

REVUE DE STATISTIQUE APPLIQUÉE

T. BENKARAACHE

Quelques tests de tendance pour les classifications hiérarchiques

Revue de statistique appliquée, tome 48, n° 4 (2000), p. 41-57

http://www.numdam.org/item?id=RSA_2000__48_4_41_0

© Société française de statistique, 2000, tous droits réservés.

L'accès aux archives de la revue « *Revue de statistique appliquée* » (<http://www.sfds.asso.fr/publicat/rsa.htm>) implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques

<http://www.numdam.org/>

QUELQUES TESTS DE TENDANCE POUR LES CLASSIFICATIONS HIÉRARCHIQUES

T. Benkaraache

Faculté d'économie de Mohammédia, BP 145, Mohammédia, Maroc

RÉSUMÉ

Dans le cadre de l'étude de la validité des structures de classification, on munit l'ensemble des structures de classification hiérarchique de même type (par exemple les hiérarchies stratifiées, hiérarchies binaires, etc.) d'une loi de probabilité uniforme et on étudie quelques variables aléatoires qui mesurent les caractéristiques classificatoires des structures de classification, comme la hauteur d'une hiérarchie stratifiée, le nombre de classes à un niveau de la hiérarchie, les tailles des classes, etc. Ces variables seront utilisées pour construire des tests de tendance d'une classification hiérarchique.

Mots-clés : classification hiérarchique, hiérarchie stratifiée, hiérarchie binaire, test de classifiabilité, chaîne de Markov

ABSTRACT

With the purpose to study the structure validity in classification, we endow the set of structures of classification (as stratified hierarchies, binary hierarchies, etc.) with a uniform distribution. We introduce some random variables which describe certain characteristic properties of random hierarchies, as the height of stratified hierarchy, the number of classes in a given level of the hierarchy, etc. All these random variables can be used in the problem of clustering tendency.

Keywords : Hierarchical classification, stratified hierarchy, binary hierarchy, classifiability testing, Markov chain.

I. Problèmes de validité et de classifiabilité en classification

L'utilisation des méthodes de classification est devenue un outil précieux d'exploration des données. Plusieurs ouvrages de référence existent maintenant et décrivent les divers aspects de la classification des données (Jain & Dubes (1988), Godehardt (1990), etc.). La diversité des travaux qui existent rend parfois bien difficile le choix des praticiens à cause du manque relatif de travaux sur la validité de ces méthodes et la significativité des résultats obtenus. La construction d'une

classification par application d'un algorithme ou méthode de classification est loin d'être la dernière étape d'une étude de classification mais offre une série de nouveaux problèmes et questions :

- La structure obtenue reflète-t-elle des groupements naturels ou n'est-elle qu'un artefact résultant de phénomènes purement aléatoires?
- Quelle est la significativité des classes obtenues?
- Comment tester l'existence d'une structure au moyen d'une hypothèse nulle de «non classifiabilité»?
- Comment juger de la significativité statistique des résultats produits? etc.

La meilleure manière pour répondre à ces questions est de construire un test statistique de l'hypothèse nulle H_0 qui consiste à supposer que la structure est tirée aléatoirement suivant une loi de probabilité uniforme contre une hypothèse alternative qui reflète l'existence de classes significatives. On peut distinguer globalement deux types d'approches pour construire de tels tests. La première, quand il s'agit d'une structure observée dont on ignore la méthode utilisée pour sa construction, et même les données de départ. Dans cette situation, on ne peut que comparer cette structure observée à une structure de même type qui soit aléatoirement tirée suivant une loi de probabilité. On peut alors construire des tests pour décider si la structure observée a une tendance aléatoire. Dans ce cas, les classes de la structure n'auront aucune significativité. La deuxième approche est celle qui se pratique sur une structure de classification construite avec une certaine méthode de classification MC connue. L'étude porte alors sur l'ensemble des structures obtenues par la méthode de classification MC à partir de données aléatoires. On pourra alors tester, en comparant la structure observée aux structures aléatoires obtenues, si cette structure observée est «significative» ou si elle est due seulement à la méthode elle-même (Jain and Dubes 1988, Gordon 1994).

Le choix de l'hypothèse nulle d'un test de tendance dépend de la nature des données utilisées et du type de structure de classification retenu. Elle se formule souvent par une distribution uniforme dans l'ensemble des structures de classifications de même type que la structure observée. Par exemple, si la structure observée est une partition, on munit l'ensemble des partitions sur n individus d'une loi uniforme.

Le but d'un test de classifiabilité (ou de tendance) est de définir une statistique qui résume quelques informations classificatoires, de calculer la distribution de cette statistique sous l'hypothèse nulle H_0 et enfin de comparer la valeur observée de la statistique à un seuil théorique (pour un niveau de confiance donné) pour accepter (ou rejeter) l'hypothèse nulle. Le rejet de H_0 signifie que la valeur observée correspond rarement à une structure aléatoire, ce qui correspond à l'acceptation de la structure observée.

La construction d'un test passe donc par la détermination d'une statistique qui doit exhiber une (ou des) informations concernant la complexité de la structure de classification (d'où l'appellation «indice de complexité»).

Nous rappelons dans le deuxième paragraphe quelques définitions sur les structures de classification hiérarchiques ainsi que quelques unes de leurs caractéristiques (indices de complexité). Dans le troisième paragraphe nous présentons quelques variables aléatoires liées aux hiérarchies stratifiées et aux hiérarchies binaires ainsi que

leurs distributions. Nous montrons enfin, dans le dernier paragraphe, comment on peut utiliser les résultats obtenus pour la construction des tests de tendance.

II. Structures de classification hiérarchique

1. Définitions de base

Le but d'une méthode de classification est de permettre, dans une population observée, de grouper les individus qui se ressemblent en fonction de leur degré de ressemblance. Ce dernier est mesuré en fonction d'un certain nombre de variables qui décrivent les n individus de la population. La plupart des méthodes utilisent comme données de départ une matrice carrée symétrique d'ordre n qui représente les dissimilarités entre les individus. Ces dissimilarités seront utilisées par l'algorithme de classification pour construire une structure de classes sur la population, comme les partitions, les hiérarchies, les hiérarchies indicées, les hiérarchies stratifiées, les pyramides, etc. (la description de ces méthodes se trouvent par exemple dans les références classiques comme Sneath & Sokal 1973 ou Jain & Dubes 1988). Introduisons quelques définitions sur les structures usuelles de classification qui nous serviront dans la suite :

Définition 1.

On note par S la population de taille n . Une hiérarchie \mathbf{H} est un ensemble de parties non vides d'individus de S (appelées classes) vérifiant :

- $S \in \mathbf{H}$;
- $\forall x \in S, \{x\} \in \mathbf{H}$;
- $\forall H_1, H_2 \in \mathbf{H}, H_1 \cap H_2 \in \{H_1, H_2, \emptyset\}$.

Une hiérarchie de classes est dite binaire si toute classe non singleton est une union de deux classes de la hiérarchie.

Définition 2.

Une hiérarchie indicée (\mathbf{H}, g) sur S est un couple formé par une hiérarchie \mathbf{H} et une fonction g définie de \mathbf{H} vers \mathbb{R}^+ , telles que :

- $\forall H_1, H_2 \in \mathbf{H}, H_1 \subset H_2$, alors $g(H_1) < g(H_2)$,
- $g(H) = 0$ si et seulement si H est un singleton.

(\subset et $<$ désignent respectivement les symboles de l'inclusion stricte et du signe inférieur strict)

L'indice g est appelé le niveau de la hiérarchie.

Une hiérarchie indicée est souvent caractérisée par la matrice ultramétrique définie par (Benzecri 1973) : $u(a, b) = \min\{g(H)/a \text{ et } b \text{ sont ensemble dans la classe } H \text{ de } \mathbf{H}\}$, ou encore par un arbre de longueur minimum obtenu à partir de la matrice ultramétrique.

Définition 3.

Une hiérarchie stratifiée sur S est une hiérarchie indicée (\mathbf{H}, g) où l'indice g est à valeurs dans un intervalle d'entiers positifs $[0, \dots, m]$, et où $m = g(S)$ est appelé la hauteur de la hiérarchie stratifiée.

Toute hiérarchie stratifiée sur S peut être définie par une suite de partitions sur S du type :

$P_0 < P_1 < \dots < P_m$, allant de la partition la plus fine P_0 en singletons à la partition la moins fine $P_m = \{S\}$. Une telle suite de partitions constitue les partitions associées à la hiérarchie stratifiée (Benkaraache 1993).

Une méthode de classification qui n'utilise que l'aspect ordinal des données produira une structure de type hiérarchie stratifiée. C'est le cas des méthodes d'agrégation du lien minimum et du lien maximum appliquées sur une matrice de dissimilarité ordinale (matrice des rangs des dissimilarités de départ). Les considérations ordinales en classification offrent la possibilité d'utiliser les outils de la théorie des graphes (Godehardt, 1990) et de conquérir de nouvelles voies pour la généralisation des structures de classification habituelles (Leclerc, 1994; Critchley et Van Cutsem, 1994; Benkaraache, 1998).

On supposera dans la suite que les dissimilarités sont sans *ex-aequo* (dissimilarité injective), ce qui implique que la matrice ordinale associée aura ses valeurs dans l'intervalle des entiers $[1 \dots N]$, avec $N = \frac{n(n-1)}{2}$. Quand la matrice de dissimilarités est injective, les dissimilarités entre classes induites par les méthodes de classification habituelles sont aussi sans *ex-aequo*. Par conséquent, les structures hiérarchiques obtenues seront dans ces conditions des hiérarchies binaires.

2. Quelques indices de complexité pour les structures hiérarchiques

Dans le but de construire des statistiques pour tester la classifiabilité, il est indispensable d'analyser la complexité des structures hiérarchiques pour connaître leurs principales caractéristiques classificatoires et de les résumer sous forme d'indicateurs (indices numériques) qui serviront plus tard comme statistiques de test.

Un indice de complexité permet d'évaluer numériquement la capacité d'une hiérarchie à vérifier un certain nombre de propriétés souhaitées. Il peut être aussi destiné à résumer certaines caractéristiques techniques de la classification hiérarchique. Un grand nombre d'indices de complexité a été proposé dans la littérature. Citons par exemple Leclerc (1985) qui présente un large choix d'indices. On se limite ici à présenter ceux que nous utiliserons dans la suite.

i) La Hauteur d'une hiérarchie stratifiée

On a déjà introduit la notion de «hauteur» d'une hiérarchie stratifiée (\mathbf{H}, g) comme étant la valeur entière $m = g(S)$ réalisée par l'indice g pour la classe grossière S . C'est aussi le plus petit niveau de la hiérarchie stratifiée où tous les éléments de S se retrouvent ensemble.

Une «bonne» classification hiérarchique produit des classes de grande taille dont les valeurs du niveau g sont faibles (c'est-à-dire de grandes classes très bas dans

l'arbre hiérarchique). Elle aura donc une hauteur faible. La hauteur peut donc être un premier indicateur sommaire de la tendance de la structure hiérarchique en question.

ii) *Un indice de complexité globale pour les hiérarchies stratifiées : L'indice IS :*

Cet indice de complexité a été introduit par Benkaraache (1993).

Soit une hiérarchie stratifiée HS de hauteur k et engendrée par la suite de partitions (P_0, \dots, P_k) . Pour tout couple (a, b) de S , on définit la quantité $M(a, b)$ par le nombre de partitions (ou de niveaux) de HS qui ne séparent pas a et b . La quantité $M(a, b)$ s'écrit en fonction de la distance ultramétrique entre a et b : $M(a, b) = k - u(a, b) + 1$.

Définition 4.

On définit un indice de complexité de HS , noté IS , par :

$$IS(HS) = \frac{1}{2} \sum_{a \neq b} M(a, b)$$

L'indice IS a été étudié initialement dans (Benkaraache, 1993; Benkaraache & Van Cutsem, 1993) où l'on montre par exemple que la valeur minimale de IS est réalisée par la hiérarchie «peigne» composée des deux partitions $(P_0, P_1 = \{S\})$. Le minimum vaut $IS \min = C_n^2$. Le maximum de IS est réalisé par la hiérarchie «Ziggourat» ou asymétrique (figure ci-dessous). Ce maximum est égal à $IS \max = \frac{n(n-1)(n+1)}{6}$.

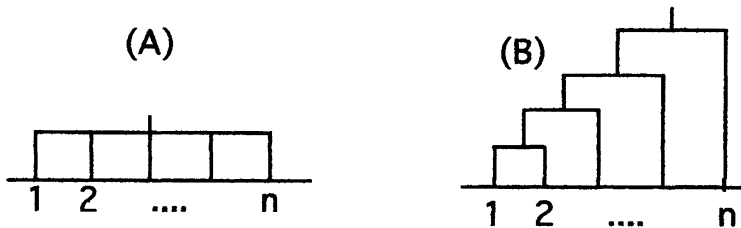


FIGURE 1
(A) : Hiérarchie peigne. (B) : Hiérarchie Ziggourat.

iii) *Temps de survie d'un élément (Leclerc, 1985)*

Soit x un élément de S . Le temps de survie de x , noté $T(x)$, est défini par le niveau de la hiérarchie où l'élément x est absorbé pour la première fois par une autre classe :

$$T(x) = m \Leftrightarrow \{x\} \in P_{m-1} \text{ et } \{x\} \notin P_m$$

La valeur $T(x)$ est un indicateur de l'isolement de l'élément x : plus $T(x)$ est grand et plus l'élément x ne ressemble pas au reste de la population.

iv) Tailles des classes contenant un élément de S :

Si on range par ordre décroissant d'inclusion les sous-ensembles de S qui sont dans la hiérarchie et qui contiennent un élément donné a , on obtient une suite de sous-ensembles emboîtés $\mathbf{A}_0 = S \supset \mathbf{A}_1 \supset \dots \supset \mathbf{A}_k = \{a\}$. On prolonge cette suite en posant $\mathbf{A}_{k+1} = \mathbf{A}_{k+2} = \dots = \mathbf{A}_n = \{a\}$. L'étude de cette suite de sous-ensembles ainsi que de la suite des cardinaux associée permet de dégager des renseignements sur la structure interne de la hiérarchie. Pour un élément donné x , on note par $t_0(x), t_1(x), \dots, t_k(x)$ (k étant la hauteur) la suite finie des tailles des classes contenant l'élément x respectivement dans les partitions P_0, P_1, \dots, P_k associées à la hiérarchie.

v) Tailles des classes obtenues à un niveau de la hiérarchie

Pour les hiérarchies stratifiées, il s'agit des classes d'une partition P_i obtenue en coupant la hiérarchie au niveau i . Dans le cas des hiérarchies binaires, il n'y a que deux classes par niveau.

Les tailles peuvent renseigner sur la tendance de la structure suivant que les classes sont plus ou moins équilibrées.

III. Variables aléatoires liées aux hiérarchies aléatoires

1. Génération aléatoire des hiérarchies

Les problèmes de la génération aléatoire uniforme et de dénombrement des structures de classification ont été réputés difficiles jusqu'à l'introduction des structures combinatoires en classification qui n'a été effective que récemment, depuis les travaux de Flajolet *et al.* (1994). En utilisant les opérations de base utilisées dans le calcul des structures combinatoires, les auteurs montrent que la plupart des structures de classification habituelles sont des structures combinatoires particulières, et peuvent donc être caractérisées par leurs fonctions génératrices.

Ces représentations combinatoires sont très adaptées pour le dénombrement et la génération aléatoire des structures de classification : le logiciel «GAIA» de Zimmermann (1994) permet maintenant d'étudier des structures combinatoires de natures très variées, en particulier les structures de classification comme les hiérarchies, les hiérarchies stratifiées, les hiérarchies stratifiées binaires, ...

2. Variables aléatoires liées aux hiérarchies aléatoires

Notons par $HS(n)$ l'ensemble des hiérarchies stratifiées sur n individus. Quand on munit $HS(n)$ d'une loi de distribution uniforme, tout indice de complexité devient une variable aléatoire. Il est alors intéressant de déterminer sa loi de distribution sous l'hypothèse d'uniformité. Reprenons les indices présentés dans le paragraphe précédent.

2.a. Distribution de la hauteur des hiérarchies stratifiées

L'idée essentielle que nous allons exploiter se développe à partir de la fameuse formule qui donne le nombre $sh(n)$ de hiérarchies stratifiées sur n individus :

$$sh(n) = \sum_{m=1}^{n-1} S(n, m)sh(m) \text{ pour } n \geq 2.$$

où $S(n, m)$ est le nombre de Stirling de second espèce (Lengyel, 1984) qui n'est autre que le nombre de partitions de n individus en m classes.

D'après l'équivalence entre une hiérarchie stratifiée et une suite de partitions ordonnées par finesse (paragraphe II.1), $sh(n)$ est aussi le nombre de suites de partitions ordonnées de la forme

$$P_0 = \{\{1\}, \dots, \{n\}\} < P_1 < \dots < P_k = \{S\}. \tag{1}$$

Étudier l'ensemble $HS(n)$ des hiérarchies stratifiées sur n individus revient donc à étudier l'ensemble $SP(n)$ des suites de partitions du type (1). Autrement dit, le tirage aléatoire uniforme d'une hiérarchie stratifiée dans l'ensemble $HS(n)$ produit une suite aléatoire de partitions ordonnée (par ordre de finesse) qui est la trajectoire d'une chaîne de Markov comme le montre le théorème suivant :

Théorème 1. – Soient P_0, P_1, \dots, P_k les partitions associées à une hiérarchie stratifiée de hauteur k . (P_0 est la partition en singletons et $P_k = \{S\}$).

Soit $M(P)$ la variable aléatoire «Nombre de classes de la partition P » et Q_n la distribution uniforme sur $HS(n)$.

i) $Q_n(P_1 = P^*) = \frac{sh(k^*)}{sh(n)}$; où k^* est le nombre de classes de P^* .

et

$$Q_n(M(P_1) = m) = \frac{S(n, m) sh(m)}{sh(n)} \text{ pour tout } m = 1..n - 1.$$

ii) Plus généralement, soit P et P' deux partitions telles que P' est moins fine que P (chaque classe de P est incluse dans une classe de P') :

$$Q_n(P_i = P' / P_{i-1} = P) = Q_{M(P)}(P_i = P') = \frac{sh(M(P'))}{sh(M(P))} \text{ pour tout } i = 1..n-1.$$

Démonstration. Utilise des arguments élémentaires de la théorie des probabilités et du calcul combinatoire. ■

Notons par K la variable aléatoire «hauteur d'une hiérarchie stratifiée sur n objets» quand l'ensemble $HS(n)$ est muni d'une distribution uniforme Q_n . Le théorème précédent permet alors de calculer la distribution de K :

– La probabilité d’avoir une hauteur égale à 1 est :

$$Q_n(K = 1) = Q_n(M(P_1) = 1) = \frac{S(n, 1) sh(1)}{sh(n)} = \frac{1}{sh(n)}$$

– La probabilité d’avoir une hauteur égale à 2 :

$$Q_n(K = 2) = \sum_{j=2}^{n-1} Q_n(M(P_1) = j)Q_j(K = 1)$$

– En général, la probabilité d’avoir une hauteur égale à k se calcule facilement par récurrence :

$$Q_n(K = k) = \sum_{j=2}^{n-1} Q_n(M(P_1) = j)Q_j(K = k - 1); \text{ pour } k = 2, \dots, n - 1.$$

En voici quelques exemples numériques :

TABLEAU 1

Distribution de la hauteur d’une hiérarchie stratifiée sur $n = 10$ individus.

$k :$	1	2	3	4	5	6	7	8	9
$P_n(K = k) :$	0.00	0.000	0.000	0.01	0.071	0.224	0.350	0.266	0.078

TABLEAU 2

Espérance mathématique de la hauteur d’une hiérarchie stratifiée pour $1 \leq n \leq 10$.

$n :$	2	3	4	5	6	7	8	9	10
$E(K) :$	1.0	1.75	2.53	3.29	4.04	4.78	5.54	6.28	7.02

De même, d’après le théorème précédent, la suite $\{M(P_0) = n, M(P_1), \dots, M(P_k) = 1\}$ des nombres de classes des partitions successives de la hiérarchie est aussi une chaîne de Markov dont le temps d’absorption par l’état «1» est égal aussi à la hauteur K . Cette approche markovienne a l’avantage de permettre d’étudier le comportement asymptotique des statistiques associées aux hiérarchies stratifiées. Van Cutsem et Ycart (1994, proposition 5.1, p. 998) montrent que la distribution asymptotique de $M(P_1)$ est une loi de Poisson tronquée :

$$\forall k \in \mathbb{N}^*, \lim_{n \rightarrow +\infty} Q_n(M(P_1) = n - k) = \frac{(\log 2)^k}{k!}.$$

et déduisent :

$$\lim_{n \rightarrow +\infty} E_n(K) = \frac{n}{2 \log 2}$$

et

$$\left(K - \frac{n}{2 \log 2} \right) / \sigma \sqrt{\frac{n}{\mu^3}} \xrightarrow{\text{loi}} N(0, 1)$$

où $\mu = 2 \log 2$ et $\sigma^2 = 2 \log(2)(1 - \log 2)$.

2.b. Distribution de l'indice de complexité ISN pour les hiérarchies stratifiées

L'indice *ISN* est la forme normalisée entre 0 et 1 (définie par $ISN = \frac{IS - IS_{\min}}{IS_{\max} - IS_{\min}}$) de l'indice *IS* présenté dans le paragraphe précédent. Une étude par simulation réalisée avec le logiciel «GAÏA» a permis d'avoir des résultats approchés de la moyenne et de l'écart-type de *ISN* ainsi que sa loi de distribution.

TABLEAU 3
Moyennes et écart-types de *ISN* obtenus avec 2 000 simulations.

<i>n</i>	moyenne	écart-type
5	0.62	0.21
10	0.53	0.16
27	0.35	0.08

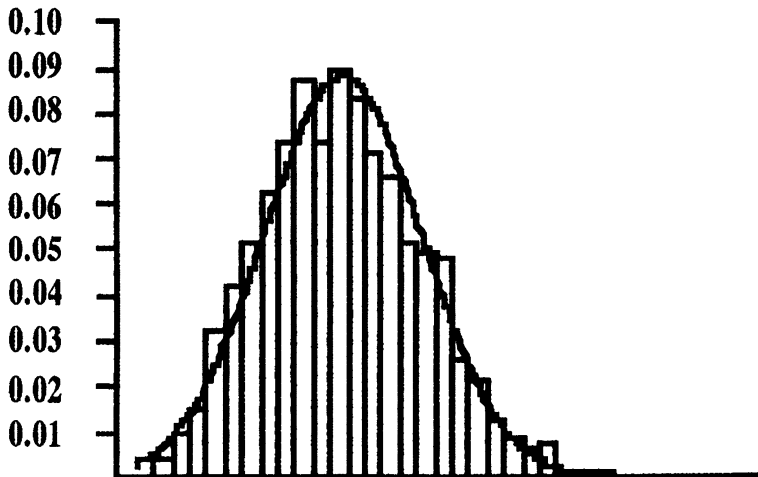


FIGURE 2
Densité de la loi de *ISN* pour $n = 27$ (sur 2 000 simulations).

Comme pour la loi de la hauteur, la loi asymptotique de l'indice ISN semble se comporter suivant une loi normale tronquée, résultat qui est loin d'être prouvé théoriquement à cause de la difficulté de manipuler la loi de ISN dans l'ensemble de toutes les hiérarchies stratifiées sur n individus.

L'indice ISN nous renseigne aussi sur la forme de la hiérarchie : la hiérarchie est d'autant plus «asymétrique» que ISN est proche de 1. Des valeurs de ISN proches de 0 laissent penser à une hauteur très faible et donc une hiérarchie proche de la hiérarchie peigne.

2.c. Cas des hiérarchies binaires.

Rappelons qu'une hiérarchie binaire est une hiérarchie où toute classe non réduite à un singleton est une union de deux autres classes de la hiérarchie. L'ensemble $HB(n)$ des hiérarchies binaires sur n individus est un ensemble fini et son cardinal $hb(n)$ vérifie l'équation :

$$\forall n \geq 2, \quad hb(n) = \frac{1}{2} \sum_{k=1}^{n-1} C_n^k hb(k) hb(n-k), \quad hb(1) = 1. \quad (\text{Murtagh1983})$$

Ce nombre est aussi égal à :

$$hb(n) = (2n-3)(2n-5)\dots 5.3.1.$$

Quand on munit l'ensemble $HB(n)$ d'une loi de probabilité uniforme, tout indice caractéristique de ce genre de structure devient une variable aléatoire. L'étude de la loi de probabilité de tels indices permettra alors de tirer des renseignements tels que les valeurs moyennes et les écarts type.

Soit B une hiérarchie binaire sur S (avec cardinal $S = n$),

Considérons la chaîne de sous-ensembles définie par un élément a de S , notée $A_0 = S \supset A_1 \supset \dots \supset A_k = \{a\}$. On prolonge cette suite en posant $A_{k+1} = A_{k+2} = \dots = A_{n-1} = \{a\}$.

Quand la hiérarchie B est aléatoire, l'entier k devient une variable aléatoire, notée K , ainsi que la suite ensembliste précédente. Plus précisément,

Théorème 2. – Soit P_n la probabilité uniforme sur l'ensemble $HB(n)$, et soit a un élément de S .

Soit B une hiérarchie binaire sur S et $A_0 = S \supset A_1 \supset \dots \supset A_K = A_{K+1} = \dots = A_{n-1} = \{a\}$ la suite des classes de B qui contiennent l'élément a .

Soit A un sous-ensemble de S qui contient a . La probabilité pour que la première partition $\{A_1, A_1^C\}$ soit la partition $\{A, A^C\}$ est donnée par

$$P_n(A_1 = A) = \frac{hb(\text{card}(A)) hb(\text{card}(A^C))}{hb(n)} \quad (2)$$

D'autre part, si A et A' sont deux sous-ensembles de S tels que $a \in A \subset A'$, alors

$$P_n(A_m = A/A_{m-1} = A') = \frac{hb(\text{card}(A)) hb(\text{card}(A' - A))}{hb(A')} \quad (3)$$

et donc

$$P_n(A_m = A/A_{m-1} = A') = P_{\text{card}(A')}(A_1 = A/A_0 = A') \quad (4)$$

Démonstration. La relation (2) découle du fait que si la première bipartition de S dans l'arbre B est définie par $\{A, A^c\}$, il y a $hb(\text{card}(A))$ arbres binaires de racine A et $hb(\text{card}(A^c))$ de racine A^c . Un raisonnement analogue permet de démontrer la formule (3). La formule (4) découle immédiatement de la formule (3). ■

Le théorème 2 montre que la suite des sous-ensembles de B contenant un élément a est une chaîne de Markov dont l'espace des états est l'ensemble des sous-ensembles de S qui contiennent a .

Considérons maintenant la suite des cardinaux des ensembles A_0, A_1, \dots, A_{n-1} . Posons, pour simplifier,

$$\forall m \in [1, \dots, n-1], X_m = \text{card}(A_m)$$

La suite $(X_m)_{1 \leq m \leq K}$ est évidemment décroissante et nous avons $X_0 = n$. L'entier K est le plus petit entier tel que $X_K = 1$. Rappelons que K est aléatoire.

Théorème 3. – *Sous les mêmes hypothèses que le théorème 2, pour tout m entier dans $[1, \dots, n]$,*

$$P_n(X_m = k/X_{m-1} = h) = C_{h-1}^{k-1} \frac{hb(k) hb(h-k)}{hb(h)} \quad (5)$$

avec $1 \leq k < h \leq n$.

En particulier,

$$P_n(X_1 = k/X_0 = n) = C_{n-1}^{k-1} \frac{hb(k) hb(n-k)}{hb(n)} \quad (6)$$

Démonstration. La bipartition de l'arbre au niveau m est définie par un sous-ensemble A de cardinal k et son complémentaire $A' - A$, où A' est le sous-ensemble du niveau $m-1$ de B qui contient l'élément a . Il y a $hb(k)$ arbres binaires de racine A et $hb(h-k)$ arbres binaires de racine $A' - A$. Il y a par ailleurs C_{h-1}^{k-1} sous ensembles de A' qui contiennent l'élément a . La formule (5) en découle immédiatement. ■

Ce théorème montre que la suite des variables aléatoires X_0, X_1, \dots, X_{n-1} , définie par l'arbre binaire aléatoire B , est une chaîne de Markov dont l'espace des états est $[1, \dots, n]$ et dont la matrice de passage est donnée par le théorème 3.

Il est aussi intéressant d'étudier la longueur de la chaîne de sous-ensembles définie par un élément a , c'est-à-dire la variable $K(a) = \inf\{m/X_m = 1\}$. Quand B est une hiérarchie aléatoire, $K(a)$ est aussi une variable aléatoire qui n'est autre que le temps d'absorption de la chaîne de Markov $\{A_m\}_{1 \leq m \leq n}$ par l'état $\{a\}$ ou celui de la chaîne $\{X_m\}_{1 \leq m \leq n}$ par l'état 1. La loi de probabilité de $K(a)$ est donnée par le théorème suivant :

Théorème 4. – Avec les mêmes notations et les hypothèses que le théorème 2,

$$P_n(K(a) = m) = \sum_{K=m}^{n-1} P_K(K(a) = m-1) P_n(X_1 = K), \text{ pour } 2 \leq m \leq n-1.$$

Démonstration. Résulte des théorèmes précédents et de la formule de Bayes, et en remarquant que pour tout $k \in [1, \dots, m-1]$, $P_k(K(a) = m-1) = 0$:

$$\begin{aligned} P_n(K(a) = m) &= P_n(X_m = 1) \\ &= \sum_{k=m}^{n-1} P_k(X_m = 1/X_1 = k) P_n(X_1 = k) \\ &= \sum_{k=m}^{n-1} P_k(K(a) = m-1) P_n(X_1 = k) \end{aligned}$$

■

L'expression du théorème peut s'écrire matriciellement. On obtient ainsi, pour tout $n \geq 2$,

$$\begin{pmatrix} P_n(K(a) = 2) \\ P_n(K(a) = 3) \\ \dots \\ P_n(K(a) = n-1) \end{pmatrix} = \begin{pmatrix} P_2(K(a) = 1) & P_3(K(a) = 1) & \dots & P_{n-1}(K(a) = 1) \\ 0 & P_3(K(a) = 2) & \dots & P_{n-1}(K(a) = 2) \\ \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & P_{n-1}(K(a) = n-2) \end{pmatrix} \begin{pmatrix} P_n(X_1 = 2) \\ P_n(X_1 = 3) \\ \dots \\ P_n(X_1 = n-1) \end{pmatrix}$$

Nous pouvons ainsi calculer de proche en proche la loi de probabilité de la variable aléatoire $K(a)$.

Nous abordons maintenant l'étude des sous-ensembles de la première partition (obtenue en partageant la classe S) d'une hiérarchie binaire B . Cette étude sera évidemment adaptable à toute autre partition en deux sous-ensembles de chacune des classes internes de B .

Théorème 5. – Soit P_n la probabilité uniforme sur l'ensemble $HB(n)$ des hiérarchies binaires. Soit B une hiérarchie binaire et soit $\{A_1, A_2\}$ la première partition obtenue en divisant S en deux classes (au niveau le plus haut de la hiérarchie). Cette étude sera évidemment adaptable à la partition en deux sous-ensembles de chacun des nœuds de l'arbre différent d'un singleton.

Supposons que $\text{card}(A_1) \leq \text{card}(A_2)$. Soit X la variable aléatoire définie par le cardinal de A_1 .

$$a) P_n(\{A_1, A_2\} = \{A, A^c\}) = \frac{hb(\text{card}(A)) hb(n - \text{card}(A))}{hb(n)}$$

$$b) \forall k \in \mathbb{N}^*, 1 \leq k \leq \frac{n}{2}$$

$$P_n(X = k) = \begin{cases} \frac{1}{2} C_n^k \frac{hb(k) hb(n - k)}{hb(n)} & \text{si } n = 2k \\ C_n^k \frac{hb(k) hb(n - k)}{hb(n)} & \text{sinon} \end{cases}$$

Démonstration. On procède comme précédemment :

a) Il y a $hb(\text{card}(A))$ hiérarchies binaires sur A et $hb(n - \text{card}(A))$ hiérarchies binaires sur A^c . Ceci entraîne la formule annoncée.

b) Le résultat se déduit de celui de a). Les nombres de combinaisons proviennent du nombre de choix du sous-ensemble A dans S . Le facteur $\frac{1}{2}$ est dû aux rôles symétriques de A et A^c quand $n = 2k$. ■

Ces formules permettent de calculer les premières valeurs numériques de $P_n(X = k)$: (voir tableau 4 page suivante).

On remarque, au vu de ces valeurs, que les probabilités décroissent de la gauche vers la droite, c'est-à-dire que les petites valeurs de X sont plus probables que celles voisinant $n/2$. Ceci montre que les hiérarchies binaires déséquilibrées (asymétriques) sont plus fréquentes dans $HB(n)$. (voir tableau 5 page suivante).

On obtient facilement une approximation asymptotique de la loi de X :

Théorème 6. – Avec les notations du théorème 5, nous avons pour tout entier positif donné non nul k ,

$$\lim_{n \rightarrow +\infty} P_n(X = k) = \frac{hb(k)}{k!2^k} \text{ et } \lim_{n \rightarrow +\infty} P_n(X = n - k) = \frac{hb(k)}{k!2^k}$$

TABLEAU 4
Probabilités de $P_n(X = k)$, $1 \leq n \leq 10$.

k	1	2	3	4	5
n = 1	1				
2	1				
3	1				
4	0.8	0.2			
5	0.714	0.285			
6	0.66	0.238	0.095		
7	0.636	0.212	0.151		
8	0.615	0.195	0.130	0.058	
9	0.6	0.184	0.117	0.097	
10	0.588	0.176	0.108	0.086	0.040

TABLEAU 5
Espérances de X , $1 \leq n \leq 10$.

n	1	2	3	4	5	6	7	8	9	10
$E_n(X)$	1	1	1	1.2	1.284	1.421	1.513	1.627	1.707	1.808

Démonstration. Tout d'abord, n tendant vers $+\infty$ et k étant un entier positif non nul, on peut toujours prendre $n > 2k$ et donc $k \neq \frac{n}{2}$.

Nous avons alors,

$$P_n(X = k) = C_n^k \frac{hb(k) hb(n-k)}{hb(n)} = \frac{hb(k)}{k!} \frac{n(n-1)\dots(n-k+1)}{(2(n-1)-1)\dots(2(n-k)-1)}$$

et il en résulte

$$\lim_{n \rightarrow +\infty} P_n(X = k) = \frac{hb(k)}{k! 2^k}.$$

La deuxième limite découle du fait que $P_n(X = k) = P_n(X = n - k)$. ■

La loi limite de X quand n est grand est donnée dans la table suivante :

TABLEAU 6
Limites des probabilités $P_n(X = k)$, $1 \leq k \leq 10$.

k	1	2	3	4	5	6	7	8	9	10
$P_n(X = k)$	0.5	0.125	0.063	0.039	0.027	0.021	0.016	0.013	0.011	0.009

IV. Tests de tendance d'une classification hiérarchique.

Les variables aléatoires étudiées dans le paragraphe précédent peuvent être utilisées comme statistique pour tester la tendance d'une classification hiérarchique. Il faut toutefois noter que la qualité du test dépend en premier lieu de la statistique utilisée et du type d'information classificatoire qu'elle résume. Autrement dit, une structure de classification observée pour laquelle la majorité des tests de tendance acceptent l'hypothèse nulle a de faibles chances pour qu'elle soit non aléatoire.

1. Test de la hauteur

Une bonne classification hiérarchique produit des classes de grande taille très bas dans la hiérarchie, donc sera de hauteur relativement faible. On peut donc se fixer, pour un niveau de significativité donné, un seuil théorique et prendre comme région de rejet de l'hypothèse «La structure est aléatoirement tirée suivant la loi uniforme» l'ensemble $\{K < t_\alpha\}$.

Si le nombre d'individus n est grand, on peut utiliser l'approximation asymptotique gaussienne de la loi de la hauteur (voir III.2.a.). En particulier, une hiérarchie aléatoire sur n individus (pour n grand) a une hauteur moyenne de $n/(2 \log 2)$.

2. Test à l'aide du nombre de classes de la partition du premier niveau

On note par $M(P_1)$ le nombre de classes de la partition du premier niveau de HS. Nous avons établi la loi exacte de $M(P_1)$ dans (III.2.a). Pour n grand, on a l'approximation asymptotique suivante :

$$\forall k \in \mathbb{N}^*, \lim_{n \rightarrow +\infty} Q_n(M(P_1) = n - k) = \frac{(\log 2)^k}{k!}$$

Voici les premières probabilités de cette loi :

TABLEAU 7

k	1	2	3	4	5	...
P_k	0.693	0.240	0.056	0.010	0.000	...

L'interprétation que l'on peut faire de ces probabilités est que, pour une hiérarchie aléatoire, la différence entre les nombres de classes de deux niveaux consécutifs est de 1 dans 69,3 % des cas, de 2 dans 24 % des cas et de plus de 3 dans 7 % des cas. Il est donc clair qu'une hiérarchie sur 50 individus par exemple, présentant dans le premier niveau une quarantaine de classes (par exemple) ne peut pas être assimilée à une hiérarchie aléatoire. Un test de détection de structure aléatoire basé sur $M(P_1)$ aura comme région de rejet $\{M(P_1) > t_\alpha\}$.

3. Test basé sur l'indice de complexité IS

L'étude de la répartition de l'indice IS (ou ISN , sous forme normalisée) montre que si la valeur observée de ISN se trouve dans la queue de la distribution la structure observée a peu de chance d'être aléatoire. La région de rejet du test sera de la forme $[t_{\frac{\alpha}{2}}, t_{1-\frac{\alpha}{2}}]$.

4. Test basé sur la taille des classes (pour les hiérarchies binaires)

Les théorèmes 6 et 7 donnent la loi de distribution de la variable aléatoire $X =$ La plus petite taille des deux classes de la première bipartition de S , sous l'hypothèse d'uniformité. On remarque que cette distribution charge les petites valeurs et décroît vite vers 0 dès que la taille atteint quelques unités. En particulier, pour de grandes valeurs de n , la loi limite (cf. table 6) montre que, dans 50 % des cas on forme des classes dans une hiérarchie binaire aléatoire en ajoutant un nouveau élément à une autre classe, dans 12,5 % l'agglomération se fait entre une classe à deux éléments et une autre classe et très peu de chance de voir deux classes de tailles dépassant 4 ou 5 individus se réunir. On peut donc proposer des tests de détection de structure aléatoire pour n fixé ou pour n grand. La région de rejet aura la forme $\{X > t_{\alpha}\}$ pour un niveau de confiance $1 - \alpha$.

Références

- BENKARAACHE T. (1993), «Problèmes de validité en classification et quelques généralisations aux ultramétriques à valeurs dans un ensemble ordonné», Thèse de doctorat, Université Joseph Fourier, Grenoble. France.
- BENKARAACHE T., VAN CUTSEM B. (1993), «Comparing hierarchical classifications», *Classification and related methods of data analysis*, in, the 4th conference of the International Federation of Classification Societies, IFCS, Paris.
- BENKARAACHE T. (1998), «L'ultramétrie inférieure maximum d'une dissimilarité à valeurs dans un inf-demi-treillis», *Math. Inf. Sci. hum.* 143, pp. 27-40.
- BENZÉCRI J. P. (1973), *L'analyse des données – I – La taxinomie*, Paris, Dunod, 1973.
- CRITCHLEY F., VAN CUTSEM B. (1994), «An order – theoretic unification and generalization of certain fundamental bijections», in Bernard Van Cutsem (éditeur), in *Classification and Dissimilarity Analysis*, Lecture Notes in Statistics, vol. 93, New York, Springer Verlag, pp. 87-148.
- FLAJOLET P., ZIMMERMANN P., VAN CUTSEM B. (1994), «A calculus for the random generation of combinatorial structures», *Theoret. Comput. Sci.* 132, pp. 1-35.
- GODEHARDT E. (1990), «Graphs as structural models. The application of graphs and multigraphs in cluster analysis», Braunschweig : Vieweg.

- GORDON A.D. (1994), «Identifying genuine clusters in a classification», *Computational Statistics and Data Analysis*, vol. 18, pp. 561-581.
- JAIN A.K., DUBES R.C. (1988), *Algorithms for clustering data*, Englewood Cliffs, NJ : Prentice Hall.
- LECLERC B. (1985), «La comparaison des hiérarchies : indices et métriques», *Math. Sc. Hum.* 92, pp. 5-40.
- LECLERC B. (1994), «The residuation model for the ordinal construction of dissimilarities and other valued objects», in Bernard Van Cutsem, (Eds.), *Classification and Dissimilarity Analysis, Lecture notes in Statistics (93)*, New York, Springer Verlag, pp. 149-172.
- LENGYEL T. (1984), «On a recurrence involving Stirling numbers», *European. J. Combin.* 5, pp. 313-321.
- MURTAGH F. (1983), «Counting dendrograms : a survey», *Discrete appl. math.* 7, pp. 191-199.
- SNEATH P.H., SOKAL R.R. (1973), *Numerical Taxinomy*, Freeman, San Francisco.
- VAN CUTSEM B., YCART B. (1994), «Renewal type behaviour of absorption times in Markov chains», *Adv. Appl. Probab.* 26, pp. 988-1005.
- ZIMMERMANN P. (1994), «Gaïa : a package for the random generation of combinatorial structures», *Maple Technical Newsletter* 1.