JOURNAL DE LA SOCIÉTÉ FRANÇAISE DE STATISTIQUE

ADALBERT F. X. WILHELM

Discussion and comments. Approche graphique en analyse des données

Journal de la société française de statistique, tome 141, n° 4 (2000), p. 87-91

<http://www.numdam.org/item?id=JSFS_2000__141_4_87_0>

DISCUSSION AND COMMENTS

Approche graphique en analyse des données

Adalbert F.X. WILHELM [1]

Graphical methods have not always been accepted in statistics. Having seen tremendous ups and downs in their history, graphics experienced a strong resurgence as a cornerstone in exploratory data analysis mainly due to the work of J.W. Tukey in the 1970s. Since then statistical graphics were more widely used not only for presentation purposes but also for exploration. The ubiquitous use of computers offers easy creation of statistical graphics and led to an explosion in the number of published info-graphics (also called propaganda graphs).

Jean-Paul Valois gives a profound summary of the history of statistical graphics pointing out the great streams of graphical evolution without getting lost in details. He is to be congratulated for his integrating neurophysiological research results as well as conclusions from the cognitive sciences to generate a typology of graphics. As he states this typology does not aim neither to cover all graphics currently in use nor to be the only possible one. There have been a couple of typologies around but Valois' point is to be strengthened that an organisation of graphical displays along the dimension of the graphic, as was often done previously, is no longer suitable. His typology takes three main principles into account : the number and types of variables, the question to solve, and the coordinate system used in the graphic. By doing this, he covers a lot of graphics for both categorical and continuous data but excludes some important plots for multi-dimensional categorical data, e.g. the class of mosaic plots which uses neither a cartesian coordinate system nor a parallel one. Mosaic plots are constructed by a hierarchical nesting scheme that can be efficiently used to visualize multivariate contingency tables.

Although the bar chart – the basic plot for univariate categorical data – was already used by Playfair two hundred years ago and although some special graphical methods have been developed for particular types of multivariate contingency tables, e.g. the fourfold display for $2 \times 2 \times k$ tables (Fienberg 1975, Friendly 1994), such plots are not readily available in standard software, and they are not yet well-known. Instead of generalizing univariate plots to two and more dimensions, recent enhancements of graphical tools have focussed on the artistic side, for example by introducing misleading three dimensional aspects in two-dimensional graphs. These attempts neither improve the ease of interpretation nor do they provide an extension to higher dimensions.

1. School of Humanities and Social Sciences, International University Bremen, P.O. Box 75 05 61, 28725 Bremen, Germany; e-mail : a.wilhelm@iu-bremen.de

Mosaic plots as introduced by Hartigan & Kleiner (1981) are a graphical analogue to multivariate contingency tables and show a contingency table's frequencies as a collection of rectangles whose areas represent the cell frequencies. The construction of mosaic plots is hierarchically and resembles the way multi-way tables are often printed on paper. For tables, rows and columns will be recursively split to include more variables. In the same way we split up horizontal and vertical axes of the mosaic plot recursively to obtain tiles that represent the cells in the contingency table. The area of each tile is chosen to be proportional to the observed cell frequency. Thus, mosaic plots are a multi-dimensional extension of divided bar charts. In their standard form mosaic plots actually combine both the parallel coordinate system and the cartesian one since every second variable is on a parallel axis and two neighboring variables are on orthogonal axes. Mosaic plots give an overview of the distribution of the total sample under investigation and at the same time individual subgroups can be compared. Cells with a high frequency will immediately strike into the analysts eye. But still it is hard to discern areas that do not differ very much. Comparisons within a row or column are straightforward since then one side will be the same for all cells, but comparing two rectangles that are neither in the same row nor in the same column might prove impossible. Additionally, a mosaic plot reflects the multi-dimensional relationship between the cases which can not be seen in the multiple bar chart view that is presented in the typology of Valois for n-dimensional categorical data.

As Valois points out matrix and array layouts have been proposed to extend uni- or bivariate displays to higher dimensions, e.g. trellis displays, co-plots and scatterplot matrices of jittered observations. By conditioning we partition the entire sample into subsamples and show a series of similar displays which we want to compare. In principle this strategy can be applied to any univariate display for categorical data. The quality of such matrices will then depend on how easy it is to make good comparisons between the cells of the matrix. A matrix of pie charts for example would only be a space saving arrangement without any multi-dimensional information. A comparison between two pie charts would be hard because the angles are varying, and, thus, corresponding categories might be drawn at different positions within the circle. A shortcoming which is not shared by other displays.

As one of the three main tasks to solve by graphics Valois names comparing the data to a model. The mosaic plot is best suited for assessing the quality of models for categorical data. If we assume that our data stems from independent variables then all tiles in the mosaic plot will align, because the side lengths of the tiles will then be completely determined by the marginal counts. In other words, deviation from an aligned pattern in the mosaic plot indicates deviation from the independence model. With three or more variables a variety of independence structures can occur. Each model-type – mutual independence, conditional independence, partial independence – shows a different and particular shape in the mosaic plot. A detailed description on how mosaic plots can be used for generating models is given in Theus & Lauer (1999). Colors or shading can be used to add residual information for such models either to the mosaic plot of the observed counts or to the mosaic

plots of the expected counts. The structure of large residuals gives a clear hint which interaction terms should be added to the model and which could possibly be dropped. Mosaic plots can become quite complex and they require a lot of experience to reveal their information. Much progress can be achieved by adding interactivity, not only to mosaic plots (Hofmann 2000), but to all kinds of plots.

Interactivity has evolved to be one of the most desirable characteristics of an up-to-date software package. Almost all developers claim that their packages are highly interactive. The close relationship between interactive graphics and the computer brings up a very important criterion which in my opinion should be added to the classifying points of Valois' typology : Which media is to be used to present the graphic and how close are creator and spectator of a graphic working together. Valois' typology assumes a strict separation of the process of producing a graphic from the process of interpreting it. This assumption is valid whenever graphics are used for presentation purposes on paper. However, the development of dynamic and interactive statistical graphics in the 1980's switched graphics from a result presenting device to an analytic tool. Plots changed their character from formerly being a final product to now being a temporary tool that can be modified and adapted according to the situation by simple mouse clicks or keyboard commands. Information overload that would prevent perception can be hidden at the first stage and made available on demand by responding to interactive user queries. Unusual observations, for example, can be easily spotted in graphics, identified by an interactive query, and then isolated for special treatment. Interactivity means that it is no longer necessary to encode all information in one plot because it is easy to receive additional information from the plot by interactive queries. In statistical consulting, the most use can be drawn from interactive statistical graphics when the client is sitting next to the consultant and the two work closely together in creating and interpreting graphics.

Valois treats the aspect of dynamic and interactive graphics rather cursory. He is not to be blamed for that because there is a huge confusion and disagreement about the definitions and meaning of interactivity, even within the statistical graphics community. Swayne & Klinke (1999) reported the results of a questionnaire that had been launched within the community about the use of interactive statistical graphics and they have expressed surprise about the different understandings of this term. They suggest using the terms direct and indirect manipulation of graphs" for describing the work in that field.

The 'Dictionary of Computing' (*Dictionary of Computing* 1991) defines '**interactive**' as "*a word used to describe a system or a mode of working in which there is a response to operator instructions as they are input. The instructions may bepresented via an input device such asakeyboard or light pen, and the effect is observable sufficiently rapidly that the operator can work almost continuously*". In the 'Computer Dictionary' *(Computer dictionary* 1994) *interactive graphics* is defined as "*a form of computer use in which the user can change and control graphic displays, often with the help of a pointing de-*

vice such as a mouse or a joystick. Interactive graphics is used in a range of computer products from games to computer-aided design (CAD) systems." Thus, two main characteristics of interactive graphics systems are the speed in which the system reacts to user instructions and the direct user control over the graphic displays. These two characteristics have been the ingredients for the definition of dynamic graphical methods given by Cleveland & McGill (1988) : "direct manipulation of graphical elements on a computer screen and virtually instantaneous change of elements". Speed is a necessary feature of an interactive system but it is in no way sufficient. Almost all software tools that are currently available react almost instantaneously to actions caused by the user and interactivity in this sense has become a standard requirement for any modern software. To base a decision on whether a software system is interactive or not only on technical speed measurements ignores the fact that human users adjust the amount of time that they are willing to wait for a response to the difficulty of the desired action. While asking for simple graphical changes a user will typically want the update within a small portion of a second. For complex tasks he/she will accept a longer response time. It is important that the reaction comes fast enough so that users do not have to interrupt their train of thought. Huber (1988) corrected the term dynamic graphic to high-interaction graphic. Highly interactive statistical graphics are not only the result of a technical development in computer science they are also the product of research and experience of statisticians and data analysts. They allow the user to grab the data, to ask questions as they arise and to search through a body of data to find interesting relationships and information.

The second characteristic of interactive graphics involves the choice of a user-interface. Although the choice of user-interface is mainly determined by the hardware used – and choosing the hardware is often more a philosophical question than a matter of quality and power – there is the general trend to unify user interfaces. More and more graphical user interfaces (GUI) are replacing command-line interfaces. Programming and batch interfaces are no longer asked for because they hamper interaction. High-interaction graphics are in majority based on GUI's but using a command-line interface does not exclude interaction per se.

To specify the general demand for speed and direct user control for interactive statistical displays I require that highly interactive statistical graphics software must be able to immediately respond to the following change and control commands created by the user :

• **Scaling :** Perception of graphical displays strongly depends on the scale. Since there are no unique choices, statistical software should provide the user with tools to flexibility change plot scales.

• **Interrogation :** Graphics should not be overloaded. On demand additional information must be available directly from the graphic.

• **Selection :** Selecting subgroups and focusing on specific data points help to reveal structure in the data set. A wide variety of tools to select groups

of points from graphical representations is needed to perform sophisticated analyses.

• **Projection Views :** Paper and screen are unfortunately restricted to two dimensions, and the human vision system is trained only for the three-dimensional world. Dimension reduction techniques are applied to produce low-dimensional views. A rapid, dynamic and smooth change of projection views is then needed to show as much of the multivariate structure as possible.

• **Linking :** Full interactivity is only achieved when selection is not restricted to a single display but propagated to other plots. This means that all displays are connected and that each view of the data shows each case consistently. Linking is the key concept of interactive statistical graphics, it builds up a relation between measurements of various variables, between different graphical representations as well as between raw data and models. These links can also perform different functions – the standard one is highlighting, others are color encoding or hiding.

Interactivity not only means that the user can interact with the data, but also that the results from the changes made by the user can be seen instantaneously. A rapid and responsive interaction facilitates active exploration in a manner that is inconceivable with static displays. Users can start to pose " What if " queries spontaneously as they work through a task. Therefore, interactive displays not only offer the possibility of comparing resulting static views of different aspects of the data, they even encourage to draw conclusions from the way things are changing. Unfortunately, this aspect cannot be shown in a written paper, this can only be seen live and on-line with a computer.

REFERENCES

CLEVELAND W. S. & MCGILL M. E., eds (1988), *Dynamic Graphics for Statistics*, Wadsworth & Brooks/Cole, Pacific Grove, CA.

Computer dictionary (1994), 2nd edn, Microsoft Press, Redmond.

Dictionary of Computing (1991), 3rd edn, Oxford University Press, New York.

FIENBERG S. E. (1975), 'Perspective Canada as a social report', *Social Indicators Research* **2**, 153-174.

FRIENDLY M. (1994), 'Mosaic displays for multi-way contingency tables', *Journal of the American Statistical Association* **89**, 190-200.

HARTIGAN J. A. & KLEINER B. (1981), Mosaics for contingency tables, *in* W. Eddy, ed., 'Computing Science and Statistics : Proceedings of the 13th Symposium on the Interface', Springer, New York, pp. 268-273.

HOFMANN H. (2000), 'Exploring categorical data : interactive mosaic plots', *Metrika* 51(1), 11-26.

HUBER P. J. (1988), Comment on 'Dynamic Graphics for Data Analysis', in Cleveland & McGill (1988), pp. 55-57.

SWAYNE D. F. & KLINKE S. (1999), 'Introduction to the special issue on interactive graphical data analysis : What is interaction?', *Computational Statistics* **14**, 1-6.

THEUS M. & LAUER S. R. (1999), 'Visualizing loglinear models', *Journal of Computational and Graphical Statistics* **8**(3), 396-412.