HAL open science

50 years of data sciences, discussion

Susan Holmes, Julie Josse

To cite this version:

Susan Holmes, Julie Josse. 50 years of data sciences, discussion. Journal of Computational and Graphical Statistics, 2017. hal-01670428

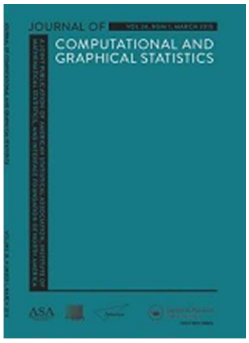
HAL Id: hal-01670428

https://hal.science/hal-01670428v1

Submitted on 21 Dec 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Discussion of 50 Years of Data Science

Journal:	*Journal of Computational and Graphical Statistics*
Manuscript ID	JCGS-17-257
Manuscript Type:	Invited Comment (as part of discussion papers)
Keywords:	Data Science, Donoho, Comment

For Review Only

SCHOLARONE™
Manuscripts

Discussion of *50 years of data-science*

Susan Holmes*
Department of Statistics, Stanford University
and
Julie Josse
Ecole Polytechnique, INRIA

September 10, 2017

For Peer Review Only

First of all, we would like to thank the author for writing such a thoughtful article. The article draws attention to so many important aspects at the intersection of data science and applied statistics. Having been raised in the French school of Data Science (Analyse des Données, see [Holmes, 2008]), this article has a particular resonance for us.

In the 1970s, a group of French Statisticians under the leadership of Benzécri revolted against the probabilistic emphasis given to the field of statistics and decided to create a new discipline that would put the applications and the data first.

Benzécri, having spent time at Bell Labs visiting Roger Shepard, shared the view that the future of applied statistics involves computational and geometric approaches (in particular, the projection of data using a variety of weighted distances). His practical vision of science and data science include the idea of *letting the data speak for themselves* and of finding *a rigorous method which extracts structures, patterns from the data* [Benzécri, 1973, p 6, Tome 2]. Data encoding, visualization and writing programs were considered crucial steps in the analysis.

Correspondence Analysis, the cornerstone method of this current, was developed jointly with Brigitte Escofier-Cordier (1965) and presented by Benzécri during six lectures at Collège de France. Correspondence Analysis was first designed to describe and visualize the associations in a contingency table crossing two categorical variables. A driving application was in linguistics with the analysis of text-word data [Murtagh, 2005] from a variety of corpus (Greek and Latin philosophy, Biblical, medieval philosophy, and Russian 20th century literature).

Benzécri attached great importance to the collaboration between disciplines and the explosion of the use of his methods in many fields such as anthropology, sociology, economics, marketing, hydrology, geography, bibliometrics, environmetrics, marked a generation of applied statisticians in France. "L'analyse des données" was especially popular in social sciences where categorical data were prevalent through surveys. Pierre Bourdieu, a pioneer, presented in "La Distinction" (1976/79), graphical maps of social spaces that can uncover complex relations [Lebaron & Le Roux, 2015]. CA was even discussed in the media "Le nouvel observateur" [1] (Derosiére in [Lebart, 2008]).

Dissemination was fostered by the availability of numerical libraries in Fortran and free software. In addition, "l'analyse des données" was taught to many students with PhD and Masters programs covering geometric projection methods ("analyses factorielles"), interpretation tools ("points supplémentaires", "parts d'inertie"), clustering, data encoding and programming. Internships in companies were also made mandatory after May 1968 [Murtagh, 2005].

*CASBS (Center for the Advanced Study of the Behavioral Sciences)

[1] you could read sentences as ..."This immense volume, redesigned flat by the computer thanks to savant calculations but which preserve at best the disparities observed between the professions... "

Other areas of data science were active and well developed by M. Tenehaus, G.Saporta amongst others. E. Diday, I.C. Lermann, M. Roux and G. Celeux developed and implemented many clustering methods and took the lead in the "Clustering societies".

Geographically, there were several teaching and research hotbeds outside of Paris: Rennes where I.C Lermann and G. Le Calve started groups; Montpellier where Y. Escoufier worked, Marseille with B. Fichet. These groups even developed their own software packages for training students and developed new methods; for instance multiway data fusion - combining multiple matrices of data formed of variables from different domains and of different nature (categorical, quantitative, counts) and multitype clustering methods. Hundreds of articles and books were published, all in French; maybe one reason why much of the work done in the 1980's did not have a large impact.

However, a broader community started to ramify as the French made contacts with the researchers in Britain such as John Gower and Frank Critchley. The Japanese and Dutch schools in multivariate nonparametrics seemed to be following somewhat similar paths and there were successful international meetings that connected these between the different groups. Michael Greenacre, one of Benzécri's best known students has written many books in English making the methods popular in social and ecological sciences.

Today, we can see the influences of the "Euclidean" school epitomized by [Cailliez and Pages, 1976] in the development of kernel methods. Clustering analyses have certainly remained center stage. Multitable coefficients such as distance correlations are now very popular [Josse, Holmes, 2016]. Visualization was an important early factor in the success of the methods. In some sense, one can even regard today's stochastic block models as a natural extension of Bertins' early graphical matrix block methods developed in *Semiologie graphique* [Bertin, 1973].

We note that we were lucky that Donoho's hero, John Chambers liked to come to France. In 1986, he showed up in Montpellier with the tape for a new software programming language called **S** [Chambers et al, 1988]. **S** became an important tool in developing and teaching French methods. It was very popular in the early 1990s with ecologists, food scientists and agronomical engineers [Chessel et al., 2004].

Data science in France is now alive and well, most of all because the young French community has become multilingual, speaking and writing in English, R and python.

For Peer Review Only

References

[Benzécri, 1973] Benzécri, J. P. (1973, 1976). L'analyse des données, Tome 1 et 2. Dunod.

[Benzćri, 1982] Benzcri, J. P. (1982). Histoire et préhistoire de l'analyse des données. (Dunod) with the texts published in 1976-77 in "les cahiers de l'analyse des données".

[Bertin, 1973] Bertin, J. "Sémiologie graphique." Flammarion, Paris (1973).

[Cailliez and Pages, 1976] Cailliez, F. and Pages., J. P. (1976). *Introduction à l'analyse des données.* SMASH, Paris.

[Chambers et al, 1988] Becker, R., Richard, A., Chambers, J.M. and Wilks, A.R. (1988) "The new S language." Pacific Grove, Ca.: Wadsworth & Brooks.

[Chessel et al., 2004] Chessel, D., Dufour, A. B., and Thioulouse., J. (2004). The ade4 package - I: One-table methods. *R News*, 4(1):5–10.

[Diday, 1973] Diday, E. (1973). The dynamic clusters method in nonhierarchical clustering. *International Journal of Computer and Information Sciences*, 2(1):62—-88.

[Escofier and Pagès, 1998] Escofier, B. and Pagès, J. (1998). *Analyse factorielles simples et multiples : Objectifs, méthodes et interprétation.* Dunod.

[Holmes, 2008] Holmes, S. (2008) *Multivariate data analysis: the French way.* Probability and statistics: Essays in honor of David A. Freedman. Institute of Mathematical Statistics. 219-233.

[Josse, Holmes, 2016] Josse J. and Holmes S. (2016) Tests of independence and beyond. Statistics Surveys, **10**, 132-167.

[Lebart, 2008] Lebart, L. About history of multivariate exploratory data analysis. *Electronic Journal for History of Probability and Statistics* **32**, 159–188.

[Lebart, Saporta,2014] Lebart, L. & Saporta, G. (2014). Historical Elements of Correspondence Analysis and Multiple Correspondence Analysis. chapter 3; book: Visualization and Verbalization of Data (Blasius et Greenacre), Chapman & Hall, 2014.

[Lebart et al., 2000] Lebart, L., Piron, M., and Morineau, A. (2000). *Statistique exploratoire multidimensionnelle.* Dunod, Paris, France.

[Le Roux, Rouanet, 2004] Le Roux, B. & Rounet, H. (2004). Geometric Data Analysis. Chapter 4 "Historical Sketch".

[Lebaron & Le Roux, 2015] Lebaron, F & Le Roux, B. (2015). La méthodologie de Pierre Bourdieu en action: espace culturel, espace social et espace social. Dunod, Paris.

[Murtagh, 2005] Murtagh, F. (2005). Correspondence Analysis and Data Coding with Java and R. Chapman & Hall.

[Saporta, 1976] Saporta, G. (1976, 2006, 2011). Probabilités, analyse des données et statistique. Edition Technip.

For Peer Review Only