

## CAD

### Ouvrage reçu

*Les cahiers de l'analyse des données*, tome 8, n° 1 (1983),  
p. 121-125

[http://www.numdam.org/item?id=CAD\\_1983\\_\\_8\\_1\\_121\\_0](http://www.numdam.org/item?id=CAD_1983__8_1_121_0)

© Les cahiers de l'analyse des données, Dunod, 1983, 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/>

## OUVRAGE REÇU

Ouvrage reçu : L'analisi delle frequenze, problemi di lessicologia ; A cura di M. Fattori e M. Bianchi ; Lessico Intellettuale Europeo XXVII ; Centro di Studio del CNR ; Edizioni dell' Ateneo ; Roma ; [Ouvrage aimablement communiqué par le Directeur du L.I.E. , le Pr. Tullio Gregory].

Des sept contributions dont se compose cet ouvrage collectif, trois utilisent l'analyse des données multidimensionnelles : il nous est agréable d'en rendre compte ici ; non seulement parce qu'elles apportent des résultats déjà intéressants mais surtout parce qu'elles témoignent d'une heureuse liberté dans l'envol vers des espaces inexplorés.

1 L'analyse factorielle dans l'élaboration des entrées de Dictionnaire ; par Ugo Berni Canani :

Il nous suffira de quelques citations :

"Supposons que nous ayons un ensemble de mots associés d'une façon paradigmatique au mot devant faire l'objet de l'article. On peut les représenter comme les noeuds d'un graphe en reliant entre eux ceux entre lesquels il y a une relation paradigmatique. S'il y a beaucoup de mots et beaucoup d'associations le résultat sera un graphe assez compliqué. Le problème est de le dessiner de la façon la plus claire possible... Il y a même un article d'un illustre spécialiste d'analyse combinatoire, W.T. Tutte, intitulé *Comment dessiner un graphe*. Or l'analyse factorielle se révèle justement un instrument très utile pour donner une bonne représentation des graphes".

L'auteur traite quelques exemples par l'analyse des correspondances et l'analyse en composantes principales (méthodes dont il s'abstient de faire une comparaison concluante). Les 6 planches qui illustrent ces exemples sont vraisemblablement les projections sur le plan  $1 \times 2$ , sans indication toutefois d'axes ni de valeurs propres. (D'ailleurs le tableau analysé lui-même, n'est pas clairement défini non plus).

La conclusion vaut d'être citée :

---

(1) *Professeur de statistique. Université Pierre et Marie Curie;*

*Professeur Tullio Gregory. Direttore Del Lessico Intellettuale Europeo Centro di Studio del C.N.R. . Roma.*

*U. Berni Canani ; J. Harnesse ; E. Brunet.*

"Sans doute s'agit-il de "bricolage factoriel". Mais il me semble que l'utilisation de procédures automatiques de ce genre pourrait faciliter la tâche de celui qui avec l'analyse de groupes de mots ou de contextes doit rédiger les articles d'un dictionnaire".

Enfin on nous permettra de rappeler que dans le Traité "*L'Analyse des Données*", la leçon [GRAPHE CORR.] TII B n° 10 est consacrée à l'analyse de la correspondance définie par un graphe.

2 Problème d'authenticité et analyse factorielle. A propos d'une oeuvre de Saint Bonaventure ; par J. Harnesse.

Ici encore citons l'auteur :

"Me trouvant face à des problèmes d'authenticité et d'étude de vocabulaire, je me suis demandé dans quelle mesure une analyse statistique pourrait m'apporter certaines lumières sur les textes que j'étudie.

Mais comment vérifier les résultats obtenus ? Comment avancer de manière sûre ? Trop de travaux actuels contiennent des conclusions hâtives, basées sur des données chiffrées qui n'ont pas été interprétées correctement ou sur des tests qui ne sont pas pertinents. De telles recherches jettent le discrédit sur l'analyse statistique en général. Certains n'hésitent pas à parler de "cancer de la statistique"...

... Il m'a semblé qu'une manière de procéder était de soumettre à l'analyse statistique un texte étudié de manière traditionnelle et d'examiner dans quelle mesure les résultats obtenus par des méthodes qui n'ont rien à voir avec les études quantitatives seraient infirmés ou au contraire étayés par l'analyse statistique.

Mon problème était le suivant : je devais rééditer une oeuvre de Saint Bonaventure, les *Collationes de septem donis Spiritus Sancti...*"

En bref, J. Harnesse explique qu'il s'agit d'un ouvrage en 9 livres dont chacun est un sermon. Après une introduction, viennent 7 livres consacrés chacun à l'un des dons du Saint Esprit. Mais le sermon sur le don de Courage est doublé d'un livre 6 qui est une homélie en l'honneur de la Vierge prononcée le même jour (en la fête de l'Annonciation, le 25 mars 1268).

Rendons la parole à J. Harnesse :

" Etant en possession de ces différents éléments, je me suis demandé si l'analyse statistique ferait apparaître une divergence entre le sixième sermon et le texte de l'oeuvre, ou bien au contraire montrerait qu'il n'existe pas de différence notable entre le sermon sur la Vierge et les autres sermons.

J'ai donc soumis le texte à deux méthodes d'analyse statistique : l'écart réduit et l'analyse factorielle...

Le tableau 1 [qui illustre l'article] présente dans l'ordre décroissant une liste des 50 lemmes ayant les écarts réduits positifs les plus importants dans le livre 6. Ces mots constituent vraiment le vocabulaire significatif et caractéristique du livre... ce sont par exemple VIRGO, MULIER, PRETIUM etc. . Mais un problème surgit immédiatement... certains [de ces] mots... présentent aussi un écart significatif dans un ou plusieurs des autres livres. Comment peut-on dès lors affirmer que le livre 6 est différent des autres et que l'étude statistique du vocabulaire nous permet de dire que ce livre

ne peut pas appartenir à l'oeuvre étudiée. La méthode d'analyse par l'écart réduit qui convient très bien pour déterminer le vocabulaire spécifique d'un livre, ne permet pas de déterminer les relations des différents livres entre eux...

L'analyse factorielle permet de décrire les relations qui existent non seulement entre les mots d'un texte, mais aussi entre les chapitres d'une oeuvre. On arrivera donc ainsi à déterminer non seulement le vocabulaire spécifique de l'un ou l'autre chapitre, mais aussi à dégager les *liens* qui existent ou non entre les différentes parties de l'oeuvre..."

Des résultats de l'analyse de correspondance nous noterons seulement ce qui suit.

"... L'axe 1 permet de trouver les deux grandes tendances de l'oeuvre : les livres 06 et 05 d'une part, les livres 04 et 08 d'autre part. L'examen des différents axes montre que 06 a un caractère tout à fait spécial. Il y a une très forte présomption qu'il appartienne à un autre ensemble de textes....

Et voici le mot de la fin :

"...Comme j'ai démontré cela par d'autres moyens, je peux dire que l'analyse statistique et l'analyse factorielle n'infirmement absolument pas les conclusions auxquelles j'étais parvenue, mais qu'au contraire elles vont dans le même sens et qu'elles permettent d'établir l'hypothèse de départ".

3 Le vocabulaire intellectuel français depuis 1789 ; par Etienne Brunet :

Cette contribution est la plus longue de l'ouvrage : pourtant, en 37 pages (dont 7 graphiques d'analyse de correspondance), l'auteur n'offre qu'un aperçu de son livre : "*Le vocabulaire français de 1789 à nos jours* (Statkine, Genève-Paris, 1981, 1824 pages). Comme pour les deux contributions précédentes, nous aimons mieux citer que paraphraser ; mais sans pouvoir dans l'espace d'un compte rendu suivre pas à pas la démarche de l'auteur qui lui même se résume!

"Point n'est besoin ... de présenter l'Institut de la langue française, où gît un trésor fabuleux de 70 millions de mots"...

Plus particulièrement, on distinguera ici un "sous-corpus technique, de près de 9 millions de mots" où "la philosophie se taille la part du lion"... sous-corpus "intellectuel" donc plutôt que technique.

"Considérons d'abord... la forme ou plutôt le volume du mot. (dont la longueur va jusqu'à 32 en tenant compte des mots composés). "... Nous avons mesuré cet espace, ou longueur moyenne du mot, dans l'ensemble du corpus et dans les sous-ensembles. Le mot est nettement plus volumineux dans la prose technique... L'analyse fine des classes de longueur (de 1 à 18) permet d'attribuer cet écart non à une moindre fréquence des mots courts (le genre technique au contraire fait un usage intensif des mots grammaticaux qui ont moins de 4 lettres) mais à une plus grande indulgence à l'endroit des mots longs, comme l'indiquent les écarts largement positifs à partir de la classe de longueur 8.

L'analyse factorielle... ne laisse aucun doute à ce sujet : les mots longs sont attirés vers la gauche du graphique au voisinage du discours intellectuel"

Ce premier exemple atteste la méthode de l'auteur : celui-ci fait divers calculs d'écarts, dans le cadre du modèle normal ; mais il recourt à l'analyse des correspondances pour avoir une vue d'ensemble. Ainsi sur l'axe 1, les classes de longueur 1 à 6 associées à *Dialogue* et *Prose littéraire* (suivis de poésie et soliloque) s'opposent aux classes 8 à 18, associées à *Prose technique* et *Reste* ; mais la classe 1 est relativement proche de l'origine, ce qui correspond à l'affection remarquée du langage intellectuel pour les mots grammaticaux.

Comme deuxième exemple nous prendrons "les choix du discours intellectuel parmi les catégories" [Substantifs, adjectifs, adverbes en *ment*, autres adverbes, verbes, divers participes...]. Ici encore l'auteur donne des écarts réduits et ajoute :

"L'analyse factorielle rend compte de ces faits avec clarté ; tout en manifestant également l'opposition des personnes, alors même que personnels et possessifs ont été retirés de l'analyse" [par quoi l'auteur fait allusion à la position extrême de *Dialogue* sur l'axe 1, à l'opposé de *Technique* et *Reste*]. Mais voici un troisième exemple.

"D'aucuns estimeront saugrenue l'idée de compter les 300 millions de caractères de notre corpus et d'étudier la répartition des 37 variétés de lettres dans le temps et dans les genres. Or le hasard n'a point ici sa place, et sur les 74 écarts réduits... [l'auteur se borne aux fréquences de 37 lettres diversement accentuées é, è, ê, ê etc., dans les deux genres du *Discours intellectuel* et de la *Prose littéraire*] 1 seul n'est pas significatif au seuil de 5% et 68 laissent au hasard moins d'une chance sur un milliard...

Les statisticiens comprendront ! Les lexicologues seront charmés de voir le hasard perdre toutes ses chances. Peut-être vaudrait-il pourtant mieux dire, que si les deux sous-corpus étaient chacun engendrés par tirages au sort successifs de lettres dans une urne, les deux urnes parentes ne pourraient être que différentes. De notre suspicion envers ces urnes nous n'avons jamais fait un secret ! Cependant l'auteur poursuit avec finesse.

"On songe immédiatement à mettre en cause l'influence des mots fréquents (... "je" sur la lettre j). Or si l'on écarte les fréquences élevées... les faits se maintiennent et les mêmes lettres n, q, p, s, x, o, é restent caractéristiques du genre technique ou intellectuel. Je livre à la perspicacité des spécialistes ces faits d'observation que l'analyse factorielle... met en évidence et dont l'interprétation peut laisser perplexes".

Sans être spécialiste nous n'hésiterons pas à avancer une interprétation : l'auteur a pu sans peine éliminer les mots les plus fréquents ; il n'a pas éliminé les morphèmes : préfixes ou suffixes ou désinences de haute fréquence, les lettres seviles dont la technique fait grand usage. Ce nous sera l'occasion de répéter qu'il faut compter non seulement des mots ou des formes, mais des morphèmes !

Citons seulement une analyse consacrée aux signes de ponctuation qui ne laisse au discours intellectuel qu'italiques et parenthèses... Il faut toutefois prendre garde ici que l'analyse d'un tableau de contingence pris tel quel, ne compare que des profils : elle montrera e.g. que les " et les () sont *relativement* aux autres signes fréquents dans le discours intellectuel : mais quant à la fréquence absolue (rapportée à la longueur du texte) elle n'apparaîtrait que si par exemple on mettait dans une même analyse

le compte des lettres (cf. *supra*) et celui des autres signes. Au reste dans toutes ces analyses il vaudrait la peine d'avoir des tranches plus étroites que "prose littéraire" ou "soliloque" voire des oeuvres isolées : l'ensemble de celles-ci pouvant être adjoind en élément supplémentaire ou même constituer le tableau principal. Ces éléments supplémentaires sont d'autant plus nécessaires que des entités telles que "prose littéraire", "prose technique" ou "soliloque" n'ont aucune existence réelle ; l'ampleur monumentale des données, reflétée par des intervalles de confiance étroits issus d'un modèle d'urne illégitime, ne doit pas faire illusion ; il ne s'agit que d'agrégats en partie arbitraires ; seules existent réellement les oeuvres individuelles ; mais l'auteur y vient lui-même.

" Pour mieux saisir les mouvements de l'histoire dans les profondeurs du vocabulaire (et non plus à la surface thématique des textes) nous avons fait appel une nouvelle fois à l'analyse factorielle".

(Ah qu'en termes galants...! mais les termes de *surface* et de *fond* ne devraient-ils pas être justifiés?) Bien que le tableau analysé ne soit pas décrit dans le langage explicite de la théorie des ensembles ( $I = \dots$  ;  $J = \dots$  ;  $k(i, j) = \dots$ ) on reconnaît qu'il s'agit d'une correspondance entre vocables et tranches de texte délimitées quant au temps et au genre. Or,

"... Quoique la machine ignore absolument les notions de temps et de genre, on voit... que les chiffres reconstituent l'un et l'autre à travers les distributions lexicales. Des deux forces d'attraction (ou de répulsion), le genre est... la plus puissante et c'est celle qui distribue à droite des graphiques les 14 tranches du discours intellectuel, avec les éléments lexicaux qui composent ce discours et qui sont exactement les mots précédemment identifiés. Quand les mots ont ainsi choisi leur camp, à gauche ou à droite, selon qu'ils appartiennent à la prose littéraire ou intellectuelle, la chronologie intervient qui les répartit de "bas en haut".

Sur cette vision tout à fait agréable au spécialiste d'analyse factorielle nous quitterons l'auteur ; en le félicitant de ses monumentales investigations ; non sans l'encourager à plus d'audace encore ; et nous excuser d'avoir parfois malignement tiré quelques plumes à ses ailes.

---

IMPRIMERIE GAUTHIER-VILLARS, France

Dépôt légal 1983 · Imprimeur : 2549 Éditeur : 032 CCCP 57373

BORDAS-Dunod 1983

avril 1983

Imprimé en France

6151-83

Le Directeur de la publication : J. M. BOURGOIS