Statistique et analyse des données

J.-B. Kazmierczak

Sur l'usage d'un principe d'invariance pour aider au choix d'une métrique

Statistique et analyse des données, tome 12, n° 3 (1987), p. 37-57

<http://www.numdam.org/item?id=SAD_1987__12_3_37_0>

Statistique et Analyse des Données
1987 - vol. 12 n° 3 p. 37-57

SUR L'USAGE D'UN PRINCIPE D'INVARIANCE POUR AIDER AU CHOIX D'UNE METRIQUE

J.-B. KAZMIERCZAK

Centre de Recherche et Développement COTELLE
13, rue Carnot
94200 IVRY SUR SEINE

Résumé : *Après avoir rappelé l'intérêt que peut présenter la donnée d'un principe d'invariance, et discuté de l'apport du principe de Yule, nous proposons un cadre géométrique et une démarche qui permettent de découvrir facilement une famille de métriques satisfaisant à un principe donné. Le cadre géométrique est celui des variétés riemanniennes ; la démarche consiste à traduire le principe d'invariance au niveau du ds^2 pour en déduire la forme du tenseur métrique.*

Une telle démarche nous permet d'englober sous le principe d'équivalence distributionnelle large des analyses telles que l'analyse factorielle des correspondances et l'analyse sphérique mais aussi l'analyse logarithmique. Elle permet également d'obtenir une quantité d'autres métriques y satisfaisant et pouvoir, par exemple, passer continûment d'une de ces analyses à une autre !

Abstract : *After recalling the interest of giving an invariance principle, and discussed the contribution of Yule's principle, we propose a geometrical scope and approach which allow us to easily discover a family of metrics satisfying to a given principle. The scope is riemannian geometry and the approach the translation in term of ds^2 of the invariance principle so that we obtain the metric tensor (first fundamental form).*

Manuscrit reçu le 15 novembre 1986
Révisé le 10 novembre 1987

With such considerations, we may include in the so-called "generalized distributional equivalence principle" different kind of data analysis : correspondence analysis and sheric analysis but also logarithmic analysis. And still many others, so it will be possible to continuously slip from one of these data analysis to an other one.

Mots clés : *Principe d'invariance, équivalence distributionnelle, principe de Yule, tenseur métrique.*

Indices de classification STMA : *06-000, 06-070, 06-110*

1. Généralités

Dans de précédentes publications (J.B. Kazmierczak, 1985a, 1985b), nous avions émis l'idée selon laquelle, en analyse de données, la recherche d'une métrique pouvait être guidée par la considération d'un "principe d'invariance".

Si une telle réflexion peut avoir quelque originalité, ce n'est pas par sa primauté - J.P. Benzécri a énoncé le principe d'équivalence distributionnelle au début des années soixante - mais plutôt par sa généralité.

Il nous semble en effet que parmi les questions que l'on est en devoir de se poser avant de soumettre des données à l'analyse statistique, l'une des plus urgentes est celle-ci :

"quelles sont les transformations portant soit sur les données elles-mêmes, soit sur les ensembles qui les sous-tendent (individus et variables), qui ne doivent pas affecter les résultats de notre analyse ?"

Habituellement, les données concernent un domaine sur lequel on possède une certaine connaissance *a priori* - ne serait-ce que celle relative à la définition des grandeurs que l'on manipule. Cette connaissance, si faible soit-elle, doit nous permettre d'énoncer un principe d'invariance.

Un exemple classique nous permet d'illustrer ce propos : un individu interrogé lors d'une enquête répond à l'aide d'une échelle d'intensité arbitrairement codée (1, 2, 3, 4). Sans plus de précision, on peut vouloir qu'une transformation monotone quelconque de cette échelle ne modifie pas les résultats de l'analyse que l'on effectuera. Un recodage sous forme disjonctive complète, en analyse des correspondances, apporte une satisfaction à ce souhait.

Dans ce travail, ce n'est pas directement cet aspect du "codage" qui sera abordé.

On se limitera au cas où les ensembles I des individus et J des variables ou caractères descriptifs sont fixés et on supposera que le tableau X des données est acquis. Ces données sont des résultats de mesures, le comptage étant considéré comme un cas particulier de mesure. Les nombres x_{ij} de ce tableau seront supposés strictement positifs ; le cas de valeurs nulles sera évoqué le moment venu.

Dans le cas où ces données sont des grandeurs essentiellement "additives extensives" (cf J.P. Benzécri, 1978) on sait l'intérêt que présente le principe d'équivalence distributionnelle (par la suite : ped).

Dans d'autres situations, par exemple lorsque la notion de "forme" domine celle de "taille", nous avons montré l'avantage qu'il pouvait y avoir à se référer au principe de Yule - principe que nous énonçons plus loin.

Aujourd'hui, rappelant que ces deux principes trouvent une unité et une généralisation dans le "ped-large", nous proposons une approche différente de la recherche de métriques satisfaisant à un principe d'invariance.

2. Trois principes d'invariance (métrique et inertielle)

2.1. A propos de profils.

La notion de profil est essentielle dans ce travail ; aussi y consacrerons-nous quelques lignes.

Nous appelons profil la classe d'équivalence définie par la relation de proportionnalité : deux colonnes (ou deux lignes) proportionnelles définissent le même profil-colonne (resp. profil-ligne).

Il est souvent commode de choisir un représentant d'une classe d'équivalence et, par abus de langage, de l'appeler profil. Cette opération se fera ici de la façon suivante :

$\underline{x} = (x^1, x^2, \ldots, x^p)$ étant un élément de $\mathbf{R}^p$,

on se donne une fonction H, homogène(*) de degré 1, définie sur $\mathbf{R}^p$; dans la classe d'équivalence de $\underline{x}$ on choisira le représentant :

$\underline{z} = (z^1, z^2, \ldots, z^p)$ défini par : $z^j = x^j / H(\underline{x})$

Notons qu'en AFC, la fonction H n'est autre que la somme des coordonnées et qu'en ACP-normée c'est, sur l'espace des variables, l'écart-type.

Dans l'analyse logarithmique (dont on reparle plus loin), si l'on se réfère à l'approche de J. Aitchison (1984a), la fonction H est la moyenne géométrique.

2.2. Ped et ped-large.

Rappelons brièvement le ped :

"on ne modifie pas la distance entre deux profils-lignes quelconques en remplaçant dans le tableau de données deux colonnes proportionnelles par une seule, celle-ci étant somme de celles-là".

(*) Rappelons la définition d'une fonction homogène de degré α . H, définie sur $\mathbf{R}^p$ est dite homogène de degré α lorsque pour tout point $\underline{x}$ de $\mathbf{R}^p$ et pour tout réel t positif on a :

$H(t.x^1, t.x^2, \ldots, t.x^p) = t^{\alpha} . H(x^1, x^2, \ldots, x^p)$

L'intérêt d'une telle attitude n'est plus à démontrer. En particulier, on connaît bien le peu de sensibilité de l'AFC quant à une modification de nomenclature (e.g. utiliser les départements français plutôt que les régions économiques).

Cependant, nous préférons modifier légèrement l'énoncé de ce principe et proposer le "ped-large" :

"on ne modifie pas la distance entre deux profils-lignes quelconques en remplaçant dans le tableau de données, deux colonnes proportionnelles par deux autres colonnes proportionnelles aux précédentes de telle sorte que la somme des deux initiales soit égale à la somme des deux finales".

Pour reprendre l'exemple de la nomenclature : le ped-large permet de déplacer la frontière séparant deux entités équivalentes, le ped (strict) autorise la suppression de cette même frontière.

Enfin, il est clair qu'une distance satisfaisant au ped satisfera aussi au ped- large. Nous avons montré par ailleurs que la réciproque est fausse ; ce qui justifie la terminologie (J.B. Kazmierczak, 1985a).

Nous voudrions, à ce stade, faire remarquer que si l'on appelle "analyse factorielle" une représentation synthétique d'un nuage de points munis de masses, il faut, non seulement se poser le problème de la distance - et de son invariance - mais aussi celui de la pondération - et de l'invariance de l'inertie.

En souhaitant, en plus de l'invariance des distances, celle des inerties, il nous faut impérativement, dans le cas du ped, affecter à chaque profil-ligne une masse qui soit une fonction homogène de degré 1 des termes de cette ligne *(cf annexe 1)* . Ainsi en AFC choisit-on de pondérer le profil-ligne par la somme des termes de la ligne.

Cette contrainte est différente pour le ped-large. S'il est toujours possible de conserver ce type de pondération, on peut maintenant y adjoindre des fonctions homogènes de degré 0. Tout système de la forme $Z + U$ où Z et U sont des fonctions homogènes de degré respectif 0 et 1, laissera les inerties invariantes.

2.3. Le principe de Yule et l'analyse logarithmique.

On peut regretter que les deux principes que nous venons de rappeler exigent, pour être appliqués, que le tableau de données possède la particularité de présenter deux colonnes proportionnelles (ou deux lignes, selon le problème ...).

Pour éviter cette difficulté, nous avons proposé le principe de Yule :

"on ne change pas la distance entre deux profils-lignes quelconques en remplaçant, dans le tableau de données, n'importe quelle colonne par une autre colonne qui lui soit proportionnelle".

Il est remarquable de constater qu'une "distance de Yule" vérifie le ped-large. La réciproque est fausse.

En ce qui concerne l'invariance des inerties, il est impératif d'affecter au profil-ligne une masse qui soit fonction homogène de degré 0 des termes de cette ligne.

Sans vouloir revenir ici sur des développements déjà exposés dans les publications citées plus haut, rappelons simplement que si dans l'énoncé du principe de Yule on souhaite conserver une symétrie parfaite entre les lignes et les colonnes du tableau de données (comme on le fait en AFC), il devient équivalent de demander que les analyses des tableaux de termes généraux :

$$(x_i^j) \qquad \text{et} \qquad (a_i \, . \, x_i^j \, . \, b^j)$$

conduisent aux mêmes résultats. Une manière simple et élégante d'y répondre consiste à effectuer une ACP simple (métrique : identité) sur le tableau "bicentré" (centré en ligne et en colonne) des logarithmes des données ; c'est ce que nous appelons analyse logarithmique.

Nous discuterons plus loin de quelques situations où une telle démarche nous paraît naturelle mais rappelons encore qu'une telle analyse possède des propriétés très voisines de celles que l'on observe en AFC. En particulier, pour un tableau de fréquences à marges constantes, au voisinage de l'indépendance, l'AFC et l'analyse logarithmique fournissent des résultats sensiblement identiques.

2.4. A propos de l'intérêt du principe de Yule.

Avant de revenir à l'essentiel de notre propos - la recherche de métriques - examinons quelques situations où la référence au principe de Yule nous semble soit indispensable, soit simplement souhaitable.

La première situation que nous présentons, rappelle celle qui nous a déjà servi à présenter l'analyse logarithmique : le tableau de "ratios économiques". Dans un tel tableau, les lignes sont des "unités statistiques économiques" telles que des entreprises, des régions économiques voire des pays ou encore, pour l'une quelconque de ces entités, un instant de son évolution (i.e. une date). Les colonnes représentent des "indicateurs économiques" qui sont définis comme des ratios : rapports entre une variable et une autre prise comme référence.

Il importe peu, si l'on se réfère au principe de Yule, de savoir quelle est la référence (ni même de savoir comment elle est définie !), il suffit de la considérer commune à toutes les variables.

L'exemple le plus simple est celui d'un tableau décrivant l'évolution du cours des monnaies ou taux de change. Une telle donnée représente la valeur d'une unité monétaire exprimée dans l'unité monétaire d'un autre pays - peu importe lequel.

Une autre situation qui ne diffère de la précédente que par la terminologie adoptée, est l'analyse de formes - en appelant "forme" la description d'un objet par un ensemble de rapports entre les mesures brutes (mensurations) et une mesure dite de "taille". Cette dernière peut être soit une mesure réellement effectuée, soit une quantité purement fictive qui n'a pas besoin d'être réellement définie puisqu'elle n'influera pas sur les résultats de l'analyse !

Une troisième situation nous est donnée par les tableaux dont une marge est fixée *a priori* . Une illustration est fournie par certaines enquêtes d'opinion.

Lorsqu'on cherche à étudier la perception d'un certain nombre de stimuli à l'aide d'un ensemble de qualificatifs, on peut procéder de la façon suivante : on présente un stimulus à un individu, puis on lui demande de choisir (dans une liste fixée *a priori*) le qualificatif qui lui semble le mieux décrire ce stimulus. Un exemple bien

connu est celui du choix d'une marque de cigarettes repris par J.P. Benzécri (1973, Tome II C, n° 3).

Un autre exemple, présenté par le même auteur et sur lequel nous reviendrons par la suite, est celui de la perception des couleurs.

Classiquement, on présente un même nombre de stimuli à chaque individu ; le total marginal est donc connu *a priori* : il n'est autre que le nombre de réponses obtenues. Notons x_{ij} le nombre d'individus qui, pour le stimulus i, ont donné la préférence au qualificatif j et $x_{i.}$ le total marginal. Si l'enquête est effectuée auprès d'un autre échantillon ayant la même représentativité, les résultats notés x^*_{ij}, conduiront à des profils

$$x_{ij} / x_{i.} \quad \text{et} \quad x^*_{ij} / x^*_{i.}$$

pratiquement identiques. Aussi une analyse utilisant une métrique de Yule sera insensible au choix de la marge $x_{i.}$, ce qui n'est pas le cas en AFC !

3. Quelques métriques et analyses particulières

Le ped-large englobe - au sens de l'invariance métrique - le ped (strict) dont on a très rapidement rappelé l'intérêt et le principe de Yule sur lequel nous nous sommes attardés plus longtemps et dont nous espérons avoir montré l'attrait qu'il présentait.

De ce fait, il nous semble intéressant de rechercher les métriques satisfaisant au ped-large.

3.1. Une approche locale.

Ecrire l'invariance de la distance entre deux (profils-) lignes lorsqu'on envisage les transformations du ped-large est certes possible, mais demeure d'exploitation difficile. En nous plaçant au niveau local, la forme euclidienne du ds^2 nous laissait espérer un développement plus facile.

La démarche consiste à munir tout d'abord l'espace des variables d'une structure de variété riemannienne (sans se préoccuper pour l'instant des problèmes de représentation). Cela revient à se donner un tenseur métrique ($g_{jj'}$) de telle sorte que :

$$ds^2 = \sum_{jj'} g_{jj'} dx^j dx^{j'}$$

soit une forme quadratique définie positive. On cherche alors à traduire le principe d'invariance au niveau du ds^2.

A ce stade, il est nécessaire d'apporter quelques précisions sur le tenseur métrique. Bien entendu, il sera fonction des coordonnées (x^j) mais sera aussi paramétré par des constantes notées (m^j). L'intérêt de ces constantes nous est révélé dès à présent par le fait que l'on s'attend à retrouver - comme cas particulier - la métrique du chi-deux qui est définie par :

$$g_{jj'} = \delta_{jj'} / m^j$$

m^j représentant la marge j. Plus généralement ces paramètres m^j seront définis par la donnée du tableau (x_i^j). On écrira :

$$m^j = M(x_1^j , x_2^j , \ldots , x_n^j)$$

On ne cherchera pas à préciser ici la fonction M, mais pour des raisons qui ne doivent qu'à la facilité des calculs que nous rencontrerons, nous supposerons que c'est une fonction homogène de degré 1 - comme c'est le cas pour la masse en AFC.

Pour être appliqué, le ped-large exige que le tableau des données possède deux colonnes proportionnelles. On pourra supposer, sans perte de généralité, qu'il s'agit des colonnes x^1 et x^2. Notons :

$$x^1 = t^1 . x^0 \qquad \text{et} \qquad x^2 = t^2 . x^0$$

Nous remplaçons ces colonnes proportionnelles par deux autres colonnes x^{*1} et x^{*2}:

$$x^{*1} = t^{*1} . x^0 \qquad \text{et} \qquad x^{*2} = t^{*2} . x^0$$

avec la condition : $t^1 + t^2 = t^{*1} + t^{*2}$
suivant en cela les règles de transformation du ped-large. Alors l'invariance du ds^2 consiste à écrire : $ds^2 = ds^{*2} = \sum_{jj'} g_{jj'} dx^j dx^{j'} = \sum_{jj'} g^*_{jj'} dx^{*j} dx^{*j'}$

Il suffit d'identifier les termes dans le développement des sommes pour obtenir les relations :

$$(t^1)^2 g_{11} + 2 t^1 t^2 g_{12} + (t^2)^2 g_{22} = (t^{*1})^2 g^*_{11} + 2 t^{*1} t^{*2} g_{12} + (t^{*2})^2 g^*_{22} \qquad (3.1.1)$$

et

$$t^1 g_{1j} + t^2 g_{2j} = t^{*1} g^*_{1j} + t^{*2} g^*_{2j} \qquad (3.1.2)$$

Pour des raison de simplicité de l'exposé, nous nous limiterons maintenant au seul cas où le tenseur métrique est de forme diagonale :

$$g_{jj'} = \delta_{jj'} . G(x^j , m^j) \qquad (3.1.3)$$

On a alors le résultat simple suivant (dont on trouvera la démonstration à l'*Annexe II*) :

$$G(x , m) = (1/m) V(x/m) + (1/m^2) W(x/m) \qquad (3.1.4)$$

où V et W sont deux fonctions quelconques ; W devant être nulle si l'on souhaite satisfaire au ped-strict.

Si, à la place du ped, on adopte le principe de Yule, les transformations à envisager s'écrivent alors :

$$x^j \quad \text{----->} \quad x^{*j} = a^j . x^j$$

L'invariance du ds^2 s'écrit alors, en tenant compte de la forme (3.1.3) :

$$ds^2 = ds^{*2} = \sum_j G(x^j , m^j) (dx^j)^2 = \sum_j G(x^{*j} , m^{*j}) (dx^{*j})^2$$

$$= \sum_j G(a^j.x^j , a^j.m^j) (a^j.dx^j)^2$$

où nous avons tenu compte de la propriété adoptée plus haut selon laquelle les paramètres m^j sont des fonctions homogènes de degré 1 des coordonnées. Ici, l'identification est immédiate :

$$G(ax , am) = (1/a^2).\ G(x , m)$$

ce qui, en prenant $a = 1/m$ et notant $W(x/m) = G(x/m , 1)$, se réécrit facilement:

$$G(x , m) = (1/m^2)\ .\ W(x/m)$$

En résumé, on retiendra que les fonctions G définissant les métriques du ped-large sont somme

* d'une fonction homogène de degré -1 : $G_1(x , m) = (1/m)\ .\ V(x/m)$

qui seule définit les métriques du ped-strict

et

* d'une fonction homogène de degré -2 : $G_2(x , m) = (1/m^2)\ .\ W(x/m)$

qui seule définit les métrique de Yule.

3.2. Retour au niveau global

Ici, deux attitudes sont envisageables :

* ou bien on introduit à ce stade le fait que les x^j sont les coordonnées d'un profil et donc qu'il existe entre eux une relation : ceci aura pour effet de rechercher la métrique induite sur la sous-variété définie par cette relation.

* ou bien on considère provisoirement que les x^j sont "libres" : ce qui conduira ensuite à "remplacer l'arc par la corde".

C'est encore pour des raisons de facilité que nous choisissons ici la seconde approche. En effet, dans ce cas, nous pouvons expliciter :

$$ds^2 = \Sigma_j\ G(x^j , m^j)\ (dx^j)^2 = \Sigma_j\ [\ (\ G(x^j , m^j)\ )^{1/2}\ .\ dx^j\ ]^2$$

et utiliser le changement de coordonnées défini par :

$$dy = [\, G(x, m) \,]^{1/2} \, . \, dx \qquad (3.2.1)$$

A ce stade, il nous semble utile de montrer, comment en choisissant pour G quelques cas particuliers, on retrouve des analyses bien connues. On a ainsi :

* $V(x/m) = 1 \quad ; \quad W(x/m) = 0$

avec $m^j = \Sigma_i \, x_i^j$: somme des termes d'une colonne, l'AFC la plus classique.

avec $m^j = (\Sigma_i \, x_i^j) / r^j$ où r^j est un coefficient défini comme la somme des masses des seules lignes i pour lesquelles les valeurs x_i^j sont non nulles, on retrouve une variante de l'AFC proposée par B. Escofier (1978). On notera qu'un tel coefficient reste invariant dans les transformations envisagées dans le ped (et accessoirement dans le ped-large) de sorte que m^j demeure de la sorte homogène de degré 1.

* $V(x/m) = m/x \quad ; \quad W(x/m) = 0$

la relation (3.2.1) conduit à la transformation : $y = x^{1/2}$ et on retrouve l'analyse sphérique telle qu'elle a été définie par M. Volle (1978), la distance entre deux profils p et q s'écrivant alors :

$$d^2(p, q) = \Sigma_j \, (\sqrt{p^j} - \sqrt{q^j})^2 \qquad (3.2.2)$$

A cette occasion il peut être intéressant de préciser, comment se serait exprimée cette distance si on avait tenu compte du fait que les coordonnées sont celles d'un profil défini par la relation : $\Sigma_j \, p^j = 1$. C'est la première attitude que nous nous sommes contentés d'évoquer au début de ce paragraphe. La manière traditionnelle - et générale - de procéder consiste à écrire : $p^r = 1 - \Sigma_{j<r} \, p^j$ (où $r = \text{card}(J)$) pour obtenir l'expression du ds^2 dans la variété de dimension (r-1) définie par les (r-1) premières coordonnées pour déterminer ensuite les géodésiques. Le calcul - lourd - n'est pas d'un grand intérêt, mais le résultat est bien connu ; on trouve :

$$d(p, q) = \text{Arccos} \, [\, \Sigma_j \, (p^j \, . \, q^j)^{1/2} \,] \qquad (3.2.3)$$

Cette dernière distance n'est autre que la longueur du grand arc de cercle joignant les points de coordonnées $(\sqrt{p^j})$ et $(\sqrt{q^j})$ situés sur la sphère unité, alors que la distance définie en (3.2.2) est celle de la corde joignant ces deux mêmes points.

* $V(x/m) = 0$; $W(x/m) = (m/x)^2$

cette fois, on tombe sur la transformation : $y = Ln(x)$ qui conduit naturellement à l'analyse logarithmique.

Bien entendu, en écrivant des expressions telles que :

$V(x/m) = a + b.(m/x)$; $W(x/m) = 0$

on arrive à définir d'autres métriques qui vont satisfaire au ped (strict !) et qui conduiront à des analyses "intermédiaires" entre l'AFC et l'analyse sphérique. D'autres combinaisons nous feront passer de l'une des trois analyses AFC, AS, AL à une autre.

4. Un exemple illustratif

Nous nous proposons, dans ce paragraphe, de comparer les résultats fournis par deux analyses : l'AFC et l'analyse logarithmique (AL par la suite).

Compte tenu du théorème d'approximation que nous avons évoqué plus haut (§ 2.3) il était plus pertinent de choisir un tableau de données pour lequel l'AFC fournisse des valeurs propres élevées.

De plus, il était souhaitable que deux points de vues puissent être adoptés : l'un justifiant l'usage de l'AFC, l'autre celui de l'AL.

Le tableau que nous avons retenu est emprunté à J.P. Benzécri (1970) et J.F. Richard et concerne un problème de perception visuelle. On présente à des sujets, huit couleurs monochromatiques projetées sur un écran blanc ; les sujets doivent apprendre à associer aux couleurs les boutons d'un clavier. La configuration de celui-ci risque de perturber les résultats (en ce sens qu'on risque de mettre en évidence la structure du clavier lors de l'analyse) aussi les auteurs ont-ils recours à la "randomisation" : on attribue à chaque sujet un code de correspondance choisi au hasard parmi l'ensemble des codes possibles.

On ne s'intéresse qu'à la fin de l'expérience, définie, pour chaque sujet indépendamment, comme la phase qui commence au premier essai où il a donné au moins quatre réponses exactes sur huit. On obtient alors la "matrice de confusion" suivante :

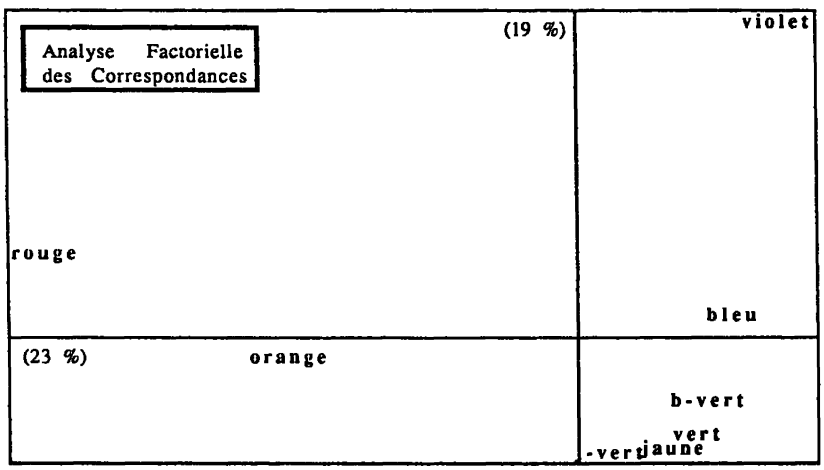
REPONSES \ STIMULI	Roug	Oran	Jaun	Jver	Vert	Bver	Bleu	Viol
Rouge	415	45	2	8	7	4	4	3
Orange	32	373	16	17	8	11	12	8
Jaune	10	12	343	70	22	20	13	10
J-vert	6	19	50	303	31	23	18	6
Vert	6	12	23	36	305	71	29	8
B-vert	10	10	15	32	91	274	38	19
Bleu	8	11	14	6	17	60	356	36
Violet	3	5	22	13	11	13	24	403

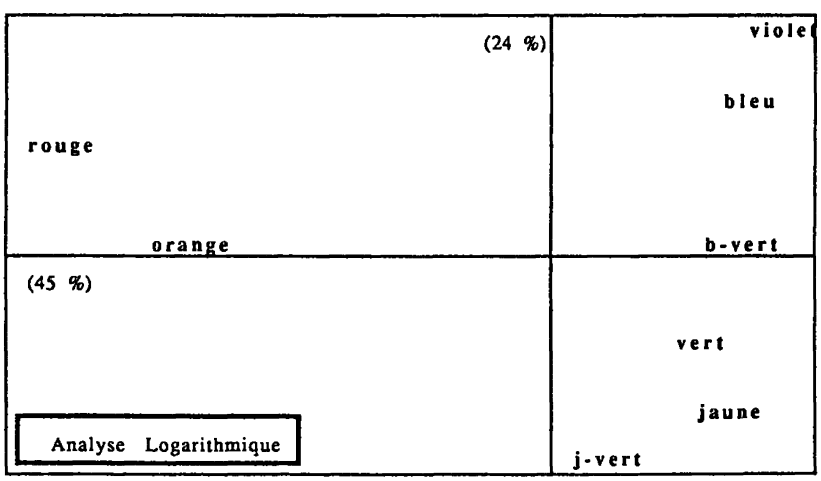
L'AFC de ce tableau fournit des valeurs propres toutes supérieures à 0.175, les deux plus grandes (0.758 et 0.624) représentent respectivement 23 et 19 % (soit 42%) de l'inertie totale.

En effectuant l'AL du même tableau, les deux plus grandes valeurs propres obtenues représentent 45 et 24 % de l'inertie. La somme approche cette fois les 70%.

Les diagrammes obtenus, dans le plan (1 , 2) sont assez ressemblants. Dans les deux cas on observe le dessin d'un triangle qui rappelle celui de la CIE (commission internationale de l'éclairage). Les points représentant les couleurs s'y répartissent dans leur ordre naturel - à une exception près le "jaune".

Toutefois, il est remarquable de constater que l'AFC distingue mal les points "jaune", "jaune-vert", "vert" et même "bleu-vert". Ces mêmes points sont bien plus régulièrement espacés sur le diagramme fourni par l'AL, ce qui semble mieux correspondre à l'idée que l'on peut avoir *a priori* , et conduit à un diagramme plus ressemblant à celui de la CIE.





4. Conclusions

A ce stade de l'étude, une remarque émerge naturellement :

le choix - certes primordial - d'un principe d'invariance, s'il conduit à une classe de métriques, ne permet pas, à lui seul, d'en définir une unique. Dès lors, le choix d'une distance n'est pas entièrement résolu.

Ce à quoi nous répondrons que :

(i) le principe d'invariance doit être vu comme un préliminaire. D'autres considérations doivent apporter, en fonction du problème étudié, de son cadre ... des éléments de choix.

Ainsi pour le ped-strict. Si le tableau étudié est une vraie table de contingence, le choix de la métrique du chi-deux devient naturel : il permet de bénéficier des résultats théoriques obtenus par ailleurs (e.g. dans le cadre inférentiel).

(ii) les raisons de simplicité ne sont pas à rejeter. Dans une famille de métriques satisfaisant à un principe d'invariance, en dehors de toute contrainte, on pourra choisir celle qui conduit à des développements particulièrement simples.

C'est ainsi que pour le principe de Yule nous préférons l'analyse logarithmique à d'autres qui utilisent des métriques plus "exotiques" (telles que celles définies à partir des coefficients d'association de Yule - cf J.B. Kazmierczak 1985 b).

(iii) enfin, certains théorèmes d'approximation existent. Ils montrent que dans des conditions particulières (e.g. au voisinage de l'indépendance) diverses métriques d'une même classe conduisent à des résultats similaires.

Annexe 1 : Invariance de l'inertie

Le ped (strict ou large) exige que le tableau possède deux lignes (ou deux colonnes) proportionnelles. Sans perte de généralité, nous supposerons que les lignes x_1 et x_2 sont proportionnelles :

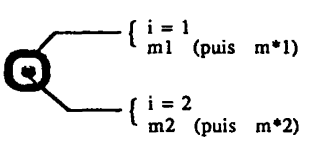
$$x_1 = t_1 . x_0 \qquad et \qquad x_2 = t_2 . x_0$$

Nous remplaçons ces lignes proportionnelles par deux autres lignes x^*_1 et x^*_2 :

$$x^*_1 = t^*_1 . x_0 \qquad et \qquad x^*_2 = t^*_2 . x_0$$

avec la condition : $t_1 + t_2 = t^*_1 + t^*_2$ (1)

Ces lignes seront affectées des masses respectives : m_1 , m_2 , m^*_1 , m^*_2 .



Les profils (-lignes) 1 et 2 sont identiques donc représentés par des points confondus. Les distances étant inchangées au cours des transformations envisagées, il suffit d'écrire :

$$m1 + m2 = m^*1 + m^*2$$

En imposant aux masses de n'être dépendantes que de la ligne du tableau, on écrira :

$m_1 = \mu(t_1 x_0)$ $\qquad m_2 = \mu(t_2 x_0)$ et de même avec *

de sorte que l'invariance de la somme des masses s'écrira :

$$\mu(t_1 x_0) + \mu(t_2 x_0) = \mu(t^*_1 x_0) + \mu(t^*_2 x_0) \qquad (2)$$

Pour x_0 donné, et compte tenu de la contrainte (1) indiquée plus haut, cette dernière relation caractérise les fonctions <u>affines</u> : $\mu(t x_0) = \beta(x_0) + \alpha(x_0).t$

Ce que nous pouvons exprimer en affirmant que le système de pondération peut s'écrire sous la forme $Z + U$, où Z et U sont des fonctions homogènes de degré respectif 0 et 1.

NB1. Dans le cas du ped (strict), deux lignes proportionnelles sont remplacées par une seule, somme de ces deux-là, de sorte que la relation (2) devient :

$$\mu(t_1\, x_0) + \mu(t_2\, x_0) = \mu((t_1 + t_2).x_0)$$

Cette fois ce sont les fonctions linéaires qui sont ainsi caractérisées. Le système de pondération se réduit aux seules fonctions homogènes de degré 1.

NB2. Enfin pour le principe de Yule, puisque l'on n'effectue pas de remplacement de lignes par d'autres lignes, il importe que les masses restent invariantes. Autrement dit : le système de pondération se limite aux seules fonctions homogènes de degré 0. Un cas particulier usuel est celui où le système de masses est uniforme.

Annexe 2 : Démonstration de la propriété : $G(x,m) = (1/m).V(x/m) + (1/m^2).W(x/m)$

En supposant que le tenseur métrique s'écrive :

$$g_{jj'} = \delta_{jj'} . G(x^j , m^j)$$

les relations générales (cf 3.1.1 et 3.1.2) :

$$(t^1)^2 \; g_{11} + 2\, t^1 t^2 \; g_{12} + (t^2)^2 \; g_{22} = (t^{*1})^2 \; g^*_{11} + 2\, t^{*1} t^{*2} \; g_{12} + (t^{*2})^2 \; g^*_{22}$$

et

$$t^1 g_{1j} + t^2 g_{2j} = t^{*1} g^*_{1j} + t^{*2} g^*_{22}$$

se résument à la seule équation :

$$(t^1)^2 \; G(x^1, m^1) + (t^2)^2 \; G(x^2, m^2) = (t^{*1})^2 \; G(x^{*1}, m^{*1}) + (t^{*2})^2 \; G(x^{*2}, m^{*2}) \qquad (1)$$

où l'on n'oubliera pas la contrainte : $t^1 + t^2 = t^{*1} + t^{*2}$ (2)

Mais, compte tenu :

(1) des relations de proportionnalité :

$$x^1 = t^1 . x^0 \; ; \; x^2 = t^2 . x^0 \; ; \; x^{*1} = t^{*1} . x^0 \; ; \; x^{*2} = t^{*2} . x^0$$

(2) du fait que la fonction $m^j = M(x_1^j , x_2^j , \ldots , x_n^j)$ est homogène de degré 1 et des relations de proportionnalité que nous venons de rappeler :

$$m^1 = t^1 . m^0 \; ; \; m^2 = t^2 . m^0 \; ; \; m^{*1} = t^{*1} . m^0 \; ; \; m^{*2} = t^{*2} . m^0$$

la relation (2) devient :

$$(t^1)^2 \; G(t^1 . x^0, t^1 . m^0) + (t^2)^2 \; G(t^2 . x^0, t^2 . m^0) = \text{idem avec (*)} \qquad (3)$$

Pour un couple (x^0 , m^0) donné, nous noterons simplement :

$$h(t) = t^2 . G(t.x^0 , t.m^0)$$

pour réécrire la relation (3) sous la forme particulièrement simple :

$$h(t^1) + h(t^2) = h(t^{*1}) + h(t^{*2}) \qquad (4)$$

ce qui, compte tenu de la contrainte (2), caractérise les fonctions affines :

$$t^2 . G(t.x , t.m) = t . v(x , m) + w(x , m)$$

soit encore, en faisant $m = 1$ et en notant $s = t.x$:

$$G(s , t) = (1/t) . v(s/t , 1) + (1/t^2) . w(s/t , 1)$$

où l'on fait apparaître deux fonctions ne dépendant que du rapport s/t . Finalement, en revenant à des variables notées de manière plus naturelle :

$$G(x , m) = (1/m) . V(x/m) + (1/m^2) . W(x/m)$$

ce qui est bien la relation attendue.

N.B. Cas du ped-strict.

Le second membre de la relation (1) se réduit au seul terme :

$$(t^1 + t^2)^2 . G((t^1 + t^2).x^0 , (t^1 + t^2).m^0)$$

et par suite la relation (4) devient :

$$h(t^1) + h(t^2) = h(t^1 + t^2)$$

ce qui réduit la famille précédente (des fonctions affines) aux seules fonctions linéaires. On doit se limiter aux fonctions de la forme :

$$G(x , m) = (1/m) . V(x/m)$$

BIBLIOGRAPHIE

Aitchison J. (1982). "The statistical analysis of compositional data". J. R. Stat. Soc. **44**, 2, 139-177.

Aitchison J. (1983). "Principal component analysis of compositional data". Biometrika. **70**, 1, 57-75.

Aitchison J. (1984). "Reducing the dimensionality of compositional data sets". Math. Geology. **16**, 6, 617-635.

Benzécri J.P. (1970). "Sur l'analyse des matrices de confusion". Rev. Stat. Appli. **XVIII**, 3, 5-63.

Benzécri J.P. (1973). "L'analyse des données" (2 tomes). Paris. Dunod.

Benzécri J.P. (1982). "Histoire et préhistoire de l'analyse des données". Paris. Bordas-Dunod.

Escofier B. (1978). "Analyse factorielle et distances répondant au principe d'équivalence distributionnelle. Rev. Stat. Appli. **XXVI**, 4, 29-37.

Kazmierczak J.B. (1985a). "Analyse logarithmique : deux exemples d'application". Rev. Stat. Appli. **XXXIII**, 1, 13-24.

Kazmierczak J.B. (1985b). "Une application du principe de Yule : l'analyse logarithmique" in "Data Analysis and Informatics, IV", Proceedings of the Fourth International Symposium on Data Analysis and Informatics. Versailles 1985 pp 393, 403. Edited by E. Diday et al., North Holland.

Volle M. (1978). "Analyse des correspondances sur la sphère". Rapport N° 252/930. Institut National de la Statistique et des Etudes Economiques. Paris.

Yule G.U. (1912). "On the method of measuring association between two attributes". J. R. Stat. Soc. 75, 579-642.