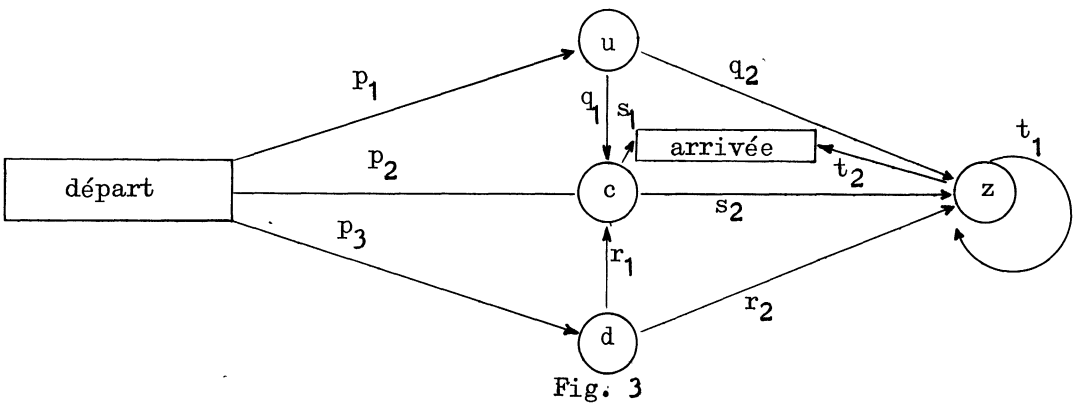


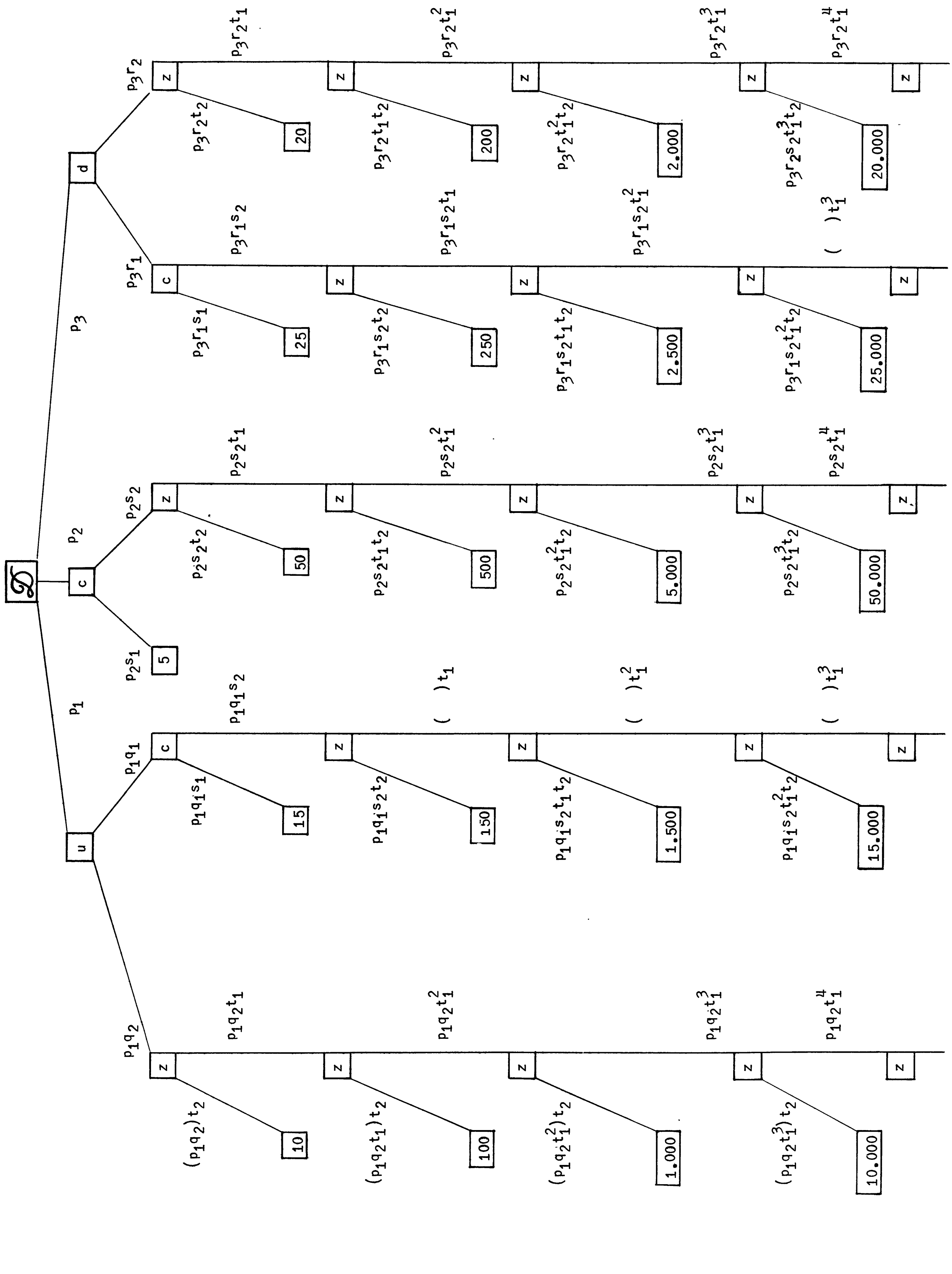
Fig. 3

constitue un réseau dont les cycles, par leur nombre et leur position caractérisent la chaîne markovienne. Leur existence permet à la "machine" de fonctionner indéfiniment et d'émettre des nombres arbitrairement grands. Pour compléter ce schéma il faut affecter aux transitions permises des probabilités de passage, c'est-à-dire estimer la probabilité pour c, par exemple, d'avoir z pour conséquent. On a au départ trois émissions possibles ayant pour probabilité p_1, p_2, p_3, avec la condition $p_1 + p_2 + p_3 = 1$ bien entendu, de même que l'on a :

$$q_1 + q_2 = r_1 + r_2 = s_1 + s_2 = t_1 + t_2 = 1$$

II - Le processus décrit, il nous reste à examiner son fonctionnement. Tout processus engendre une distribution dont on peut ici donner deux représentations. La première rappelle les aspects linguistique du phénomène en cause; l'autre, graphique, nous permettra d'en souligner le caractère parétien.

L'arbre des nombres émis (Fig. 4) peut être considéré comme le lexique des "mots" écrits avec les lettres p_1 p_2 p_3 q_1.....s_2...etc... si on définit le mot comme la probabilité d'un nombre obtenue en vertu de l'hypothèse markovienne par


5
10
100
1.000
10.000
15
150
1.500
15.000
50
500
5.000
50.000
25
250
2.500
25.000
20
200
2.000
20.000

multiplication des probabilités de passage des chiffres le composant. Ces mots sont en même temps la description d'un cheminement dans l'arbre des nombres émis vers des embranchements et la désignation codée de ceux-ci. Car on a ici en effet un code unitaire et net, c'est-à-dire qu'il est possible de mettre une étiquette à chaque arrêt, et dans lequel la scansion est marquée par t_2. Au départ on constate quelques irrégularités, puis l'arbre se stabilise très rapidement parce que le modèle étant extrêmement simple et l'alphabet très limité, par le jeu de la boucle s'instaure un régime permanent où n'apparaissent plus que t_1 et t_2. Pour souligner le caractère périodique du phénomène, relevons une partie du dictionnaire (Fig. 5) de ces deux façons de dire les nombres en rangeant ceux-ci dans l'ordre naturel, arithmétique. On constate aisément qu'ils sont en progression géométrique de même que leur probabilité :

$$\Pr(1000) = \Pr(100)\, t_1$$

la progression des probabilités ayant pour raison la probabilité t_1 de la boucle.

100:	$(p_1\ q_2\ t_1)$	$t_2 = \alpha t_2$
1 000:	$(p_1\ q_2\ t_1)$	$t_1 t_2 = \alpha t_1 t_2$
10 000:	$(p_1\ q_2\ t_1)$	$t_1 t_1 t_2 = \alpha t_1^2 t_2$
100 000:	$(p_1\ q_2\ t_1)$	$t_1 t_1 t_1 t_2 = \alpha t_1^3 t_2$

Fig. 5 : Lexique des nombres et des "mots".

Cela nous amène à la deuxième représentation: au graphique en coordonnées logarithmiques où apparaît le caractère parétien du phénomène. On a deux progressions géométriques, c'est-à-dire deux tables des puissances successives de la base. Si nous désignons les puissances successives de $10 = u_o$ par $y = u^n u_o$, à chacun de ces nombres correspond une probabilité $x = p^o p^n$ où $p_o = \alpha t_2$. Si on applique à chacune de cette correspondance la transformation logarithmique, on peut écrire :

$$\begin{array}{ccc} y = u^n u_o & \longleftrightarrow & x = p_o p^n \\ \log \downarrow & & \downarrow \log \\ Y = nU + U_o & \longleftrightarrow & X = nP + P_o \end{array}$$

d'où l'on tire

$$n = \frac{Y - U_o}{U} = \frac{X - P_o}{P}$$

et l'on peut écrire

$$Y = \frac{U}{p} X + C$$

$$= aX + b$$

Or a étant la pente constante de la droite: on peut écrire

$$a = \frac{U}{p} = \frac{\log 10}{\log p} = \frac{1}{\log p} \qquad \text{avec} \quad o < p < 1$$

ce qui entraîne que log p est négatif.

On a donc

$$a = \frac{1}{\log p} < o$$

dans l'équation de la droite $y = C\, x^a$, d'où :

$$y = \frac{C}{x^k} \quad \text{avec} \quad k > o$$

ce qui est l'équation de la loi de Pareto. En portant sur du papier bi-logarithmique les nombres émis par le processus selon leur rang et en abcisses leur probabilité nous devons voir les points obtenus s'ajuster selon une droite où la puissance de x se lit directement comme la pente constante (Fig. 6).

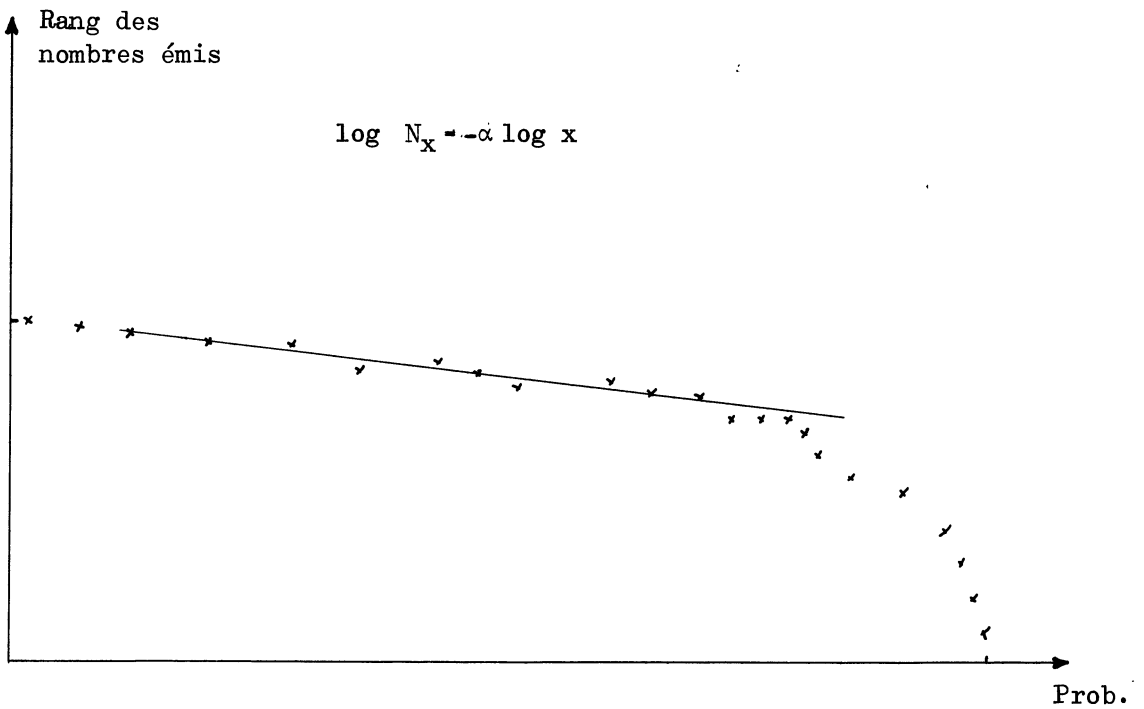


Fig. 6 : Graphique parétien.

Pour porter ces probabilités sur l'axe des x, il faut estimer la valeur de chacun des "mots". D'autre part, pour classer ces mots il y avait un ordre objectif mais il n'en est pas toujours ainsi, par exemple lorsque le phénomène étudié participe davantage de la linguistique. Alors s'impose le choix d'un autre critère: les fréquences d'apparition. C'est par leur intermédiaire que nous allons, pour l'une des séries présentées par J. Durand, estimer les probabilités de passage nous permettant de calculer la valeur des "mots".

*
* *

On peut, à partir des données recueillies et reproduites ci-dessous, dresser l'arbre des ventes (Fig. 7). L'échantillon est constitué par 313 ventes se répartissant ainsi :

107 dont le premier chiffre est 1

dont 77 = 10
13 = 15
13 = 100
4 = 150

122 dont le premier chiffre est 2

dont 2 = 20
111 = 25
5 = 200
4 = 250

84 dont le premier chiffre est 5

dont 59 = 5
25 = 50

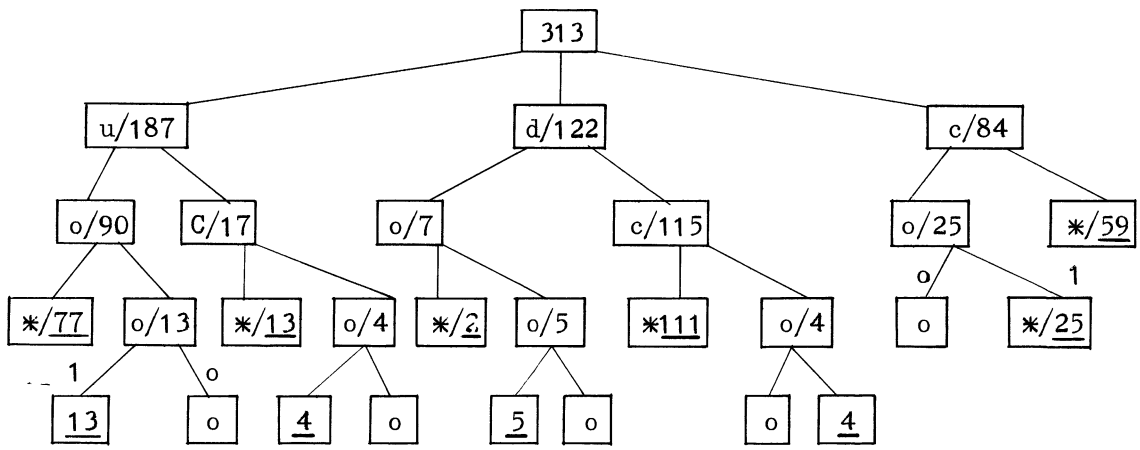


Fig. 7 : Arbre des ventes.

Ayant inscrit les effectifs dans les cases correspondantes cela va nous permettre d'estimer les probabilités élémentaires. Considérons la case c/84 par exemple; ces 84 cas constituent les 100 % des nombres dont le premier chiffre est 5 qui se décomposent en 59 commandes effectives de 5 (les nombres se terminant par stop ont leurs effectifs soulignés) qui en représentent les 70 %. On estime ainsi à 0,70 la valeur de s_1. On en déduit que les 25 autres nombres commençant par 5 représentent 30 % des cas, d'où la valeur 0,30 pour s_2. Cet effectif de 25 correspond à la totalité des réalisations de 50 d'où le chemin conduisant à la case */25 est affecté de la probabilité 1, l'autre chemin ne pouvant recevoir que la probabilité 0. Le raisonnement peut être refait pour chacune des branches de l'arbre des ventes et nous permet d'établir les probabilités de passage dont on trouve les valeurs numériques au tableau I.

$p_1 = 0,34$	$q_1 = 0,16$	$r_1 = 0,94$	$s_1 = 0,85$	$t_1 = 0,12$
$p_2 = 0,39$	$q_2 = 0,84$	$r_2 = 0,06$	$s_2 = 0,15$	$t_2 = 0,88$
$p_3 = 0,27$				

Tableau I

Par composition de ces probabilités élémentaires, on peut attribuer à chacun des nombres émis par la machine une valeur. Ces valeurs figurent en troisième colonne du dictionnaire (Fig. 9) qui range les nombres dans l'ordre arithmétique. Pour aboutir à la représentation parétienne il faut les ordonner suivant leur valeur en fréquence, puis on porte en ordonnées les rangs des nombres ou leur population cumulée et en abscisses les fréquences qui ont toutes été multipliées par 100.000 pour s'étaler de 0,1 à 39.000. Cet étalement est considérable par rapport à celui de la population, ce qui donne au graphique une allure inhabituelle, car dans les graphiques de revenus par exemple, l'étalement maximum est situé sur l'axe des ordonnées (Fig. 6: graphique parétien).

Que penser de cet ajustement? Il faut se poser la question à deux points de vue. On peut s'interroger sur la valeur de l'ajustement à la loi empirique par un échantillon assez restreint; il aurait été intéressant de disposer de relevés plus étendus. D'autre part, l'ajustement, au moyen des probabilités estimées, de la loi de progression géométrique dégagée en première partie mérite examen. S'il est satisfaisant, du moins quand on en juge par rapport aux ajustements que l'on peut voir dans le domaine des revenus, il n'est pas parfait et met en évidence les effets des manipulations numériques. En effet, au cours des calculs, il a fallu "arrondir" plusieurs fois: dans le calcul des pourcentages de fréquence pour respecter la condition $p_1 + p_2 + p_3 = 1$ et aussi lors du calcul des valeurs numériques des mots codés qui comportaient plus de dix chiffres après la virgule.

D'où les écarts constatés par rapport au strict alignement attendu, écarts que l'on rencontre aussi dans des ajustements de revenus, mais dont on rend compte différemment par des considérations d'ordre économique et qui ne sauraient être avancées ici.

*

* *

On a ici un modèle de processus markovien très simple qui en fonctionnant produit un arbre. Si sur cet arbre on porte des valeurs numériques après un régime transitoire irrégulier, on voit apparaître le caractère "périodique" du phénomène qui est de ce fait justiciable d'un ajustement parétien. Si sur l'arbre on porte des lettres on a un code unitaire et net banal. Le phénomène étudié participant de la linguistique et ce mot figurant dans le texte, il ne faudrait pas croire que le modèle soit linguistique. Pour qu'il puisse y prétendre il serait nécessaire de le perfectionner et de le compliquer considérablement. Tel qu'il est présenté le modèle semble cependant rendre compte d'un phénomène en soi mineur, mais situé dans son contexte socio-économique.

1...	p_1	0,34
2...	p_3	0,27
5...	p_2	0,39
5 *	$p_2 s_1$	0,33
10...	$p_1 q_2$	0,28
10 *	$(p_1 q_2) t_2$	0,25
15...	$p_1 q_1$	0,54
15 *	$p_1 q_1 s_1$	0,04
20...	p_3	0,016
20 *	$p_3 r_2 t_2$	0,014
25...	$p_3 r_1$	0,25
25 *	$p_3 r_1 s_1$	0,21
50...	$p_2 s_2$	0,058
50 *	$p_2 s_2 t_2$	0,051
100...	$p_1 q_2 t_1$	0,03
100 *	$(\alpha) t_2$	0,03
150...	$p_1 q_1 s_2$	0,008
150 *	$(\ ) t_2$	0,007
200...	$p_3 r_2 t_1$	0,0019
200 *	$(\ ) t_2$	0,0017
250...	$p_3 r_1 s_2$	0,0038
250 *	$(\ ) t_2$	0,0033
500...	$p_2 s_2 t_1$	0,007
500 *	$(\ ) t_2$	0,006
1.000...	$p_1 q_2 t_1^2$	0,004
1.000 *	$(\alpha\ t_1) t_2$	0,003
1.500...	$p_1 q_1 s_2 t_1^2$	0,0009
1.500 *	$(\ ) t_2$	0,0008
2.000...	$p_3 r_2 t_1^2$	0,00023
2.000 *	$(\ ) t_2$	0,00020
2.500...	$p_3 r_1 s_2 t_1$	0,0045
2.500 *	$(\ ) t_2$	0,0040
5.000...	$p_2 s_2 t_1^2$	0,0008
5.000 *	$(\ ) t_2$	0,0007
10.000...	$p_1 q_2 t_1^3$	0,0004
10.000 *	$(\alpha\ t_1^2) t_2$	0,0003
15.000...	$p_1 q_1 s_2 t_1^2$	0,00010
15.000 *	$(\ ) t_2$	0,00011
20.000...	$p_3 r_2 t_1^3$	0,000027
20.000 *	$(\ ) t_2$	0,000023
25.000...	$p_3 r_1 s_2 t_1^2$	0,0005
25.000 *	$(\ ) t_2$	0,0004
50.000...	$p_2 s_2 t_1^3$	0,0001
50.000 *	$(\ ) t_2$	0,00009 $-\varepsilon$
100.000...	$p_1 q_2 t_1^4$	0,00006
100.000 *	$(\alpha\ t_1^3) t_2$	0,000005 $+\varepsilon$
150.000...	$p_1 q_1 s_2 t_1^3$	0,00001
150.000 *	$(\ ) t_2$	0,00001 $-\varepsilon$
200.000...	$p_3 r_2 t_1^4$	0,000003
200.000 *	$(\ ) t_2$	0,000003 $-\varepsilon$
250.000...	$p_3 r_1 s_2 t_1^3$	0,00006
250.000 *	$(\ ) t_2$	0,00006 $-\varepsilon$
500.000...	$p_2 s_2 t_1^4$	0,00001
500.000 *	$(\ ) t_2$	0,00001 $-\varepsilon$

Fig. 9: Dictionnaire