LES CAHIERS DE L'ANALYSE DES DONNÉES

T. K. GOPALAN
SAGOMBAYE NODJIRAM

Sur l'application des méthodes multidimensionnelles à une anthologie de données. (4) Analyse des proximités

Les cahiers de l'analyse des données, tome 18, n° 4 (1993), p. 469-476

<http://www.numdam.org/item?id=CAD_1993__18_4_469_0>

Les Cahiers de l'Analyse des Données
Vol. XVIII 1993 n°4 pp.469 476.

SUR L'APPLICATION DES MÉTHODES MULTIDIMENSIONNELLES À UNE ANTHOLOGIE DE DONNÉES

(4) ANALYSE DES PROXIMITÉS

[MÉTH. ANTH. DONNÉES (4)]

T. K. GOPALAN
SAGOMBAYE NODJIRAM

4 Représentation d'un ensemble muni de distances

Sous ce titre, nous considérons deux exemples proposés par Br. F.J. MANLY. Seul de deuxième de ceux-ci présente des données qui ne se mettent pas directement sous la forme d'un tableau de correspondance. À propos de cet exemple, nous rappellerons sommairement des principes de la représentation d'un ensemble muni de distances.

4.1 Concordances entre les votes de 15 parlementaires

H.C. ROMESBURG: *Cluster Analysis for Researchers*; Lifetime Learning Publications; Belmont, California; (1984).

Les données traduisent les attitudes de votes d'un ensemble I de 15 représentants de l'état de New Jersey, au Congrès des États Unis, dans 19 scrutins relatifs à l'environnement. La matrice publiée est celle des discordances de vote; avec k(i, i') = nombre de scrutins où les parlementaires i et i' n'ont pas eu la même attitude; (e.g.: i s'est abstenu et i' a voté contre).

Br. F.J. MANLY entreprend d'analyser cette matrice symétrique comme s'il s'agissait d'une matrice de distances. Il cherche d'abord une représentation de I, dans un espace euclidien de dimension finie aussi faible que possible, respectant bien les distances données elles-mêmes. Puis préfère imposer des conditions moins strictes: respecter les inégalités entre distances. Ayant

considéré trois représentations distinctes obtenues, suivant ce dernier critère, dans des espaces de dimension 2, 3 et 4, il s'arrête à la représentation de dimension 3, qui est seule publiée.

D'après l'expérience acquise depuis l'analyse, par A. HAMROUNI, des scrutins à l'Assemblée des Nations Unies, (*CAD*, Vol.I, pp.161-195 et 259-286; 1976), nous estimons qu'il convient de soumettre à l'analyse de correspondance le tableau complet C × A décrit ci-après.

C : ensemble de tous les membres du Congrès (plutôt que les seuls représentants de la N.J.);

A : ensemble des 57=3 × 19, attitudes de vote possibles dans les scrutins relatifs au problème auquel on s'intéresse; soit {Oui, Non, Abs} à chaque scrutin;

k(c, a) = 1 si le sujet c a adopté l'attitude a, et zéro sinon.

Par l'analyse factorielle, en effet, un tel tableau peut être rapidement traité de façon univoque.

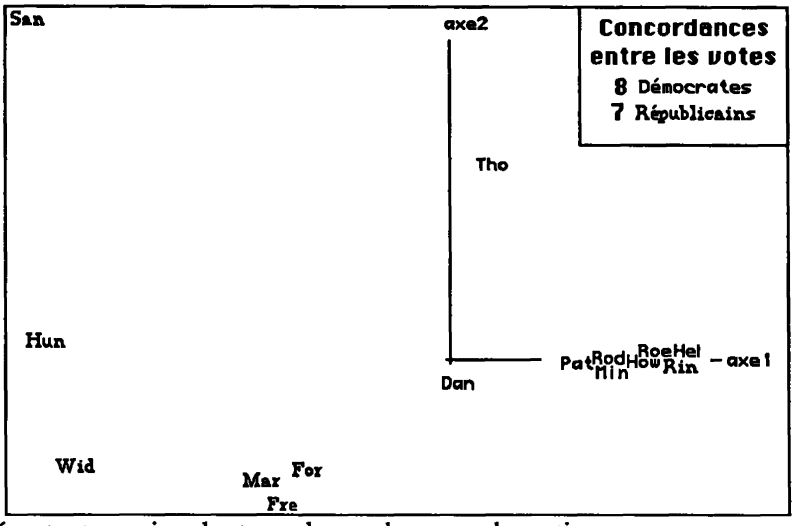
Mais s'il faut nous contenter des données publiées, nous préférons les mettre sous la forme, complémentaire, d'une matrice de concordance; avec k(i,i') = nombre des scrutins, parmi les 19 considérés, où i et i' ont adopté la même attitude; (la diagonale étant ainsi mise à 19). En effet, la concordance

```
parlementaires de la Nouvelle Jersey
trace   :   2.165e-1
rang    :      1     2     3     4     5     6     7     8     9    10
lambda  :   1435   352   185    83    40    31    17    10     6     4 e-4
taux    :   6629  1624   854   382   186   143    78    48    29    19 e-4
cumul   :   6629  8253  9107  9489  9675  9818  9896  9944  9973  9992 e-4
```

représente une sorte de correspondance.

Le plan (1,2) issu de cette matrice ne diffère pas sensiblement de celui publié dans le *Primer*. Comme l'a noté Br. F.J. M., l'axe 1 rend compte de l'opposition entre les deux partis (distingués, sur notre graphique par la typographie des abréviations des noms): presque tous les Démocrates (D) sont étroitement groupés; les Républicains (R) sont plus dispersés, l'un d'eux, Rin, rejoint même l'autre parti; mais à cette exception près, les deux partis se séparent suivant l'axe 1.

Sur l'axe 2, se détachent San (R) et Tho (D): Br. F.J. M. note que ces deux



représentants se signalent par de nombreuses abstentions.

Notre facteur 3 ne concorde pas avec celui publié dans le Primer; mais le facteur n'étant pas interprété, il n'y a rien à conclure de cette différence.

L'analyse de la correspondance C × A, par la représentation simultanée des votants et de leurs attitudes, servirait l'interprétation des facteurs; la CAH étant un recours dans le cas, improbable ici, où il y aurait de multiples dimensions.

4.2 Représentation des distances routières entre 13 villes de l'île du Sud de la Nouvelle Zélande

4.2.0 Des données au problème

L'auteur du *Primer* a lui-même compté, d'après une carte, les distances routières entre les 13 villes d'un ensemble I. Du fait des particularités du relief et des contingences du développement économique, ces distances ne sont pas exactement proportionnelles aux distances mesurées à vol d'oiseau. On a donc, *a priori*, un système de distances qui ne peut être représenté par les distances euclidiennes sur une image plane de I, même distincte de celle de la carte. Le problème est donc posé d'obtenir, par une telle image, une représentation approchée aussi fidèle que possible.

4.2.1 Méthodes d'analyse des distances

Sous le titre de *Multidimensional Scaling*, est publiée, en langue anglaise,

une abondante littérature consacrée à la représentation des distances. Dans le Traité de *L'Analyse des Données*, la leçon: 'Représentation euclidienne d'un ensemble fini muni de masses et distances', [Repr. Eucl.], T.IIB, n°2, traite de ce problème.

Comme le précise cette leçon, d'un ensemble I de n points muni de distances, il existe une représentation unique dans un espace muni d'une forme quadratique de distance. Dans le cas général, l'espace support du nuage est de dimension n-1. La représentation n'est proprement euclidienne que si la forme quadratique de distance est définie positive; ce qui n'est réalisé que sous certaines conditions plus restrictives que la simple inégalité du triangle entre les distances.

Pratiquement, on s'intéresse à des représentations euclidiennes approchées de dimension aussi faible que possible. Si l'on attribue aux points de I des masses, (éventuellement, la même masse à tous les points,) on peut, par simple diagonalisation d'une matrice carrée symétrique, déterminer les axes principaux d'inertie de I; et représenter I dans l'espace engendré par les premiers axes; le nombre des axes retenus étant fixé selon les principes usuels de l'analyse factorielle.

L'ajustement des distances de la figure ainsi obtenue à celles imposées initialement peut être amélioré par une méthode itérative.

Beaucoup d'auteurs ne recourent pas à l'analyse de l'inertie: leurs algorithmes itératifs sont relativement coûteux; et on ne les emploie communément que pour $n<50$.

Dans la plupart des applications, à la différence du problème routier considéré ici, les distances elles-mêmes n'ont pas de valeurs qui s'imposent: elles représentent seulement, en termes numériques, une certaine notion de proximité; qui peut être adéquatement exprimée par une ordonnance, i.e., un système complet d'inégalités entre les distances.

Ceci a suggéré à R.N. SHEPARD une intéressante découverte géométrique: pourvu que le cardinal n de I soit assez élevé relativement à la dimension p de l'espace ambiant, supposé euclidien, la figure I peut (à une similitude près) être reconstruite, avec une approximation satisfaisante, d'après la seule connaissance de l'ordonnance.

On a pu donner aux conditions d'approximation une forme assez précise, et démontrer que l'ensemble des distances (au carré) entre points de I peut être assimilé à un échantillon d'effectif n.(n-1) issu d'une loi de chi2 à p dimensions. Ainsi, l'ordonnance est traduite en un système de distances auquel s'applique l'algorithme, rapide, de la recherche des axes principaux d'inertie. Une construction au compas étant possible si p=2 et n≈10.

Tout système de distances peut même, utilement, être réduit à son ordonnance, d'après laquelle, par la loi du χ^2, on calcule un nouveau système de distances, plus approprié à la recherche d'une représentation euclidienne.

4.2.2 Essai d'analyse de correspondance

Br. F.J. MANLY applique aux données routières un algorithme itératif sans recours à une diagonalisation. Il obtient, en dimension 2, une représentation satisfaisante; en ce que, d'une part, la constellation des villes y est semblable à celle de la carte; et que, d'autre part, les distances imposées sont assez bien respectées.

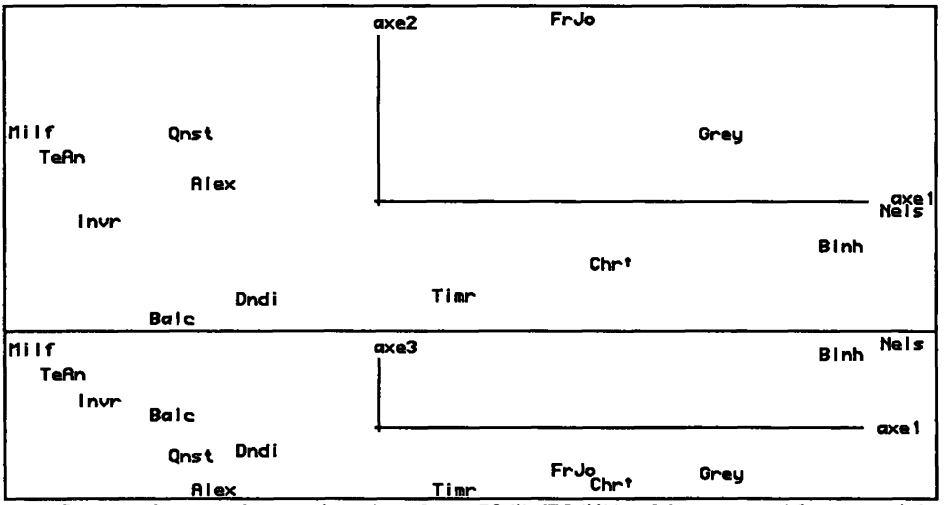
Nous ne reprendrons pas ces données avec les diverses méthodes présentées au §4.2.1, mais rendrons compte d'un essai, suggéré par l'analyse de correspondance du §4.1.

Partons d'une matrice 13 × 13 dont toutes les cases ont la même valeur, par exemple 10000. En retranchant de cette matrice la matrice des distances données par le *Primer* (distances en miles, allant de 50 à 756; et généralement ≈300), on obtient une matrice de proximité, analogue à la matrice de concordance du §4.1.

```
Nouvelle Zélande : Les routes de l'île du Sud
trace  :   2.822e-4
rang   :      1     2     3     4     5     6     7     8     9    10
lambda :   2450   209    91    34    14     8     6     3     2     2 e-7
taux   :   8683   742   324   119    49    30    22    12     7     6 e-4
cumul  :   8683  9425  9749  9868  9918  9947  9969  9981  9989  9995 e-4
```

L'analyse de cette matrice de proximité fournit un plan (1,2) très semblable au graphique plan publié dans le Primer: seuls sont moins fidèles les rapports de distances entre villes proches; ce qui s'explique par le fait qu'on n'a eu recours à aucune itération pour améliorer l'ajustement.

L'axe 3, seul non négligeable après l'axe 2, ne peut modifier sensiblement la représentation des distances; parce que, dans le plan (1,3) les points dessinent une courbe; ce qui implique qu'en prenant en compte F3, on ne



modifie pas la représentation locale: (F3(i)-F3(i')) n'étant notable que si la différence (F1(i)-F1(i') est elle-même forte.

Cette remarque montre que l'analyse a bien reconnu la dimension de la configuration; ce qui, en *Multidimensional Scaling*, se fait généralement par tâtonnement.

4.2.3 Analyse de correspondance et matrice de proximité

A posteriori, il faut expliquer le succès de l'analyse de la matrice de proximité.

Le nombre 10000 a été choisi pour être grand vis-à-vis des distances. Pour la clarté des développements mathématiques, on peut se rapporter à la matrice divisée par 10000, où toutes les cases sont de la forme:

$$k(i, i') = 1 - d(i, i').e\text{-}4 \quad ;$$

et poser désormais:

$$k(i, i') = 1 - \Delta(i, i') \quad ;$$

avec une matrice de distance Δ, infinitésimale.

Les lignes et colonnes de la matrice de proximité ainsi construite ont toutes un total qui, en valeur relative, ne diffère de n=cardI, que par un terme d'ordre 1 en Δ; à cette approximation, la distance distributionnelle, au carré,

entre i et i' peut s'écrire comme une somme indicée par $j \in I$.

De façon précise, notons $\Delta ii'$ pour $\Delta(i, i')$ et,

$\Delta i = (1/n) \sum\{\Delta ij \mid j \in I\}$;

on a pour la distance distributionnelle:

$$d^2(i, i') = \sum\{((k(i,j)/k(i))-(k(i',j)/k(i')))^2 . (k/(k(j)) \mid j \in I\}$$

$$\approx (1/n) . \sum\{((1-\Delta ij)/(1-\Delta i))-(1-\Delta i'j)/(1-\Delta i')))^2 \mid j \in I\}$$

$$\approx (1/n) . \sum\{((\Delta ij-\Delta i'j)-(\Delta i-\Delta i'))^2 \mid j \in I\};$$

cette expression n'est autre que la variance, pour j in I, de la différence (Δij-$\Delta i'j$) des distances de j à i et à i'. La différence varie entre $-\Delta ii'$ et $+\Delta ii'$, et son écart type $\approx d(i, i')$ est de l'ordre de $\Delta ii'$.

Il semble même que les écarts de proportionnalité entre d(i,i') et $\Delta ii'$ soient favorables à la représentation de I: en effet, si I est un nuage à peu près confiné à un plan P, pour (i, i') transverse à ce plan, (Δij-$\Delta i'j$) a une variance très faible; et, en général, la composante de (i, i') transverse à P, compte peu dans la différence (Δij-$\Delta i'j$). Ainsi le passage de Δ à d réduit les dimensions transverses relativement aux dimensions principales.

Si, dans la matrice de proximité, on désire prendre en compte des masses μ_i, il suffit de poser:

$$k(i, i') = \mu_i . \mu_{i'} (1 - \Delta(i, i'))$$;

et les considérations précédentes se généralisent intégralement.

Références bibliographiques

A) Sur l'analyse des attitudes de vote

A. HAMROUNI: "Les scrutins de 1967 à l'Assemblée des Nations Unies; 1° partie: Les analyses factorielles; 2° partie: Classification automatique et analyses complémentaires."; in *CAD*, Vol.I, pp.161-196 et 259-286; (1976).

(Un extrait de ce travail est repris dans Pra1:CORR; IV §8; et traduit dans *Correspondence Analysis Handbook*; M. Dekker ed.)

FEHRI ALEYA: *Analyse des votes des pays africains à l'O.N.U., entre les années 1976 et 1984*; Thèse; Université Pierre et Marie Curie; Paris; (1987).

B) Sur l'analyse des proximités

Roger N. SHEPARD: "The analysis of Proximities: Scaling with an unknown distance function, I & II"; in *Psychometrika*; Vol.27, nos 2 & 3; (1962).

J.-P. BENZÉCRI: "Analyse factorielle des proximités"; in *Publications de l'Institut de Statistique de l'Université de Paris*; 1964-65.

ejusdem: "Statistical Analyis as a tool to make patterns emerge from data"; in *Methodologies of Pattern Recognition*; ed. WATANABE; Academic Press; N.-Y.; (1969). en français dans:

ejusdem: "Les méthodes de l'analyse des données"; [Honolulu], TIAn°2; in *L'Analyse des Données*, (I) : La Taxinomie; Dunod; Paris; (1973).

ejusdem: "Représentation euclidienne d'un ensemble fini muni de masses et distances": [Repr. Eucl.], T.IIBn°2; in *L'Analyse des Données*, (II) : L'Analyse des correspondances; Dunod; Paris; (1973).