

CAD

**Mémoire reçu : Statistical patterns in
Hebrew and Arabic roots**

Les cahiers de l'analyse des données, tome 11, n° 4 (1986),
p. 497-500

http://www.numdam.org/item?id=CAD_1986__11_4_497_0

© Les cahiers de l'analyse des données, Dunod, 1986, tous droits réservés.

L'accès aux archives de la revue « Les cahiers de l'analyse des données » implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
<http://www.numdam.org/>

MÉMOIRE REÇU : STATISTICAL PATTERNS IN HEBREW
AND ARABIC ROOTS
(M. WEITZMAN - UNIVERSITY COL. LONDON)

par J.P. Benzécri

Mémoire reçu : Statistical Patterns in Hebrew and Arabic Roots,
par Michael Weitzman (University Col. London).

Faisant écho, dans la lettre d'envoi de cet intéressant mémoire, à d'amicales discussions dont le colloque de Louvain-la-Neuve a été le cadre (cf. [TEXTE BIBLE] in CAD Vol XI n° 1 pp. 7-30 1986), M.W. écrit:

"I see that you think factor analysis the only useful approach",
et s'enquiert de ce que je pense de ses :

"own proposals based on multidimensional scaling". En rendant compte du mémoire, nous nous appliquerons principalement à préciser similitudes et différences entre plusieurs méthodes d'analyse multidimensionnelle.

1 Rappel de la structure de la racine sémitique : On sait que dans l'état actuel de ces langues, la grande majorité des formes nominales ou verbales de l'arabe ou de l'hébreu apparaissent formées suivant des règles morphologiques strictes et bien inventoriées à partir de racines comprenant trois lettres. Il n'est toutefois pas exact que tout triplet de lettres constitue une racine ; ce qui pose la question de caractériser en termes généraux les triplets figurant dans les lexiques vis à vis de ceux non attestés. Dans la mesure où l'ensemble des racines attestées est contingent où les emprunts et mots peu usités ont des places incertaines, où les tendances phonétiques varient au cours de la très longue histoire des langues, ... on conçoit que la caractérisation cherchée ait, en partie au moins, forme statistique.

La première référence statistique citée par M.W. est Greenberg, (in *Word* 6 (1950) pp. 162-181) ; puis vient J. Kurytowicz (*Studies in Semitic Grammar and Metrics*, 1973) ; et pour l'hébreu seulement K. Koskinen (*Zeitschr. D. Morg. Gesell.* ; 114 ; (1964) pp.16-58).

Nous nous bornerons à considérer ce que M.W. dit de l'hébreu (bien que nous disposions également d'analyses de corr. pour l'arabe dans la thèse de Ibrahim Hamouda Kamal ad dîn),

2 Affinités entre lettres et place dans la racine : M.W. emprunte d'abord à K. Koskinen, un tableau à 3 lignes donnant pour 1099 racines l'inventaire des consonnes se trouvant respectivement en position I, II et III : e.g. 42 racines ont pour deuxième consonne un aleph ... Il faut noter que dans ce tableau manquent les lettres faibles ou semi-voyelles yod et vav ; ce qui pose la question du choix des racines retenues : Koskinen, prend à la fois racines verbales et nominales ; G.E. Weil (cf.

.... G.E. Weil (cf.

rajouter : Trilitéralité fonctionnelle ou bilitéralité fondamentale des racines hébraïques. Un essai d'analyse quantifiée. PUF - PARIS). prend exclusivement

() Professeur de statistique. Université Pierre & Marie Curie.

prend exclusivement les racines verbales de l'hébreu biblique ; choix qui, à l'analyse, nous est apparu trop restrictif ; et pouvait paraître tel *a priori* dans la mesure où les potentialités phonétiques sont l'objet de l'étude statistique.

M.W. remarque d'abord que les trois lignes du tableau ne coïncident pas : il calcule entre elles des coefficients de corrélation ce qui a tout au plus valeur indicative, dans la mesure où on ne sait à quel modèle de production aléatoire rapporter l'ensemble des racines... La différence apparaît la plus nette entre les positions I et II . L'originalité de M.W. est ici de tirer immédiatement des conséquences linguistiques intéressantes de l'examen de ces deux lignes. Il relève que

"Les consonnes {^pg z ḥ k n s ^cq š s} sont plus fréquentes en première position, et {b d h t ḷ m p r t} en deuxième ; {š} étant également fréquente en 1-ère et 2-ème ."

Et il décrit les consonnes affectionnant la position I comme "sifflantes (sibilants) velaires et gutturales" ; et celles de la portion II comme "labiales dentales et liquides (n excepté)". Il formule l'hypothèse explicative que lorsque deux consonnes se rencontrent sans voyelle interposée (ce qui est souvent le cas pour I et II en sémitique) l'hébreu préfère le mouvement d'arrière en avant et de fricatif à instantané (occlusif ou liquide) plutôt que l'inverse.

Ces conclusions s'accordent avec celles que nous avons formulées, mais seulement après analyse de la correspondance entre 1-ère et 2-ème radicale ; nous n'aurions pas osé proposer une interprétation fondée sur une classification des phonèmes, si cette classification n'était pas sortie elle-même de l'analyse...

3 Correspondance entre première et deuxième radicale et indice de dissimilarité entre consonnes : Bien qu'il ne publie pas explicitement le tableau des correspondances, formé des nombres :

$k(c, c')$ = nombre de racines dont la 1-ère radicale est la consonne c , et la 2-ème est c' ,

M.W. a manifestement ce tableau en vue quand il écrit :

"Afin de montrer sur un exemple comment mesurer la dissimilarité entre deux consonnes, considérons en hébreu a et b : la table (à trois lignes) est basée sur 1099 racines sont 66 ont a en 1-ère place, et 89 ont b en 2-ème place. On s'attend donc à trouver :

$$1099 \times (66/1099) \times (89/1099) = 5,344 \text{ racines}$$

ayant a en 1-er et b en 2-ème ; il y en a en fait 9 ; soit un rapport $(9/5,344) = 1,684$. Semblablement on attend $((56 \times 42)/1099) = 2,14$ racines avec b en 1-er et a en 2-ème, et il y en a 3, soit 1,402 fois ce qui est attendu. Nous pouvons définir comme indice de dissimilarité la moyenne de ces deux rapports soit $(1,684 + 1,402)/2 = 1,543$."

M.W. insiste ensuite sur l'importance essentielle des calculs en terme de fréquence, la présence de très nombreuses racines comportant une certaine paire de consonnes pouvant seulement résulter de la fréquence élevée de chacune de celles-ci sans qu'interviennent affinité ou répulsion.

Ce calcul d'indice mérite une exégèse ...

D'abord M.W., (comme il est également de règle en analyse des correspondances prend pour terme de référence le modèle d'indépendance :

$$k(c, c') = k(c) k(c') / k$$

et l'objet explicite de son analyse est, comme en analyse des correspondances, de manifester sous forme géométrique la structure des écarts existant entre les données réelles et le modèle d'indépendance : nous nous accordons pleinement avec lui en cela.

Ensuite, il n'hésite pas à prendre en compte des différences telles que $3 \neq 2,14$ (ou même $5,344 \neq 9$) qui prises individuellement n'ont pas de signification statistique : nous l'approuvons de faire ainsi, mais il importait de souligner que seule l'analyse multidimensionnelle permet de tirer des conclusions d'un système de nombres dont chacun pris individuellement ne dit rien.

Enfin, et c'est ici qu'il faut formuler les plus expresses réserves !, M.W. décrète que cet écart par rapport au modèle d'indépendance doit être compris comme résultant d'une affinité entre les contraires; deux consonnes étant d'autant plus volontiers associés en position I et II dans une même racine qu'elles sont plus dissemblables et il postule que l'on retrouvera par ses calculs de fréquence la notion de similitude familière aux phonéticiens.

En l'occurrence cette hypothèse est rendue vraisemblable par les remarques déjà faites par M.W. sur les positions préférées des diverses classes de lettres ... mais il est clair qu'elle n'a aucun caractère général. Si par exemple au lieu de consonnes on considérait des pays, avec un tableau

$k(p, p')$ = valeur des exportations en 1985 du pays p vers le pays p' ,

des valeurs élevées de $k(p, p')$ et $k(p', p)$ (relativement au modèle d'indépendances) seraient la marque d'une proximité entre p et p' et non d'un éloignement (ou d'une dissimilarité). Or il importe de souligner que dans la mesure où (cf. § 4) on cherche à placer sur une carte plane les points "c" de telle sorte que ceux-ci soient d'autant plus proches qu'ils sont plus semblables (ou d'autant plus éloignés qu'ils sont plus dissemblables) il est essentiel de décider si on regarde l'indice quotient calculé comme croissant ou comme décroissant en fonction de la distance.

Il est donc hautement souhaitable que la procédure d'analyse se déroule automatiquement sur tout tableau de nombres positifs, sans hypothèse préalable d'association par similitude ou par contraste.

4 Représentation géométrique d'un ensemble d'éléments munis de masses et de distance : Au tableau qu'il a calculé de l'indice de dissimilarité entre consonnes, M.W. applique un programme de "multidimensional scaling" (MDS) ; par quoi les auteurs anglosaxons entendent la méthode issue de R.N. Shepard (cf. TIA n° 2 ; et ENS2, I n° 7 § 2.6) qui vise à construire dans un espace de petite dimension (par exemple un plan), une représentation d'un ensemble qui respecte au mieux une

"ordonnance" donnée, c'est-à-dire un système d'inégalités entre distances.

Les résultats publiés par M.W. diffèrent assez peu de ceux que nous avons obtenus, (cf. *Prat. Ling.*, p. 39 ; ou [TEXTE BIBLE] déjà cité) pour que nous puissions nous dispenser d'en rendre compte (même si M.W. préfère la structure qu'il obtient lui-même à l'intérieur des 4 groupes de consonnes qu'il trouve comme nous).

Il importe en revanche de préciser les positions relatives des méthodes MDS, analyse factorielle (AF), analyse des correspondances (A d C).

D'abord, même si la tradition de l'analyse en composante principale (AFP) ne parle pas de nuage de points, il reste possible d'interpréter aussi bien AFP que AdC et MDS comme des méthodes de réduction de la dimension pour un ensemble décrit initialement dans un espace euclidien de dimension élevée ; ou ce qui mathématiquement est analogue, par la matrice des distances de point à point. Une première question est donc celle du mode de calcul de cette matrice des distances et de l'algorithme de réduction de la dimension.

On a vu qu'en MDS la définition de l'indice de similarité est chaque fois un problème nouveau, et peut reposer sur un choix crucial. En analyse factorielle, on fait de simples calculs euclidiens.

Il en résulte qu'en analyse factorielle l'algorithme de réduction de la dimension (qui aboutit à une carte plan etc.) est beaucoup plus rapide qu'en MDS : où les coûts de calcul seraient prohibitifs si l'on appliquait tels quels les programmes usuels à des centaines d'éléments.

Ensuite l'analyse des correspondances prenant pour données de base un tableau rectangulaire de nombres positifs ($I \times J$) fournit une représentation *simultanée* des deux ensembles (ensemble I des lignes et ensemble J des colonnes), alors que le MDS ne représente qu'un seul ensemble et l'AFP ne peut attribuer à deux ensembles distincts I et J des rôles équivalents. Il en résulte, dans le cas présent, que tandis que M.W. doit symétriser ses données raisonnant en termes de "consonne" et non de "1-er élément" et "2-ème élément", et que le graphique plan qu'il obtient ne montre aucunement quels principes régissent les associations entre ces 2 éléments (en fait on a vu qu'il a dû au contraire *postuler* une association par antithèse) ; l'analyse des correspondances montre sur le plan (1,2) chaque consonne, considérée comme initiale entourée des consonnes qui la suivent le plus volontiers en II ; et réciproquement à côté de chaque consonne considérée comme 2-ème, les initiales qui lui correspondent de façon prédominante ; ce qui permet de tirer de l'analyse même la conclusion que l'association se fait par antithèse.

En résumé représentation simultanée des deux ensembles (qui n'ont pas à être identiques : on analyse ordinairement des tableaux rectangulaires et non carrés), relative facilité des calculs, pour ne rien dire des aides à l'interprétation (qui permettent de nuancer les suggestions des graphiques) sont cause que nous regardons le tableau rectangulaire de nombres positifs, comme la donnée quasi universelle à construire pour un problème d'analyse de données, et l'a. des corr. comme la méthode naturelle pour traiter de tels tableaux.

Nous laissons à Michaël Weitzman la liberté d'apprécier dans quelle mesure ces avantages font de l'analyse des correspondances "the only useful approach" ...