# Les pièges de l'analyse des correspondances

Philippe Cibois

**Abstract**

Abstract: The Pitfalls in correspondence Analysis. Built into correspondence analysis are several ill-understood effects whose consequences are dangerous for the user who ignores them, and may as a result become pitfalls nevertheless possible to avoid. Two of these effects are considered here. The homothetic effect makes comparable two correspondence analyses brought to bear on data having the same deviance structure at independence but of strongly differentiated intensities. The distinction effect creates a factor from the simple fact that several individuals have the same rare thus distinctive behavior. Finally, there is a type of case in which the order structure is pertinent and the structure of deviance is not. Correspondence analysis should not be used under those circumstances.

*Histoire & Mesure, 1997, XII-3/4, 299-320*

**Philippe Cibois***

# Les pièges de l'analyse des correspondances

**Résumé :** L'analyse des correspondances possède quelques effets peu connus qui peuvent avoir des conséquences dangereuses pour l'utilisateur s'il les ignore, et qui peuvent se transformer en un véritable piège qu'il lui est cependant possible d'éviter. Deux effets dangereux sont repérés : l'effet d'*homothétie* qui rend comparables deux analyses des correspondances faites sur des données ayant même structure d'écarts à l'indépendance mais d'intensité très différente ; et l'effet de *distinction* qui crée un facteur du simple fait que quelques individus associent les mêmes comportements rares donc distinctifs. Il existe, par ailleurs, une classe de cas où la structure d'ordre est pertinente et où la structure des écarts à l'indépendance ne l'est pas : il faut alors s'abstenir d'utiliser l'analyse des correspondances.

**Abstract: The Pitfalls in correspondence Analysis.** Built into correspondence analysis are several ill-understood effects whose consequences are dangerous for the user who ignores them, and may as a result become pitfalls nevertheless possible to avoid. Two of these effects are considered here. The *homothetic* effect makes comparable two correspondence analyses brought to bear on data having the same deviance structure at independence but of strongly differentiated intensities. The *distinction* effect creates a factor from the simple fact that several individuals have the same rare thus distinctive behavior. Finally, there is a type of case in which the order structure is pertinent and the structure of deviance is not. Correspondence analysis should not be used under those circumstances.

* Université de Picardie Jules Verne, Chemin du Thil, 80025 - Amiens.

Une méthode statistique peut conduire à des résultats inattendus si l'utilisateur maîtrise mal certaines de ses propriétés et, dans certains cas, cette ignorance peut devenir un piège. Le but de cet article est de montrer quelques effets peu connus de l'analyse des correspondances qui peuvent avoir des conséquences dangereuses pour l'utilisateur ignorant le piège mais qui sont sans grand risque pour qui le repère. J'en ai recensé deux, l'effet d'*homothétie* et l'effet de *distinction* et l'on verra comment on peut les éviter sans grande difficulté ; enfin je montrerai une classe de cas où il ne faut pas du tout employer l'analyse des correspondances.

## 1. L'effet d'homothétie

***Principe***

Une homothétie c'est une modification d'un ensemble (de nombres par exemple) où le facteur de modification (de réduction par exemple) est le même pour tous les éléments. L'effet d'homothétie en analyse des correspondances survient quand les écarts à l'indépendance de deux tableaux sont homothétiques : dans ce cas on risque de ne pas voir la modification dans les résultats numériques ou graphiques.

Pour faire voir l'effet, nous allons faire subir une homothétie à un tableau d'écarts à l'indépendance et comparer les analyses et les graphiques factoriels avant et après homothétie. Le tableau d'origine est, comme toujours, égal à la somme du tableau d'indépendance et des écarts à l'indépendance. Le tableau dérivé est égal à la somme du tableau d'indépendances resté inchangé et des nouveaux écarts à l'indépendance rendus extrêmement faibles par une homothétie de réduction très importante. On verra que les résultats sont très semblables et que si on ne fait pas attention à ce phénomène on risque de commenter des tableaux en écart infime à l'indépendance et qui donc ne devraient pas être examinés sans précaution.

Pour en juger, imaginons une population de 100 étudiants classés selon leur série du bac et selon leur destination l'année suivante : université, classes préparatoires aux grandes écoles et autres orientations (IUT et autres formations à finalités professionnelles).

| *Série* | *Université* | *Classes prép.* | *Prof.* | *Total* |
|---|---|---|---|---|
| Littéraire | 13 | 2 | 5 | 20 |
| Éco. et soc. | 20 | 2 | 8 | 30 |
| Scientifique | 10 | 5 | 5 | 20 |
| Tech. et pro. | 7 | 1 | 22 | 30 |
| *Total* | *50* | *10* | *40* | *100* |

Faire l'analyse de ce tableau, c'est faire la décomposition du tableau des écarts à l'indépendance obtenu en faisant pour chaque case la différence entre l'observation et la situation correspondant à l'indépendance (produit des marges divisé par le total).

Soit le tableau correspondant à l'indépendance :

| *Série* | *Université* | *Classes prép.* | *Prof.* | *Total* |
|---|---|---|---|---|
| Littéraire | 10 | 2 | 8 | 20 |
| Éco. et soc. | 15 | 3 | 12 | 30 |
| Scientifique | 10 | 2 | 8 | 20 |
| Tech. et pro. | 15 | 3 | 12 | 30 |
| *Total* | *50* | *10* | *40* | *100* |

Par exemple la case Littéraire-Université est le produit des marges $20 \times 50 = 1\,000$ divisé par le total général de 100 soit un effectif théorique sous l'hypothèse d'indépendance de 10.

Par soustraction entre l'observation et le tableau correspondant à l'indépendance on obtient le tableau des écarts à l'indépendance :

| *Série* | *Université* | *Classes prép.* | *Prof.* |
|---|---|---|---|
| Littéraire | 3 | 0 | − 3 |
| Éco. et soc. | 5 | − 1 | − 4 |
| Scientifique | 0 | 3 | − 3 |
| Tech. et pro. | − 8 | − 2 | 10 |

Par exemple pour la case Littéraire-Université, l'observation est de 13, l'indépendance de 10 et l'écart à l'indépendance de + 3. Il y a attraction entre le fait d'avoir un bac littéraire et le fait d'aller ensuite à l'Université.

Faisons subir une homothétie réductrice à ce tableau d'écarts à l'indépendance en divisant tous ses éléments par un facteur 10. On a le tableau suivant :

| *Série* | *Université* | *Classes prép.* | *Prof.* |
|---|---|---|---|
| Littéraire | 0,3 | 0,0 | – 0,3 |
| Éco. et soc. | 0,5 | – 0,1 | – 0,4 |
| Scientifique | 0,0 | 0,3 | – 0,3 |
| Tech. et pro. | – 0,8 | – 0,2 | 1,0 |

Nous pouvons constituer à partir de ces écarts un tableau hypothétique de données en ajoutant au tableau d'indépendance initial les écarts à l'indépendance divisés par 10. Le tableau résultant a mêmes marges que le précédent, même structure d'écarts à l'indépendance mais ses écarts ont subi une homothétie réductrice. Ce tableau hypothétique est le suivant :

| *Série* | *Université* | *Classes prép.* | *Prof.* | *Total* |
|---|---|---|---|---|
| Littéraire | 10,3 | 2,0 | 7,7 | 20 |
| Éco. et soc. | 15,5 | 2,9 | 11,6 | 30 |
| Scientifique | 10,0 | 2,3 | 7,7 | 20 |
| Tech. et pro. | 14,2 | 2,8 | 13,0 | 30 |
| *Total* | *50* | *10* | *40* | *100* |

Par exemple pour la case Littéraire-Université, le résultat 10,3 est la somme de l'indépendance 10 et de l'écart divisé par le coefficient réducteur : 3/10 = 0,3. Comme on peut le voir facilement, ce tableau hypothétique est très proche de l'indépendance.

Comparons maintenant les résultats de l'analyse des correspondances du tableau d'origine avec celle du tableau modifié.

*Données d'origine*

Le $\varphi^2$ est de: 0.249167
*Facteur 1*
Valeur propre = 0.199806
Pourcentage du total = 80.2
*Facteur 2*
Valeur propre = 0.049360
Pourcentage du total = 19.8
Coordonnées factorielles (F=) et contributions pour le facteur (CPF)

*Données modifiées*

Le $\varphi^2$est de : 0.002492
*Facteur 1*
Valeur propre = 0.001998
Pourcentage du total = 80.2
*Facteur 2*
Valeur propre = 0.000494
Pourcentage du total = 19.8
Coordonnées factorielles (F=) et contributions pour le facteur (CPF)

*Lignes du tableau*

| Ligne | F=1 | CPF | F=2 | CPF |
|---|---|---|---|---|
| Littéraire | 297 | 88 | 114 | 52 |
| Éco.&Soc. | 254 | 97 | 216 | 283 |
| Scientifique | 344 | 118 | – 404 | 661 |
| Technique | – 681 | 697 | – 22 | 3 |
| * | | 1 000 | | 1 000 |

*Modalités en colonne*

| Colonne | F=1 | CPF | F=2 | CPF |
|---|---|---|---|---|
| Université | 341 | 290 | 144 | 210 |
| Classe Prép | 478 | 115 | – 623 | 785 |
| Prof. | – 545 | 595 | – 24 | 5 |
| * | | 1 000 | | 1 000 |

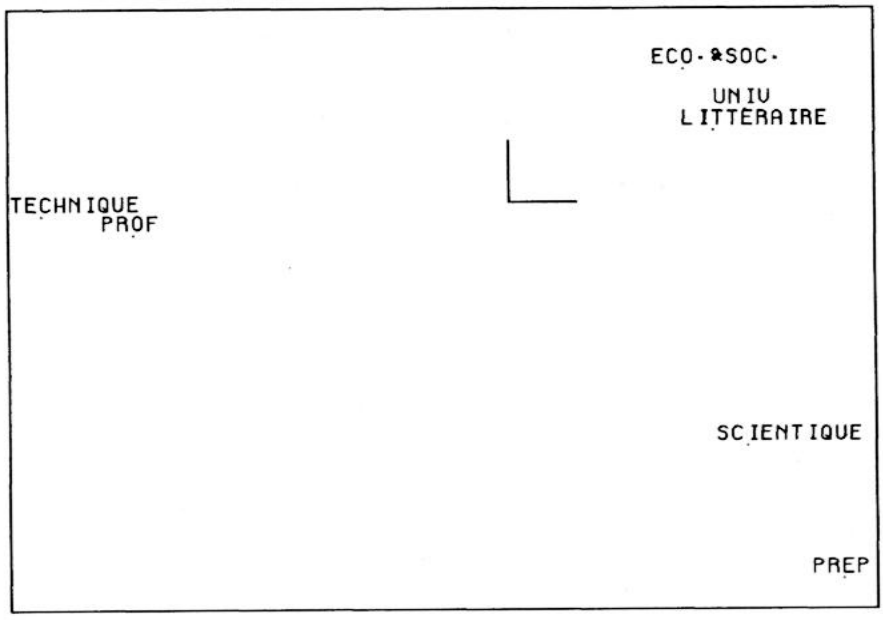
*Lignes du tableau*

| Ligne | F=1 | CPF | F=2 | CPF |
|---|---|---|---|---|
| Littéraire | 30 | 88 | 11 | 52 |
| Éco.&Soc. | 25 | 97 | 22 | 283 |
| Scientifique | 34 | 118 | – 40 | 661 |
| Technique | – 68 | 697 | – 2 | 3 |
| * | | 1 000 | | 1 000 |

*Modalités en colonne*

| Colonne | F=1 | CPF | F=2 | CPF |
|---|---|---|---|---|
| Université | 34 | 290 | 14 | 210 |
| Classe Prép | 48 | 115 | – 62 | 785 |
| Prof. | – 55 | 595 | – 2 | 5 |
| * | | 1 000 | | 1 000 |

Comme on peut le voir, les différences sont discrètes : tous les pourcentages d'explication et les contributions aux facteurs sont identiques, le $\varphi^2$ et les valeurs propres sont divisés par 100 et les coordonnées factorielles par 10. Comme bien souvent les programmes usuels cadrent leur graphique au maximum de la page, les deux graphiques factoriels des deux tableaux risquent d'être strictement identiques au graphique suivant :



On voit sur ce graphique que le premier facteur oppose le pôle formé par les séries techniques et les formations à finalités professionnelles, d'une part, aux deux autres pôles que sépare le 2$^{e}$ facteur à savoir les pôles des séries scientifiques, d'autre part, en conjonction avec les classes préparatoires aux grandes écoles et un pôle universitaire alimenté par les deux séries littéraires et sciences économiques et sociales.

Cette même interprétation vaut pour les deux tableaux qui ont même structure d'écarts à l'indépendance mais, pour le deuxième, les écarts sont négligeables et il est très dangereux de faire des commentaires sur des écarts aussi infimes à l'indépendance. Il y a là un risque grave de tirer des conclusions d'un graphique *qui ne fait qu'exprimer la structure des écarts, non leur intensité.* Car c'est là que se situe la leçon de cette expérience : deux structures homothétiques seront exprimées par des graphiques homothétiques en théorie et souvent identiques en pratique. Les contributions qui correspondent à des valeurs propres très faibles peuvent être d'un niveau tout à fait acceptable et, de ce fait, l'utilisateur peut prendre des risques de commentaires inconsidérés.

Pour éviter ces risques, il faut toujours regarder les valeurs propres elles-mêmes et le $\varphi^2$ dont elles sont la décomposition. Par exemple dans le premier tableau, le $\varphi^2$ est de 0,2492 ce qui correspond à un $\chi^2$ de 24,92 puisque l'effectif du tableau est de 100. Si les données étaient issues d'un échantillon, avec un degré de liberté de 6, le seuil du $\varphi^2$ est à 1 % de 16,8 et l'hypothèse d'indépendance peut être rejetée. Il n'en est pas de même avec le deuxième tableau où le $\varphi^2$ observé est lui aussi divisé par 100 et est donc égal à 0,168 : les données deviennent compatibles avec l'hypothèse d'indépendance. Les écarts à l'indépendance sont trop faibles pour pouvoir être interprétés.

On voit donc le danger que l'on court si l'on ignore cet effet d'homothétie et si l'on interprète des graphiques factoriels sans regarder de trop près les résultats. D'une manière pratique, le signal d'alarme est la présence de valeurs propres très faibles qui supposent un $\chi^2$ très faible lui aussi, et donc des données proches de l'indépendance. Il semble bien que certains soient tombés dans ce piège comme le manifeste l'exemple réel que nous allons maintenant traiter.

***Astrologie et statistique ou le zodiaque vu de Sirius***

Ce titre est celui d'un article [1] paru dans le *Journal de la Société de statistique de Paris*, et dont le but est résumé par l'auteur de la façon suivante : « L'esprit scientifique ne peut s'accommoder du rejet d'une hypothèse si tous les moyens d'investigation dont il dispose n'ont pas été épuisés. Cette règle vaut aussi pour l'astrologie. Rebelle à toute étude directe, elle ne peut échapper à la loi des grands nombres. ».

Le propos de l'auteur est résolument polémique : l'astrologie ne peut échapper à la statistique pas plus que le lapin ne peut échapper au fusil du chasseur. À cette fin, l'auteur, en utilisant les données du recensement de 1975, croise les signes astrologiques d'environ 68 000 couples non-agriculteurs. Son hypothèse est que si les signes ont une influence, cela devrait se voir sur les choix réciproques des conjoints et qu'avec des effectifs très importants il est possible de trancher.

L'auteur construit donc un tableau croisé de 12 lignes et 12 colonnes, les lignes correspondent aux signes astrologiques des hommes et les colonnes à ceux des femmes. On note à l'intersection d'une ligne et d'une colonne, le nombre de couples ayant des signes identiques. Dans son article, il ne nous donne pas les données d'origine mais les moyens de les reconstituer puisqu'on a un tableau donnant les écarts à l'indépendance et les effectifs des marges.

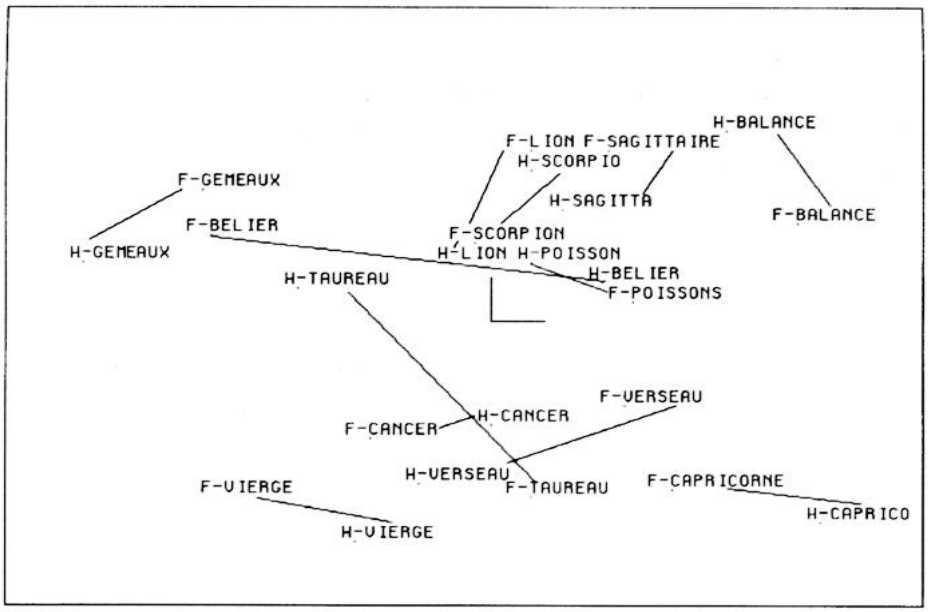
| *Femme* / *Homme* | *Vers* | *Pois* | *Bel* | *Taur* | *Gem* | *Canc* | *Lion* | *Vier* | *Bal* | *Scor* | *Sagi* | *Capr* | *Marge* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Verseau | 33 | – 36 | – 13 | – 4 | 20 | 7 | – 27 | 29 | – 3 | – 19 | – 9 | 24 | 5 848 |
| Poiss. | – 37 | 61 | – 12 | – 15 | – 3 | – 2 | – 10 | 0 | 6 | 30 | – 9 | – 11 | 6 034 |
| Bélier | 15 | 8 | 26 | 5 | – 25 | – 21 | 29 | – 30 | 8 | – 3 | – 20 | 8 | 6 280 |
| Taureau | – 14 | – 29 | 27 | 44 | 17 | 8 | 27 | – 7 | – 13 | – 3 | – 20 | – 39 | 6 108 |
| Gém. | – 11 | – 14 | 30 | – 15 | 69 | 6 | – 6 | 15 | – 45 | 4 | – 15 | – 17 | 5 809 |
| Cancer | 3 | 12 | 0 | 11 | – 15 | 29 | – 15 | 3 | – 2 | – 19 | – 3 | – 5 | 5 636 |
| Lion | – 22 | 12 | 17 | – 31 | – 9 | – 8 | 11 | 4 | – 3 | 29 | – 15 | 17 | 5 566 |
| Vierge | – 8 | – 1 | 10 | 14 | – 36 | 19 | – 19 | 52 | – 25 | – 8 | – 1 | 3 | 5 276 |
| Balance | 13 | 6 | – 22 | – 19 | – 5 | – 18 | 28 | – 20 | 37 | – 18 | 33 | – 14 | 5 553 |
| Scorpi. | – 5 | – 19 | – 12 | 5 | 10 | – 4 | 2 | – 16 | 7 | 9 | 43 | – 23 | 5 184 |
| Sagitt. | 11 | – 10 | 6 | – 40 | 4 | – 13 | – 12 | – 2 | 18 | – 5 | 31 | 13 | 5 170 |
| Capric. | 23 | 10 | – 56 | 45 | – 27 | – 4 | – 9 | – 26 | 15 | 3 | – 16 | 43 | 5 521 |
| *Marge* | *5 843* | *5 979* | *6 173* | *6 263* | *5 950* | *5 729* | *5 561* | *5 221* | *5 595* | *5 018* | *5 098* | *5 555* | *6 7985* |

1. AUBRY, B., 1978, p. 380.

Comme le $\chi^2$ de ce tableau est de 138,01 et que le seuil standard à 5 % avec 121 degrés de liberté est de 147, l'auteur conclut qu'on ne peut pas rejeter l'hypothèse d'indépendance et « que tout se passe comme si les caractères statistiquement observés se distribuaient en ignorant la composante astrale » [2].

Cependant l'auteur qui veut utiliser l'analyse des correspondances note que les résultats d'une analyse sur le même tableau « sont très différents de ceux obtenus sur une distribution-témoin composée d'une façon aléatoire sur les mêmes effectifs marginaux (c'est-à-dire en ajoutant aux valeurs croisées théoriques une composante aléatoire). Cette remarque, qui n'a pas d'incidence sur la conclusion générale de l'étude, a simplement pour but de montrer, sur un cas particulier, les limites de l'application d'un test statistique » [3].

Ce texte a une curieuse consonance : l'auteur a fait une analyse factorielle du tableau et, plutôt que d'en rendre compte, l'a comparée avec d'autres analyses qu'il a générées aléatoirement. Pour comprendre ce qui s'est passé, refaisons l'analyse factorielle dont voici le premier plan factoriel.




2. AUBRY, B., 1978, p. 384.
3. AUBRY, B., 1978, p. 386.

On comprend que l'auteur ait été surpris à la vision d'un tel plan et qu'il ne l'ait pas retenu puisqu'on y découvre des proximités angulaires (conjonctions), donc des attractions, entre tous les signes des hommes et ceux des femmes, sauf pour bélier et taureau. Un tel graphique semble donner raison à la croyance de l'influence des signes astrologiques sur le choix du conjoint. Est-ce bien la conclusion que l'on doit en tirer ?

En réalité nous sommes devant un cas d'effet d'homothétie comme nous le révèle la décroissance des premières valeurs propres :

| *Facteur* | *Valeur propre* | *% d'explication* |
|---|---|---|
| 1 | 0,00057 | 28 |
| 2 | 0,00040 | 17 |
| 3 | 0,00035 | 14 |
| 4 | 0,00018 | 9 |
| 5 | 0,00015 | 7 |
| -- | -- | -- |
| *Total* | *0,00203* | *100* |

En effet le $\chi^2$ du tableau divisé par l'effectif correspond à un $\varphi^2$ de 138,01/67985 = 0,00203. Bien que les taux d'explication des premiers facteurs correspondent à la normale, ce $\varphi^2$ extrêmement faible, décomposé en plusieurs facteurs, nous signale que nous travaillons sur un tableau très proche de l'indépendance. Cependant ces écarts à l'indépendance ont une structure particulière, comme on le voit dans le tableau ci-dessous où l'on n'a retenu que les écarts positifs égaux ou supérieurs à 9.

| *Femme* / *Homme* | *Vers* | *Pois* | *Bel* | *Taur* | *Gem* | *Canc* | *Lion* | *Vier* | *Bal* | *Scor* | *Sagi* | *Capr* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Verseau | ***33*** | | | | 20 | | | 29 | | | | 24 |
| Poisson | | ***61*** | | | | | | | | 30 | | |
| Bélier | 15 | | ***26*** | | | | 29 | | | | | |
| Taureau | | | 27 | ***44*** | 17 | | 27 | | | | | |
| Gémeau | | | 30 | | ***69*** | | | 15 | | | | |
| Cancer | | 12 | | 11 | | ***29*** | | | | | | |
| Lion | | 12 | 17 | | | | ***11*** | | | 29 | | 17 |
| Vierge | | | 10 | 14 | | 19 | | ***52*** | | | | |
| Balance | 13 | | | | | | 28 | | ***37*** | | 33 | |
| Scorpion | | | | | 10 | | | | | ***9*** | 43 | |
| Sagittaire | 11 | | | | | | | | 18 | | ***31*** | 13 |
| Capric | 23 | 10 | | 45 | | | | | 15 | | | ***43*** |

On constate, en effet, en premier lieu, que tous les effectifs diagonaux sont positifs ce qui explique les attractions notées sur le graphique ; mais aussi que ces écarts diagonaux sont loin d'être les seuls. Sur le graphique nous avons mis en relief les attractions entre mêmes signes par un artifice graphique, mais ce ne sont pas les seules conjonctions. On peut également en repérer de nombreuses autres comme en haut à droite entre homme scorpion et femme sagittaire (écart de 43), entre femme sagittaire et homme balance (33) ; à gauche entre homme gémeaux et femme bélier (30) ; en bas entre femme vierge et homme verseau (29) etc. En comptant exactement les cumuls d'écarts positifs on s'aperçoit que les 2/3 sont hors diagonale et 1/3 sur cette diagonale. Ce sont ces derniers qui apparaissent le plus immédiatement sur le graphique mais ils sont tout à fait comparables en intensité à ceux qui sont hors diagonale.

Si la diagonale est privilégiée par l'intérêt que nous lui portons et si cela explique notre perception du graphique, il n'empêche que tous les effectifs de la diagonale sont positifs bien qu'avec des valeurs de liaison faibles mais significatives. Par exemple le plus fort PEM (Pourcentage de l'écart maximum [4]) correspond à la case diagonale des gémeaux : l'écart à l'indépendance est le plus fort du tableau avec un effectif de 69 mais le PEM n'est que de 1,3 % ($\chi^2$ de 11,2 significatif à 1 % pour 1 DL).

Si les écarts positifs à l'indépendance, tous positifs, de la diagonale sont du même ordre de grandeur que les autres écarts positifs du tableau, il n'empêche que le fait qu'ils soient tous positifs peut difficilement être considéré comme un effet du hasard. Il faut tenter de rendre compte de ce phénomène et pour cela faire un peu d'astrologie.

Que dit en effet l'astrologie à propos des signes des conjoints ? Que les conjoints sont plutôt de même signe ? Les choses sont beaucoup plus compliquées car l'astrologie fait intervenir la position du ciel au moment de la naissance. Croire que l'astrologie dit que les gens de même signe sont faits pour s'entendre, c'est une croyance grossière en l'astrologie, celle qui existe lorsqu'on en ignore tout, celle que l'on a encore même quand on ne croit pas à l'astrologie. C'est le cas par exemple de l'auteur de l'article, qui lutte contre l'astrologie, mais qui pense que la croyance astrologique signifie le rapprochement des signes zodiacaux : c'est ce qu'il a cherché à quantifier. En ce sens l'auteur est un bon reflet de la croyance

4. *Cf.* P. Cibois, 1993.

populaire en l'astrologie, croyance suffisamment générale puisqu'elle est partagée même par ses adversaires, et qui peut suffire à expliquer les légers écarts tous positifs sur la diagonale. Comme l'auteur lui-même l'a perçu (p. 383-384), la croyance issue de l'astrologie que des gens de même signe sont plus aptes que les autres à s'unir, peut suffire, par un effet de prophétie auto-réalisatrice, à expliquer le phénomène qui nous occupe. L'effet est faible (PEM de l'ordre de 1 %) mais perceptible (significatif pour 7 signes sur 12) du fait des grands effectifs mis en cause.

## 2. L'effet de distinction

L'analyse des correspondances, du fait de la distance du $\varphi^2$ qu'elle utilise, pondère les petits effectifs et les prend ainsi en compte : c'est même là une de ses qualités reconnues. Cependant cette qualité peut se transformer en piège. Il suffit pour cela, que quelques modalités soient prises en même temps par un tout petit nombre d'individus et qu'ainsi ce regroupement apparaisse dans le premier facteur de l'analyse.

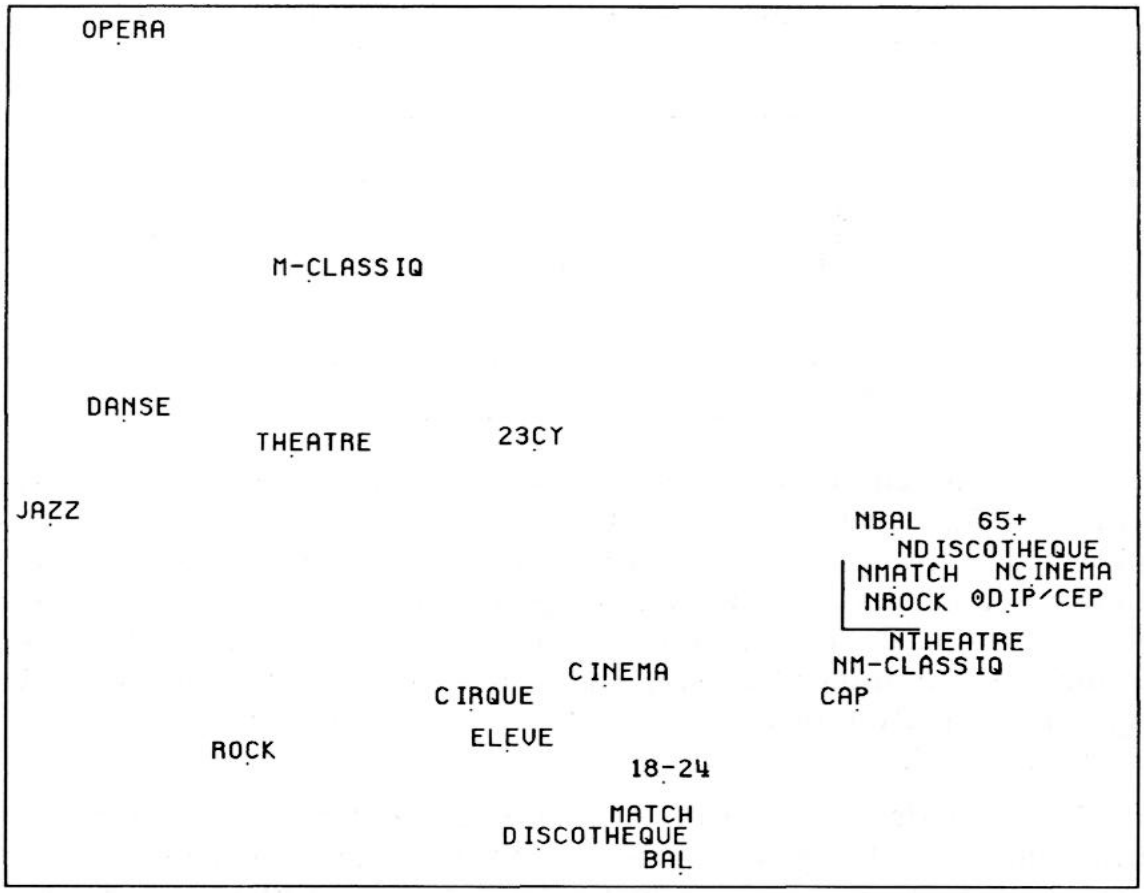
Pour montrer cet effet, nous allons immédiatement travailler sur des données réelles, celles de l'enquête sur les pratiques culturelles des Français de 1989 [5], dans laquelle on trouve un ensemble de questions portant sur les sorties suivantes effectuées ou non dans l'année précédente. Les sorties sont classées par ordre d'importance décroissante ; la population est celle des 4 722 adultes de l'enquête ; ce qui est pris en compte est le fait d'avoir effectué la sortie indiquée dans les 12 derniers mois.

| | | % |
|---|---|---|
| Cinéma | 2 106 | 44,6 |
| Bal | 1 131 | 24,0 |
| Discothèque ou boîte | 1 104 | 23,4 |
| Match | 1 014 | 21,5 |
| Théâtre | 621 | 13,2 |
| Concert musique classique | 457 | 9,7 |
| Concert de rock | 427 | 9,0 |
| Cirque | 378 | 8,0 |
| Spectacle de danse | 294 | 6,2 |
| Concert de jazz | 281 | 6,0 |
| Opéra | 156 | 3,3 |

5. DONNAT, O. & COGNEAU, D., 1990.

On a donc 11 activités (et les 11 non-activités aux effectifs complémentaires). On éclairera la compréhension du graphique en mettant en variables supplémentaires le sexe, l'âge et le niveau de diplôme du répondant.

Regardons le premier plan factoriel en prenant en compte toutes les contributions au premier ou au deuxième facteur, supérieures à 10 (la moyenne est de 1 000/22 modalités actives soit 45 pour mille).




Trois pôles apparaissent. En haut à gauche une « culture de sorties » [6] faite d'activités à forte charge culturelle : opéra, musique classique, spectacle de danse, théâtre qui sont pratiquées par un public diplômé de 2e ou 3e cycle universitaire (ou grandes écoles). On

6. L'expression est de Denis Cogneau (DONNAT, O. & COGNEAU, D., 1990, p. 173) qui examine les mêmes données avec plus de modalités, des recodages plus complexes, mais qui arrive à peu près au même premier facteur.

a en bas un pôle jeune avec le jazz qui fait la jonction entre le haut et le bas : cinéma, rock, match, discothèque ou bal. Enfin, on a droite un pôle d'absence de sorties, de « réclusion chez soi »[7] propre aux gens âgés et de faible niveau d'instruction.

Si on les prend pour des conjonctions, comme dans l'expression « culture de sorties », les rapprochements manifestés par le graphique sont fallacieux. Par exemple cette culture de sorties culturelles n'existe guère comme le manifeste les chiffres suivants qui donnent le nombre d'individus qui ont pratiqué simultanément dans la même année les 5 sorties (jazz, musique classique, spectacle de danse, opéra, théâtre) :

| | | % |
|---|---|---|
| 5 sorties | 12 | 0,3 |
| 4 sorties | 51 | 1,1 |
| 3 sorties | 114 | 2,4 |
| 2 sorties | 283 | 6,0 |
| 1 sortie | 637 | 13,5 |
| 0 sortie | 3 625 | 76,8 |
| *Total* | *4 722* | *100* |

Au vu de ces indications deux conclusions s'imposent : le pôle correspondant aux sorties culturelles est, d'une part, largement minoritaire et il est, d'autre part, constitué par des correspondances multiples à deux sorties qui sont de peu d'importance (6 %) et encore plus faibles au-delà (4 %). On voit sur cet exemple que des pratiques culturelles très minoritaires, et de ce fait distinctives, sont mises en avant par l'analyse des correspondances d'une manière très flatteuse par rapport à leur importance statistique.

Il ne faudrait pas croire que l'examen des contributions aux différents axes change quoi que ce soit à l'analyse précédente. Pour le premier facteur, du côté négatif, les cinq plus fortes contributions (supérieures à deux fois la moyenne) correspondent au théâtre, au jazz, au spectacle de danse, aux concerts de rock et de musique classique. Les autres contributions supérieures à la moyenne sont cinéma, discothèque et opéra. Du côté positif, une seule contribution est supérieure à la moyenne, c'est l'absence de cinéma. Pour le deuxième facteur, du côté positif, musique classique et opéra ont des contributions supérieures à deux fois la moyenne, le théâtre et le fait

7. DONNAT, O. & COGNEAU, D., 1990.

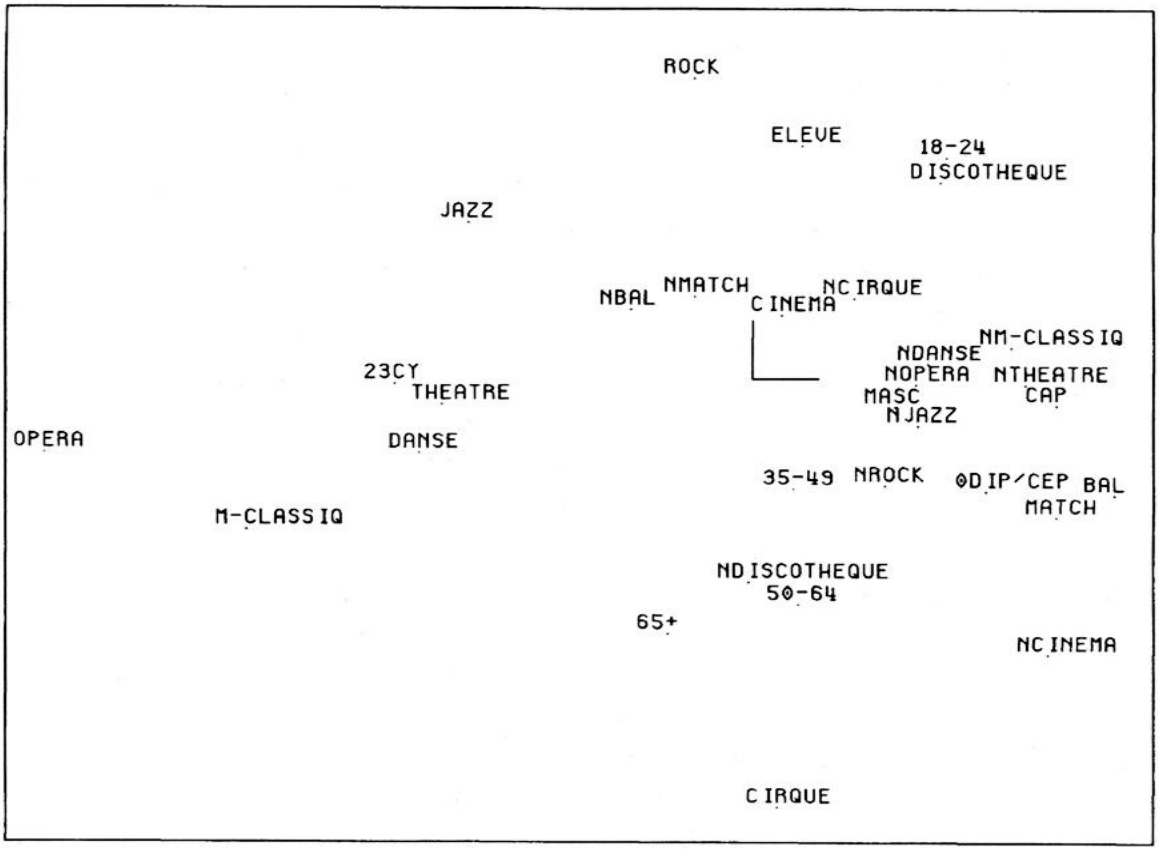
de ne pas aller au bal, des contributions supérieures à la moyenne. Du côté négatif, bal, discothèque et match sont les activités aux contributions supérieures à la moyenne (et même à deux fois la moyenne).

Le signal d'alarme de cet effet de distinction est ici la distance au centre. On sait que des points qui contribuent bien à un axe mais qui sont de faible effectif sont, de ce fait, éloignés du centre de gravité (effet de levier). En général quand un seul point est dans ce cas, le chercheur considère qu'il s'agit d'un phénomène perturbateur et le met en élément supplémentaire. Dans le cas de l'effet de distinction, les quelques modalités prises ensemble par quelques individus suffisent à faire croire à la réalité du facteur. Celui-ci correspond bien à un phénomène mais son importance numérique doit être vérifiée, même si son importance sociale ne fait pas de doute.

Enfin ce graphique est simplificateur pour une autre raison : c'est qu'on y a traité *présences* et *absences* de sorties comme des variables actives ce qui fait que le premier facteur oppose globalement les *présences* aux *absences*. À droite du plan on trouve toutes les absences liées au faible niveau social et aux âges élevés. Il y a là une exagération qui est le pendant des deux autres pôles puisque l'un était constitué par des sorties de niveau culturel élevé et l'autre par des sorties de jeunes. Le négatif de ces deux pôles engendre un pôle artificiel de non-sorties lié au faible niveau culturel et à l'âge élevé. Cependant il est assez facile de lutter contre cette simplification abusive, simplement en mettant en élément supplémentaire la non-sortie d'une manière tout à fait analogue à ce qu'on pratique souvent pour les non-réponses à des questions dans une enquête. Avec cette simple modification on a le premier plan qui figure à la page suivante.

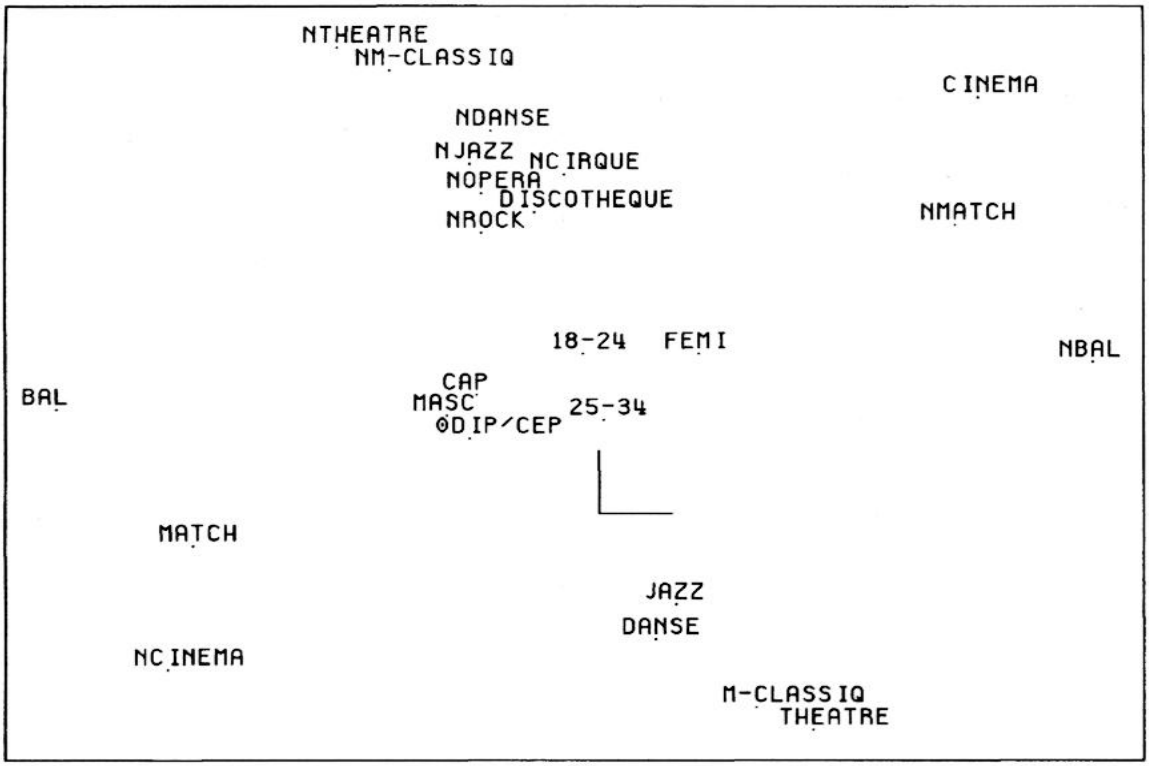
Dans ce plan, il n'y a plus opposition entre présences et absences. Les absences de sorties, traitées en éléments supplémentaires se mélangent avec les sorties. Par exemple l'ancien pôle d'exclusion se différencie : on distingue à droite une culture de niveau de diplôme peu élevé associant sortie au bal et au match, une culture correspondant à des âges croissants n'excluant pas toute sortie (cirque comme accompagnateur d'enfants), une culture jeune de boîte et de rock et, enfin, et toujours, le même pôle de sorties culturelles avec ses refus. Le cinéma enfin a une position centrale sur l'axe horizontal (de niveau social) mais légèrement du côté des plus jeunes sur l'axe vertical (d'âge).

La modification a évité les simplifications abusives du premier plan mais n'a pas modifié l'effet de distinction : comme l'éloignement des points nous signale une faiblesse numérique, il serait




possible de s'en tenir là. Si l'on tient à éviter de prendre des risques, une solution simple consiste à faire, sur le tableau des écarts à l'indépendance, une analyse en composantes principales. La distance euclidienne employée fera que l'éloignement du centre dépendra de l'importance numérique des effectifs des modalités. Comme on travaille sur le simple tableau des écarts à l'indépendance, on évite l'effet de taille lié au tableau brut. Les données des deux analyses (correspondances et composantes principales) sont directement comparables car elles ont le même nombre de facteurs.

On découvre alors un paysage légèrement différent (on a pris les mêmes critères d'apparition des points avec le même niveau de contribution que dans le graphique précédent). Les oppositions sont maintenant diagonales : en haut à droite le cinéma figure seul, pratique la plus massive de toutes, la plus éloignée du centre et proche des variables supplémentaires d'absence de match et de bal. On a le symétrique en bas à gauche. En haut, légèrement à gauche, la discothèque est entourée des non-pratiques culturelles. En symétrique



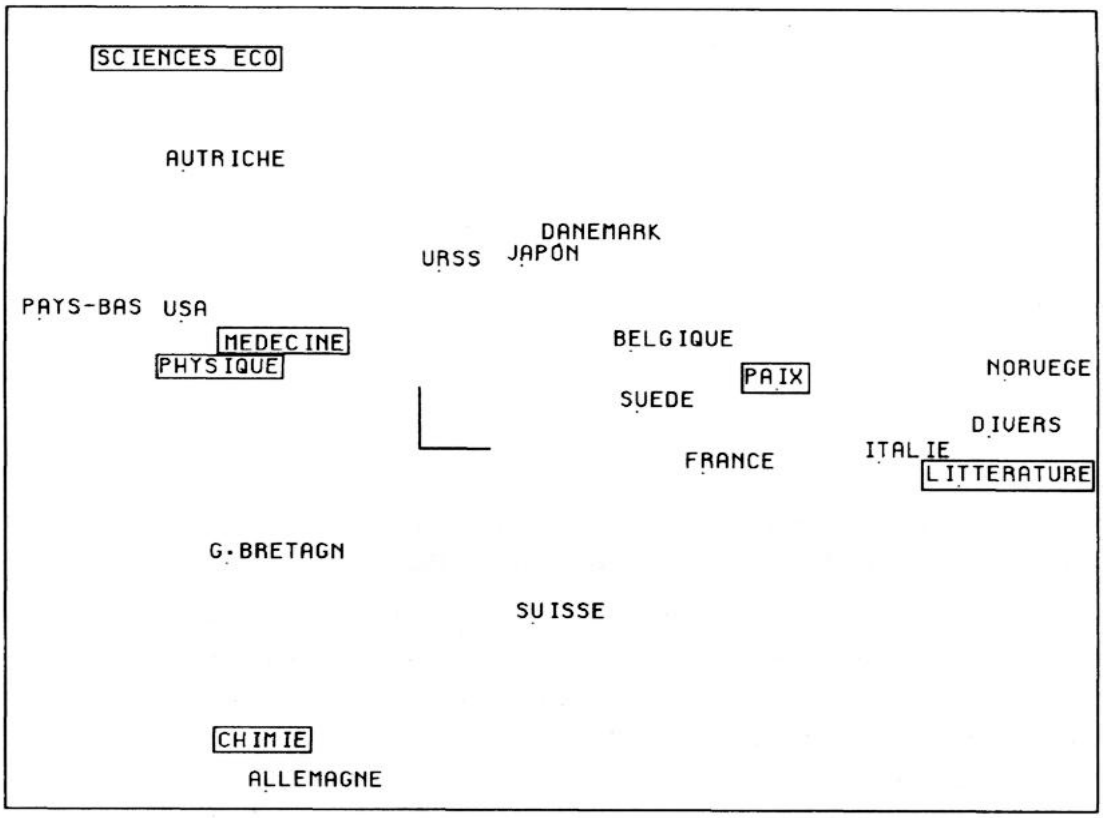

en bas à droite, les pratiques culturelles n'apparaissent que dans leurs éléments les plus importants numériquement.

On remarque que les variables supplémentaires de sexe, âge et diplôme sont beaucoup moins « explicatives » qu'en analyse des correspondances, c'est-à-dire qu'elles sont plus proches du centre. Ceci correspond davantage à la réalité en ce sens que les liens entre modalités actives (comportements) sont plus forts que les liens entre modalités actives et supplémentaires (entre comportements et variables de statut) mais les exagérations de l'analyse des correspondances sont bien appréciées par les chercheurs car elles visualisent des phénomènes discrets.

En conclusion que faut-il faire ? Il faut être alerté du risque d'effet de distinction aussitôt que plusieurs points bien éloignés du centre forment un facteur ou un pôle du graphique. Il suffit de revenir aux données par l'intermédiaire de quelques comptages ou tris croisés pour se rendre compte de l'importance numérique de l'effet observé. Il est souvent prudent de comparer dans ce cas avec les résultats d'une analyse en composantes principales sur les écarts à l'indépendance. Elle présente l'avantage d'être sensible à l'effet de taille et de montrer les associations des grandes masses de réponses.

## 3. Où il faut s'abstenir

Prenons une analyse factorielle qui est parue dans le journal *Le Monde* [8] et qui, de ce fait, n'est pas passée inaperçue. Elle a même été reprise dans des ouvrages de vulgarisation. Il s'agit de l'analyse d'un tableau répartissant les titulaires des diverses disciplines du prix Nobel selon leur nationalité. Le plan factoriel publié était le suivant :




Il était assorti de ce commentaire : « Un résultat frappant apparaît après un simple coup d'œil sur le graphe (et qui est d'ailleurs confirmé par l'analyse fine des calculs qui ont permis de dessiner celui-ci) : le dualisme sciences-littérature, que l'on aurait pu croire vérifié au seul niveau individuel, se manifeste clairement comme un phénomène de groupe. Ainsi, les pays situés à gauche de l'axe des ordonnées sont très nettement à dominante scientifique : les Pays-Bas, l'Autriche, les États-Unis, la Grande-Bretagne, l'Allemagne.

8. Doré, J.C. & Gordon, É., 1981.

D'autres (à droite sur le graphique), peuvent être qualifiés de 'littéraires' : l'Italie, la Norvège, les 'pays divers'. Quant à la France, la Belgique, au Danemark, à l'Union Soviétique, la Suède et la Suisse, ils présentent une double tendance ».

Pour voir ce qu'il en est, revenons aux données (issues du *Quid* de 1980) qui sont les suivantes :

| | *Sc.éco.* | *Physique* | *Chimie* | *Médecine* | *Paix* | *Littérature* |
|---|---|---|---|---|---|---|
| Pays-Bas | 1 | 5 | 2 | 2 | 1 | 0 |
| USA | 9 | 43 | 24 | 55 | 16 | 8 |
| Autriche | 1 | 3 | 1 | 4 | 2 | 0 |
| G.Bretagne | 2 | 20 | 21 | 19 | 7 | 6 |
| Allemagne | 0 | 14 | 24 | 11 | 4 | 7 |
| URSS | 1 | 7 | 1 | 2 | 1 | 4 |
| Japon | 0 | 3 | 0 | 0 | 1 | 1 |
| Suisse | 0 | 0 | 4 | 5 | 3 | 2 |
| Danemark | 0 | 3 | 0 | 4 | 1 | 3 |
| Belgique | 0 | 0 | 1 | 4 | 3 | 1 |
| Suède | 2 | 3 | 4 | 4 | 4 | 6 |
| France | 0 | 9 | 6 | 7 | 9 | 11 |
| Italie | 0 | 2 | 1 | 2 | 1 | 5 |
| Norvège | 1 | 0 | 1 | 0 | 2 | 3 |
| Divers | 0 | 5 | 6 | 10 | 20 | 21 |

Supposons un instant l'existence d'un pôle scientifique et d'un pôle littéraire et, pour vérifier la présence de ces pôles, regroupons les disciplines dans le sens proposé par les auteurs, c'est-à-dire en additionnant les résultats des Prix Nobel de la Paix et de Littérature d'une part (pôle « littéraire ») et les autres prix d'autre part (pôle « scientifique »). On a, en prenant les cinq premiers dans chaque catégorie, les résultats suivants :

| | *Sciences* | *Lettres* |
|---|---|---|
| USA | 131 | 24 |
| G. Bretagne | 62 | 13 |
| Allemagne | 49 | 11 |
| France | 22 | 20 |
| Suède | 13 | 10 |
| *Total* | *277* | *51* |

Les trois premiers pays « scientifiques » correspondent bien à ce qui était dit dans le commentaire : USA, Grande-Bretagne et Allemagne étaient bien dits à « dominante scientifique ». On trouve ensuite la France, mais on nous disait qu'elle présentait une « double tendance ». Ce qui ne va plus, c'est ce qui concerne les pays

« littéraires » : en effet, ce sont les mêmes cinq pays « scientifiques » qui sont aussi les premiers dans ce domaine. La seule différence est que la France passe du 4$^{e}$ au 2$^{e}$ rang. Ce sont les mêmes cinq pays qui ont le plus grand nombre de prix Nobel tant pour les prix « littéraires » que pour les prix « scientifiques » : tous ont cette « double tendance » qui n'était dans le commentaire attribuée qu'à la France et à la Suède.

Deux conclusions peuvent être tirées de cette expérience : tout d'abord que le Nobel est un club anglo-saxon où la France est bien admise (et où la présence de la Suède, à la fois juge et partie, pose quelques problèmes) ; ensuite, et c'est le point central ici, que la structure d'ordre du tableau, si importante dès qu'il y a compétition, n'est pas prise en compte par l'analyse des correspondances. La raison en est simple et tient au fait que ce sont les écarts à l'indépendance qui sont représentés par l'analyse des correspondances. Pour mieux le voir examinons le modèle ci-dessous où l'on suppose trois pays, A, B et C qui ont eu des prix Nobel de lettres et de sciences.

| | *Sciences* | *Lettres* | *Total* |
|---|---|---|---|
| A | 8 | 4 | 12 |
| B | 4 | 3 | 7 |
| C | 2 | 2 | 4 |
| *Total* | *14* | *9* | *23* |

Le pays A est le meilleur en sciences et en lettres, le pays B est le deuxième pour les deux disciplines. La décroissance des sciences, comme dans les données réelles, est plus rapide que celles des lettres et le nombre des prix y est plus grand.

Comme l'analyse des correspondances visualise les écarts à l'indépendance, il suffit de repérer ceux-ci sur ce tableau, par exemple en comparant les pourcentages en ligne avec le pourcentage toutes lignes confondues.

| | *Sciences* | *Lettres* | *Total* |
|---|---|---|---|
| A | 66,7 | 33,3 | 100 |
| B | 57,1 | 42,9 | 100 |
| C | 50,0 | 50,0 | 100 |
| *Total* | *60,9* | *39,1* | *100* |

Seul le pays A est en écart positif pour les sciences et en écart négatif pour les lettres. Sur le plan factoriel, il sera en conjonction avec les sciences mais en opposition avec les lettres, bien qu'il soit là aussi le premier, car la structure des écarts à l'indépendance manifeste simplement les différences entre les deux distributions. Ce qui n'est pas pertinent a été visualisé ; la structure d'ordre, qui est pertinente, ne l'a pas été. La conclusion est simplement qu'il ne fallait pas se servir de l'analyse des correspondances puisqu'on ne s'intéresse pas aux écarts à l'indépendance. Toute série de nombres, présentée sous forme de tableau n'est pas susceptible d'être traitée par l'analyse des correspondances.

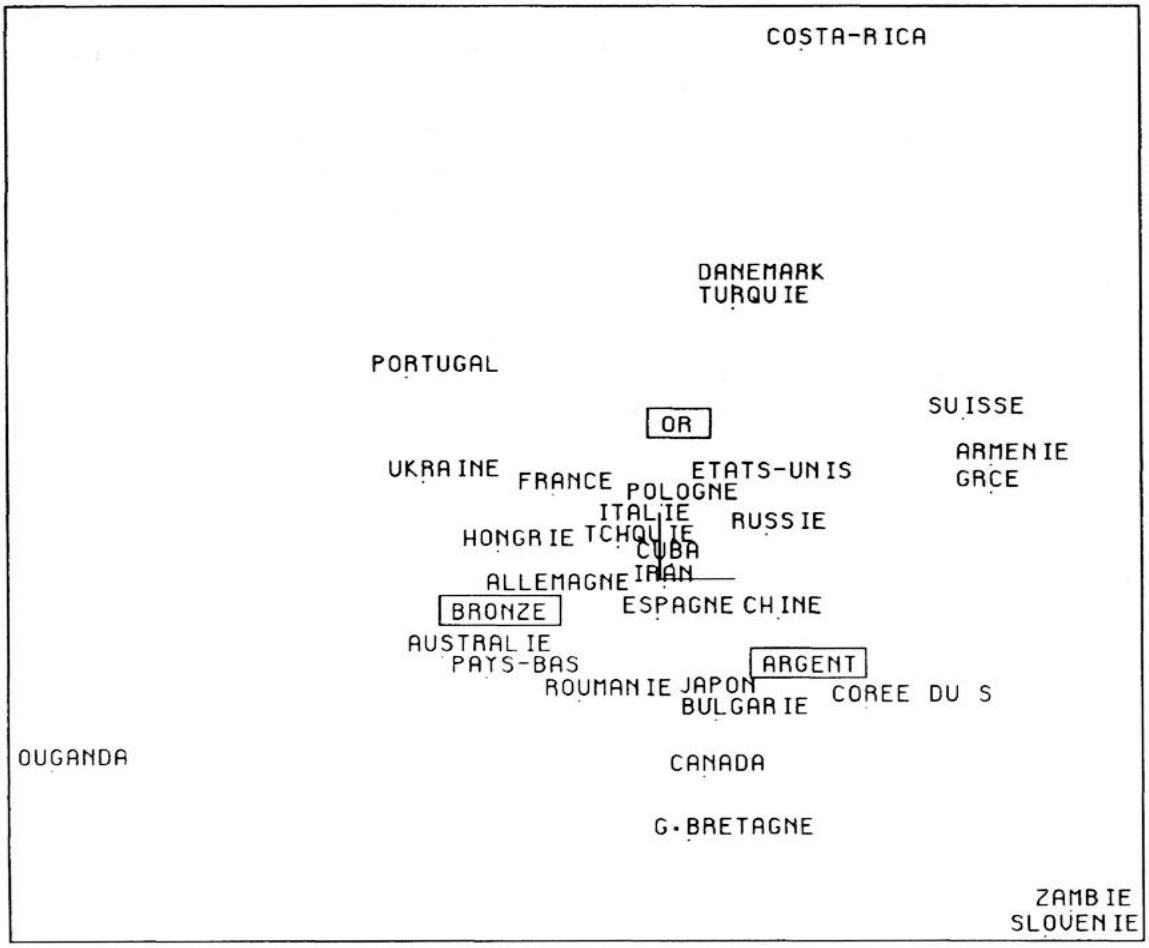
Pour montrer le ridicule de cette utilisation, poussons le raisonnement à son comble et traitons un tableau analogue mais plus grand, celui des résultats à des Jeux olympiques où l'on donne le nombre de médailles de chaque catégorie (or, argent, bronze) par pays participant.

Soit un extrait d'un tel tableau pour les J.O. d'Atlanta (1996) :

| | *Or* | *Argent* | *Bronze* | | *Or* | *Argent* | *Bronze* |
|---|---|---|---|---|---|---|---|
| États-Unis | 44 | 32 | 25 | Grèce | 4 | 4 | 0 |
| Russie | 26 | 21 | 16 | Tchéquie | 4 | 3 | 4 |
| Allemagne | 20 | 18 | 27 | Suisse | 4 | 3 | 0 |
| Chine | 16 | 22 | 12 | Danemark | 4 | 1 | 1 |
| France | 15 | 7 | 15 | Turquie | 4 | 1 | 1 |
| Italie | 13 | 10 | 12 | Canada | 3 | 11 | 8 |
| Australie | 9 | 9 | 23 | Bulgarie | 3 | 7 | 5 |
| Cuba | 9 | 8 | 8 | Japon | 3 | 6 | 5 |
| Ukraine | 9 | 2 | 12 | G.-Bretagne | 1 | 8 | 6 |
| Corée du Sud | 7 | 15 | 5 | Iran | 1 | 1 | 1 |
| Pologne | 7 | 5 | 5 | Arménie | 1 | 1 | 0 |
| Hongrie | 7 | 4 | 10 | Portugal | 1 | 0 | 1 |
| Espagne | 5 | 6 | 6 | Costa-Rica | 1 | 0 | 0 |
| Roumanie | 4 | 7 | 9 | Slovénie | 0 | 2 | 0 |
| Pays-Bas | 4 | 5 | 10 | Zambie | 0 | 1 | 0 |
| | | | | Ouganda | 0 | 0 | 1 |

Si l'on fait l'analyse de ce tableau on a le plan qui figure page suivante (il n'y a que deux facteurs puisqu'il n'y a que trois colonnes dans le tableau).

Le premier facteur oppose bronze et argent car il se trouve que du fait des *ex aequo* le nombre de médailles de bronze (298) est plus grand que celui des médailles d'argent (273), lui-même plus grand que celui des médailles d'or (271). Les deux pays les plus au centre sont l'Iran et Cuba puisque l'Iran n'a aucun écart à l'indépendance : il a seulement une médaille de chaque catégorie et puisque Cuba, au




8e rang mondial, est presque dans la même situation, avec 9 médailles d'or, 8 d'argent et 8 de bronze. Ces deux pays ont un tiers de leur(s) médaille(s) dans chaque catégorie, comme la moyenne et n'ont donc pas d'écart à l'indépendance. Cette similitude de position montre bien que la structure des écarts à l'indépendance, qui rapproche deux pays très différents, n'est pas l'élément pertinent du tableau.

Le premier pays du tableau, les États-Unis, n'est pas celui qui est le plus en attraction avec l'or puisque c'est le Costa-Rica : en effet ce pays a 100 % de médailles d'or (mais il n'en a qu'une) tandis que les États-Unis n'en ont que 44 pour 101 médailles au total. Ils sont en cela proches de la Pologne qui en a 7 sur 17, soit aussi une valeur proche de 40 %. Les pays extrêmes sont ceux qui n'ont qu'une catégorie de médailles ; davantage vers l'intérieur se trouvent ceux qui ont un couple de médailles comme l'Arménie (une d'or, une d'argent) ou la Grèce (4 et 4). Vers le centre se trouvent les pays dont les médailles sont réparties de manière plus ou moins équilibrée.

Ces quelques exemples montrent que la logique de la représentation des écarts à l'indépendance n'est pas ce qui est intéressant dans un tel cas de figure. C'est dans le tableau que la structure d'ordre est correctement exprimée, là où les pays sont classés d'abord par rapport aux nombres de médailles d'or, puis à égalité au nombre de médailles d'argent, puis de bronze. On démontre par l'absurde que dans le cas où la structure d'ordre est pertinente, il faut s'abstenir de faire une analyse des correspondances qui, représentant la structure des écarts à l'indépendance, est totalement inadéquate.

En analyse des correspondances, « tout est permis, mais tout n'est pas profitable [9] ». On peut évidemment soumettre à un programme tout tableau de chiffres et il en sortira toujours quelque chose. Le piège ici est celui que Lazarsfeld soulignait déjà à propos des méthodes quantitatives en général en disant que leur usage ressemblait parfois à celui qu'un jeune enfant fait d'un marteau que l'on vient de lui offrir : tout dans ce cas devient digne de recevoir un coup. Il est des cas où il vaut mieux s'abstenir d'utiliser l'analyse des correspondances comme il est d'autres cas où il faut vérifier que l'on n'est pas tombé dans le piège de l'effet d'homothétie ou de l'effet de distinction.

## BIBLIOGRAPHIE

AUBRY, Bernard, « Astrologie et statistique ou le zodiaque vu de Sirius », *Journal de la société de statistique de Paris*, 1978, 119 (4), pp. 380-386.

CIBOIS, Philippe, « Le PEM, pourcentage de l'écart maximum : un indice de liaison entre modalités d'un tableau de contingence », Bulletin de Méthodologie Sociologique, 1993, n° 40, pp. 43-60.

DONNAT, Olivier et COGNEAU, Denis, *Les pratiques culturelles des français 1973-1989*, Paris, La découverte / La documentation française, 1990.

DORÉ, Jean-Christophe et GORDON, Elisabeth, « Nobélisés et Nobélisables », *Le Monde*, 14 octobre 1981.

9. 1 Corinthiens 10/23.