A Strategy for Automating the Presentation of Statistical Graphics for Users without Data Visualization Expertise
A Position Paper

Pere Millán-Martínez[1] and Pedro Valero-Mora[2]
[1]*Universitat de València, València, Spain*
[2]*Department of Methodology of the Behavioural Sciences, Universitat de València, València, Spain*

Keywords: Statistical Graphics Taxonomy, Data Visualization, Automatic Presentation, Visual Data Analysis, Graphic Literacy.

Abstract: The growing need to convert the data in databases into knowledge for a public without data visualization expertise requires the ever more precise selection of graphics to be presented to the user for consideration. This can be achieved through a more detailed characterization of the data as well as the data visualization task that the user wishes to accomplish. One way to limit the number of possible graphics based on the data is to characterize the multiple properties that can be described for each variable represented by a column of data. This paper presents seven dimensions with their respective levels that can serve as a framework for classifying statistical graphics such that their effectiveness in performing a given task may then be evaluated.

1 INTRODUCTION

Open data policies have made an enormous amount of data available that needs to be converted into knowledge, and the most effective way of doing it is with graphics that show the properties and relationships of the variables in a dataset. If the goal is to help users unfamiliar with data visualization to better interpret the data, it is necessary to define a strategy to adequately limit the number of graphic representations that are automatically presented for the users consideration, without requiring users to predetermine the characteristics of the graphic they are looking for.

There have been many attempts to construct an automated system to present statistical graphics to users, but none of the strategies employed have received broad acceptance because they suggest only one supposedly ideal solution or a broad, unranked selection of possible graphical solutions. These strategies can be classified in terms of how they address the characteristics of the data, the characteristics of the user, the limitations of the hardware and the characteristics of the sought after graphic. Figure 1 summarizes the types of inputs considered by graphic's automation systems.

The automated selection of graphics via the characterization of the data that Kamps (1999) calls "functional design", is capable of considering different aspects of the data: the characteristics of the variables taken separately; the relationships between variables; the structure of the data; and the number of variables to graphically relate and represent. If the characteristics of the variables taken separately cannot be implicitly deduced from the data, the users ability to specify these characteristics is constrained by the number of variables in a dataset and the dimensions of the variables being considered. A much greater effort on the part of the user is required to describe the relationships between the variables if these cannot be implicitly deduced form the data, given that the number of relationships increases exponentially with the number

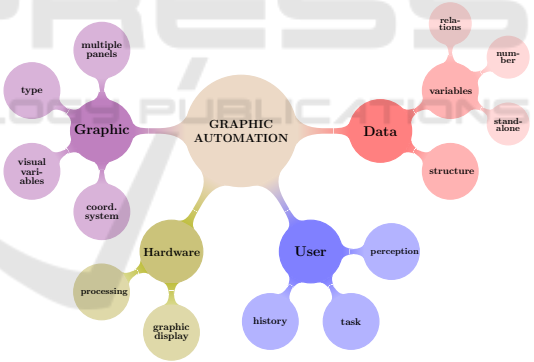

Figure 1: Types of inputs considered by graphic's automation systems.

MillÃąn-MartÃnez P. and Valero-Mora P.
A Strategy for Automating the Presentation of Statistical Graphics for Users without Data Visualization Expertise - A Position Paper.
DOI: 10.5220/0006220702940298
In *Proceedings of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2017)*, pages 294-298
ISBN: 978-989-758-228-8

of variables to be related. Additionally, selecting the graphic types to be presented based on the structure of the data may not be useful given that the structure of data may be easily changed without losing the information.

With respect to the characteristics of the user, we may consider: human perceptual capabilities, or what Kamps (1999) calls "perceptual design"; the questions the user is seeking answers to, or what Casner (1990) calls "task-based graphic design"; and finally user preferences as deduced from a users graphic selection history, which is known as a "recommender system". Graphic automations that rely on any of these user characteristics require that the gamut of possible graphic representations be previously limited based on the characteristics of the data; otherwise the system might evaluate and suggest graphics that cannot be constructed from the data.

Hardware limitations have more of an impact on aesthetics and usability than in the selection of the graphic types to be presented, especially considering that the users of databases tend to work on desktop computers with a broadband internet connection; therefore, these limitations can be overlooked a priori when classifying statistical graphics.

With respect to the characteristics of the sought after graphic, an automated system could require the user to define the coordinate system, the visual variables to be used, the more or less generic graphic type, or its possible decomposition into multiple panels. Defining any of these, however, requires a certain degree of data visualization knowledge; therefore, this strategy is not suited for users who are not all that familiar with data visualization.

When thinking about an automated graphic selection system for users who are not that familiar with data visualization, the strategy is to identify those graphics that could possibly work based on the data and thus restrict the selection to only those that can perform a specific purpose with the greatest effectiveness.

Of the various aspects of data characteristics, the characterization of variables taken separately is a very effective method because it allows us to breakdown the problem of what types of graphics to present into three parts; the number of variables to be represented; the dimensions to be considered to characterize those variables; and the mutually exclusionary levels to be considered for each dimension. However, the strategies used thus far by automated graphic systems have not tried to classify statistical graphics based on a multidimensional characterization of the variables taken separately.

The aim of this paper is to identify the various dimensions of characteristics that can be described for a variable as represented in a data column, and that can then be used to limit the gamut of graphics to be evaluated prior to presenting it to the user. Before doing so, we will review the state of the art in automated graphic selection and then present a list of dimensions with mutually exclusive levels that make it possible to limit the set of possible graphic types.

2 PREVIOUS WORKS

Bertin (1967, p.34) refers to the components of the various variable measurement scales as levels of organization, and he distinguishes three such levels: qualitative for those concepts that can simply be differentiated; ordered for those variables that have an inherent sequence; and quantitative for those with a quantifiable quality. Another characteristic Bertin uses to limit the gamut of acceptable graphics is the length of a variable, defined by Bertin (1967, p.33) as the number of divisions that make it possible to identify them as short variables if their length is equal to or less than four, long variables if their length is greater than 15, and medium variables for those with lengths between five and 15. This classification yields two dimensions with three levels each. Ware (2004, p.24) relates Bertins measurement scales with those of Stevens (1946) and distinguishes 4 levels of variable attributes: nominal, ordinal, interval and ratio.

Another classification is proposed by Bachi (1968, p.10), who characterizes variables according to sequence type, such as linear, circular, geographical and unordered qualitative sequences. Further, Bachi identified subcategories for linear sequences, differentiating quantitative, temporal and qualitative linear sequences. Bachi also identified subcategories for geographic sequences, distinguishing between distribution and movement. This classification yields one dimension with seven levels.

The BHARAT system (Gnanamgari, 1981), a pioneer in the automated presentation of graphics, uses multiple dimensions to characterize variables, such as: continuity, totality, cardinality (defined as the number of unique values for a variable), units and range. From these five dimensions, Gnanamgari identifies levels for only the first two, which are dichotomous, and for the other three he establishes ad hoc rules to evaluate the graphics to be presented to the user.

Other systems, such as APT (Mackinlay, 1986), BOZ (Casner, 1990) and Vista (Senay and Ignatius, 1994) also use Bertins levels of organization, but not his variable length classification. Thus, their charac-

terization is unidimensional with three levels. This characterization has variants, such as that used in SAGE (Roth and Mattis, 1990), which subdivides the ordinal and cardinal levels according to whether they refer to amounts or reference values, and adds a second dimension that refers to the fundamental physical magnitudes of time, space, temperature and mass. The NSP system (Robertson, 1990) distinguishes variables according to whether they are nominal or ordinal. Further, for nominal variables, it distinguishes between those with multiple values and those with one single value, and for ordinal variables, it distinguishes between discrete and continuous values, thus yielding only one dimension with 4 levels. The Polaris system (Stolte et al., 2002) only distinguishes between ordinal and quantitative levels, while the Tableau system (Mackinlay et al., 2007), which derived from Polaris, distinguishes between the categorical level, with three sublevels based on whether the values are normal, dates or geographical units, and the quantitative level with two sublevels, based on whether the variables are predictor or response variables, thus yielding one dimension with 5 levels. Lastly, the VizRec recommender system (Mutlu et al., 2015) distinguishes 3 levels based on whether the variables are categorical, temporal or numeric.

3 CHARACTERIZATION OF VARIABLES TAKEN SEPARATELY

Usually the variables stored in a database are characterized based on the memory reserved by the computer system to store variable values. Thus we end up with, for example, datatype classifications such as Boolean, text strings of a predefined maximum length, numeric values that can be whole or real numbers with an established decimal point precision, etc. The characterization we propose is based on the properties that can be described for each column of values for a variable, which may affect the selection of one graphic over another.

3.1 Characterization Dimensions

Graphic Measurement Scale. Variables can be classified according to measurement scale, and we can distinguish between qualitative and quantitative variables, the difference being that the values of the former cannot be summed up. Among the qualitative variables, we can distinguish between unordered and ordered, the difference being that the values of the latter maintain a lesser-to-greater relationship. Among the quantitative variables, we can distinguish three levels: conventionally bounded scalar values, which are measured on an interval scale, such as the temperature in Celsius degrees; scalars bounded on one end, which are measured on a ratio scale, such as a persons age; and lastly scalars bounded on two ends, which are measured on an absolute scale, such as the probability of a certain event occurring, which is bounded between zero and one. A variable can transform its measurement scale in the direction of greater to lesser restrictions; for instance, scalars bounded on two ends can be transformed into scalars bounded on one end, and these can in turn be transformed into conventionally bounded scalars, which can then be transformed into ordered categories that can then be transformed into unordered categories. The graphic measurement scale allows us to distinguish, for example, between a simple point graph, a simple bar chart and a dichotomous pie chart.

Cardinality Factor. This dimension relates the cardinality of a variables data (meaning the actual number of unique values in the data) with the cardinality of the variable (this being the potential number of observable values). This dimension distinguishes between sample type values (when the values in the data are few compared to those that can be potentially observed or of interest) and population type variables (when the values in the data coincide or practically coincide with those that can be potentially observed or of interest). Population type values are typically factors, categories or equidistant intervals of quantitative variables within an interval of interest. The cardinality factor allows us to distinguish, for example, between a histogram and a point graph with drop lines that connect points to one of its axes.

Sequentiality. This dimension addresses the possibility that the order in which a variables values appear in a data column contains information about the sequence in which they were observed. Typically, this order is found in the “date” variable of a temporal series, but information on sequence can be contained in any data column with sequentially ordered values. Sequentially ordered data columns can be transformed into non-sequentially ordered data columns when the position of the data in the column is irrelevant in the graphic representation to be suggested. Sequentiality allows us to distinguish, for instance, between a scatter plot and a scatter plot with points connected by a line that follows a sequence.

Cyclicality. Cyclicality, meaning the domains cyclic or non-cyclic character, concerns only quantitative and ordered qualitative variables, since unordered qualitative variables lack an inherent order. Periodic variables can sometimes be transformed into aperiodic variables and vice versa, as is the case, for example, with the variable "time" depending on the data analysis being performed. The characterization of a variable as cyclic is especially important when determining whether to suggest graphics with polar, cylindrical and spherical coordinate systems.

Explicitness. A variables explicitness helps us distinguish the explicit level, when the value scale must be represented graphically, from the ambiguous level, when the scale should not be represented graphically, but other characteristics should be, such as the number of values or the unique values contained in a data column. An explicit variable can be transformed into an ambiguous variable by simply omitting the scale values, and vice versa. A clear example of a variable represented in an ambiguous manner is the various observations of value pairs for two variables in a scatter plot, where each point can be discerned, but not the order that it occupies in the data column nor the name of the corresponding informant.

Variable Length. This dimension was defined by Bertin (1967) as the number of unique values that it is useful to identify, and that influences the use of one visual variable over another and the use of translations, rotations and reflections. As previously noted, Bertin distinguishes between short, medium and long variables, but automated graphic systems tend to establish ad hoc rules when evaluating what graphics to present. The levels that we believe are the most relevant are: variables with a unique value, such as the name of the winner of a race or an observed temperature, since many graphical methods were developped to represent single observations; dichotomous variables that are very common in datasets and may suggest the use of reflections to facilitate the comparison between pairs of observations; and variables with lengths greater than two. A length threshold of between 5 and 12 could also be established to identify, for example, variables susceptible to being represented via multiple panels or a retinal visual variable instead of a spatial visual variable. The ideal number of levels in this dimension depends on the accuracy that is intended in a graphics automation system.

Georeferencing. This dimension is applicable only to qualitative variables and considers the possibility that these categories can be linked with geospatial points, lines and polygons. The characterization of a variable as georeferenced allows us to present values as postal, census and political units on a map.

4 DISCUSSION

The dimensions presented above allow us to divide the domain of known graphics in ever more limited subgroups, but if we consider the possible transformations between levels for each dimension, the gamut of possible graphics expands. Using the graphic measurement scale, for example, enables us to divide single-variable graphics into five groups, two-variable graphics into fifteen groups, and three-variable graphics into 35 groups. If we also use the cardinality factor, the possible combinations for single-variable graphics increases to 10, for two-variable graphics it grows to 55, and for three-variable graphics it jumps to 220. Thus, the use of successive dimensions allows us to more precisely narrow the set of appropriate graphics.

The multidimensional characterization of variables taken separately has few comparable antecedents. The characterization implemented by the BHARAT system comes closest. It considers continuity, totality, cardinality, units and range, but identifies dichotomous levels for only the former two dimensions and establishes ad hoc thresholds for the other dimensions when evaluating which graphics to present to the user. Because the automated presentation of one graphic or another requires the classification of graphics based on the dimensions of the variables represented, we believe that it is necessary to unambiguously define each level and limit them to a reduced number. It is for this reason that we have not introduced dimensions such as, for example, units of measurement; doing so would produce as many levels as combinations of fundamental magnitudes.

Once graphics have been classified according to a multidimensional characterization of the data, we can then further reduce the gamut of graphics to be presented to the user by using a multidimensional characterization of the tasks to be performed, as proposed by Schulz et al. [2013]. In order to evaluate the task performance efficiency of a graphic drawn from a gamut of possibilities based on the data, it is necessary to include information provided by the user, whether via experiments that measure a users task performance efficiency with respect to a gamut of graphics, the score users give to the graphic-task binominal, or a users graphic selection history when that information

is available for the task being performed.

5 CONCLUSION

This article critiques the strategies employed by different automated graphic systems and focuses specifically on their characterization of variables taken separately. We also identify as many as seven dimensions of attributes that can be used to describe a column of data, and that help determine the appropriateness of one graphic over another, and, consequently, can be used to limit the gamut of graphics to evaluate prior to presenting them to the user. Of these seven dimensions, one has five levels, another four, and the rest are dichotomous; thus, it is not difficult to characterize the variables if their properties cannot be implicitly deduced from the data.

This paper presents a framework for classifying statistical graphics without requiring that the user predetermine the characteristics of the graphic being sought and without considering hardware limitations. Ideally, this classification should be complemented with information about the effectiveness with which the user performs various tasks. In this way, the gamut of graphics presented to the user can be reduced to just one or only a few possible solutions.

ACKNOWLEDGEMENTS

This work has been developed during a research residency facilitated by Michael Friendly in his DataVis laboratory at the Department of Psychology of York University (Toronto, ON, Canada).

REFERENCES

Bachi, R. (1968). *Graphical rational patterns: A new approach to graphical presentation of statistics*. Transaction Publishers.

Bertin, J. (1967). *Sémiologie graphique*. Mouton, Paris.

Casner, S. M. (1990). Task-analytic design of graphic presentations. Ph.D. dissertation.

Gnanamgari, S. (1981). Information presentation through default displays. Ph.D. dissertation.

Kamps, T. (1999). *Diagram Design: A Constructive Theory*. Springer Berlin Heidelberg.

Mackinlay, J. (1986). Automating the design of graphical presentations of relational information. *ACM Trans. Graph.*, 5(2):110–141.

Mackinlay, J., Hanrahan, P., and Stolte, C. (2007). Show me: Automatic presentation for visual analysis. *Visualization and Computer Graphics, IEEE Transactions on*, 13(6):1137–1144.

Mutlu, B., Veas, E., Trattner, C., and Sabol, V. (2015). Vizrec: A two-stage recommender system for personalized visualizations. In *Proceedings of the 20th International Conference on Intelligent User Interfaces Companion*, IUI Companion '15, pages 49–52, New York, NY, USA. ACM.

Robertson, P. (1990). A methodology for scientific data visualisation: choosing representations based on a natural scene paradigm. In *Visualization, 1990. Visualization '90., Proceedings of the First IEEE Conference on*, pages 114–123.

Roth, S. F. and Mattis, J. (1990). Data characterization for intelligent graphics presentation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 193–200. ACM.

Senay, H. and Ignatius, E. (1994). A knowledge-based system for visualization design. *Computer Graphics and Applications, IEEE*, 14(6):36–47.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, 103:677–680.

Stolte, C., Tang, D., and Hanrahan, P. (2002). Polaris: A system for query, analysis, and visualization of multidimensional relational databases. *Visualization and Computer Graphics, IEEE Transactions on*, 8(1):52–65.

Ware, C. (2004). *Information Visualization: Perception for Design*. Interactive Technologies. Elsevier Science, 2nd edition.