Une méthode simple pour analyser rapidement de grands tableaux de nombres *

Michel Volle

27 avril 2006

La méthode décrite ici a déjà été publiée dans *Économie et statistique* [6] en 1974. Elle a été utile. Comme cet article est maintenant difficile à trouver, je la publie de nouveau mais en la complétant par des exemples et par un projet d'extension aux hypercubes.

* *

La statistique et la comptabilité produisent en quantité de grands tableaux de nombres que la masse des données rend opaques à l'interprétation.

Je présente ici une méthode qui aide à analyser et commenter rapidement une grande quantité de gros tableaux. Elle relève logiquement de l'analyse des données [1]. Toutefois les calculs qu'elle nécessite sont plus simples que ceux qu'exigerait une analyse factorielle : on les réalise facilement sur un tableur (voir annexe 2).

Cette méthode ne s'applique qu'aux tableaux de contingence, que l'on appelle aussi « tableaux carrés » (même quand ils sont rectangulaires), c'est-à-dire aux tris croisés qui représentent la ventilation d'une quantité ou d'une « population » (nombre de personnes, nombre d'euros comptabilisés etc.) selon deux caractères qualitatifs [1].

Exemple : le tableau qui répartit le chiffre d'affaires d'une entreprise par produit et par mois sur une période d'une ou plusieurs années.

Les tableaux de nombres ne sont pas tous des tableaux de contingence : cette méthode n'est donc pas universelle. Cependant elle est utile car les tableaux de contingence représentent une part importante des résultats que produisent la statistique et la comptabilité.

Nous montrerons que cette méthode peut s'appliquer au croisement de plus de deux caractères. Le tableau multidimensionnel est appelé un « cube » s'il croise trois caractères et, s'il en croise plus de trois, un « hypercube ». Parmi les collections de tableaux on rencontre souvent des cubes (par exemple lorsqu'on publie une série de tableaux représentant le croisement de deux caractères sur des périodes successives : le découpage du temps est alors le troisième caractère du cube). Les hypercubes sont d'usage fréquent dans les *datawarehouses* [8], où ils sont une façon de présenter les « tables de faits » [3].

1 Principe de la méthode

Considérons le tableau que fournit un tri croisé. On peut lui associer une mesure de la « quantité d'information » qu'il apporte, en donnant au mot « information » le sens que lui attribue Shannon ([5], [4] ; voir annexe 1). On peut aussi associer à chacune de ses cases une mesure de la contribution de la case à cette quantité d'information.

1. Nous utiliserons ici les mots « population » et « individu » en un sens large, et même lorsqu'il s'agit de la ventilation d'une quantité.

On est alors naturellement conduit à classer les cases dans l'ordre des contributions décroissantes, puis à concentrer son attention sur celles qui apportent le plus d'information. Si en effet nous trouvons que dans un tableau de vingt lignes et dix colonnes, comportant donc 200 cases, 5 cases apportent 90 % de l'information, ce sont ces cases que nous devrons examiner et commenter en premier et il sera peut-être inutile de parler des autres.

En annexe 2 on analysera à titre d'exemple un tableau de 260 cases donnant la répartition de la population française en 1999 par région, sexe et classe d'âge. Les deux tiers de l'information qu'apporte ce tableau sont concentrés dans 7 régions sur 26, la moitié de l'information est concentrée dans 10 % des cases. Si l'on veut interpréter et commenter ce tableau, c'est sur ces régions-là et sur ces cases-là qu'il faudra focaliser l'attention, puis attirer celle du lecteur.

2 Formulaire

Cette méthode est fondée sur la théorie de l'information : les démonstrations sont données en annexe 1. Ici nous nous contentons d'indiquer son formulaire afin de montrer comment elle fonctionne [7].

* *

Considérons le tableau de contingence qui ventile une population selon deux caractères qualitatifs I et J repérés respectivement par les indices i et j, et ayant pour cardinaux respectifs N et K.

Notons n_{ij} le nombre des individus qui possèdent à la fois les modalités i de I et j de J, $n_j = \sum_i n_{ij}$ le nombre des individus qui possèdent la modalité j, n_i le nombre de ceux qui possèdent la modalité i, n l'effectif total de la population.

Les données se présentent alors comme sur le tableau 1 : les modalités des caractères I et J sont rangés respectivement en lignes et en colonnes ; les marges en haut et à gauche contiennent respectivement, comme titres des colonnes et des lignes, les intitulés des modalités de J et de I ; les marges en bas et à droite contiennent respectivement les nombres n_j et n_i ; le coin en bas à droite contient le total n, le coin en haut à gauche ne contient rien.

		j		
i		n_{ij}		n_i
		n_j		n

TAB. 1 – *Présentation du tableau à analyser*

On note f_{ij} la part, ou « fréquence », de la case (i,j) dans le tableau :

$$f_{ij} = \frac{n_{ij}}{n}.$$

Dans la population, les fréquences des modalités i et j se notent respectivement :

$$f_i = \frac{n_i}{n} \ , \ f_j = \frac{n_j}{n}.$$

* *

Nous montrons en annexe 1 que l'information apportée par le tableau peut se mesurer par la quantité :

$$Lien(I,J) = \sum_{ij} \frac{(f_{ij} - f_i f_j)^2}{f_i f_j}.$$

Nota Bene : du point de vue géométrique, $Lien(I,J)$ s'interprète comme le carré de la distance entre la distribution[2] f_{IJ} et la distribution « produit des marges » $f_I f_J$, mesurée selon la métrique du χ^2 centrée sur la distribution $f_I f_J$: $Lien(I,J) = \|f_{IJ} - f_I f_J\|^2_{f_I f_J}$, que nous noterons $\|f_{IJ} - f_I f_J\|^2$ tout court. Si la façon dont le tableau est construit résulte du tirage au sort de n individus parmi une population distribuée selon deux caractères I et J indépendants, la quantité $nLien(I,J)$ suit une loi du χ^2 à $(N-1)(K-1)$ degrés de liberté. Cette propriété fonde un test d'indépendance des deux caractères[3].

Si l'on ne disposait pas des données qui résultent du croisement des caractères I et J, la seule estimation possible de la fréquence de la case (i,j) dans la population considérée serait le « produit des marges » $f_i f_j$. $Lien(I,J)$ mesure donc l'information apportée par le croisement de ces deux caractères.

La contribution de la case (i,j) à cette information est :

$$c_{ij} = \frac{(f_{ij} - f_i f_j)^2}{f_i f_j}.$$

$Lien(I,J)$ et c_{ij} sont ici les deux outils essentiels : pour analyser un tableau de contingence, nous calculerons d'abord $Lien(I,J)$, puis la *contribution relative* de chaque case c'est-à-dire la part de sa contribution dans $Lien(I,J)$:

$$cr_{ij} = \frac{c_{ij}}{Lien(I,J)}.$$

* *

En repérant les cases les plus « éloignées du produit des marges » au sens des expressions ci-dessus, on ne fait rien d'autre que de systématiser la pratique professionnelle du statisticien : en effet celui-ci, lorsqu'il examine un tableau de contingence, y recherche les cases dont la fréquence s'écarte du « produit des marges » : ou bien elles sont erronées, ou bien elles signalent un phénomène réel. Il faut donc d'abord les vérifier, puis les interpréter.

Quantifier l'information permet de systématiser cette pratique, de l'appliquer à des tableaux de grande taille, de classer les cases dans l'ordre des cr_{ij} décroissants, enfin de calculer le cumul des cr_{ij}. Si on obtient un cumul élevé avec un petit nombre de cases on pourra se dispenser d'examiner les autres.

Ainsi le calcul apporte une aide à l'interprétation du tableau et à la rédaction des commentaires : il faudra attirer l'attention du lecteur, de façon sélective, sur les cases qui apportent le plus d'information.

2.1 Information apportée par une ligne ou une colonne

Il est utile de calculer les quantités $cr_i = \sum_j cr_{ij}$ et $cr_j = \sum_i cr_{ij}$, car elles signalent les lignes et les colonnes qui contribuent le plus à l'information.

2. On note f_I le vecteur $\{f_i | i \in I\}$, f_{IJ} le tableau $\{f_{ij} | (i,j) \in (I \times J)\}$ etc.

3. En pratique, le test du χ^2 est convenablement approché en utilisant la règle suivante : si

$$\frac{nLien(I,J) - (N-1)(K-1)}{\sqrt{(N-1)(K-1)}} > 3,$$

on peut rejeter l'hypothèse d'indépendance de I et J avec un risque d'erreur inférieur à 5 %. Si cette quantité est supérieure à 6, le risque d'erreur est inférieur à 1 %.

Il peut être intéressant de représenter par un histogramme, nommé « spectre », l'information qu'apporte l'ensemble des cases d'une même colonne ou d'une même ligne [6]. On peut par exemple, dans le tableau « région, classe d'âge » que nous présentons en annexe, associer un spectre à chaque région (figure 1).

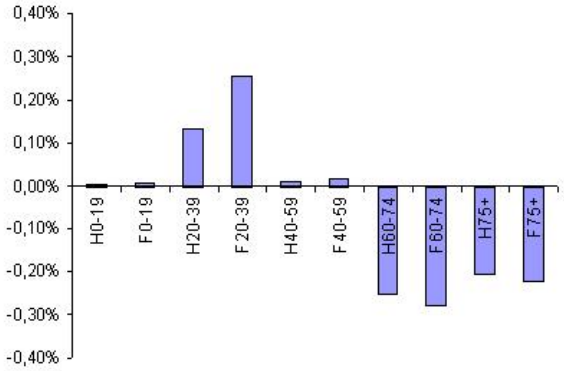

FIG. 1 – *Spectre de l'Île-de-France*

Pour faciliter la lecture des spectres, on convient d'orienter vers le haut le tuyau d'orgue qui représente cr_{ij} si $f_{ij} > f_i f_j$ (c'est-à-dire si la case (i,j) est « plus chargée que le produit des marges ») et de l'orienter vers le bas dans le cas contraire. La surface de l'histogramme est égale à la somme des contributions de la ligne, ou colonne, représentée.

La fréquence conditionnelle du caractère i dans la colonne j se prononce « f de i si j » et se note :

$$f_i^j = \frac{n_{ij}}{n_j} \text{ ; de même, } f_j^i = \frac{n_{ij}}{n_i}.$$

Comme

$$c_{ij} = f_j \frac{(f_i^j - f_i)^2}{f_i},$$

la contribution de la colonne j peut s'interpréter comme le carré de la distance (au sens de la métrique du $\chi 2$) de la colonne j à la marge verticale, pondérée par le poids de la colonne j dans le tableau : $c_j = f_j \| f_I^j - f_I \|^2$.

* *

On ne suivra pas la même démarche selon que l'on s'intéresse au tableau considéré dans son ensemble ou que l'on s'intéresse plus particulièrement à certaines lignes ou colonnes. Dans l'exemple que nous considérons en annexe 2 la région Île-de-France, qui représente 18,2 % de la population, « pèse » lourd ; la Guyane, avec 0,3 % de la population, est beaucoup plus « légère ». Il en résulte que la contribution de cette dernière est plus faible que celle de l'Île-de-France, alors que la distribution de sa population est beaucoup plus éloignée de la distribution marginale.

Si l'on s'intéresse aux proportions il faut faire abstraction de la taille des colonnes (ou lignes) et associer par exemple à chaque colonne un spectre bâti selon les quantités q_i^j :

$$q_i^j = \frac{(f_i^j - f_i)^2}{f_i}.$$

3 Démarche

Voici les étapes selon lesquelles la méthode peut se dérouler en pratique :

1) vérifier s'il y a lieu que les caractères I et J ne sont pas indépendants en utilisant le test du χ^2.

2) construire le tableau f_{IJ}.

3) examiner les distributions marginales f_I et f_J : c'est le point de départ de la méthode, qui examinera par la suite les écarts à ces distributions marginales.

4) construire le tableau des cr_{ij} *signées* (car il importe de savoir si f_{ij} s'écarte de $f_i f_j$ par excès ou par défaut), complété par les sommes cr_i et cr_j qui indiquent les contributions de chaque ligne et chaque colonne ;

5) repérer les lignes et colonnes pour lesquelles les cr_i et cr_j sont les plus élevées : ce sont celles dont le spectre apportera le plus d'information ;

6) classer les cases dans l'ordre des cr_{ij} décroissantes. Le plus souvent (mais pas toujours) elles appartiennent aux lignes et colonnes repérées dans l'étape (5).

7) si l'on s'intéresse aux proportions dans chaque colonne (resp. ligne) pour rechercher les modalités j (resp. i) qui s'écartent le plus de la moyenne, il faut construire le tableau des q_i^j (resp. q_j^i) signées et le compléter par les sommes $q_j = \sum_i q_i^j$ (resp. $q_i = \sum_j q_j^i$).

3.1 De l'observation à l'interprétation

Nous n'avons fait jusqu'à présent que *lire* le tableau, nous n'avons pas *expliqué* les phénomènes que cette lecture révèle. Pour interpréter le tableau, il faudra compléter la lecture en recourant à des hypothèses sur les causes de ces phénomènes. L'interprétation supposera souvent un recours à des informations que le tableau ne comporte pas.

En effet, l'analyse du tableau a fait apparaître la corrélation entre les caractères I et J. Or si le constat d'une corrélation amorce l'inférence vers des hypothèse causales, il ne la remplace pas.

À elle seule la lecture du tableau ne peut pas fournir ces hypothèses ; mais elle conduit à poser des questions qui orienteront la recherche des explications. La méthode que nous proposons consiste à rendre cette lecture plus sélective, à la concentrer sur les cases (ou lignes et colonnes) les plus « originales » au sens de la théorie de l'information, et ainsi à orienter la recherche des explications vers les pistes *a priori* les plus fécondes.

Présenter les données sous une forme sélective qui oriente et facilite leur interprétation, c'est faire un travail semblable à celui qu'effectue un typographe sur un texte manuscrit : le texte imprimé contient moins d'information que le manuscrit dont les corrections portent la trace des hésitations de l'auteur, mais comme sa lecture est plus facile le lecteur aura moins de peine à en dégager le sens.

4 Application aux hypercubes

Nota Bene : Dans beaucoup de cas, on peut se contenter d'analyser des tableaux carrés. Lorsque s'accroît le nombre des caractères que l'on croise, la complexité de la démarche croît plus que proportionnellement. Cette partie de l'article n'intéressera donc que les spécialistes.

4.1 Du tableau carré au cube

Considérons d'abord le « cube » qui représente le croisement des trois caractères I,J et K sur une population.

On peut sur ce cube calculer comme ci-dessus une distance au « produit des marges » :

$$\|f_{IJK} - f_I f_J f_K\|^2 = \sum_{ijk} \frac{(f_{ijk} - f_i f_j f_k)^2}{f_i f_j f_k}.$$

Si l'on cherche à évaluer le gain d'information apporté par l'*intérieur* du cube, on doit toutefois remarquer que ce gain s'additionne dans l'expression ci-dessus à celui qu'apportent les trois *faces* du cube, en nommant ainsi les tableaux carrés croisant deux caractères et dont la case courante contient, par exemple, le nombre $n_{ij} = \sum_k n_{ijk}$.

Le gain d'information apporté par l'intérieur du cube est donc :

$$Lien(I,J,K) = \|f_{IJK} - f_I f_J f_K\|^2 - [Lien(J,K) + Lien(K,I) + Lien(I,J)],$$

et la contribution de la case (i,j,k) à ce gain d'information est :

$$c_{ijk} = \frac{(f_{ijk} - f_i f_j f_k)^2}{f_i f_j f_k} - [f_i c_{jk} + f_j c_{ki} + f_k c_{ij}].$$

La contribution relative est :

$$cr_{ijk} = \frac{c_{ijk}}{Lien(I,J,K)}.$$

Notons $Lien_k(I,J)$ l'information apportée par le tableau carré que constitue le segment (ou « tranche ») découpé dans le cube par la modalité k du caractère K :

$$Lien_k(I,J) = \sum_{ij} \frac{(f_{ij}^k - f_i^k f_j^k)^2}{f_i^k f_j^k}.$$

Avec un peu de patience, il est facile de démontrer que :

$$Lien(I,J,K) = \sum_k f_k Lien_k(I,J) - Lien(I,J).$$

Il en résulte, par permutation des lettres i, j et k, deux relations analogues comportant une sommation sur les tranches i ou j.

Le gain d'information apporté par l'intérieur du cube est donc égal à la moyenne des gains d'information apportés par des tranches parallèles, diminué du gain qu'apporte la face sur laquelle ces tranches s'empilent.

La démarche pour étudier un cube sera la suivante :

1) examiner les distributions marginales f_I, f_J et f_K ;

2) examiner les « faces » du cube f_{JK}, f_{KI} et f_{IJ}, ce qui revient à évaluer $Lien(J,K)$, $Lien(K,I)$ et $Lien(I,J)$, puis la contribution des lignes, colonnes et cases de ces tableaux au gain d'information qu'ils apportent.

3) examiner les contributions relatives cr_{ijk} signées, repérer les cases donc la contribution est la plus forte, évaluer le cumul des contributions etc. comme lors de l'analyse d'un tableau.

Nota Bene : si l'une des trois faces apporte nettement plus d'information que les deux autres, il pourra être utile d'analyser l'intérieur du cube en considérant les tranches qui s'empilent au-dessus de cette face.

Tout comme lors de l'étude d'un tableau carré, les contributions les plus fortes attirent l'attention et orientent l'interprétation du cube. La recherche de la cause du phénomène peut déclencher une opération de *datamining*, impliquant un retour aux données élémentaires dont le cube assure la présentation ainsi qu'un recoupement éventuel avec d'autres données.

4.2 Du cube aux hypercubes

Considérons maintenant l'hypercube résultant du croisement de plus de trois caractères (I,J,...Z) (nous dirons que le nombre de caractères est le « rang » de l'hypercube).

À partir de l'examen du cube, on voit comment on va étudier un hypercube :

1) examiner les marges f_I, f_J,...f_Z ;

2) examiner les tableaux croisés f_{JK},f_{KL},...f_{ZI}, en se concentrant sur ceux qui apportent le plus d'information ;

3) examiner de même les cubes f_{JKL}, f_{KLM} etc.

4) puis, en procédant par ordre de dimension croissante, examiner tous les sous-hypercubes que l'on peut construire en combinant les diverses dimensions de l'hypercube considéré, en se concentrant sur ceux qui apportent le plus d'information ;

5) enfin, calculer comme pour le cube l'information apportée par l'intérieur de l'hypercube et identifier les cases qui apportent le plus d'information.

L'analyse d'un hypercube de rang x implique d'analyser des hypercubes de rang $x-1$ jusqu'à ce que l'on parvienne aux marges. La complexité est exponentielle, puisqu'un hypercube de rang x contient 2^x-1 sous-hypercubes (en comptant les cubes, tableaux carrés et marges). Orienter l'analyse en sélectionnant, à chaque étape, ceux des hypercubes qui apportent le plus d'information permet toutefois de réduire cette complexité. La démarche, étant récursive, se prêterait bien en principe à la programmation en Scheme [2].

Il me reste à mettre au point ce programme, puis à tester la méthode sur un exemple afin de vérifier son efficacité.

Annexe 1 : Théorie de l'information

Considérons une population de n individus répartis selon un caractère qualitatif I, la part de la modalité i étant $p_i = n_i/n$.

La connaissance du caractère I apporte sur cette population une information dont la mesure est :

$$\mathcal{J}(I) = \sum_i p_i Log_2 \frac{1}{p_i} \text{ (formule de Shannon)}$$

(en effet, alors qu'il fallait $Log_2 n$ bits pour identifier un individu de cette population il n'en faut plus en moyenne, si l'on connaît le caractère I, que $\sum_i \frac{n_i}{n} Log_2 n_i$).

$\mathcal{J}(I)$ est également, dans cette population, la quantité d'information nécessaire pour identifier une modalité du caractère I.

Considérons dans la population un sous-ensemble de m individus dont la répartition selon le caractère I obéit à la distribution $q_i = m_i/m$.

Si l'on sait que l'individu que l'on veut repérer appartient à ce sous-ensemble, cela apporte sur le caractère I l'information suivante :

$$\sum_i q_i Log_2 \frac{q_i}{p_i} \text{ (gain d'information de Kullback)}$$

(en effet, il faut en moyenne $\sum_i q_i Log_2 \frac{1}{q_i}$ pour repérer une modalité du caractère I si on sait que l'individu appartient au sous-ensemble, et $\sum_i q_i Log_2 \frac{1}{p_i}$ si on l'ignore).

Considérons une population répartie selon les deux caractères I et J. Si nous savons qu'un élément possède la modalité j de J, cela apporte sur I l'information :

$$\mathcal{J}(I/j) = \sum_i f_i^j Log_2 \frac{f_i^j}{f_i}.$$

En moyenne, le fait de connaître *a priori* le caractère J apporte sur I une information égale à :

$$\mathcal{J}(I/J) = \sum_j f_j \mathcal{J}(I/j)$$

On vérifie aisément que cette quantité est égale à :

$$\mathcal{J}(I,J) = \sum_{ij} f_{ij} Log_2 \frac{f_{ij}}{f_i f_j}.$$

Cette expression étant symétrique selon les lettres i et j, on trouve un résultat remarquable : l'information apportée par I sur J est la même que celle apportée par J sur I. Comme $\mathcal{J}(I/J) = \mathcal{J}(J/I)$, l'information mutuelle des deux caractères peut comme nous l'avons fait ci-dessus se noter $\mathcal{J}(I,J)$.

On peut interpréter $\mathcal{J}(I,J)$ comme le gain d'information lorsque l'on passe de la connaissance des distribution marginales f_I et f_J à celle du tableau croisé f_{IJ}. $\mathcal{J}(I,J) = 0$ si et seulement si $f_{IJ} = f_I f_J$: il est donc équivalent de dire que les deux caractères I et J sont indépendants dans la population considérée, ou que leur information mutuelle est nulle. On peut dire aussi que $\mathcal{J}(I,J)$ mesure la corrélation des deux caractères qualitatifs I et J.

* *

Supposons que f_{IJ} soit peu différent de $f_I f_J$ et posons $f_{ij} = f_i f_j(1 + z_{ij})$. En faisant un développement limité de $\mathcal{J}(I,J)$ selon z_{ij}, on trouve :

$$\mathcal{J}(I,J) = \frac{1}{2Log2} \sum_{ij} f_i f_j (z_{ij}^2 + \epsilon z_{ij}^2).$$

D'où, en posant :

$$Lien(I,J) = \sum_{ij} \frac{(f_{ij} - f_i f_j)^2}{f_i f_j},$$

$$\mathcal{J}(I,J) \sim \frac{1}{2Log2} Lien(I,J).$$

L'expression $Lien(I,J)$ sur laquelle s'appuie notre méthode est donc, à une constante multiplicative près, une mesure approchée de l'information mutuelle des deux caractères I et J sur la population considérée.

Annexe 2 : Exemple d'application

Nous avons appliqué la méthode au tableau de 260 cases qui représente les résultats du recensement de la population de 1999 (source : `www.insee.fr`). La population est répartie par région (26 régions en considérant les DOM comme des « régions »), sexe et classe d'âge (deux sexes, cinq classes d'âge).

Le tableau et les résultats de l'analyse se trouvent dans le fichier Excel `www.volle.com/statistiques/pop1999.xls`. La dernière feuille de ce fichier contient le résultat de l'analyse factorielle des correspondances appliquée au même tableau (voir `www.volle.com/travaux/afc.htm`).

Nous ne faisons ici qu'esquisser le commentaire des résultats : il serait plus fouillé si nous faisions une étude « en vraie grandeur ».

4.3 Test du χ^2

Le test du χ^2 donne 42 000 : l'hypothèse d'indépendance des deux caractères est donc largement rejetée.

4.4 Examen des distributions marginales

La population considérée comporte 60 190 000 individus. Les régions sont de taille diverse. La plus grande est l'Île-de-France, avec 11 millions ; la plus petite est la Guyane, avec 160 000. La taille moyenne est de 2 300 000, l'écart-type est de 2 170 000.

La pyramide des âges (figure 2) présente une base étroite (faible nombre de jeunes), et dans les classes âgées une forte proportion de femmes.

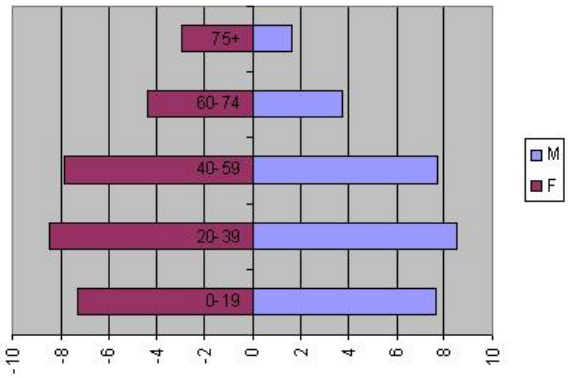

FIG. 2 – *Pyramide des âges*

4.5 Spectres les plus significatifs

Les deux spectres qui apportent le plus d'information sont ceux de l'Île-de-France (23,6 %, figure 3) et de la Réunion (14,4 %, figure 4). L'Île-de-France est caractérisée par une forte proportion de jeunes adultes (notamment de jeunes femmes) et une faible proportion de personnes de plus de 60 ans. La Réunion se distingue par une forte proportion de jeunes. Ces deux spectres, à eux seuls, représentent 38 % de l'information contenue dans le tableau.

* *

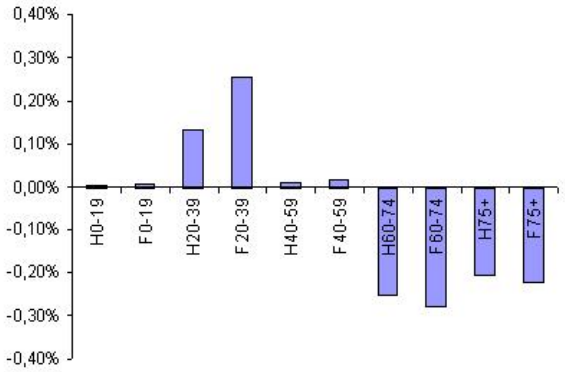

FIG. 3 – *Spectre de l'Île-de-France*

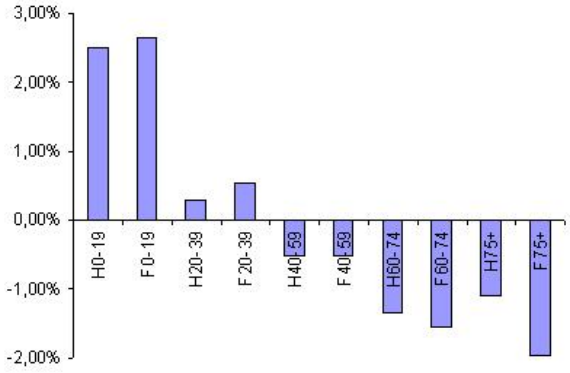

FIG. 4 – *Spectre de la Réunion*

4.6 Cases les plus significatives

Les cases les plus significatives sont, dans l'ordre, celles qu'indique le tableau 2. Les premières d'entre elles appartiennent - ce n'est pas une surprise - aux deux régions dont les spectres sont les plus significatifs.

Ligne	Colonne	cr_{ij} signée	cumul
F60-74	IdF	-4,75%	4,75%
F20-39	IdF	4,37%	9,12%
H60-74	IdF	-4,33%	13,45%
F75+	IdF	-3,80%	17,25%
H75+	IdF	-3,53%	20,77%
F0-19	Réunion	2,95%	23,73%
H0-19	Réunion	2,80%	26,53%
H20-39	IdF	2,27%	28,80%
F75+	Réunion	-2,17%	30,97%
etc.			

TAB. 2 – *Les cases qui apportent le plus d'information*

4.7 Modalités les plus « originales »

Dans le cas considéré ici, les spectres des régions sont les plus faciles à lire, et sans doute les plus intéressants, en raison de la nature ordinale du caractère « classe d'âge ». La région dont les proportions s'éloignent le plus de la moyenne est la Guyane (figure 5), avec $q_j = 26{,}6$ % : elle est caractérisée par une très forte proportion de jeunes.

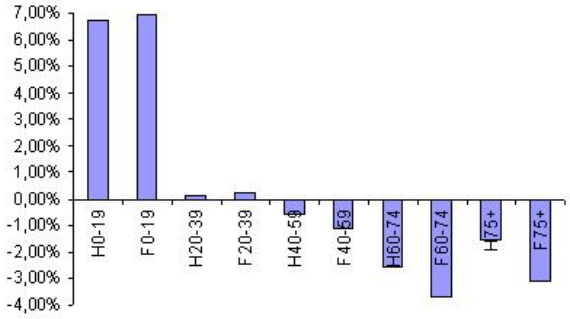

FIG. 5 – *Spectre de la Guyane*

On retrouve la Réunion (13 %), on découvre le Limousin (5,2 %), région la plus âgée de France (figure 6) etc. Les régions dont les proportions sont les plus proches de celles de la France entière sont Champagne-Ardennes (0,03 %) et Franche-Comté (0,04 %).

4.8 Interprétation

La structure de la population de l'Île-de-France, par exemple, s'explique par les flux migratoires : cette région fonctionne comme une pompe qui aspire de jeunes adultes et les refoule plus tard. La structure de la population de la Guyane s'explique, au moins pour

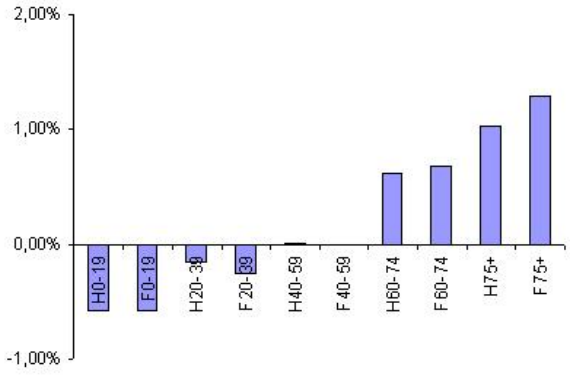

FIG. 6 – *Spectre du Limousin*

partie, par la perméabilité de la frontière avec le Surinam, qui n'a pas la même législation sociale, etc.

Références

[1] Jean-Paul Benzécri. *Analyse des données*. Dunod, 1973.

[2] Laurent Bloch. *Initiation à la programmation avec Scheme*. Technip, 2001.
Introduction au langage qui procure le plus de plaisir au programmeur.
Voir `www.laurent-bloch.org/Livre-Scheme/TDM.html`.

[3] Nenad Jukic. Modeling strategies and alternatives for data warehousing projects. *Communications of the ACM*, (49-4), avril 2006.

[4] Alfred Renyi. *Calcul des probabilités*. Dunod, 1966.

[5] Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, juillet - octobre 1948.

[6] Michel Volle. Une méthode pour lire et commenter automatiquement de grands tableaux statistiques. *Économie et Statistique*, (52), janvier 1974.

[7] Michel Volle. *Analyse des données*. Economica, quatrième edition, 1997.

[8] Michel Volle. Fonctionnement d'un système informatique d'aide à la décision (siad). In *Revue des Nouvelles Technologies de l'Information, EGC 2004*, volume I. CEPAD, 2004.
Voir `www.volle.com/travaux/siad.htm`.