REVUE DE STATISTIQUE APPLIQUÉE

J.-M. TRICOT
R. LATOUR

Une approche de la comparaison de profils à l'aide du coefficient de concentration généralisé

Revue de statistique appliquée, tome 43, n° 3 (1995), p. 35-53
<http://www.numdam.org/item?id=RSA_1995__43_3_35_0>

Rev. Statistique Appliquée, 1995, **XLIII** (3), 35-53

UNE APPROCHE DE LA COMPARAISON DE PROFILS À L'AIDE DU COEFFICIENT DE CONCENTRATION GÉNÉRALISÉ

J.-M. Tricot (1), R. Latour (2)

(1) Département d'économétrie Université de Genève
102 boulv Carl Vogt, 1211 Genève 4 (Suisse)
(2) HEC Université de Montréal 5255 ave Decelles
Montréal Québec H3T 1V6 (Canada)

RÉSUMÉ

Nous développons une méthode de comparaison de deux profils à l'aide d'un coefficient de concentration des points d'un nuage de l'un des profils par rapport à l'autre. Un tel coefficient possède asymptotiquement, tout comme le coefficient de Gini, un équivalent théorique lui-même équivalent à la distance L_1. L'étude d'un tableau de contingence $q \times r$ peut alors être améliorée en tenant compte des comparaisons possibles entre ses lignes à l'aide du coefficient de concentration. On déduit ainsi d'un tel tableau, un tableau de contingence étendu $3q \times r$ par une décomposition adéquate en trois parties de chaque profil-ligne d'origine. On obtient, par une analyse des correspondances, une représentation graphique plus significative.

Mots-clés : *profil, distance, densité, coefficient de concentration, coefficient de Gini, analyse des correspondances.*

ABSTRACT

A profile comparison method is developed using a concentration coefficient of a point cloud related to a first profile, given a second profile. Such a coefficient asymptotically gets, like Gini's coefficient, his own theoretical corresponding formula which is equivalent to L_1 distance. Then, a $q \times r$ contingency table study can be improved by considering possible comparisons between rows by means of the coefficient. Thus, we are able to derive a $3q \times r$ contingency table from the previous table, having made a three caracteristic parts decomposition of each row. In a correspondence analysis, we obtain a more detailed graphical representation.

Keywords : *Profile, distance, density, concentration coefficient, Gini's coefficient, correspondence analysis.*

1. Introduction

On veut montrer comment il est possible de transformer un tableau de contingence binaire par une décomposition adéquate des profils-lignes. Le nouveau tableau de données ainsi obtenu est alors soumis à l'analyse des correspondances et les résultats graphiques sont plus détaillés et donc plus significatifs.

Pour ce faire, il est nécessaire d'introduire les outils permettant d'effectuer, d'une façon générale, la comparaison de deux profils ce qui rend possible, dans le nouveau tableau de données, de comparer chaque profil-ligne au centre de gravité des autres profils-lignes.

A la section 2 on définit le coefficient de concentration utilisé pour mesurer un degré d'agrégation des points d'un nuage unidimensionnel; et ceci permet d'établir certains parallèles avec différents coefficients rencontrés dans la littérature. Dans la section 3 on indique comment généraliser la notion de concentration introduite précédemment, d'abord dans le cas continu, établissant pour le coefficient, un équivalent théorique, à savoir la distance L_1; et ensuite, dans le cas discret, définissant un coefficient de concentration généralisé pour comparer des profils 2 à 2. Un test d'adéquation peut alors être introduit. Une étude empirique d'un tableau de contingence est développée à la section 4. Dans cette section, les éléments de comparaison entre deux profils étant établis, chaque profil-ligne du tableau de contingence peut être représenté par trois nouveaux profils construits à partir de la donnée globale des autres profils-lignes. Le tableau de contingence $q \times r$ initial induit ainsi un tableau de contingence étendu $3q \times r$ soumis à l'analyse des correspondances. La représentation des profils-lignes initiaux par des triangles de $\mathbb{R}^r$ et non plus des points, facilite l'interprétation graphique.

2. Concentration d'un nuage

Le type de coefficient de concentration auquel on s'intéresse, est déjà présent dans la littérature (on peut se référer, par exemple, à [8]). On en rappelle la construction ci-dessous. En outre on indiquera le lien existant entre ce coefficient et d'autres coefficients, en particulier, le coefficient de concentration de Gini.

Soit X une variable aléatoire réelle de loi p^X. On suppose que le support de X est l'intervalle semi-ouvert à droite $[a, b[(a, b \in \mathbb{R},\ a < b)$ et que l'on dispose d'un n-échantillon $\underline{x} = (x_1, \ldots, x_n)(x_i \in [a, b[)$ assimilable à un nuage unidimensionnel de n points.

On mesure la concentration des points du nuage en considérant un recouvrement G (appelé aussi grille) de $[a, b[$, telle que G soit l'ensemble des intervalles semi-ouverts à droite $I_k = [a + (k-1)(b-a)/n,\ a + k(b-a)/n[$ (appelé aussi maille), pour $k = 1, n$. La terminologie est empruntée à [2]. On a : $[a, b[= \bigcup_{k=1}^{n} I_k$.

La grille G va permettre de mesurer une «intensité des espacements» entre points du nuage $\underline{x}$. La concentration du nuage $\underline{x}$ s'obtient en calculant la proportion de mailles de $[a, b[$ ne contenant aucun point du nuage, c'est-à-dire «inoccupées». Ainsi, on définit un coefficient de concentration du nuage $\underline{x}$, noté c_{Tn} (expression tronquée du coefficient défini en Annexe) comme suit :

$$c_{Tn}(\underline{x}) = (1/n) \sum_{k=1}^{n} 1_{\{0\}}\Big(\sum_{i=1}^{n} 1_{I_k}(x_i)\Big),$$

où $1_{\mathcal{L}}(\cdot)$ est la fonction indicatrice réelle qui vaut 1 sur $\mathcal{L}$ et 0 ailleurs.

Le coefficient de concentration $c_{Tn}(\underline{x})$ s'interprète comme un équivalent asymptotique lorsque le nombre de points augmente indéfiniment du coefficient défini en Annexe. D'après les notations en Annexe, G est mis pour G_2 et I_k pour I_{k2}.

On a les inégalités $0 \leqslant c_{Tn} \leqslant (n-1)/n$. A propos du recouvrement G à l'aide duquel est construit $c_{Tn}(\underline{x})$, on vérifie qu'il permet de définir une dissimilarité entre les points du nuage $\underline{x}$. En effet on peut considérer la dissimilarité qui, à tout couple (x_i, x_j) $(1 \leqslant i \leqslant n,\ 1 \leqslant j \leqslant n)$ associe le nombre de mailles inoccupées et incluses soit dans $[x_i, x_j]$ si $x_i \leqslant x_j$, soit dans $[x_j, x_i]$ sinon. Cette dissimilarité est d'ailleurs un écart (cf. [4]) puisque l'inégalité triangulaire est vérifiée.

Le coefficient de concentration se calcule, en d'autres termes, comme la proportion de mailles du recouvrement G en «déficit» par rapport à la présence d'un point du nuage. On va montrer qu'il se calcule aussi comme la somme des «excédents» de points du nuage par rapport à 1, dans les différentes mailles non inoccupées de G. En effet, posons :

$$p_k(\underline{x}) = (1/n) \sum_{i=1}^{n} 1_{I_k}(x_i),$$

qui est la proportion de points x_i tombant dans I_k. Cette expression est l'estimation usuelle de $p^X(I_k)$. En notant $p_k = 0$, l'ensemble $\{\underline{x}' \in [a, b[^n | p_k(\underline{x}') = 0\}$ et $p_k > 0$ l'ensemble $\{\underline{x}' \in [a, b[^n | p_k(\underline{x}') > 0\}$, on a :

$$\begin{aligned} c_{Tn}(\underline{x}) &= (1/n) \sum_{k=1}^{n} 1_{p_k=0}(\underline{x}) \\ &= (1/n) \sum_{k=1}^{n} (1 - 1_{p_k>0}(\underline{x})) \\ &= (1/n) \sum_{k=1}^{n} (np_k(\underline{x}) - 1) 1_{p_k>0}(\underline{x}) \end{aligned}$$

ce qui achève la preuve.

Finalement, considérons l'hypothèse H_0 suivante :

$$H_0 : X \quad \text{suit une loi uniforme sur } [a, b[,$$

et notons $p_0 = 1/n = p^X(I_k | H_0)$. On vérifie que $\forall \underline{x}' \in [a, b[^n, \sum_{k=1}^{n} (p_k(\underline{x}') - p_0) = 0$ et que $p_k > 0$ est l'ensemble $\{\underline{x}' \in [a, b[^n | p_k(\underline{x}') - p_0 \geqslant 0\}$; et que, par conséquent :

$$c_{Tn}(\underline{x}) = \sum_{k=1}^{n} |p_k(\underline{x}) - p_0|_+ = \sum_{k=1}^{n} |p_0 - p_k(\underline{x})|_+ = (1/2) \sum_{k=1}^{n} |p_k(\underline{x}) - p_0|.$$

Le coefficient de concentration c_{T_n} peut être comparé à certains indices rencontrés dans la littérature économique et recensés par exemple dans [6]. Pour

cela adoptons les quelques notations suivantes : posons $n_k = np_k(\underline{x})$ (fréquence absolue des points dans I_k) et considérons les statistiques d'ordre $x_{[1]} > \ldots > x_{[n]}$ associées au nuage $\underline{x}$. Enfin, notons F_n la fonction de répartition empirique de la loi de X et F_0 la fonction de répartition de la loi uniforme sur $[a, b[$. Nous pouvons énumérer les statistiques suivantes :

1) Le coefficient de concentration :

$$c_{Tn}(\underline{x}) = \sum_{k=1}^{n} \left(\frac{n_k}{n} - \frac{1}{n} \right)_+ .$$

2) Le coefficient de Shutz :

$$S(\underline{x}) = n \sum_{k=1}^{n} \left(\frac{x_k}{\Sigma x_k} - \frac{1}{n} \right)_+ .$$

3) Le coefficient de Gini :

$$G(\underline{x}) = 1 - \left(2 \sum_{k=1}^{n} \frac{k}{n} \frac{x_{[k]}}{\Sigma x_k} - \frac{1}{n} \right).$$

4) La statistique de Kolmogorov pour la loi uniforme :

$$K = \sup |F_n - F_0|.$$

En termes économiques, considérant le nuage $\underline{x}$ comme un ensemble de revenus ($[a, b[\subset \mathbb{R}^+$), le coefficient de concentration est obtenu en fonction d'un système de pondérations sur des classes de revenu (n_k/n) tandis que le coefficient de Shutz est obtenu en fonction d'un système de pondérations sur les revenus eux-mêmes $(x_k / \sum x_k)$. Le coefficient de Gini qui est répertorié comme un coefficient de concentration, se rapporte à un barycentre de l'ensemble des pondérations possibles d'une classe de revenu, $\{(k/n,\ x_{[k]}/\sum x_k) | k = 1,\ n\}$, les coefficients barycentriques étant des pondérations sur des revenus ordonnés. En ce qui concerne la statistique de Kolmogorov, on trouve dans [9] l'inégalité :

$$K - \frac{1}{n} < c_{Tn}(\underline{x}),$$

de sorte que la région critique d'un test d'uniformité de la répartition des points du nuage basé sur la statistique de Kolmogorov, est incluse (à $1/n$ près) dans celle d'un test basé sur le coefficient de concentration. Nous aurons plus loin des précisions supplémentaires sur un tel test.

3. Généralisation sur la notion de concentration

Le coefficient de concentration $c_{Tn}(\underline{x})$ est, comme on l'a vu, une mesure de la déviation entre deux distributions empiriques dont l'une est uniforme. On peut, bien sûr, le considérer comme une estimation du paramètre de population $(1/2)\sum_{k=1}^{n} |p^X(I_k) - 1/n|$ dépendant d'une taille d'échantillon prédéterminée. En fait, on se limitera dans le paragraphe suivant, à un autre type d'approche théorique concernant les distances.

3.1. Le cas continu

On suppose X absolument continue de densité f définie sur $[a, b[$ et on note g_0 la densité de la variable Y_0 uniformément distribuée sur $[a, b[$. Dans ce cas, on peut dire que $c_{Tn}(\underline{x})$ possède, tout comme le coefficient de Gini, un équivalent théorique $c(X, Y_0)$ défini par :

$$c(X, Y_0) = \int_a^b (f - g_0)_+.$$

Une généralisation est alors immédiate : la concentration $c(X, Y)$ de X par rapport à Y où X et Y sont supposées absolument continues de densités respectives f et g sur un sous-ensemble $\mathcal{E}$ de $\mathbb{R}$, est définie par :

$$c(X, Y) = \int_{\mathcal{E}} (f - g)_+.$$

De cette manière on distingue les notions de densité et de concentration par le fait qu'une densité est définie en elle-même et engendre une mesure sur une partie de $\mathbb{R}$ (ici, $[a, b[$). Tandis qu'un coefficient de concentration apparaît comme un indice de déviation entre deux densités. En fait, $c(X, Y)$ est exactement égal à la demi-norme L_1 entre f et g. Le coefficient de concentration théorique est donc une distance. Il donne une indication sur les excédents de masse attribués à des parties du support de f par rapport à la répartition de la masse induite par g sur ce support. Le schéma suivant (figure 1) nous donne une idée du comportement réciproque de f et g. Dans ce schéma on a adopté la notation simplifiée $\Delta = c(X, Y) = c(Y, X)$.

On voit bien sûr, qu'au niveau théorique, une analyse de concentration peut être effectuée dans le cas d'un support non borné des variables X et Y.

3.2. Retour au cas discret

Un problème pratique va faire intervenir non pas des densités mais des profils. Et plutôt qu'un support réel, on aura des classes d'observation. Plus précisément, on suppose avoir deux séries $\underline{x} = (x_1, \ldots, x_n)$ et $\underline{y} = (y_1, \ldots, y_t)$ extraites respectivement des variables X et Y, et que, pour ces deux variables, soit déterminé un même ensemble de r classes d'observation disjointes, $J = \{J_k,\ k = 1,\ r\}$.

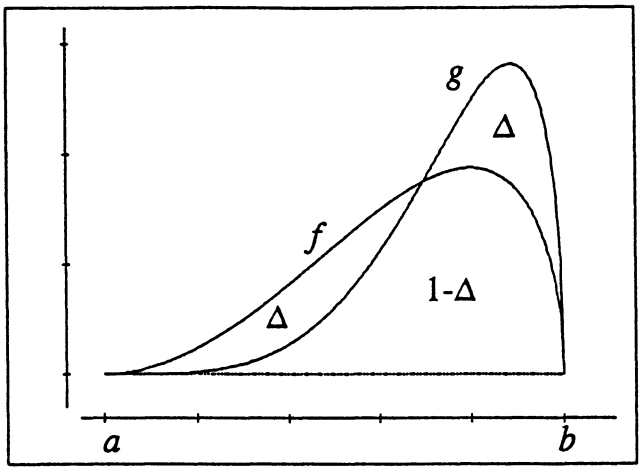


FIGURE 1

Positions respectives de deux courbes de densité pour un calcul de concentration

Alors, par analogie avec le cas continu, on définit la concentration $c(\underline{x},\ \underline{y}, J)$ du profil $\left((1/n)\sum_{i=1}^{n} 1_{J_k}(x_i)\right)_{k=1,\,r}$ par rapport au profil $\left((1/t)\sum_{j=1}^{t} 1_{J_k}(y_i)\right)_{k=1,\,r}$ par :

$$c(\underline{x},\ \underline{y},\ J) = \sum_{k=1}^{r} \left((1/n)\sum_{i=1}^{n} 1_{J_k}(x_i) - (1/t)\sum_{j=1}^{t} 1_{J_k}(y_j)\right)_+ ,$$

expression qui peut s'écrire : $\sum_{k=1}^{r}(n_k/n - t_k/t)_+ = \sum_{k=1}^{r}(f_k - g_k)_+$, où n_k (resp. t_k) est le nombre de points x_i (resp. y_j) tombant dans J_k, et f_k (resp. g_k), la proportion associée.

L'expression de $c(\underline{x},\ \underline{y},\ J)$ généralise celle de $c_{Tn}(\underline{x})$ qui correspond au cas où $r = t = n$ et où $\forall k \in \{1, \ldots, n\}$, $y_k \in J_k$ auquel cas $t_k/t = 1/n$ car $t_k = 1$.

A partir de là, une comparaison détaillée de deux profils et plus précisément, du premier profil (profil 1) au deuxième profil (profil 2), est obtenue à l'aide de trois vecteurs $1 \times r$:

1) un vecteur d'«excédents» du profil 1 par rapport au profil 2 défini en prenant pour $k^{\text{ème}}$ composante le $k^{\text{ème}}$ terme de $c(\underline{x},\ \underline{y}, J)$;
2) un vecteur de «déficits» du profil 1 par rapport au profil 2 défini en prenant pour $k^{\text{ème}}$ composante le $k^{\text{ème}}$ terme de $c(\underline{y}, \underline{x}, J)$;

3) un vecteur de «parties communes» des deux profils défini en prenant pour $k^{\text{ème}}$ composante $\text{Min}((1/n)\sum_{i=1}^{n} 1_{J_K}(x_i),\ (1/t)\sum_{j=1}^{t} 1_{J_k}(y_j)) = \text{Min}\,(n_k/n,\ t_k/t)$
(la somme des composantes de ce dernier vecteur de parties communes correspond dans le cas continu à l'aire notée $1 - \Delta$; cf. la figure 1).

On remarque que dans le calcul des vecteurs d'excédents et de déficits du profil 1, les profils 1 et 2 afférents à $\underline{x}$ et $\underline{y}$ respectivement jouent des rôles dissymétriques.

Une telle comparaison entre les profils 1 et 2 est identifiable à une «décomposition» du profil 1 en trois parties, à l'aide du profil 2.

L'aspect test dégagé à la fin du §2 peut être repris maintenant. En effet, rappelons que l'on a considéré $n_k = np_k(\underline{x})$, effectif échantillonnal de la $k^{\text{ème}}$ classe du découpage de $[a, b[$ à l'aide de G. On considère maintenant $(N_1, \ldots, N_r)$ le vecteur multinomial défini, pour le système de classes J, par $N_k = np_k(\underline{X})$, $k = 1, r$ ($\underline{X}$ étant le vecteur aléatoire de réalisation $\underline{x}$). Puisque l'inégalité rapportée dans [3],

$$\text{Prob}\{(1/2)\sum_{k=1}^{r}\left|\frac{N_k}{n} - \varphi_k\right| \geqslant C_0\} \leqslant 2^{r+1}\exp\,(-2nC_0^2),$$

est vraie lorsque φ_k est la probabilité d'appartenance à J_k, soit $p^X(J_k)$, on peut donc définir une valeur critique C_0 du test d'hypothèse : $\forall k \in \{1, \ldots, r\}$, $p^X(J_k) = \varphi_k$, en prenant $2^{r+1}\exp(-2nC_0^2)$ comme erreur α de $1^{\text{ère}}$ espèce du test.

On dira, si φ_k est en fait calculée à partir de $\underline{y}$ par : $\varphi_k = g_k = t_k/t$, que la concentration du profil 1 par rapport au profil 2 est significative en cas de rejet de l'hypothèse : $\forall k \in \{1, \ldots, r\}$, $p^X(J_k) = g_k$, c'est-à-dire si :

$$c(\underline{x},\ \underline{y}\ J) \geqslant \sqrt{(1/n)\text{Log}\sqrt{2^{r+1}/\alpha}}.$$

Dans l'exemple simulé en Annexe, on a $n = r = 20$ et $g_k = 1/n = 1/20$. Avec $\alpha = 0.05$ on obtient $C_0 = 0.662$. On déduit que la concentration des points de $\underline{x}$ n'est pas significative puisque l'on a $c_{T20}(\underline{x}) = 0.450$. D'autre part, ce qui est cohérent avec le calcul précédent, la statistique de Kolmogorov admet l'hypothèse d'uniformité au seuil 0.05 (la valeur critique est égale à 0.294 et $K = 4.333/20 = 0.216$).

4. Un tableau de contingence étendu

Le but de ce §4 est de montrer l'intérêt que présente l'étude de concentration concernant les différents profils-lignes d'un tableau de contingence binaire pour ensuite déduire de ce tableau un tableau de contingence étendu plus riche en information que le précédent. Un tel tableau est soumis à l'analyse des correspondances.

Pour ce faire on va directement et succintement se référer aux données extraites d'un recensement canadien pour une étude de distribution de revenus.

4.1. Source et description des données

On dispose de données collectées lors du recensement de 1986 au Canada et sélectionnées par l'Institut Québécois de Recherche sur la Culture (IQRC) sous forme d'un échantillon systématique. Cet échantillon a été extrait de l'ensemble des revenus d'emploi déclarés non nuls des particuliers des 12 provinces et territoires canadiens.

Pour des raisons d'insuffisance d'effectif, on a écarté de l'étude une province (l'Ile du Prince-Edouard) et deux territoires. La taille de l'échantillon représentant 2% de la population des individus (particuliers) à revenus d'emploi positifs, est $n = 216090$.

On a considéré $r = 10$ classes de revenu d'effectifs égaux en vue de prolonger ultérieurement l'étude en comparant les années 1971 et 1986 : en conservant des classes d'effectifs égaux en coupe longitudinale, on élimine en grande partie les effets de l'inflation et de la restructuration du marché.

Dans le tableau 1 ont été rassemblées différentes statistiques associées directement à l'échantillon de données.

TABLEAU 1

Statistiques de base sur les classes de revenu pour 9 provinces du Canada

Rev.	b≤	≤B	e	$	$/e	E.-T.	% $
.1	1	3000	21609	32504832	1504.23	836.48	0.73
.2	3000	6278	21609	99566713	4607.65	980.25	2.24
.3	6278	10000	21609	179538770	8308.52	1139.38	4.04
.4	10000	14000	21609	259856079	12025.36	1166.78	5.84
.5	14000	18000	21609	344226552	15929.78	1123.23	7.74
.6	18000	21676	21609	424862416	19661.36	1046.56	9.55
.7	21676	26000	21609	517579309	23952.03	1312.06	11.64
.8	26000	31596	21609	623980607	28875.96	1476.69	14.03
.9	31596	40000	21609	765967489	35446.69	2516.37	17.22
1.	40000	+	21609	1200249019	55543.94	20047.64	26.98
2% de la population de 9 provinces du CANADA R>0			TOT 216090	TOT 4448331786	MOY 20586.	E-T 16756.	TOT 100.00

La variable revenu comprend 10 modalités-revenus $\{.1, .2, \ldots, .9, 1.\}$; et aux bornes des classes de revenu séparant deux classes, les revenus ont été redistribués de manière à égaliser les effectifs de classe. Les sigles $b \leqslant$, $\leqslant B$, e, \$, \$/e, E.-T. et % \$ désignent, respectivement pour chaque classe de revenu, une borne inférieure, une borne supérieure, un effectif, un cumul des revenus, un revenu moyen, un écart-type des revenus et une part de la richesse. En marge du bas du tableau, les caractéristiques précédentes sont obtenues pour la réunion des classes.

4.2. Tableau de contingence binaire et analyse (Étude A)

Disposant des différentes valeurs d'échantillon on a croisé les variables qualitatives province (à $q = 9$ modalités correspondant aux 9 provinces sélectionnées)

et revenu (à $r = 10$ modalités correspondant aux 10 classes de revenus). Pour les modalités-provinces, on a utilisé les symboles (SYMB.) indiqués en 2ème colonne du tableau 2.

TABLEAU 2

Tableau de contingence croisant les variables province et revenu

		Revenu										
PROV. CAN. 1986	SYMB.	.1	.2	.3	.4	.5	.6	.7	.8	.9	1.	TOTAL
Terre-Neuve	T	390	438	383	396	419	260	252	346	290	205	3379
Nouvelle-Ecosse	E	802	787	699	703	849	609	635	659	562	463	6768
Nouveau-Brunswick	B	676	583	541	569	599	499	438	450	437	314	5106
Quebec	Q	4441	4580	4831	6004	5804	5858	5557	5283	4743	4600	51701
Ontario	O	8787	8621	8265	8103	8280	8548	8720	8706	8883	9618	86531
Manitoba	M	985	1014	1115	1014	943	847	924	855	821	685	9203
Saskatchewan	S	993	955	1067	924	745	758	764	765	721	663	8355
Alberta	A	2027	2249	2278	1955	2024	1954	2067	2077	2264	2600	21495
Colombie-Britanique	C	2508	2382	2430	1941	1946	2276	2252	2468	2888	2461	23552
2%de la pop. rev.>0	TOTAL	21609	21609	21609	21609	21609	21609	21609	21609	21609	21609	216090

L'utilisation du logiciel SPAD.N a permis d'effectuer une analyse des correspondances, que l'on appelle Étude A, sur le tableau binaire. On donne, à titre indicatif, la représentation des nuages dans le plan (1, 2) suivi du spectre des valeurs propres.

On désire ne faire qu'un bilan rapide, à ce stade de l'étude, des configurations respectives des deux nuages car cette analyse ne doit servir que de simple élément de comparaison au regard de celle qui suivra et qui, dans une certaine mesure, apportera graphiquement des conclusions plus parlantes voir plus complètes.

Précisons que le plan (1, 2) représente environ 86.5% de l'inertie totale et que la suite ordonnée des modalités de la variable revenu fait apparaître la manière dont elle explique chacun des axes F1 et F2.

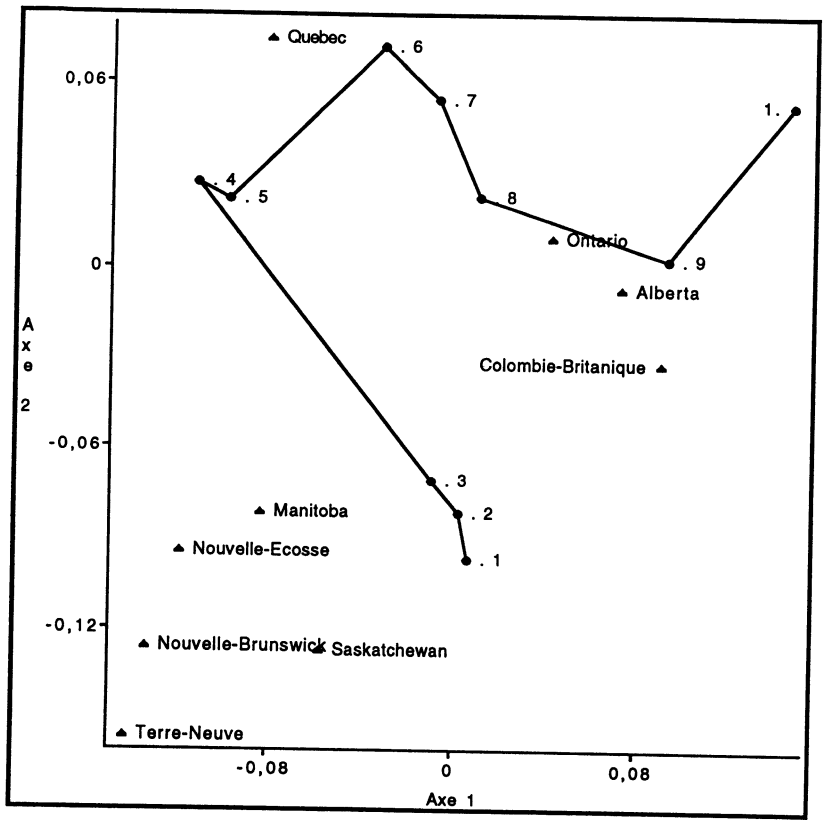
En fait, l'examen des aides à l'interprétation, conduit à différents constats. Tout d'abord le premier axe traduit une opposition entre revenus moyens {.4, .5} côté (F1< 0), et revenus élevés {.9, 1.} côté (F1> 0). Tandis que le deuxième axe traduit une opposition entre revenus faibles {.1, .2, .3} côté (F2< 0) et revenus moyens-élevés {.6, .7} côté (F2 > 0). D'ores et déjà nous pouvons dire qu'il existe trois groupes de provinces à comportements distincts, soient les groupes {C, A, O}, {Q} et {B, E, T, S, M} qui se rapprochent respectivement des classes de revenus élevés, moyens, et faibles; mais les différences entre ces groupes sont assez globales et peu subtiles et l'on ne peut guère se prononcer sur les comportements intragroupes (le Québec excepté). Voilà pourquoi on va opérer une «décomposition» des profils-lignes afin de pouvoir affiner l'analyse portant sur les provinces sélectionnées.

4.3. Décomposition des profils-lignes

Chaque profil-ligne du tableau 1 de contingence (cf. §4.2), caractérisant une province particulière, est comparé au profil caractérisant l'ensemble des autres provinces et obtenu en marge du tableau de contingence après suppression de la ligne afférente à la province considérée au départ : soit $n_{ik} (1 \leqslant i \leqslant q,\ 1 \leqslant k \leqslant r)$, l'effectif de cooccurrence de la $i^{\text{ème}}$ modalité-province et de la $k^{\text{ème}}$ modalité-revenu. Suivant les notations usuelles de l'analyse des données, on pose $n_{i.} = \sum_k n_{ik}$ et

FIGURE 2

Représentation des points provinces et classes de revenu dans le plan principal



HISTOGRAMME DES 9 PREMIERES VALEURS PROPRES

NUMERO	VALEUR PROPRE	POURCENT.	POURCENT. CUMULE	
1	.0053	52.80	52.80	**********
2	.0034	33.72	86.52	**********
3	.0006	5.60	92.12	*********
4	.0005	4.81	96.93	********
5	.0002	2.11	99.04	****
6	.0001	.57	99.61	*
7	.0000	.36	99.96	*
8	.0000	.04	100.00	*
9	.0000	.00	100.00	*

$n_{\cdot k} = \sum_i n_{ik}$. On veut comparer le profil de la $i^{\text{ème}}$ province :

$$\text{PROV}_i = (n_{i1}/n_{i\cdot}, \ldots, n_{ir}/n_{i\cdot}),$$

$(r = 10)$, avec le profil du Canada moins la $i^{\text{ème}}$ province :

$$\text{CAN}_{-i} = ((n_{\cdot i} - n_{i1})/(n - n_{i\cdot}), \ldots, (n_{\cdot r} - n_{ir})/(n - n_{i\cdot})),$$

(d'après l'exemple, $n_{\cdot 1} = \ldots = n_{\cdot r} = 21609$, sont des fréquences calculées pour $q = 9$ provinces).

En se référant au §3.2, on introduit pour plus de clarté, les notations suivantes : on note $f_{ik} = n_{ik}/n_{i\cdot}$ (resp. $g_{ik} = (n_{\cdot k} - n_{ik})/(n - n_{i\cdot})$) la $k^{\text{ème}}$ composante de PROV_i (resp. CAN_{-i}). Le vecteur d'excédents de PROV_i par rapport à CAN_{-i} est noté $\text{EX}_i = ((f_{i1} - g_{i1})_+, \ldots, (f_{ir} - g_{ir})_+)$. Le vecteur de déficits de PROV_i par rapport à CAN_{-i} est noté $\text{DF}_i = ((g_{i1} - f_{i1})_+, \ldots, (g_{ir} - f_{ir})_+)$. Et le vecteur de parties communes des deux profils est noté $\text{PC}_i = (\text{Min}(f_{i1}, g_{i1}), \ldots, \text{Min}(f_{ir}, g_{ir}))$. Enfin la concentration de PROV_i par rapport à CAN_{-i} est notée c_i. On a :

$$(f_{ik} - g_{ik})_+ = n_{ik}/n_{i\cdot} - (n_{\cdot k} - n_{ik})/(n - n_{i\cdot}) \quad \text{si } f_{ik} - g_{ik} > 0 \text{ et } 0 \text{ sinon,}$$

$$(g_{ik} - f_{ik})_+ = (n_{\cdot k} - n_{ik})/(n - n_{i\cdot}) - n_{ik}/n_{i\cdot} \quad \text{si } f_{ik} - g_{ik} < 0 \text{ et } 0 \text{ sinon,}$$

$$c_i = \sum_{k=1}^{r} (f_{ik} - g_{ik})_+ = \sum_{k=1}^{r} (g_{ik} - f_{ik})_+ = (1/2) \sum_{k=1}^{r} |f_{ik} - g_{ik}|.$$

On déduit de EX_i un vecteur $(i > -i)$ d'«excédents absolus» à l'échelle des individus de la $i^{\text{ème}}$ province :

$$(i > -i) = \text{EX}_i n_{i\cdot} = (\text{PROV}_i - \text{CAN}_{-i})_+ n_{i\cdot} .$$

La somme des composantes de $(i > -i)$ compte, avec le facteur d'échelle $n_{i\cdot}$, et à un arrondi près, le nombre d'individus de l'échantillon afférents à la $i^{\text{ème}}$ province qu'il faudrait retrancher de cet échantillon dans les classes de revenu adéquates (et redistribuer dans les autres classes), pour obtenir l'égalité des profils PROV_i et CAN_{-i}.

De même on déduit de DF_i un vecteur $(i < -i)$ de «déficits absolus» à l'échelle des individus du Canada moins la $i^{\text{ème}}$ province :

$$(i < -i) = \text{DF}_i (n - n_{i\cdot}) = (\text{CAN}_{-i} - \text{PROV}_i)_+ (n - n_{i\cdot}).$$

La somme des composantes de $(i < -i)$ compte, avec le facteur d'échelle $n - n_{i\cdot}$, et à un arrondi près, le nombre d'individus de l'échantillon afférents au Canada moins la $i^{\text{ème}}$ province qu'il faudrait retrancher de cet échantillon dans les classes de revenu adéquates (et redistribuer dans les autres classes), pour obtenir l'égalité des profils PROV_i et CAN_{-i}.

Enfin, on déduit de PC_i un vecteur $(i = -i)$ de «parties communes absolues» à l'échelle du Canada (toutes provinces confondues) :

$$(i = -i) = \text{PC}_i n.$$

En posant MARGE$= (n_{\cdot 1}, \ldots, n_{\cdot r})$, on vérifie aisément que l'on a la relation (R) :

$$\text{MARGE} = (i > -i) + (i = -i) + (i < -i) \qquad \text{(R)}$$

La modalité-province d'indice i peut se décomposer en trois sous-modalités $\{(i > -i),\ (i = -i),\ (i < -i)\}$, chacune du nom des vecteurs associés (par abus de notation). Sachant que $\text{PROV}_i = (i = -i)/n + (i > -i)/n_i$. (parties communes PC_i+ excédents EX_i), on reconstitue le tableau de contingence, par exemple, à l'aide de $(i = -i)$ et $(i > -i)$, $i = 1,\ r$, et pour une telle reconstitution, le triplet $\{(i > -i),\ (i = -i),\ (i < -i)\}$ apparaît comporter une information redondante mise en évidence par la relation (R). Cependant, un intérêt se dégage à conserver tous les composants du triplet pour «expliquer» la modalité-province d'indice i puisque tous ont une interprétation primordiale en terme de comparaison de profils, c'est-à-dire en terme de concentration et c'est ce que l'on illustrera par la suite.

4.4. Caractéristiques du tableau de contingence étendu et analyse (Étude B)

À partir de la décomposition en trois de la $i^{\text{ème}}$ modalité-province, on peut construire un tableau de contingence $3 \times r$ croisant le triplet $\{(i > -i),\ (i = -i),\ (i < -i)\}$ avec les classes de revenu. L'effectif de cooccurrence $(i > -i) \times k$ de la sous-modalité $(i > -i)$ et de la $k^{\text{ème}}$ classe de revenu, est un excédent d'individus de la $i^{\text{ème}}$ province classés dans la $k^{\text{ème}}$ classe de revenu : $n_{i\cdot}(n_{ik}/n_{i\cdot} - (n_{\cdot k} - n_{ik})/(n - n_{i\cdot}))_+$. L'effectif de cooccurrence $(i < -i) \times k$ est un excédent de canadiens hors province, autrement dit : $(n - n_{i\cdot})((n_{\cdot k} - n_{ik})/(n - n_{i\cdot}) - n_{ik}/n_{i\cdot})_+$. Et l'effectif de cooccurrence $(i = -i) \times k$ est un complément aux deux précédents par rapport à l'effectif de la $k^{\text{ème}}$ classe de revenu (différence entre $n_{\cdot k}$ et l'un des deux précédents effectifs, l'autre étant automatiquement nul).

De là, par simple juxtaposition, on déduit, pour toutes les provinces, un tableau de contingence étendu $3q \times r$ interprétable comme un tableau ternaire ou comme une bande de Burt : c'est un tel type de tableau (tableau 3) que l'on souhaite analyser (cf., par exemple, [5] ou [10]).

On peut noter que la marge du bas de ce tableau étendu est égale à q fois la marge correspondante du tableau initial (*i.e.* q MARGE avec les notations du §4.3) et que, dans l'analyse des correspondances du tableau étendu, le centre de gravité des profils des trois lignes $\{(i > -i),\ (i = -i)$, et $(i < -i)\}$, est identique au centre de gravité global, lui-même identique au centre de gravité des profils des lignes du tableau initial dans l'analyse des correspondances de ce dernier tableau.

Dans la marge d'extrême droite, on a rajouté les pondérations des profils-lignes (totaux des lignes divisés par 216090). Elles ont toutes une interprétation en terme de concentration. En effet, si S est le vecteur $1 \times r$ dont toutes les composantes valent 1 et sachant d'autre part que le coefficient de concentration c_i peut s'écrire : $c_i = (\text{PROV}_i - \text{CAN}_{-i})_+ S' = (\text{CAN}_{-i} - \text{PROV}_i)_+ S'$, on a :

$$c_i = \frac{(i > -i)S'}{n_{i\cdot}} = \frac{(i < -i)S'}{n - n_{i\cdot}} = \frac{[(i > -i) + (i < -i)]S'}{n}$$

et comme $(\text{MARGE})S' = n$, on a aussi :

$$1 - c_i = \frac{(i = -i)S'}{n}.$$

TABLEAU 3

Tableau de contingence croisant les triplets $\{(i > -i),\ (i = -i),\ (i < -i)\}$ *avec le revenu*

	.1	.2	.3	.4	.5	.6	.7	.8	.9	1.	TOTAL	%
T>-T	53	102	46	59	82	0	0	8	0	0	350	.0016
T=-T	21556	21507	21563	21550	21527	16627	16116	21601	18546	13110	193703	.8964
T<-T	0	0	0	0	0	4982	5493	0	3063	8499	22037	.1020
E>-E	129	114	23	27	178	0	0	0	0	0	471	.0022
E=-E	21480	21495	21586	21582	21431	19444	20274	21041	17944	14783	201060	.9304
E<-E	0	0	0	0	0	2165	1335	568	3665	6826	14559	.0674
B>-B	169	74	31	60	91	0	0	0	0	0	425	.0020
B=-B	21440	21535	21578	21549	21518	21118	18537	19044	18494	13289	198102	.9168
B<-B	0	0	0	0	0	491	3072	2565	3115	8320	17563	.0813
Q>-Q	0	0	0	1096	833	904	509	148	0	0	3490	.0162
Q=-Q	18562	19143	20192	20513	20776	20705	21100	21461	19824	19226	201502	.9325
Q<-Q	3047	2466	1417	0	0	0	0	0	1785	2383	11098	.0514
0>-0	223	0	0	0	0	0	112	88	383	1609	2415	.0112
0=-0	21386	21529	20640	20235	20677	21347	21497	21521	21226	20000	210058	.9721
0<-0	0	80	969	1374	932	262	0	0	0	0	3617	.0167
M>-M	68	98	203	98	24	0	4	0	0	0	495	.0023
M=-M	21541	21511	21406	21511	21585	19888	21605	20076	19277	16084	204484	.9463
M<-M	0	0	0	0	0	1721	0	1533	2332	5525	11111	.0514
S>-S	164	124	241	92	0	0	0	0	0	0	621	.0029
S=-S	21445	21485	21368	21517	19268	19605	19760	19786	18648	17148	200030	.9257
S<-S	0	0	0	0	2341	2004	1849	1823	2961	4461	15439	.0714
A>-A	0	110	143	0	0	0	0	0	127	500	880	.0041
A=-A	20378	21499	21466	19654	20347	19644	20780	20880	21482	21109	207239	.9590
A<-A	1231	0	0	1955	1262	1965	829	729	0	0	7971	.0369
C>-C	171	30	84	0	0	0	0	127	598	119	1129	.0052
C=-C	21438	21579	21525	17809	17855	20882	20662	21482	21011	21490	205733	.9521
C<-C	0	0	0	3800	3754	727	947	0	0	0	9228	.0427
CANADA	21609	21609	21609	21609	21609	21609	21609	21609	21609	21609	216090	1.000

Donc la somme des pondérations des profils des modalités $(i > -i)$ et $(i < -i)$ donne c_i et la pondération du profil de la modalité $(i = -i)$ vaut le complément à 1 de c_i.

Quant à la marge du bas, il s'agit du vecteur MARGE qui est la marge du bas de chacun des tableaux de contingence $3 \times r$ concernant une province donnée. Elle est constante pour chacun de ces tableaux par construction. De plus chacun de ces tableaux comporte toujours sur ses colonnes un zéro soit en 1ère soit en 3ème position. On constate donc en définitive que, par le biais de la décomposition des modalités-provinces (en trois sous-modalités), le tableau de contingence étendu $3q \times r$ peut être considéré comme résultant d'un codage de type barycentrique (se référer à [1]) des sous-modalités $(i > -i)$, $(i = -i)$, et $(i < -i)$ pour chaque modalité-province jouant le rôle de nouvelle variable.

4.4.1. Quelques considérations géométriques

À ce stade de l'analyse, deux remarques s'imposent.

La première consiste à dire qu'en effectuant la décomposition précédente des modalités et en se ramenant au cas d'un codage de type barycentrique tel qu'indiqué, les profils-colonnes de la bande de Burt ne diffèrent pas sensiblement, les uns par rapport aux autres, des profils-colonnes du tableau de contingence initial (tableau 2)

malgré le changement de dimension opéré : sans s'étendre en explications formelles, ceci provient de la manière dont a été construite cette bande de Burt et l'on va donc retrouver graphiquement une configuration du nuage des modalités-revenus proche de l'ancienne configuration (celle de l'Étude A).

En deuxième remarque il faut mentionner que le centre de gravité des profils-lignes n'a pas changé d'un tableau à l'autre. Il représente le Canada (à travers les provinces sélectionnées). Un profil de modalité $(i = -i)$ sera donc représenté proche du centre de gravité dans la mesure où la $i^{\text{ème}}$ province associée sera très semblable, dans son profil, à l'ensemble canadien.

4.4.2. Représentation par triangles

On fait maintenant l'étude du tableau 3 à l'aide d'une analyse des correspondances classique. Comme pour la brève analyse du §4.2 (Étude A), on se limite aux deux premiers facteurs explicatifs de plus de 84% de l'inertie totale.

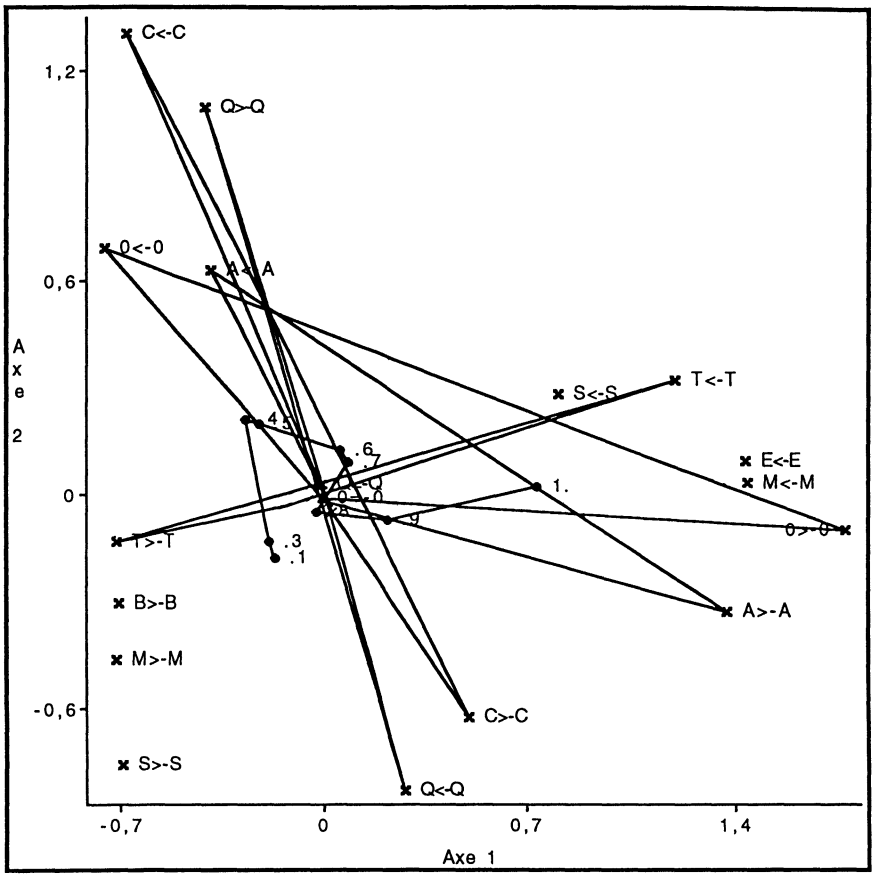
En accord avec la première remarque du §4.4.1, les modalités-revenus ont sensiblement la même configuration dans les deux Études A et B. Une faible différence apparente provient d'une légère rotation des axes d'une Étude à l'autre et d'un changement d'échelle. Autrement dit, on a une similitude entre les Études A et B du point de vue des profils-colonnes. Mentionnons également la superposition de certains points, à savoir .1 et .2 pour les classes de revenu; (E> $-$E) et (T> $-$T) d'une part et (B< $-$B) et (M< $-$M) d'autre part pour les sous-modalités des décompositions des modalités- provinces.

Comme pour l'Étude A, les aides à l'interprétation permettent d'expliquer les axes. L'axe F1 oppose revenus élevés $\{.9, 1.\}$ et revenus moyens $\{.4, .5\}$. Quant à F2, il oppose revenus faibles $\{.1, .2, .3\}$ et revenus moyens $\{.4, .5, .6\}$.

En ce qui concerne la deuxième remarque du §4.4.1, précisons que toutes les modalités $(i = -i)$ pour $i = 1, q$ ont un poids très important et un profil homogène (sans discontinuités contrairement aux profils des modalités $(i > -i)$ et $(i < -i)$ très discontinus du fait de la présence de valeurs nulles). Voilà pourquoi elles sont regroupées au voisinage du centre de gravité : n'ont été indiquées que $(Q = -Q)$ et $(O = -O)$ pour des raisons de clarté graphique; ce qui signifie donc que chaque province se comporte de manière très semblable au Canada globalement. De plus il faut noter que les classes .7 et .8 qui ont une faible inertie et se situent également proches de l'origine, apparaissent expliquer le mieux la distribution du revenu pour l'ensemble des provinces. Le comportement homogène des provinces est confirmé par le fait que les concentrations sont faibles : la plus grande prévaut pour Terre-Neuve avec .104. Elles sont néanmoins toutes supérieures à la valeur critique égale à .005 (cf. la fin du §3.2).

Dans cette Étude, la $i^{\text{ème}}$ province est maintenant représentée par un triangle de sommets les modalités $(i > -i)$, $(i = -i)$ et $(i < -i)$. On voit nettement s'opposer le Québec et la Colombie- Britannique, indiquant que les classes moyennes sont très représentées au Québec comparativement au reste du Canada alors qu'elles sont peu représentées en C.$-$B. : attirance de $(C < -C)$ et de $(Q > -Q)$ dans le quadrant (F1 <0; F2 >0). On peut déceler l'absence comparative au Québec (toujours en 1986)

FIGURE 3
Représentation des triplets $\{(i > -i),\ (i = -i), (i < -i)\}$
et des classes de revenu dans le plan principal



HISTOGRAMME DES 9 PREMIERES VALEURS PROPRES

NUMERO	VALEUR PROPRE	POURCENT.	POURCENT. CUMULE	
1	.0818	67.82	67.82	**********
2	.0198	16.39	84.21	**********
3	.0070	5.82	90.03	*******
4	.0042	3.50	93.53	*****
5	.0028	2.32	95.86	***
6	.0021	1.76	97.62	***
7	.0017	1.44	99.06	**
8	.0009	.71	99.77	*
9	.0003	.23	100.00	*

de faibles revenus puisque $(Q < -Q)$ est associée à {.1, .2, .3} côté (F2 <0). En corollaire, par opposition au Québec, nous obtenons une prépondérance de revenus faibles et de revenus élevés en C.-B..

Passant rapidement en revue les autres provinces, on peut voir un comportement similaire, particulièrement par attirance vers les classes de revenus élevés, des

provinces de l'Ontario (avec la métropole Toronto) et de l'Alberta (pétrolière). Concernant les provinces maritimes {B, E, T} et du centre {S, M}, on a une opposition plus marquée avec {O, A} qu'avec {Q, C} mais davantage sur les classes de revenus élevés {.9, 1.} que sur les classes de revenus faibles.

5. Conclusion

On a mis en évidence un raffinement possible d'une analyse des correspondances en partant d'un tableau de données usuel et en réalisant une transformation de ces données faisant intervenir empiriquement la distance L_1 entre les profils d'origine. Cette distance possède une interprétation en terme de coefficient de concentration. Une telle transformation revient, en fait, à construire un nouveau tableau de données très spécifique découlant du croisement de deux variables qualitatives déduites des variables initiales.

Une étude de comparaison de profils à l'intérieur d'un même tableau peut évidemment être généralisée en considérant une suite chronologique de tableaux et en comparant des profils ou profils marginaux d'un tableau à l'autre.

Concernant les données utilisées à titre expérimental, il est apparu qu'elles présentaient une certaine homogénéité ce qui est normal pour une analyse de distribution de revenu limitée à une entité géographique homogène, ici, un pays particulier. La représentation graphique proposée était destinée à donner une grande amplitude à un phénomène faiblement hétérogène.

Remerciements : Nous voulons exprimer notre gratitude aux spécialistes du Centre International de Statistique et d'Informatique Appliquées (CISIA, 1, av Herbillon, 94160 St-Mandé, France) qui ont apporté leur soutien logistique dans la conception graphique.

Références bibliographiques

[1] CAZES P. (1990), Codage d'une variable continue en vue de l'analyse des correspondances, *Revue de Statistique Appliquée*, Vol 38, n°3, pp. 35-51.

[2] DE LA VALLÉE POUSSIN C. (1950), *Intégrales de Lebesgue, Fonctions d'Ensemble, Classes de Baire*, Gauthier-Villars, Paris.

[3] DEVROYE L. (1987), *A Course in Density Estimation*, Birkhaüser, Boston.

[4] DIDAY E., LEMAIRE, J., POUGET, J. & TESTU, F. (1982), *Éléments d'Analyse de Données*, Dunod, Paris.

[5] JAMBU M. (1989), *Exploration Informatique et Statistique des Données*, Dunod, Paris.

[6] MARSHALL A. & OLKIN, I. (1979), *Inequalities : Theory of Majorization and its Applications*, Academic Press, Londres.

[7] SAPORTA G. (1990), *Probabilités, Analyse des Données et Statistique*, Technip, Paris.

[8] TRICOT J.-M. (1990), *Méthode des Réseaux en Analyse de Données, Application à l'Analyse de Concordance*, Thèse de doctorat, Imprimerie Nationale, Genève.

[9] TRICOT J.-M. (1991), Indice de concentration et statistique de Kolmogorov, application à des distributions de revenu, *Les Cahiers du GERAD*, G-91-35, Montréal.

[10] YOUBI A. (1989), Comparaison entre les états de sujets après l'effort sous divers traitements, *Les Cahiers de l'Analyse des Données*, Vol XIV, n°4, pp. 439-448.

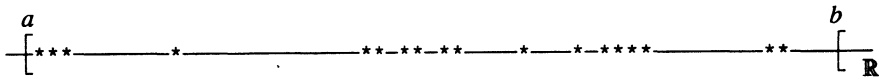
Annexe

On rappelle la définition du coefficient de concentration telle qu'on la trouve, par exemple, dans [8]. Soit une variable aléatoire réelle X de support $[a, b[(a,\ b \in \mathbb{R},\ a < b)$ et un nuage de n points $\underline{x} = (x_1, \ldots, x_n)(x_i \in [a, b[)$ obtenu en tirant un n-échantillon de X.

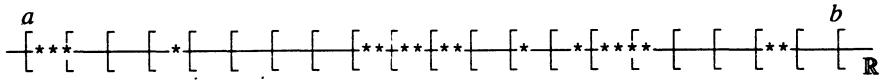
On considère une suite (réseau) de recouvrements de $[a, b[$, $(G_m)_{m=1,2,\cdot\cdot}$ telle que G_m (grille m) soit l'ensemble des intervalles de $[a, b[$ semi-ouverts à droite $I_{km} = \left[a + \frac{k-1}{n^{m-1}}(b-a),\ a + \frac{k}{n^{m-1}}(b-a)\right[$ (maille-m), pour $k = 1$, $n^{m-1}(m \geqslant 1)$. La terminologie est empruntée à [2]. Par construction on a donc : $[a, b[= \bigcup_{k=1}^{n^{m-1}} I_{km}$ et $G_m = \{I_{km} | k = 1,\ n^{m-1}\}$.

Le schéma suivant nous montre les premières grilles sur $[a, b[$ pour un nuage simulé de 20 points où l'on suppose (par exemple) que le point à l'extrême droite est de multiplicité égale à 3 :

Grille 1, constituée de l'unique maille-1 $[a, b[$:



Grille 2, ensemble de $n = 20$ mailles-2 :



Grille 3, ensemble de $n^2 = 400$ mailles-3 :

etc...

On remarque que pour des raisons de lisibilité, il n'a pas été possible de représenter dans l'exemple ci-dessus, toutes les mailles-3 de la grille 3 ni leur diamètre réel égal à $(b-a)/20^2$.

On remarque également que chaque grille est un ensemble à «similitude interne» (cf. la théorie fractale) en ce sens qu'elle est contruite à l'aide d'un procédé récursif en partant d'un motif initial qui est l'intervalle $[a, b[$ ou la grille 1 (ensemble de n^0 intervalle).

D'après la définition, le nombre de mailles-m est égal à n^{m-1}, $m = 1, 2, \ldots$. Étant donnés $W_{km}(\underline{x})$ la proportion de points de $\underline{x}$ appartenant à I_{km} et $N_{km}(\underline{x})$ le nombre de mailles $-(m+1)$ ne contenant aucun point du nuage («inoccupées»), inclues dans I_{km}. On définit alors un nombre moyen de mailles- $(m+1)$ inoccupées dans une maille de la grille m, par :

$$n_m = \sum_{k=1}^{n^{m-1}} W_{km}(\underline{x}) N_{km}(\underline{x})$$

De cette manière, on fait abstraction, dans le calcul de n_m, des mailles-m inoccupées de la grille m puisque pour une maille inoccupée d'indice k de cette grille, $W_{km} = 0$. Cette quantité n_m mesure, par le biais de la grille $m+1$, une intensité des espacements entre points ou un degré d'agrégation des points, en moyenne sur des sous-intervalles $[a, b[$ contenant au moins un point de $\underline{x}$, c'est-à-dire pertinents du point de vue d'un calcul de concentration.

Pour tenir compte du réseau de grilles, on calcule un barycentre des n_m pour $m = 1, 2, \ldots$. Ainsi on définit un coefficient de concentration $c_n(\underline{x})$ du nuage $\underline{x}$ par :

$$c_n(\underline{x}) = \sum_{m=1}^{\infty} \frac{n_m}{n^m}.$$

Le choix des coefficients barycentriques $1/n^m$; $m = 1, 2, \ldots$ (dont la somme est égale à $1/(n-1)$) est guidé par deux considérations :

Premièrement, par construction, $0 \leqslant n_m \leqslant n-1$ puisque $W_{km} > 0 \Rightarrow 0 \leqslant N_{km} \leqslant n-1$. Ainsi, $c_n(\underline{x}) \leqslant 1$. En fait il est aisé de voir que $1/n \leqslant c_n(\underline{x}) \leqslant 1$, la valeur 1 étant atteinte lorsque tous les points de $\underline{x}$ sont superposés (auquel cas, pour tout $m \geqslant 1$, $n_m = n-1$ et la concentration est maximale), et la valeur $1/n$ étant atteinte lorsque chacune des mailles-2 I_{k2}, $k = 1, n$, contient un point de $\underline{x}$ (auquel cas, $n_1 = 0$; pour tout $m \geqslant 2$; $n_m = n-1$ et la concentration est minimale).

Deuxièmement $c_n(\underline{x})$ est asymptotiquement équivalent à son premier terme n_1/n puisque $n_m \leqslant n-1 \Rightarrow \sum_{m=2}^{\infty} n_m/n^m \leqslant 1/n$. Autrement dit, si n est «assez grand», il est suffisant d'étudier la concentration des points du nuage $\underline{x}$ à l'aide du recouvrement G_2, ce qui revient à approximer le coefficient de concentration c_n par son expression tronquée c_{Tn} définie par :

$$c_{Tn}(\underline{x}) = \frac{n_1}{n}$$

Pour ce qui est de l'exemple précédent, le calcul numérique donne aisément $n_1/n = 9/20$ et $n_2/n^2 = (1/n^2)[(3/n)(n-1) + (1/n)(n-1)(n-3)] =$

$(n-1)/n^2 = 19/400$ (les termes n_m/n^m pour $m \geqslant 3$ sont nuls); d'où : $c_n(\underline{x}) = .497$ et $c_{Tn}(\underline{x}) = 0.450$. Comme le note G. Saporta (cf. [7]), il est toujours insuffisant de résumer une série statistique par un seul indicateur. Dans le cas présent on peut remarquer que la méthode d'analyse de concentration abordée ici, a l'avantage de tenir compte d'une suite d'indicateurs $(n_m/n^m)_{m=1,2,\ldots}$ dont les différents termes comportent un intérêt en eux-mêmes.

Enfin, considérons la fonction indicatrice réelle $1_{\mathcal{L}}(\cdot)$ qui vaut 1 sur $\mathcal{L}$ et 0 ailleurs. Compte tenu de ce que les intervalles $I_{v\ m+1}$ contenus dans I_{km} sont tels que l'indice v varie de $(k-1)n+1$ à kn, une expression formalisée pour $c_n(\underline{x})$ est la suivante :

$$c_n(\underline{x}) = \sum_{m=1}^{\infty} \frac{1}{n^m} \sum_{k=1}^{n^{m-1}} \left[\frac{1}{n} \sum_{i=1}^{n} 1_{I_{km}}(x_i) \right] \left[\sum_{v=(k-1)n+1}^{kn} 1_{\{0\}} \left(\sum_{j=1}^{n} 1_{I_{v\ m+1}}(x_j) \right) \right],$$

où les contenus des crochets représentent respectivement $W_{km}(\underline{x})$ et $N_{km}(\underline{x})$ d'après les définitions qu'on en a donné.