JOURNAL DE LA SOCIÉTÉ STATISTIQUE DE PARIS

BRANISLAV IVANOVITCH

Examen de la qualité d'une classification hiérarchique

Journal de la société statistique de Paris, tome 126, n° 2 (1985), p. 48-54

<http://www.numdam.org/item?id=JSFS_1985__126_2_48_0>

EXAMEN DE LA QUALITÉ D'UNE CLASSIFICATION HIÉRARCHIQUE (*)

Branislav IVANOVITCH
Professeur-docteur, Université de Belgrade, Yougoslavie

Dans la pratique, avant d'adopter une classification hiérarchique type, il est nécessaire d'examiner si le partage en classes de chaque niveau est uniformément fin et si la classification est équilibrée et cohérente par rapport à un caractère mesuré sur les éléments de l'ensemble de base.

Dans ce but, on donne deux tests simples (test d'équilibre et test de cohérence) utiles pour : a) *estimer si la classification type proposée est acceptable ;* b) *estimer si la classification révisée est meilleure que l'originale, ou* c) *choisir entre plusieurs propositions alternatives d'une classification type.*

In the practice, before to adopt some hierarchical standard classification, it is necessary to examine if the division in classes of its each hierarchical level is uniformly fine and if the classification is well balanced and coherent with regard to some variable observed.

Herewith, two simple tests are given (test of equilibrium and test of coherence) that may be useful to estimate if: a) *the proposed standard classification is acceptable, if* b) *the revised classification is better than its original version or* c) *to make choice between several alternatives of the standard classification.*

I – INTRODUCTION

Il existe dans un grand nombre de pays une nécessité croissante de faire, sur le plan gouvernemental ou scientifique, des analyses comparatives, nationales ou internationales, des phénomènes qui se présentent sous une forme structurale dont les parties sont des agrégations des éléments d'un ensemble de base.

Pour que ces analyses soient efficaces et les résultats comparables au cours du temps ou entre différents pays ou régions, par rapport à certaines catégories économiques ou sociales, elles devraient se baser toujours sur une classification type de cet ensemble d'éléments.

La majorité des classifications types internationales ont été élaborées après la deuxième guerre mondiale par le Secrétariat et les différentes agences des Nations Unies et recommandées à tous les pays membres de l'O.N.U.

En général, une classification type est hiérarchique, à plusieurs niveaux (étages), et la présentation se fait sous la forme d'une nomenclature, c'est-à-dire d'une liste de dénominations codées des classes d'éléments de cette classification.

Dans la pratique, avant d'adopter une classification type, il faut se poser la question si le partage des classes de chaque niveau est uniformément fin et si la classification est équilibrée et cohérente par rapport à un caractère qu'on mesure sur les éléments de l'ensemble de base. Ce sont quelques propriétés indispensables pour qu'une classification puisse être considérée comme valable.

Notre intention est de présenter au cours de cette conférence deux procédés d'examen de ces propriétés, deux procédés d'ailleurs très simples mais qui pourraient nous aider à estimer si la proposition d'une classification type est acceptable, si la révision d'une classification type est meilleure que sa version précédente ou à faire le choix parmi plusieurs propositions alternatives d'une classification type.

(*) Communication faite le 23 janvier 1985 devant les Sociétés de statistique de Paris et de France.
Journal de la Société de statistique de Paris, tome 126, n° 2, 1985.

II – Propriétés d'une classification hiérarchique par rapport a la finesse du partage des classes

Soit S un ensemble statistique à N éléments. Toute partition de S représente une classification dont les classes sont des parties de S. En subdivisant ces classes nous aurons une nouvelle classification de S qui sera plus fine que la première. En continuant avec la subdivision nous formerons une suite de classifications, chacune d'elles plus fine que la précédente. Ce procédé est possible jusqu'au moment où toutes les classes ne possèdent qu'un seul élément.

Ainsi, nous obtenons une chaîne de classifications :

$$K_0 = \{ S \} , K_1, \ldots, K_r, \ldots, K_s, \ldots, K_n , \qquad n \leqslant N , \tag{2.1}$$

la première contenant une seule classe, l'ensemble S, et la dernière des classes unitaires.

Si tous les éléments de n'importe quelle classe de K_s sont toujours dans une même classe de chaque K_r, $r<s$, la chaîne (2.1) représente une classification hiérarchique $\mathcal{H}(K)$ de S.

$$\mathcal{H}(K) = \{ K_0, K_1, \ldots, K_p \} \tag{2.2}$$

La classification la plus fine K_p de la hiérarchie $\mathcal{H}(K)$ s'appelle classification terminale, qui n'est pas obligatoirement composée des classes unitaires, surtout dans le cas de N très élevé. L'indice p indique la longueur de la chaîne de $\mathcal{H}(K)$.

En mesurant le caractère X sur les éléments de S, nous pouvons définir aussi la structure hiérarchique de X par rapport à la classification hiérarchique $\mathcal{H}(K)$ de l'ensemble S, en la désignant par $\{ X, S, \mathcal{H}(K) \}$.

En principe toute classe A_{ir} de la classification K_r devrait être plus « importante » que toute classe A_{js} d'une classification plus fine K_s de $\mathcal{H}(K)$.

Nous pouvons mesurer cette « importance » soit par son cardinal,

$$z_1 = | A_{ir} | , \tag{2.3}$$

soit par le nombre de terminaux qu'elle contient

$$z_2 = | \{ A_{jp} \mid A_{jp} \subseteq A_{ir} \} | , \tag{2.4}$$

soit par la valeur totale du caractère X, mesuré sur les éléments qu'elle contient,

$$z_3 = x_{ir} = \sum_{j}^{z_2} x_{jp} , \tag{2.5}$$

x_{jp} étant la valeur de X du j-ème terminal contenu dans A_{ir}.

Dans le cas où la classification terminale est unitaire, z_1 est identique avec z_2. Si $x_{ir} = z_2(A_{ir})$, z_2 est identique avec z_3.

Si

$$\forall A_{ir}, A_{js} \, [A_{ir} \in K_r \quad et \quad A_{js} \in K_s \quad et \quad r<s \Rightarrow z_2(A_{ir}) \geqslant z_2(A_{js})],$$

la classification hiérarchique $\mathcal{H}(K)$ est *graduelle*.

Si

$$\forall A_{ir}, A_{js} \, [A_{ir} \in K_r \quad et \quad A_{js} \in K_s \quad et \quad r<s \Rightarrow z_3(A_{ir}) \geqslant z_2(A_{js})],$$

la classification hiérarchique est *équilibrée* par rapport à X.

Si

$$\forall A_{ir} \mid A_{ir} \in K_r \quad et \quad r \in \{ 1, \ldots, p \} \Rightarrow \frac{z_2(A_{ir})}{z_3(A_{ir})} = \text{Const.},$$

c'est-à-dire si le total de X de chaque classe de K_r est proportionnel au nombre de ses terminaux, la classification hiérarchique $\mathcal{H}(K)$ est *cohérente* par rapport à X.

Enfin, si $\mathcal{H}(K)$ est à la fois graduelle, équilibrée et cohérente par rapport à X, nous dirons qu'elle est *parfaite* par rapport à X.

Remarquons que si X est le nombre de terminaux, la classification hiérarchique $\mathcal{H}(K)$ sera graduelle si elle est équilibrée par rapport à X. C'est à cause de cette raison qu'il suffit, dans l'examen de la qualité d'une classification hiérarchique par rapport à X, d'examiner seulement son équilibre et sa cohérence par rapport à X.

III – ÉQUILIBRE DE LA HIÉRARCHIE $\mathcal{H}(K)$ PAR RAPPORT A X

Soit $x_{ir} = X(A_{ir})$ et

$$a_r = \min_r \{ x_{ir} \}, \quad b_r = \max_r \{ x_{ir} \},$$

où $r \in \{ 1,\ldots,p \}$, $i \in \{ 1,\ldots,k_r \}$ et k_r le nombre de classes de K_r.

L'intervalle $I_r = [a_r ; b_r]$ contient les points des valeurs de X de toutes les classes de K_r.

Si $\bigcap_{r=1}^{p} I_r = \emptyset$, la hiérarchie $\mathcal{H}(K)$ est équilibrée par rapport à X.

Si $\bigcap_{r=1}^{p} I_r \neq \emptyset$, le déséquilibre de $\mathcal{H}(K)$ est d'autant plus fort que le mélange des points des différents K_r est plus grand.

Dans le cas $p = 2$, nous avons deux ensembles de points $K_1 = \{x_{i1}\}$ et $K_2 = \{x_{j2}\}$ et le degré d'équilibre de la hiérarchie $\mathcal{H}(K)$, par rapport à X, pourrait être représenté comme coefficient de séparation de deux ensembles de points K_1 et K_2, c'est-à-dire :

$$\sigma_{12} = \frac{k_1 k_2 (\bar{x}_1 - \bar{x}_2)}{\sum_{i=1}^{K_1} \sum_{j=1}^{K_2} | x_{i1} - x_{j2} |} \tag{3.1}$$

Etant donné que

$$K_2 \text{ plus fine que } K_1 \Rightarrow k_1 \leqslant k_2 \Rightarrow \bar{x}_1 \geqslant \bar{x}_2 \Rightarrow \sigma_{12} \geqslant 0$$

Si les classifications K_1 et K_2 sont identiques, il sera

$$k_1 = k_2 \Rightarrow \bar{x}_1 = \bar{x}_2 \Rightarrow \sigma_{12} = 0$$

Si $\mathcal{H}(K)$ est équilibrée par rapport à X, les ensembles des points K_1 et K_2 seront séparés et le coefficient d'équilibre sera égal à un.

Dans le cas d'une hiérarchie à p étages, désignons par σ_{rs} le coefficient de séparation des ensembles de points K_r et K_s, où $r<s$ et $\{r,s\} \subseteq \{1,\ldots,p\}$.

Nous pouvons maintenant définir comme mesure d'équilibre de $\mathcal{H}(K)$ par rapport à X, le produit

$$\sigma_{1..p} = \prod_{r,s>r}^{p} \sigma_{rs} \tag{3.2}$$

Il est évident que

$$0 \leqslant \sigma_{1..p} \leqslant 1$$

et que ce coefficient sera égal à 1 si $\bigcap_{r=1}^{p} I_r = \emptyset$.

Exemple : Prenons comme exemple la révision de la Classification type de catégories professionnelles qui a été effectuée en Yougoslavie en 1976.

C'est une classification hiérarchique à quatre étages et elle est généralement utilisée par l'Office fédéral de statistique pour établir des structures professionnelles des travailleurs yougoslaves.

On a utilisé des données statistiques de l'année 1974 afin d'examiner si la classification révisée est plus équilibrée par rapport au nombre de travailleurs (X) que sa version précédente.

Le tableau 1 donne des valeurs des coefficients d'équilibre σ_{rs} de la classification antérieure. On remarque tout de suite que les mélanges existent partout, entre tous les quatre niveaux de classification. Il est vrai que les mélanges entre le premier et le quatrième et entre le premier et le troisième niveau peuvent être considérés comme négligeables mais c'est surtout le déséquilibre de K_3 par à rapport à K_4 qui est considérable et qui met en doute la bonne qualité de cette classification.

TABLEAU 1

Matrice des coefficients d'équilibre de la C.T.C.P. yougoslave pour l'année 1974

Niveau	1	2	3	4
1	0,0000	0,9743	0,9987	0,9996
2		0,0000	0,7188	0,8440
3			0,0000	0,3207
4				0,0000

TABLEAU 2

Matrice des coefficients d'équilibre de la C.T.C.P. yougoslave, révisée, pour l'année 1974

Niveau	1	2	3	4
1	0,0000	0,9399	0,9879	0,9971
2		0,0000	0,6239	0,8939
3			0,0000	0,6158
4				0,0000

Tandis que la classification K_3 est devenue, après la révision, bien plus équilibrée, les classifications K_1 et K_2 ont perdu quelque peu de leur équilibre.

Etant donné que les coefficients d'équilibre de la C.T.C.P. initiale $\mathcal{H}(K)$ et de la C.T.C.P. révisée $\mathcal{H}^*(K)$ sont respectivement

$$\sigma_{1234} = 0{,}189 \qquad \text{et} \qquad \sigma^*_{1234} = 0{,}318,$$

on pourrait conclure que la C.T.C.P. révisée est, globalement considéré, plus équilibrée que la C.T.C.P. initiale, mais que l'amélioration n'est pas générale.

L'inconvénient du coefficient global d'équilibre est qu'il est nul si un seul des σ_{rs} est nul. Cependant, $\sigma_{rs} = 0$ signifie l'identité des classifications K_r et K_s et il n'y a pas de raison de dire que la hiérarchie est dans ce cas totalement déséquilibrée.

Il serait peut-être plus approprié d'utiliser le coefficient

$$\sigma'_{1..p} = \frac{\sum_{r,s>r}^{p} k_r k_s (\bar{x}_r - \bar{x}_s)}{\sum_{r,s>r}^{p} \sum_{i}^{K_r} \sum_{j}^{K_s} | x_{ir} - x_{js} |}, \quad r<s, \tag{3.3}$$

c'est-à-dire, après la réduction,

$$\sigma'_{1..p} = \frac{\sum_{r,s>r}^{p} k_r k_s (\bar{x}_r - \bar{x}_s)}{\sum_{r,s>r}^{p} k_r k_s (\bar{x}_r - \bar{x}_s)/\sigma_{rs}}. \tag{3.4}$$

Ce coefficient n'est nul que si toutes les classifications de $\mathcal{H}(K)$ se réduisent en une seule. Si $\mathcal{H}(K)$ est équilibrée, sa valeur sera égale à 1, car

$$\bigcap_{r=1}^{p} I_r = \varnothing \Rightarrow \forall_{rs} [\{r,s\} \subseteq \{1,\ldots p\} \Rightarrow \sigma_{rs} = 1] \Rightarrow \sigma'_{1..p} = 1.$$

IV – Coefficient d'équilibre de $\mathcal{H}(K)$ au cas où le nombre de classes est grand

Si le nombre de classes de la classification terminale de $\mathcal{H}(K)$ est très élevé, le calcul du coefficient d'équilibre devient encombrant. Ainsi, pour examiner l'équilibre de la Classification type pour le commerce international (la C.T.C.I. de l'O.N.U.), par rapport aux valeurs du commerce international (X), il faut calculer 2 359 163 différences absolues.

D'autre part, la sensibilité de ce coefficient pourrait nous paraître très forte, car il suffit qu'une seule des différences $(x_{ir} - x_{js})$, $r<s$, soit négative pour qu'il soit inférieur à 1.

C'est pour cette raison que nous allons utiliser, si le nombre de terminaux de $\mathcal{H}(K)$ est très élevé, un raisonnement plus statistique et dire que le mélange entre les ensembles de points de K_r et K_s est négligeable si

$$\bar{x}_r - \bar{x}_s \geqslant \alpha\,(\sigma_r + \sigma_s)\,, \tag{4.1}$$

où σ_r et σ_s sont des écarts types de X, respectivement dans K_r et K_s, $r<s$ et $\alpha>0$ un nombre arbitraire qui dépend de notre critère de tolérence.

Une mesure statistique de l'équilibre entre les classifications K_r et K_s pourrait être maintenant exprimée par le coefficient

$$\sigma''_{rs} = \frac{\bar{x}_r - \bar{x}_s}{\sigma_r + \sigma_s}\,, \tag{4.2}$$

où $\alpha = 1$, $r<s$. C'est un coefficient non-négatif et (4.1) est vrai et $\alpha = 1 \Rightarrow \sigma''_{rs} \geqslant 1 \Rightarrow$

$\Rightarrow K_r$ est équilibré par rapport à K_s

Dans le cas du déséquilibre : $\sigma'_{rs}<1$.

Ce coefficient est une mesure de distance, puisque même au cas de la séparation totale, sa valeur varie avec la distance des moyennes de X, entre les classifications K_r et K_s.

Le coefficient global d'équilibre de la hiérarchie $\mathcal{H}(K)$, par rapport à X, sera

$$\sigma''_{1\ p} = \frac{\sum_{r=1}^{p} \sum_{s>r}^{p} (\bar{x}_r - \bar{x}_s)}{\sum_{r=1}^{p} \sum_{s>r}^{p} (\sigma_r + \sigma_s)}\,, \tag{4.3}$$

ou, après la réduction,

$$\sigma''_{1\ p} = \frac{\sum_{i=1}^{p} (p - 2i + 1)\,\bar{x}_i}{(p-1)\sum_{i=1}^{p} \sigma_i} \tag{4.4}$$

Il est évident que ce coefficient est non-négatif et si $\sigma''_{1\ p} \geqslant 1$, nous dirons que la hiérarchie $\mathcal{H}(K)$ est équilibrée par rapport à X.

V – Cohérence d'une classification hiérarchique par rapport a X

Soit $n_i(r,s)$ le nombre des classes A_{js} de K_s appartenant à A_{ir} de K_r, c'est-à-dire

$$n_i(r,s) = |\ \{ A_{js} \mid A_{js} \subseteq A_{ir} \}\ |. \tag{5.1}$$

Les classifications K_r et K_s seront cohérentes par rapport au caractère X si la finesse du partage des classes de K_r en classes de K_s est proportionnelle aux valeurs correspondantes de X, c'est-à-dire

$$\forall_i \left[i \in \{1,\ldots,k_r\} \quad et \quad r<s \Rightarrow \frac{n_i(r,s)}{x_{ir}} = \text{Const.} \right],$$

où k_r est le nombre de classes de K_r.

Si cette proposition est vraie, le coefficient de corrélation de la suite des couples

$$\{ < n_i(r,s) ; x_{ir} > \} , \qquad i \in \{ 1,\ldots,k_r \} ,$$

sera égal à 1.

Si elle n'est pas vraie, la valeur de ce coefficient sera d'autant plus petite que 1 que l'incohérence est plus forte.

C'est pour cette raison que nous avons pris le coefficient

$$R [n(r,s) ; X_r] \tag{5.2}$$

comme mesure de cohérence de la classification K_r vis-à-vis de la classification K_s et par rapport au caractère X.

Si $s = p \Rightarrow n_i(r,p) = z_2(A_{ir}) \quad et \quad x_{ir} = z_3(A_{ir})$

de sorte que

$$R [K_r ; X] = R [z_2(r) ; z_3(r)] \tag{5.3}$$

Alors, nous dirons simplement que R représente la mesure de cohérence de la classification K_r de $\mathcal{H}(K)$ par rapport à X.

Le coefficient de cohérence de $\mathcal{H}(K)$ par rapport par X sera présenté par le produit

$$R [\mathcal{H}(K) ; X] = \prod_{r=1}^{p-1} R [z_2(r) ; z_3(r)] \tag{5.4}$$

D'après cette définition, une hiérarchie sera cohérente par rapport à X si toutes ses classifications en sont cohérentes.

Remarquons, cependant, que ce coefficient est rarement utilisé dans la pratique.

VI – Cohérence de la Classification Type pour le Commerce International de l'O.N.U.

Afin d'assurer la comparabilité des statistiques du commerce international, un comité d'experts statisticiens de la Société des Nations a publié, en 1938, une liste minimum de marchandises pour les statistiques du commerce international.

Depuis, le progrès technologique dans la production industrielle, les développements des moyens de transport et d'information ainsi que l'accroissement de la diversification du goût et des besoins du consommateur ont sensiblement changé la structure du commerce international et déjà, à plusieurs reprises, il a fallu réformer ou réviser la classification internationale des marchandises.

Au moment où je me suis préoccupé de cette classification, on utilisait la C.T.C.I., Rév. 1, qui représentait une synthèse de la C.T.C.I. initiale, adoptée par l'Assemblée Générale de l'O.N.U. en 1950, et de la Nomenclature Douanière de Bruxelles (la N.D.B.), adoptée par le Conseil de Coopération Douanière, en 1955. C'était une classification hiérarchique, à cinq étages, de l'ensemble de marchandises entrant dans le commerce international. Le partage en 10 sections est le moins fin et chaque section est identifiée par le premier chiffre du code de la C.T.C.I. Les sections se désagrègent en 56 divisions et les divisions en 177 groupes fournissant les données dont les utilisateurs ont le plus souvent besoin dans leurs analyses économiques.

A leur tour, les groupes se désagrègent en 625 sous-groupes dont 257 d'entre eux se divisent en 944 positions subsidiaires.

Ainsi, la C.T.C.I., Rév. 1, contenait 1 312 positions (classes terminales), chacune d'elles étant codée par cinq chiffres.

Chaque position de la C.T.C.I. permet l'identification automatique de la N.D.B. du fait qu'il existe la clé de conversion entre les deux classifications.

Tandis que les dix sections sont définies d'après la nature des grands groupes de marchandises, le partage par divisions, ou par groupes, devrait apporter une distribution plus équitable par rapport au nombre de positions et à la valeur du commerce international.

Cependant, la cohérence de la classification à deux chiffres (par divisions) de la C.T.C.I., Rév. 1, par rapport aux valeurs des exportations du commerce international était assez faible. Ainsi, pour l'année 1971 elle n'était que de 0,206. Le coefficient de cohérence de la classification à trois chiffres (par groupes) n'était pas bien plus élevé ($R_3 = 0,479$) pour que l'on puisse dire qu'il existe une assez forte tendance stochastique entre le nombre de positions par groupes et la valeur des exportations mondiales.

Une deuxième révision de la C.T.C.I. a été élaborée au cours des années 1969-1972 par un groupe d'experts *ad hoc*, désigné par l'Office de Statistique de l'O.N.U., et du Conseil de Coopération Douanière de Bruxelles. Entre autres recommandations d'ordre général, était celle de l'établissement d'un rapport plus fonctionnel entre la finesse du partage et l'importance dans le commerce mondial de chaque division et de chaque groupe. Malheureusement, la pression de certaines forces, comme par exemple celle des chimistes sur le plan sectoriel ou celle des pays développés vis-à-vis du groupe des pays en voie de développement ont bien baissé la qualité de la nouvelle révision de la C.T.C.I.

Ainsi, pour les exportations, la valeur du coefficient de cohérence de la classification à deux chiffres a baissé en 1976 à 0,155 et celle de la classification à trois chiffres à 0,324.

Les valeurs correspondantes pour la France (0,302 et 0,476) sont relativement assez bonnes par rapport aux moyennes mondiales, mais elles ne sont pas si satisfaisantes par rapport aux autres pays développés (voir le tableau 3).

TABLEAU 3

Coefficients de cohérence des classifications de la C.T.C.I., Rév. 2, de quelques pays développés

Pays	Niveau de classification par		
	groupes	divisions	sections
Autriche	0,460	0,637	0,836
Suisse	0,382	0,548	0,539
Italie	0,359	0,543	0,594
Grande-Bretagne	0,332	0,515	0,495
Pays-Bas	0,319	0,401	0,459
France	0,302	0,476	0,567
Portugal	0,251	0,426	0,913
Yougoslavie	0,238	0,416	0,807
Suède	0,220	0,347	0,548
Danemark	0,195	0,279	0,329
Finlande	0,157	0,157	0,868
Irlande	0,123	0,143	0,348
Norvège	0,059	0,229	0,593
Islande	−0,040	−0,089	0,150

Pour les pays en voie de développement ces chiffres sont plutôt négatifs. Ainsi, les chiffres correspondants pour l'Iraq sont : (−0,055; −0,039; −0,305).

Dans les analyses économiques des problèmes liés au commerce international d'un pays, il est important que la C.T.C.I. corresponde bien à ses structures des exportations et des importations. En d'autres termes, les classes les plus restrictives de la C.T.C.I. comprenant les principaux groupes de marchandises exportées ou importées par ce pays ne devraient pas contenir d'autres marchandises, afin que les données statistiques, pour lesquelles ce pays a un intérêt spécial, soient explicites et pas noyées dans d'autres chiffres.

C'est pour cette raison qu'on pourrait supposer qu'il y a des pays où l'on commence déjà à penser à la nouvelle (troisième) révision de la C.T.C.I., qui aura lieu dans quelques années, et à préparer des propositions, avec des arguments valables et qui s'accorderaient bien avec leurs propres intérêts.