JOURNAL DE LA SOCIÉTÉ FRANÇAISE DE STATISTIQUE

FERNANDO TUSELL

Discussion and comments. Approche graphique en analyse des données

Journal de la société française de statistique, tome 141, n° 4 (2000), p. 83-85

<http://www.numdam.org/item?id=JSFS_2000__141_4_83_0>

DISCUSSION AND COMMENTS

Approche graphique en analyse des données

Fernando TUSELL [1]

It is a pleasure to comment the paper by Jean-Paul Valois; the author has to be commended for a very fine exposition.

In a work of this nature there are many aspects deserving comment. I will select just a few miscellaneous and largely unrelated comments which are of particular interest to me and may be to others.

Let us begin with the historical issues. The author emphasises that the revival of graphical methods in the decade of the seventies predates the use of computers. Such revival is attributed to the cultural context :

> « Suggérons un possible mouvement culturel de revalorisation des perceptions visuelles sous l'influence des médias qui se répandent à partir des années 60 [...] »

No doubt the cultural context played a role, but even in the seventies computers were widespread, if not at every desk top. The landmark work [1] already echoes this availability, even if manual methods are still advocated for e.g. permuting matrices. It is clear on the other hand that graphics such as Chernoff's faces (see [2]) were never considered practical or even feasible by purely manual means. Much of what we see today cannot be understood – even conceived – without fast computers. No matter how influential the cultural context may have been, it appears to me that the computer revolution remains the main driving force in the revival of graphics.

It seems to me that much of the relative importance given to graphics can be explained by the fluctuating gap between existing graphical tools and the nature of problems addressed. Simple comparison of two time series could be addressed by means of graphs going back at least to Playfair's time – see [7] for instance. More complex problems, particularly involving moderate to large numbers of variables and/or cases were not so easily treated graphically. They had to wait until proper tools were available.

This raises what I think is now a fundamental question : whether graphical methods are able to cope with present problems, or else the gap has widened.

1. Departamento de Economía Aplicada III (Estadística y Econometría). Facultad de CC.EE. y Empresariales, Universidad del País Vasco, Avda. del Lehendakari Aguirre, 83, 48015 BILBAO. E-mail : etptupaf@bs.ehu.es.

In my view, this means whether graphical methods can play a role with today's huge data sets. It strikes me that very few of the graphs in the author's typology in Figure 8 are useful to meet the present challenges in data mining, for instance. There is much need to develop graphics that will guide our intuition in the search of "interesting" patterns, present in perhaps only a tiny minority of cases lost in a huge data set. Some useful tools are being developed, but the need for more is sorely felt. This is specially important because in the typical data mining situation "interesting" is not predefined, and exploratory (perhaps, graphical) tools must bear the brunt of the task.

I liked very much the industrial example. I have a minor qualm with parallel box plots supplemented with an "envelope", as in the author's Figures 6 and 8. It seems to me that the use of a shaded envelope embracing a number of cases in the population hides an important feature : the number of cases present and their likeness. For instance, the envelope in Figure 6 could be the outcome of $k - 1$ cases clustered ant the bottom edge and a single one accounting for the top edge, or else the k cases could populate the envelope more uniformly.

This is information that we do not want to loose. It seems better to me to resort to ordinary linked box plots, in which each case in the envelope is represented on its own. If there is a large number of them, the visual effect degrades to a shaded envelope, but otherwise some structure may still be visible.

A minor comment while still looking at graphs with parallel coordinates is that not only the order (of coordinate axes, box plots, etc.) can be changed. Permutation is basic to enhancing the ability of the graph to convey the right information, but spacing may help too. [3] contains a whole chapter devoted to designing profiles with variable order *and* spacing between the parallel coordinate axes, aimed at enhancing the linearity of the profiles or reducing the number of crossings. It is a pity that such work is largely ignored in applied work when the algorithm used is fairly simple and easy to program.

Incidentally, much the same can be said regarding the *bond energy* algorithm (see [5]), conceivably of great help in permuting a graphical matrix such as those advocated by [1]. While a human can always do a better job than the computer, in realistic situations with time constraints such algorithms can be life savers to the graphical analyst.

I wish to add only three short additional comments, largely unrelated to Valois article. First, there is now a wealth of software allowing even the unsophisticated user to do a fairly good job. There are no longer any excuses if we fail to meet Tukey's excellence requirement, quoted by Jean-Paul Valois in his introduction. Further, these software tools are within the reach of everyone, with some respectable packages – like R; see [4] – free to the user.

Second, it seems to me that, important as the static graphics are – for they will always be the cornerstone of successful communication of ideas –, the emphasis is now switching to dynamic graphics, which afford unprecedented ease in looking at data in a variety of ways that we could only dream of a few

years back. Again, fairly capable software exists, even free for the grab, like Xgobi (see [6] for instance) or its successor Ggobi.

To close I would like to add that increasing awareness of the role to be played by graphics in data analysis seems to me of the utmost importance. Consequently I acclaim the editor's decision to publish the article by Jean-Paul Valois, and the implicit judgement that this matters deserve a greater share of attention from the profession.

REFERENCES

BERTIN J., *La Graphique et le traitement graphique de l'information.* Flammarion, 1977.

CHERNOFF H., The use of faces to represent points in k-dimensional space graphically. *J. Amer. Statist. Ass.*, 1973.

HARTIGAN J.A., *Clustering Algorithms.* Wiley, New York, 1975.

IHAKA R. and GENTLEMAN R. R : a language for data analysis and graphics. *J. of Comp. and Graphical Stats.*, 5 :299–314, 1996.

MCCORMICK W.T., SCHWEITZER P.J., and WHITE T.W., Problem decomposition and data reorganization by a clustering technique. *Operations Research*, 1972.

SWAYNE D.F., COOK D., and BUJA A., XGobi : Interactive dynamic data visualization in the X-Window system. En http ://www.att.research.com/areas/stat/xgobi.

WAINER H., *Visual Revelations.* Copernicus, New York, 1997.