



Lawrence Berkeley Laboratory

UNIVERSITY OF CALIFORNIA

Physics, Computer Science & Mathematics Division

Presented at the International Congress on Applied Systems Research and Cybernetics, Acapulco, Mexico, December 12-14, 1980

AN APPLICATION OF FUZZY SET THEORY TO DATA DISPLAY

William H. Benson

December 1980

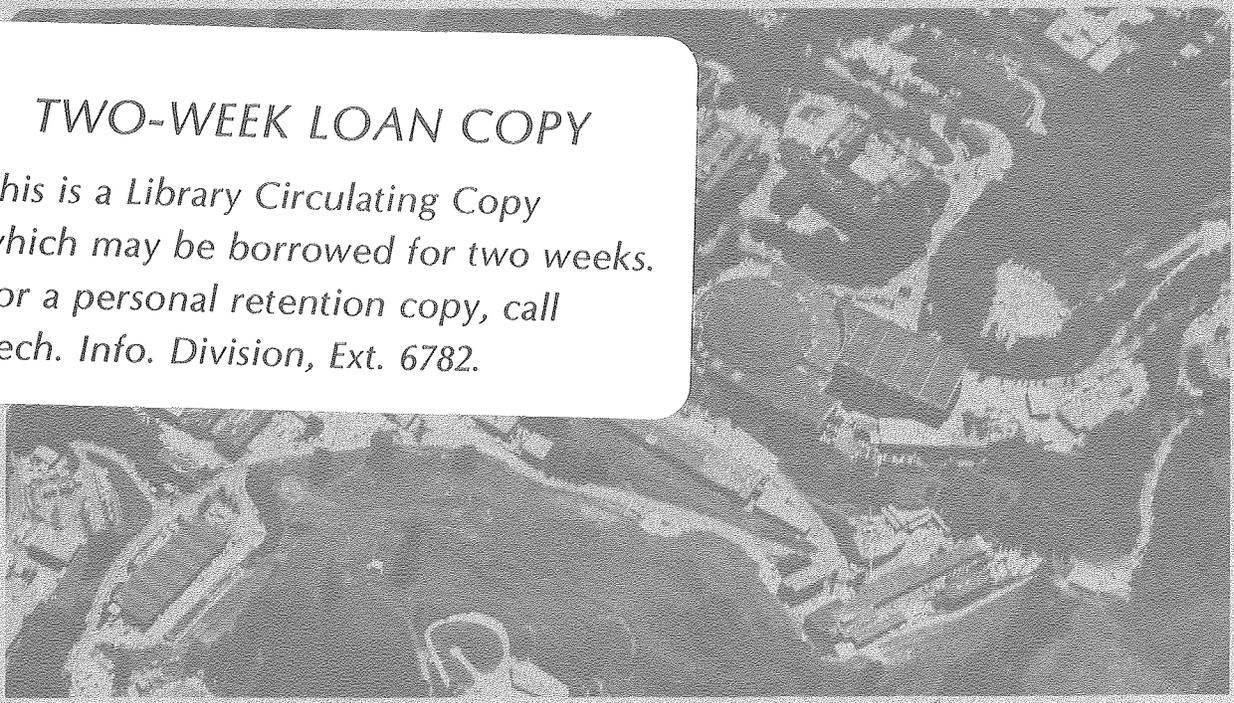
RECEIVED
LAWRENCE
BERKELEY LABORATORY

FEB 17 1981

LIBRARY AND
DOCUMENTS SECTION

TWO-WEEK LOAN COPY

This is a Library Circulating Copy which may be borrowed for two weeks. For a personal retention copy, call Tech. Info. Division, Ext. 6782.



LBL-11590 c.2

DISCLAIMER

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor the Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or the Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or the Regents of the University of California.

AN APPLICATION OF FUZZY SET THEORY TO DATA DISPLAY

William H. Benson
Computer Science and Applied Mathematics Department
Lawrence Berkeley Laboratory
University of California
Berkeley, California 94720

ABSTRACT

Categorization supports decision making, letting an analyst look at data from different perspectives and different levels of detail. An approach to data analysis is described in which membership in subjectively defined categories is modeled by the fuzzy nature of color categories and presented via computer graphics for visual inspection by the analyst.

KEYWORDS

Subjective categories; color computer graphics; decision support.

INTRODUCTION

An interactive computer graphics program is described in which basic notions from fuzzy set theory are used to help data analysts formulate and visualize subjective categories. The program has been developed in the application area of labor-related statistics. Where analytic tasks are not well defined, or data is incomplete or imprecise, it is often helpful to conceptualize data variables such as population or unemployment rates by subjective level. For example, an analyst interested in unemployment but concerned about statistical fluctuations due to small sample sizes might want to know where "unemployment levels are HIGH but population is NOT LOW". This phrase describes a subjective category. The term subjective refers to the analyst organizing the data relevant to the purpose at hand, and the deliberate blurring of category boundaries by linguistic expressions (such as HIGH, NOT LOW).

The value of simple graphic forms to represent statistical data has been apparent since the invention of the bar chart in 1786 (Beniger and Robyn, 1978). Labor statistics are regularly presented and analysed with graphic aids, and the question above can be quickly answered with a little mental effort, such as comparing lengths on bar charts. But for even simple tasks this effort can quickly become burdensome. H.A. Simon (1977) cites weaknesses and limitations in selecting and remembering information, and argues (as summarized in Kling, 1980) that "data and methods that help focus attention and evaluate choices improve the technical performance of a decision maker."

The present paper is concerned with a mode of analysis in which identification of category members and recognition of degrees of membership is the primary information. This information is presented graphically so that an analyst can focus attention on a region of interest in data space.

EXAMPLE

A brief example is presented in Fig. 1 to illustrate such an analysis. Figures are shown here in black, gray, and white but are described throughout this paper as if seen in color (red, orange, and yellow respectively).

The idea for developing this application of fuzzy set theory grew out of experience with management information reporting and data analysis needs at a U.S. Department of Labor regional office. One important need concerns monitoring the performance of employment and training programs, which try to match job seekers with job openings listed with the programs by local employers.

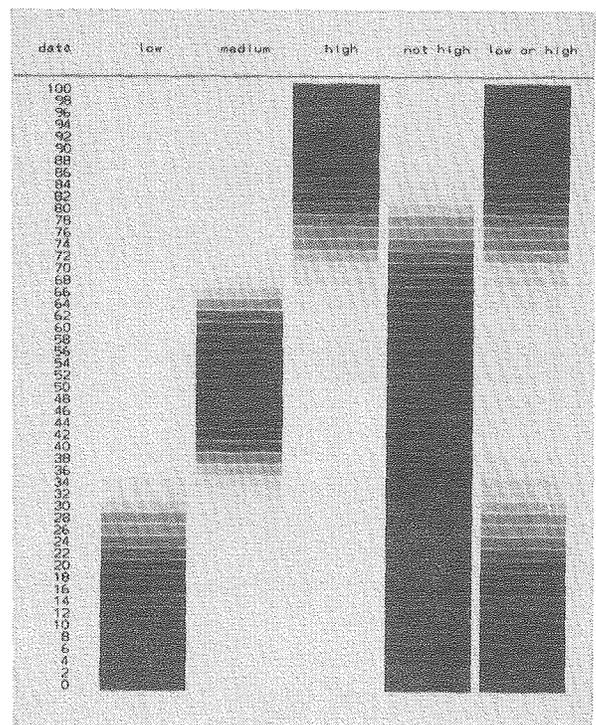
A commonly accepted measure of performance is the overall fill rate - the percentage of total job openings actually filled. In Fig. 1, data for job openings listed, openings not filled, and fill rate is broken down by job type so that industries and occupations where there is the greatest potential for improving the overall fill rate can be identified for followup (are employers insincere in listing jobs?, are training programs inadequate?, etc.).

An intuitive analysis, considering both ratios and counts of job positions, points to two kinds of jobs for followup: 1) those with low fill rates, but enough openings to make a difference in overall fill rate; and 2) those with many openings not filled, regardless of fill rate. The above conditions define a disjunctive category characterizing greatest potential for improved performance. Where and how well the data fit this characterization stands out in red and degrees of red. Electrical, communication, food service, and banking jobs all fit well, insurance less so, and quarrying just barely. Following are several observations about this example:

Jobs	Degree of fit	Openings	Not Filled	Fill Rate (%)
printing		275	110	60
machinery		362	146	60
electrical	■	687	471	31
pipelines		1	1	0
communication	■	771	695	10
food service	■	2989	1196	60
banking	■	625	539	14
insurance	■	399	335	15
postal service		1	0	100
air transport		32	10	69
metal mining		8	5	38
coal mining		2	1	50
oil/gas		8	3	63
quarrying		31	29	26
building		1429	122	91
construction		237	30	97
contractors		950	194	80
food products		732	45	94
textiles		92	44	52
apparel		187	48	74
wood products		890	124	80

XBB 800-14242

Fig. 1. Job service placement performance



XBB 800-14040

Fig. 2. Membership

The category is imposed on the data by the analyst, rather than determined by a cluster of related attributes in a particular data set (for example, a notion of "poor performance" may be relevant regardless how many fit.) The linguistic expression describing the region of interest in data space can be easily formulated, understood, and reformulated as needed. The category boundaries are necessarily imprecise by virtue of the mapping from linguistic terms onto the data (Hersch and Caramazza, 1976). In this example, just a rough idea of poor performance is sufficient to identify situations for followup. It is not necessary that poor performance be well defined and every case put in rank order. In general, deliberate blurring is a useful strategy for at least three reasons: undue precision is not needed for the purpose at hand; the data itself is imprecise; and the level of anxiety in decision making is reduced (Kochen, 1979).

The situation is one of decision support rather than decision making. Even though a great number of important considerations are absent from the chart (experience, intuition, politics, legal requirements, etc.) as well as economic and demographic variables affecting placement performance, attention is focused where improved overall performance is most possible (red and orange). For jobs where the overall fill rate cannot be significantly affected, data variation is irrelevant and distracting. This variation is collapsed and hidden in a single color (yellow).

Since degree of fit is explicitly represented, there is latitude for subjective interpretation of the display. Insurance and quarrying might be included or left out according to time and resources available for followup. An analyst could notice "any degree of red", or only "the best reds". Subjective interpretation is also discussed below for another example of performance analysis.

It may seem that the color scale information is superfluous and can just as well be recovered from a bar chart. Certainly the rows could be re-arranged to assist reading (by ranking on Fill Rate, checking those considered low for sufficient Openings Received, and ranking the remainder on Not Filled). In general, however, there is considerable cognitive effort and burden on short term memory involved in estimating the range of each data column; estimating a sub range such as HIGH or NEAR 100; matching data with sub ranges; and integrating matches across columns to form a composite impression of performance, making tradeoffs in importance and noting exceptions. Further, the same process must be repeated whenever the display is read again.

FUZZY SET MODEL FOR SUBJECTIVE CATEGORIES

A fuzzy set model has been incorporated into an existing program for analysis and display of tabular data. Data is viewed in terms of the standard statistical data structure of cases and variables, where each column is a homogeneous data set such as population or unemployment rate, and each row is a case, such as a county, with scores for each column variable.

From the analyst's point of view, categories can be formulated by expressions in linguistic variables. A linguistic variable, such as population, takes words rather than numbers as values. Primary terms such as HIGH, MEDIUM, and LOW are represented by fuzzy sets in the context of the particular data being considered. The vocabulary can be extended using linguistic hedges such as VERY, which are interpreted as operators on fuzzy sets. Expressions can be combined by logical operators of conjunction, disjunction and negation.

Expressions both determine a new fuzzy set and define a new attribute or variable for each data case. Given an expression, the program computes for each data item (case) the degree of membership in the fuzzy set. This new variable is thus the membership function for the category described by the expression. Membership is

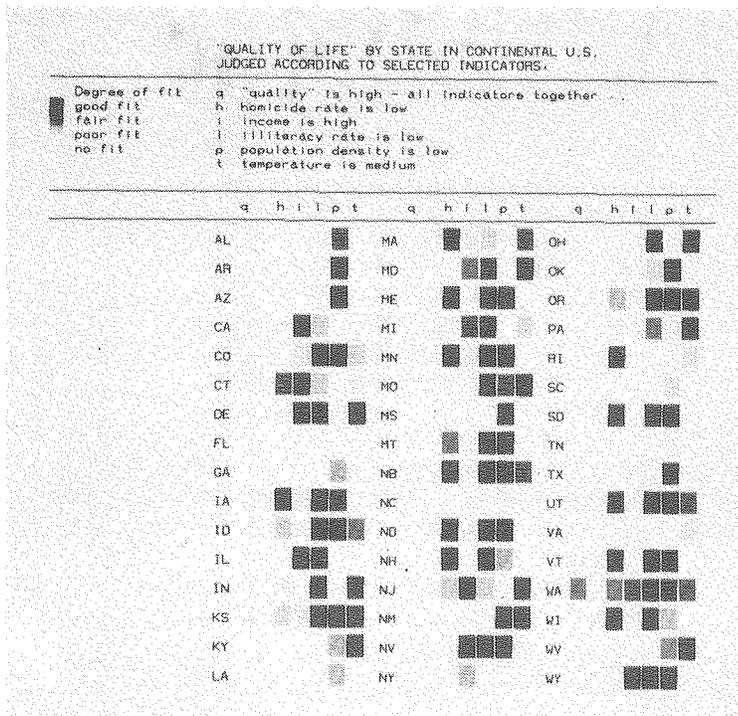
displayed according to a color scale so that where and how well the cases fit the category description can be determined by visual inspection.

Membership functions for the primary terms LOW, MEDIUM, and HIGH are illustrated in Fig. 2 by color scale instead of the usual function curves. These terms apply to the lower half, middle half, and upper half of the data range respectively to provide a coarse and partially overlapping coverage. Degree of red corresponds to degree of membership. For example, the highest values are the most red for the category HIGH in column 4. The values in the 90s are well described by term HIGH, those in the 50s and 60s can just barely be so described, and those less than 50 not at all. Definitions for logical operators are taken from Zadeh (1973). For membership functions f and g ,

NOT $f = 1-f$ f AND $g = \min(f,g)$ f OR $g = \max(f,g)$.

AN EXAMPLE OF PERFORMANCE ANALYSIS

Perceptual properties of color suggest that color scales can be constructed which effectively present information about category membership and support operations such as prototype matching. These topics are discussed below in terms of Fig. 3, in which states in the continental U.S. are evaluated according to five objectives (indicators of "quality of life"). Membership functions for the five indicator categories are derived from raw data as described above. Degree of fit to category membership is shown in black, gray, and white but is described below as if scaled by degrees of redness in the spectral sequence from red through orange to yellow. The display is organized so that it can be scanned like newspaper columns.



XBB 800-14041

Fig. 3. Evaluating multiple objectives by visual inspection

An analyst can compare states with one another and with an implied prototype row of all red chips to judge "quality is high". Since this term is not well defined, it is legitimate and often easier to evaluate choices in the context of a particular data set, rather than set up a model of rational behavior involving many judgments about exceptions and tradeoffs in importance for every conceivable situation.

If all objectives are of equal importance, so that no tradeoffs or exceptions are allowed, it is appropriate to combine them with the linguistic AND. The result is shown in the q columns. The row WA fits best, but others can also fit well if exceptions and tradeoffs can be made. CO,IA,IN,KS,OR also satisfy all the conditions, at least to some degree. At this point, the income objective (i) might seem less important, since except for WA, the above states just marginally satisfy it. The analyst might make an exception of income, provided a state met the other objectives well enough. On this basis NB and UT look better than any of CO,IA,IN,KS,OR. On the other hand, if the income objective is considered essential but a good score can compensate for any single exception, then CT,NJ,NV look best.

COLOR SCALES FOR CATEGORY MEMBERSHIP

Data can be coded for display by graphic variables such as size, texture, color, gray scale, shape, and orientation. How graphic variation is perceived can reveal relations in the data as well as imply relations that do not exist. Bertin (1967) describes four perceptual properties of graphic variables and their implications for data display.

- * Associative - a variable is associative when graphic items differentiated by it can be grouped together spontaneously and seen as similar. Color has this property. Most vegetation is seen as simply green and slight differences in hue ignored. Size is not associative because small items lack visibility. Since the largest stand out spontaneously, size variations are almost impossible to ignore.
- * Selective - a variable is selective when variations can be spontaneously identified and isolated from the background. For example, on standard tests for color blindness, a familiar pattern all in the same color (a letter) can be seen in what is otherwise a splatter of multi-colored dots. Shape does not have this property.
- * Ordered - when a natural order is obviously apparent. For example, differences in size are ordered from smallest to largest. Shape is not an ordered variable.
- * Quantitative - when variation can be compared on a ratio scale. Only variations of size can convey the relation of proportion in quantitative data.

Color allows a viewer to shift easily between two perceptual attitudes: association-disregarding variation in order to see similarities; and selection-distinguishing variation to isolate similar instances. These same attitudes are expected to figure in the analyst's consideration of the data. The first attitude enables the rows and columns to be seen as a homogeneous field of colored chips, all equally visible. Since rows are seen as essentially similar, they can be compared with one another or the prototype for color content.

The second attitude allows the viewer to attend to only some of the chips, say those sufficiently red, even though separated within a row or between rows. The ability to disregard spatial location, paying attention selectively to what is to be compared, significantly aids the task of comparing rows with one another, especially rows widely separated in the display.

A viewer shifts between these two attitudes in selecting various ranges of color to compare. For example, all colors with any degree of red can be associated together and confounded in perception so that color can be integrated visually across a row to form an impression of overall amount of red. This impression, qualified by noting exceptions in yellow, can measure how well the row matches the prototype. Again, to enhance contrast, the same orange can be associated with red in one context to complete a row of all red chips, and with yellow in another row already lacking enough red to match the prototype.

Although the color sensations of blue, green, yellow, and red are seen as very different and clearly unordered, it is also clear that there is a continuous gradation of hues from one to the next along the visible spectrum. It is commonly recognized in both perception and language that this gradation leads to categories of color and degrees of membership in color categories. Expressions such as a good red, an off red, slightly red, yellowish red, reddish yellow, etc. indicate the degree the corresponding colors approximate an ideal example of red. Kay and McDaniel (1975, 1978) have explicitly used fuzzy set theory to model the continuity of membership in color categories. The basic distinction between degree and kind of fit (some fit vs. no fit) is expressed well by color. For example, red contrasts with yellow while degrees of both red and yellow can be seen in intermediate colors.

Despite the precision with which linguistic expressions are evaluated, it seems more reasonable to regard values of degree of fit as rough indicators rather than accurate measurements. That is, it is inappropriate to distinguish small differences. With this in mind, degree of fit is clearly ordered, but lacks a definition of intensity which would give meaning to a comparison such as "twice as good a fit". Since ratio comparisons are implicit in variations of size, it would be misleading to use this graphic variable. Use of color can suggest a rough ordinal scale but discourage overly precise comparisons between nearby values.

SUMMARY

Examples have been presented to illustrate how subjective categories can support decision making. A color scale has been advocated to display category membership.

ACKNOWLEDGMENT

This work was supported by the Applied Mathematical Sciences Program of the Office of Energy Research, U.S. Department of Energy, under contract No. W-7405-ENG-48, and the U.S. Department of Labor.

REFERENCES

- Beniger, J. A., and D. L. Robyn (1978). Quantitative graphics in statistics: a brief history. American Statistician, 32, No. 1, 1-11.
- Bertin, J. (1967). Semiologie Graphique. Gauthier-Villars, Paris.
- Hersch, H. H., and A. Caramazza (1976). A fuzzy set approach to modifiers and vagueness in natural language. J. of Exp. Psych.: General, Vol. 105, No. 3, 254-276.
- Kay, P., and C. K. McDaniel (1975). Color categories as fuzzy sets. Working paper No. 44, Language Behavior Research Laboratory, University of California, Berkeley.
- Kay, P., and C. K. McDaniel (1978). The linguistic significance of the meanings of basic color terms. Language, Vol. 54, No. 3, 610-646.
- Kling, R. (1980). Social analyses of computing: theoretical perspectives in recent empirical research. Computing Surveys, Vol. 12, No. 1, 61-110.
- Kochen, M. (1979). Enhancement of coping through blurring. Fuzzy Sets and Systems, 2, 37-52.
- Simon, H. A. (1977). The New Science of Management Decision Making. Prentice-Hall, Englewood Cliffs, N.J.
- Zadeh, L. A. (1973). Outline of a new approach to the analysis of complex systems and decision processes. IEEE Trans. Syst., Man & Cybern., Vol. SMC-3, No. 1, 28-44.