AMADO online, une application multilingue pour la visualisation et l'analyse graphique de matrices de données

Nguyen-Khang Pham*, Jean-Hugues Chauchat**

*Can Tho University, College of Information & Communication Technology
Campus II, 3/2 street, Ninh Kieu district, Can Tho city, Viet Nam
pnkhang@cit.ctu.edu.vn
http://www.cit.ctu.edu.vn/ pnkhang/index-en.html

**Université de Lyon, Laboratoire ERIC, 69676 Bron Cedex, France
jean-hugues.chauchat@univ-lyon2.fr
http://eric.ish-lyon.cnrs.fr

Résumé. *AMADO online* permet de visualiser et d'analyser des matrices de données (présence/absence, ou comptages, ou réponses sur des échelles, ou mesures de variables hétérogènes) selon les principes de Jacques BERTIN. Les nombres sont représentés par des rectangles dont la surface leur est proportionnelle. Pour mettre en évidence la structure des données, les lignes (et/ou les colonnes) peuvent être réordonnées à la main, ou automatiquement selon leurs coordonnées sur le premier axe factoriel de l'Analyse des Correspondances, ou en Composantes Principales, selon la nature des données ; une double Classification Ascendante Hiérarchique est disponible. Selon le cas, on obtient une sériation (chronologique par exemple), ou des blocs exacts ou approximatifs, ou des classes relativement homogènes. Ces graphiques sont fidèles aux données et faciles à lire. Le *Guide de l'Utilisateur*, en français et en anglais, détaille les commandes sur de nombreux fichiers d'exemples fournis avec l'outil.

1 Introduction

AMADO online permet de représenter graphiquement un tableau croisé de nombres, puis de permuter les lignes et les colonnes pour faire apparaître la structure des données : - soit une structure diagonale (sériation) si elle existe, - soit une structure en classes croisées des lignes et des colonnes, voire en blocs. De nombreuses options de mise en forme sont disponibles dans les menus.

L'objectif est de faciliter la lecture des données en mettant en évidence leur structure par des opérations très simples de permutation de lignes (et/ou de colonnes) selon les principes de Jacques Bertin (2017, 1999).

Développé dans le cadre du consortium "Paris Time Machine", l'outil est accessible librement sur la plateforme https://paris-timemachine.huma-num.fr/amado-online/ ; il complète et améliore le travail de Chauchat et Risson (1998). Actuellement les menus sont proposés en 7 langues : anglais, français, espagnol, italien, russe, ukrainien et vietnamien.

Les *Guides de l'Utilisateur* sont disponibles en français et en anglais, chacun faisant une trentaine de pages ; ils présentent plusieurs types de tableaux avec, pour chacun, les données sources et les suites de commandes des menus d'*AMADO online* permettant d'obtenir les graphiques tels que ceux présentés ici ; ces graphiques peuvent être sauvegardés en format image PNG ou vectoriel SVG.

2 Exemple d'un petit tableau croisé de comptage

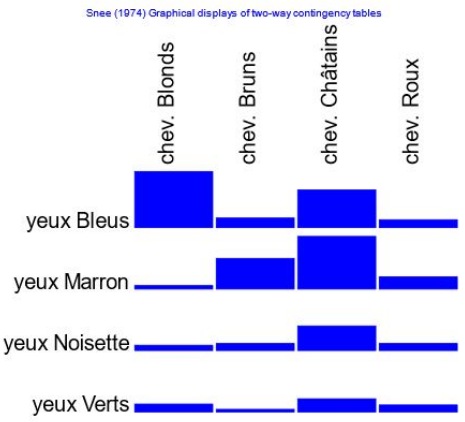
Le traitement des données commence par leur importation - soit par un simple copier-coller depuis un tableur, - soit par l'ouverture d'un fichier TXT ou CSV. La Fig. 1 montre une suite de traitements d'un tableau classique issu de Snee (1974).

Snee (1974) Graphical displays of two-way contingency tables	chev. Blonds	chev. Bruns	chev. Châtains	chev. Roux
yeux Bleus	94	20	84	17
yeux Marron	7	68	119	26
yeux Noisette	10	15	54	14
yeux Verts	16	5	29	14

Après avoir :
- réordonné les lignes et colonnes selon leurs coordonnées sur le 1er facteur de l'Analyse des Correspondances,
- mis les hauteurs proportionnelles aux % en colonne,
– mis les bases proportionnelles à leurs effectifs totaux, on obtient le graphique ci-dessous.

Graphique visualisant le tableau ci-dessus :

Sur l'ensemble du graphique, les surfaces des rectangles restent proportionnelles aux effectifs.



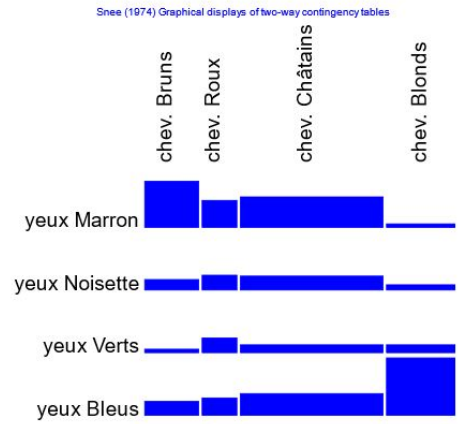


FIG. 1 – *Tableau croisant les couleurs des yeux et des cheveux de 592 personne, et graphiques produits par AMADO online.*

Les données doivent contenir les intitulés de lignes et de colonnes, mais pas les totaux marginaux ; on importe les données en valeurs absolues ; la première cellule peut contenir un titre.

Dès l'importation des données, *AMADO online* les affiche sous forme d'histogrammes, la surface de chaque rectangle étant proportionnelle au nombre qu'il représente. Ensuite, les commandes permettent la mise en forme du graphique : insertion, correction ou suppression du titre, permutations des lignes (et des colonnes), dimensions du graphique, transposition, taille et couleur des libellés et des valeurs, format d'affichage des valeurs, représentation en % dans chaque colonne, normalisation des lignes ou colonnes, suppression de lignes ou de colonnes, insertion de séparateurs entre des lignes ou des colonnes, largeurs des colonnes proportionnelles à leurs totaux, application de l'Analyse des Correspondances {Benzécri (1982), Greenacre (1984), Tenenhaus et Young (1985)} ou de classifications ascendantes hiérarchiques des lignes et colonnes, etc.

Les graphiques produits par *AMADO online* sont simples à lire ; ils donnent au lecteur un accès direct au résultat : chaque élément d'information - chaque nombre du tableau de données est restitué dans sa forme originelle : les hauteurs des rectangles sont proportionnelles aux valeurs du tableau original, soit en nombres absolus, soit en %.

3 Exemple de tableau de mesures avec des variables d'unités différentes

Dans la Fig. 2, les colonnes ont des unités différentes (cm3, Hp, Km/h, Kg, cm) et ne sont pas directement comparables.

Des hauteurs de rectangles proportionnelles aux nombres du tableau n'auraient pas de sens ici. Pour tout calcul ultérieur, il faut les normaliser : chaque valeur du tableau est centrée sur le minimum de colonne (de façon à ce que tous les résultats soient positifs ou nuls) puis divisée par l'écart-type de colonne ; on obtient alors des nombres purs, c'est-à-dire "sans dimension") X_{ij} devient $\frac{X_{ij}-Min_j}{\sigma_j}$. Ensuite, les procédures de calcul sont effectuées sur ces "nombres purs". Comme *AMADO online* ne peut représenter que des nombres positifs, la plus petite valeur devient zéro ; dans notre exemple, la "Smart Fortwo Coupé" est la plus petite voiture parmi les 6 variables, les 6 valeurs deviennent zéro pour elle dans le graphique.

Sur l'arbre de classification, on distingue les classes de voitures :

— la Smart Fortwo Coupé est seule, la plus petite pour toutes les variables ;
— les Citroën C2, Nissan Micra, Citroën C3 et la Peugeot 307 forment un groupe homogène de 4 petites voitures ;
— petites (mais plus sportives) les Mini, Renault Clio, BMW Z4 et Audi TT ;
— les grandes voitures familiales Land Rover Defender, Nissan X-Trail, Volkswagen Touran, Renault Scenic et Audi A3 ;
— la Land Rover Discovery est spécifique, étant longue, large et lourde, relativement peu puissante pour sa taille et plutôt lente ;
— les 6 berlines grandes, nerveuses et rapides : Mercedes Classe C, Jaguar S, BMW 530d, Peugeot 407, BMW 745i, Mercedes Classe S ;
— enfin les grandes, très puissantes (et extrêmement chères) : Ferrari, Bentley et Aston Martin.

6 caractéristiques de 24 voitures

Voitures (Tenenhaus 2004)	Cylindrée cm3	Puissance ch	Vitesse Km/h	Poids Kg	Largeur cm	Longueur cm
Citroën C2 1.1 Base	1124	61	158	932	1659	3666
Smart Fortwo Coupé	698	52	135	730	1515	2500
Mini 1.6 170	1598	170	218	1215	1690	3625
Nissan Micra 1.2 65	1240	65	154	965	1660	3715
Renault Clio 3.0 V6	2946	255	245	1400	1810	3812
Audi A3 1.9 TDI	1896	105	187	1295	1765	4203
Peugeot 307 1.4 HDI	1398	70	160	1179	1746	4202
Peugeot 407 3.0 V6 B	2946	211	229	1640	1811	4676
Mercedes Classe C 27	2685	170	230	1600	1728	4528
BMW 530d	2993	218	245	1595	1846	4841
Jaguar S-Type 2.7 V6	2720	207	230	1722	1818	4905
BMW 745i	4398	333	250	1870	1902	5029
Mercedes Classe S 40	3966	260	250	1915	2092	5038
Citroën C3 Pluriel 1	1587	110	185	1177	1700	3934
BMW Z4 2.5i	2494	192	235	1260	1781	4091
Audi TT 1.8T 180	1781	180	228	1280	1764	4041
Aston Martin Vanquis	5935	460	306	1835	1923	4665
Bentley Continental	5998	560	318	2385	1918	4804
Ferrari Enzo	5998	660	350	1365	2650	4700
Renault Scenic 1.9 d	1870	120	188	1430	1805	4259
Volkswagen Touran 1.	1896	105	180	1498	1794	4391
Land Rover Defender	2495	122	135	1695	1790	3883
Land Rover Discovery	2495	138	157	2175	2190	4705
Nissan X-Trail 2.2 d	2184	136	180	1520	1765	4455

Après normalisation et classification

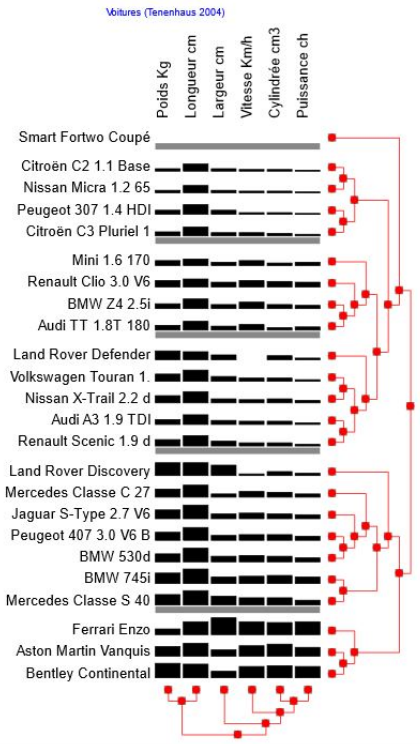


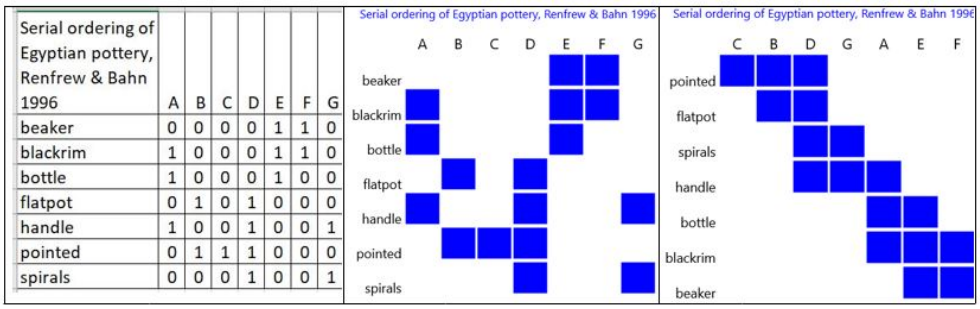
FIG. 2 – *Tableau de mesures hétérogènes : 24 voitures et 6 caractéristiques, inspiré de M. Tenenauss, et graphique de classification automatique en 6 groupes produit par AMADO online.*

Du côté des variables, le poids et la longueur sont fortement corrélés, tout comme la cylindrée et la puissance. La vitesse d'une part et la largeur d'autre part, sont moins corrélées aux autres caractéristiques, et on voit pourquoi sur le graphique : la Land Rover Discovery est large, lourde et peu rapide, tandis que la Ferrari Enzo est légère mais très large, tout en étant très puissante et très rapide.

4 Jacques BERTIN et l'histoire des représentions graphiques de tableaux

L'idée de permuter les lignes et les colonnes d'une matrice dans le but de révéler une structure cachée dans une matrice de données est ancienne : Sir Flinders Petrie (1899) a présenté il y a un siècle une "séquence dans les vestiges préhistoriques", c'est-à-dire une "sériation" chronologique des formes et éléments de décor d'objets trouvés lors de fouilles en Egypte ; le

petit exemple représenté Fig. 3 est proposé dans Renfrew et Bahn (1991) pour illustrer cette démarche. Les tombes (C, B, D, G, A, E et F) sont réordonnées, selon un axe de présence-absence des éléments de décor repérés par l'archéologue dans les sites de cette région ; cet ordre correspond probablement à l'ordre chronologique (direct ou inverse) d'invention puis d'abandon de ces créations artistiques.



Serial ordering of Egyptian pottery, Renfrew & Bahn 1996	A	B	C	D	E	F	G
beaker	0	0	0	0	1	1	0
blackrim	1	0	0	0	1	1	0
bottle	1	0	0	0	1	0	0
flatpot	0	1	0	1	0	0	0
handle	1	0	0	1	0	0	1
pointed	0	1	1	1	0	0	0
spirals	0	0	0	1	0	0	1

FIG. 3 – *Données de Renfrew-1991, inspirées par celles de Sir Flinders PETRIE pour dater des tombes égyptiennes. Les colonnes (A, B, ...) représentent des tombes où ont été trouvés les objets contenant, ou non, les éléments de décor nommés en lignes. AMADO online produit la sériation en un clic.*

Cette idée a eu une influence croissante dans les mathématiques appliquées, en particulier dans les sciences du comportement. (Bertin (2017, 1999) a mis côte à côte des histogrammes, en utilisant une échelle appropriée, et a permuté les éléments pour révéler les structures sous-jacentes dans les données. Depuis lors, cette approche a connu un essor considérable en France et dans le monde : Heer, Pal, Dhieb (2018), Arabie et al. (1978) et Perin et al. (2018). Ensuite, la diffusion des méthodes d'analyse des données multidimensionnelles [Escofier-Cordier (1969), Benzécri (1982), Greenacre (1984), Tenenhaus et Young (1985), Hoffman et DeLeeuw (1992)] a quelque peu éclipsé cette approche purement visuelle.

Certes, les techniques numériques de l'analyse des données permettent de découvrir rapidement les grands traits de la structure du tableau et on économise ainsi un temps considérable dans la recherche du meilleur couple de permutations des n lignes et des p colonnes du tableau parmi les $n!\,p!$ solutions possibles.

Mais, en analyse factorielle, les listes de coordonnées et autres aides numériques à l'interprétation sont utiles au statisticien mais souvent incompréhensibles pour le chercheur en sciences sociales. Leur interprétation demande un œil averti, et ils doivent peut-être une partie de leur succès auprès du grand public à leur ésotérisme même... De leur côté, les arbres de classification donnent une représentation utile mais déformée ("ultramétrique") du tableau originel.

Au contraire, les graphiques construits par *AMADO online* peuvent utiliser l'analyse factorielle ou la classification tout en donnant au lecteur un accès direct au résultat : chaque élément d'information - chaque nombre du tableau de données - est restitué dans sa forme originelle,

soit en nombre absolu, soit en pourcentage. C'est uniquement l'ordre des lignes et des colonnes qui a changé, et cela suffit pour visualiser la structure des données.

5 Conclusion

AMADO online est un outil libre, multilingue, qui permet de transformer un tableau croisé en graphiques puis de les transformer pour mettre en évidence la structure des données. Ces graphiques sont fidèles aux données et faciles à lire.
AMADO online est adapté aux tableaux petits ou moyens (jusqu'à une cinquantaine de lignes et colonnes, ou plus selon le moniteur utilisé), tels que ceux qui sont construits en Sciences Humaines et Sociales où chaque élément a été défini précisément et doit être restitué facilement dans l'ensemble.

Références

Arabie, P., A. Scott, Boorman, et P. R. Levitt (1978). Constructing blockmodels : How and why ? *Journal of Mathematical Psychology 17*(1), 21–63.

Benzécri, J.-P. (1967-1982). *L'Analyse des Données, t. I : Taxinomie ; t. II : L'Analyse des Correspondances*. Paris : Bordas.

Bertin, J. (1967-1973-1999). *Sémiologie Graphique. Les diagrammes, les réseaux, les cartes*. Paris : Mouton, Gauthier-Villars.

Bertin, J. (1967-2017). *La graphique et le traitement graphique de l'information*. Paris : Zones sensibles.

Chauchat, J.-H. et A. Risson (1998). Bertin's graphics and multidimensional data analysis. In J. Blasius et M. Greenacre (Eds.), *Visualization of Categorical Data*, pp. 37 – 45. Academic Press.

Dhieb, M. (2018). Traduire encore Bertin aujourd'hui : pourquoi faire ? Les nouvelles faces cachées de la sémiologie graphique. *Proceedings of the ICA 1*, 27.

Escofier-Cordier, B. (1969). L'analyse factorielle des correspondances. *Cahiers du Bureau Universitaire de Recherche Opérationnelle Série Recherche 13*, 25–59.

Greenacre, M. (1984). *Theory and applications of Correspondence Analysis*. Academic Press.

Heer, J. M. A brief history of data visualization. stanford human computer interaction seminar. https ://www.youtube.com/watch ?v=N00g9Q9stBo.

Hoffman, D. et J. DeLeeuw (1992). Interpreting multiple correspondence analysis as a multidimensional scaling method. *Marketing Letters 3*, 259–272.

Perin, C., J.-D. Fekete, et P. Dragicevic (2018). Jacques Bertin's legacy in information visualization and the reorderable matrix. *Cartography and Geographic Information Science*.

Petrie, W. M. F. (1899). Sequences in prehistoric remains. *The Journal of the Anthropological Institute of Great Britain & Ireland 29*, 295–301.

Renfrew, C. et P. Bahn (1991). Archaeology : Theories, methods and practice. *Archaeological Journal 148*(1), 329–330.

Snee, R. (1974). Graphical display of two-way contingency tables. *The American Statistician 28*(1), 9–12.

Tenenhaus, M. et F. W. Young (1985). An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data. *Psychometrika 50*, 91–119.

A Remerciements

Les auteurs remercient chaleureusement :
— l'équipe du consortium *Paris TimeMachine* qui a soutenu ce travail : Jean-Luc Pinol, Hélène Noizet, Paul Rouet, Laurent Costa, Julien Avinain, Éric Mermet et le conseil scientifique ;
— nos collègues universitaires qui nous ont aidé dans la traduction des menus : Annie Morin et Jairo Cugliari pour l'espagnol, Linda Gattuso pour l'italien, Iryna Zolotariova pour l'ukrainien, Olena Orobinska-Goncharova pour le russe ;
— Basu Tallur qui a partagé son expérience en création de vidéo ;
— Sylvain Clément qui a prêté sa voix pour la vidéo de démonstration en anglais ;
— Jean Dumais pour ses relectures attentives et ses nombreux conseils d'amélioration du *guide de l'utilisateur* en français et sa traduction en anglais ;
— Alban Risson qui, alors étudiant, avait réalisé la première version d'AMADO.

Summary

AMADO online is an application for visualizing and analysing data matrices (presence-absence, or cross-tabulation, or responses on scales, or measures of heterogeneous variables) according to the principles of Jacques Bertin. Numbers are represented by rectangles whose area are proportional to them. To highlight the structure of the data, the rows (and/or columns) can be reordered, manually, or automatically according to their coordinates on the first factorial axis of Correspondence Analysis, or in Principal Components Analysis, according to the nature of the data; a double Hierarchical Agglomerative Clustering is available. Depending on the case, a seriation (chronological for example), or exact or approximate blocks, or relatively homogeneous classes are obtained. These graphs are faithful to the data and easy to read. The *User's Guide*, in French and English, describes the sequence of commands for many example files provided with the tool.