

Lawrence Berkeley National Laboratory

Lawrence Berkeley National Laboratory

Title

AN APPLICATION OF FUZZY SET THEORY TO DATA DISPLAY

Permalink

<https://escholarship.org/uc/item/7w2409zn>

Author

Benson, William H.

Publication Date

1980-02-01



Lawrence Berkeley Laboratory

UNIVERSITY OF CALIFORNIA

Physics, Computer Science & Mathematics Division

To be published as a chapter in RECENT DEVELOPMENTS
IN FUZZY SET AND POSSIBILITY THEORY, Ed. R.R. Yager,
Pergamon Press, 1981

AN APPLICATION OF FUZZY SET THEORY TO DATA DISPLAY

William H. Benson

February 1981

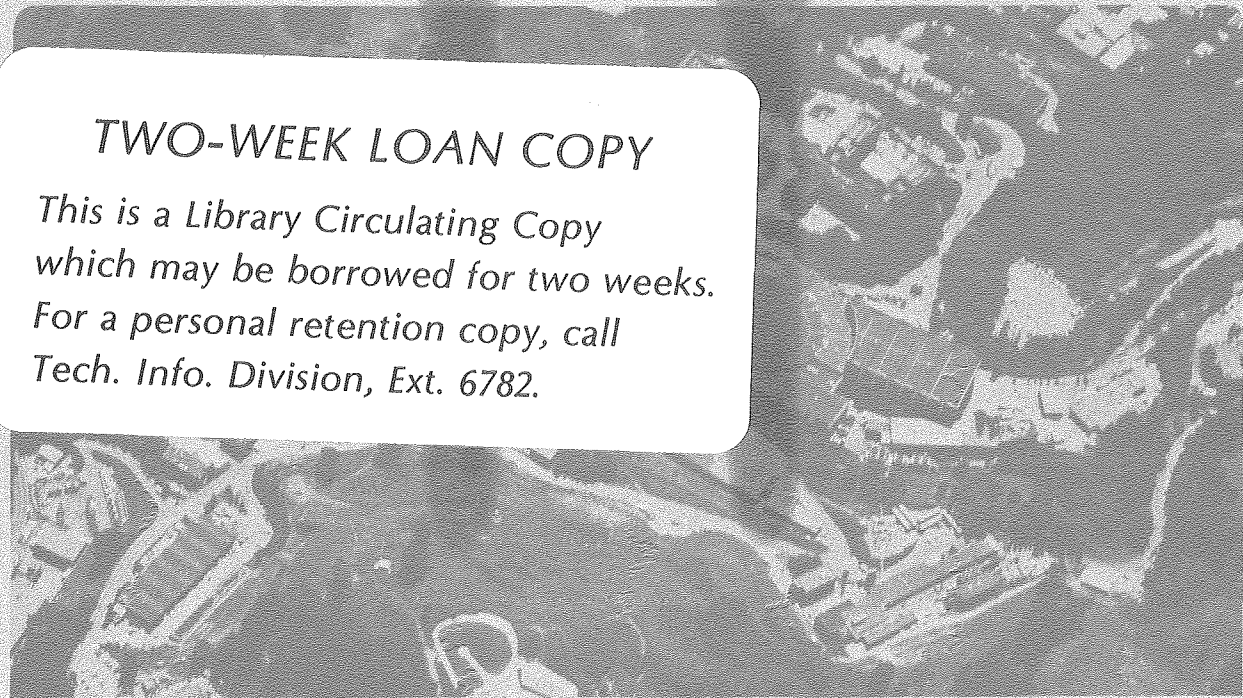
RECEIVED
LAWRENCE
BERKELEY LABORATORY

APR 7 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 Rev. 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, CA.

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 (adapted from a U.S. Department of Labor regional report) 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 (the spectral sequence from red through orange to yellow).

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:

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.

The viewers attention is attracted first to the primary message, degree of fit, in color. Supporting detail about Openings, Not Filled, and Fill Rate is shown by the accompanying bar charts in neutral grays. Linguistic terms such as HIGH, MEDIUM, LOW are defined within the context of each column variable. Each column has been scaled to a different range to normalize bar size with respect to HIGH and LOW, etc. Outliers have been truncated to the edge. The same size bar is as HIGH or LOW in one column as in any other. For example, for electrical jobs the number of Openings is about as LOW as the Fill Rate. For any graph or chart, viewers often want to compare their own characterization of the data against the display. To include detailed data in the display seems especially important in this application where the color scale indicates only a fixed approximation to one of several shades of meaning of the linguistic expression.

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.

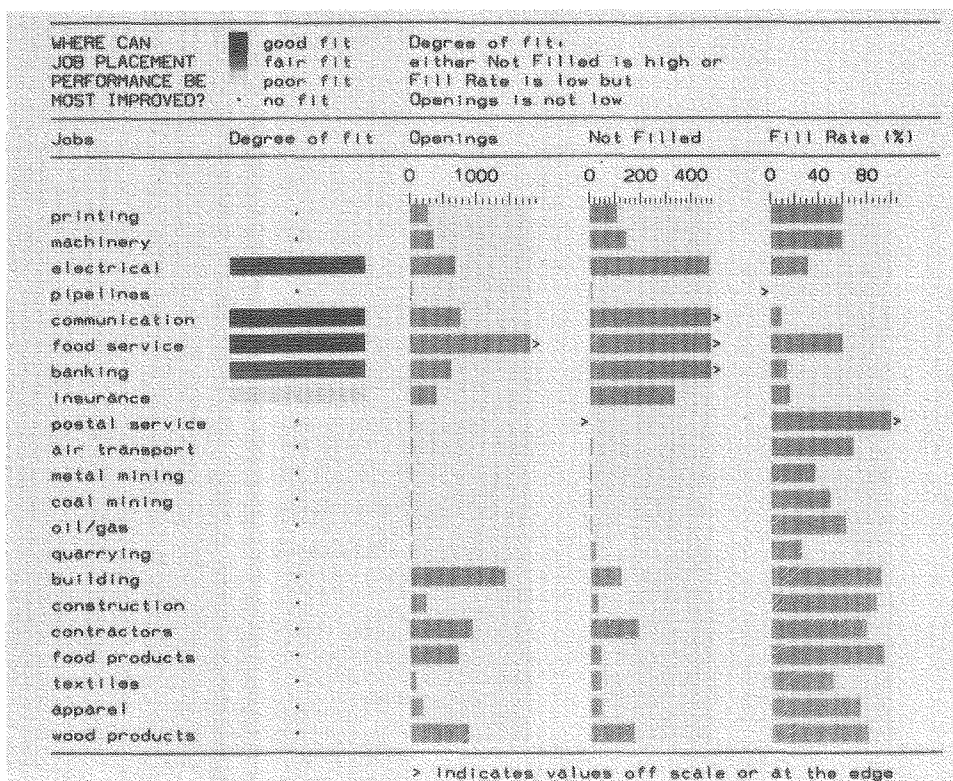


Fig. 1. Job service placement performance. XBB812-975

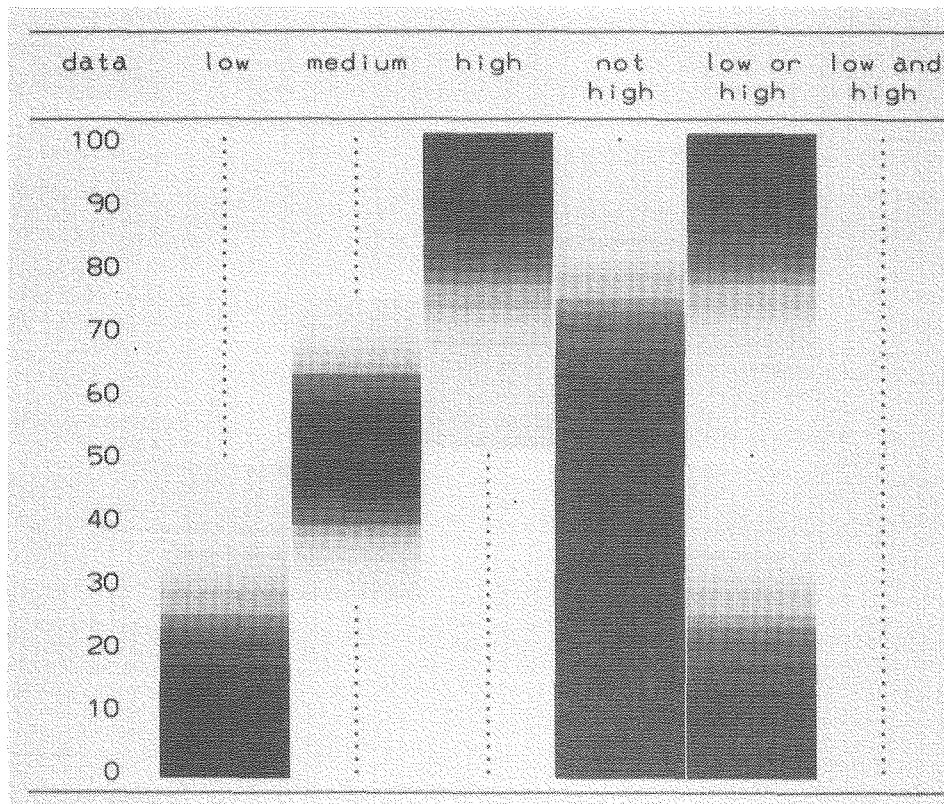
FUZZY SET MODEL FOR SUBJECTIVE CATEGORIES

A fuzzy set model has been incorporated into an existing program for analysis and display of tabular data (Benson and Kitous, 1977). 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.



XBB812-973

Fig. 2. Membership functions for sample linguistic terms and expressions.

Three other primary terms NEAR x, OVER x, and UNDER x define a fuzzy interval for an arbitrary ten percent of the range on either side of a particular value of interest. Membership functions for primary terms are derived from the normalized range of each data variable using the S shaped curve in Zadeh (1975).

$$S(u;a,b) = \begin{cases} 0 & \text{for } u \leq a \\ 2[(u-a)/(b-a)]^2 & \text{for } a \leq u \leq (a+b)/2 \\ 1-2[(u-b)/(b-a)]^2 & \text{for } (a+b)/2 \leq u \leq b \\ 1 & \text{for } u \geq b \end{cases}$$

For data u and x normalized to the interval [0,1],

$$\begin{array}{ll} \text{MEDIUM} = S(u;.25,.5) & \text{for } u \leq .5 \\ \text{MEDIUM} = 1-S(u;.5,.75) & \text{for } u \geq .5 \\ \text{NEAR } x = S(u;x-.1,x) & \text{for } u \leq x \\ \text{NEAR } x = 1-S(u;x,x+.1) & \text{for } u \geq x \end{array} \quad \begin{array}{l} \text{LOW} = 1-S(u;0,.5) \\ \text{HIGH} = S(u;.5,1) \\ \text{UNDER } x = 1-S(u;x-.1,x+.1) \\ \text{OVER } x = S(u;x-.1,x+.1) \end{array}$$

With this definition, "x is OVER x" is as good a characterization as "x is NOT OVER x", while "x is NEAR x" has full membership. Primary terms can be modified using hedges and combined with logical operators. Definitions are taken from the literature (Zadeh, 1975). For membership functions f and g,

$$\text{NOT } f = 1-f \quad f \text{ AND } g = \min(f,g) \quad f \text{ OR } g = \max(f,g) \quad \text{VERY } f = f^2 .$$

Implementation is guided by the expectation that users will prefer to compose concise and clearly understandable expressions which can serve as colloquial though somewhat stilted descriptions of categories of interest. A typical example is 'DEGREE "RATE" IS OVER 8 AND "POPULATION" IS NOT VERY VERY LOW'. Expressions are introduced by the word DEGREE; variables are in double quotes, and IS evaluates the membership function for the fuzzy set on the right at each value of the data variable on the left.

Overly complex expressions will be too hard to understand to be useful. For this reason, expressions are parsed using a simple operator precedence grammar model. Precedence is 1) NOT, hedges, OVER, UNDER, NEAR, 2) IS, 3) AND, OR. No parentheses markers are used; AND and OR have equal precedence. Expressions are evaluated from right to left so that ambiguities may be resolved by association to the right.

So that one variable may be compared with another, each range is normalized to the interval [0,1]. Definition of primary terms such as HIGH and NEAR is made with respect to normalized values. Since this definition is sensitive to outliers, data are usually normalized to a more robust range (Chambers, 1977). Essentially, the central portion of the raw data range is first determined from the order statistics, typically the middle two quartiles. This portion is then extended toward, but not beyond, the raw data minimum and maximum on either side by a chosen factor times the extent of the central portion.

CATEGORY STRUCTURE

Categorization is a powerful tool for organizing memory and thought in everyday life. The preceding sections have suggested a similar role for subjective categories in data analysis. This role is developed further by considering category structure in terms of prototypes, or best examples, which serve as references against which membership is judged. The structure of human category systems has been investigated in cognitive psychology, notably by Rosch (1977, 1978), who proposes two basic and general principles underlying categorization.

The first principle, cognitive economy, expresses the function of category systems in maximizing information while minimizing cognitive effort. To categorize an object is to consider it equivalent to other objects in the same category, and different from objects not in that category. Information is increased when shared properties can be predicted from knowing any one property (including the name). Effort is reduced by not differentiating objects when differentiation is irrelevant to the purpose at hand. The second principle recognizes that attributes do not appear independently of each other; that there is correlational structure in the perceived world.

These principles have implications for two dimensions of category systems - level of inclusiveness, and structure within a level. The implication for the latter is that "to increase the distinctiveness and flexibility of categories, categories tend to become defined in terms of prototypes or prototypical instances that contain the attributes most representative of items inside and least representative of items outside the category" (Rosch, 1978, p. 30).

Cognitive economy dictates distinctiveness of categories, which is achieved by emphasizing and exaggerating existing structure through prototypes. Prototypes can maintain categories as discrete even when attributes are continuous, or correlational structure is incomplete. Prototypes allow greater use of representational codes, such as imagery. The most economical code is just a concrete image of a typical category member.

The exaggeration of structure in prototypes leads to an organization of categories in terms of a core of central members similar to each other, with atypical members on the periphery, not only unlike the prototype but unlike each other (Glass, Holyoak, and Santa, 1979, p. 340). Since degree of membership arises from degree of similarity to the prototype, the processing mechanism of prototype matching - comparing an instance to the prototype and summing up the evidence - is an efficient strategy for evaluating category membership.

The example in Fig. 1 illustrates how simple questions about subjective categories can be asked using linguistic expressions and answered by visually matching to a prototypical best fit. Other tasks involving prototype matching (Glass, Holyoak, and Santa, 1979) are

- a) evaluating logical arguments - generating or rejecting hypotheses by exemplifying supporting cases or finding counter examples;
- b) reasoning by analogy, or from incomplete knowledge to generate plausible hypotheses. An instance matching a prototype on some (relevant and sufficient number of) attributes is assumed to match on all.

AN EXAMPLE OF PERFORMANCE ANALYSIS BY PROTOTYPE MATCHING

Performance evaluations are commonly made by combining and integrating performance measures across several variables. An example, viewed in terms of categories and prototypes, is given below to illustrate how this approach aids analysis.

Performance evaluation, similar to multi-objective decision making, is often not a well defined task. It may be that ideal performance is easy to describe, but no cases meet it, so that exceptions and compromises must be made. More commonly, it may be impossible to construct a decision function to the analyst's complete satisfaction. Yager (1980a) cites difficulties in completely listing all the objectives, in combining them, and in stating tradeoffs.

Yager (1980b) has also discussed the significance of the concept of importance, specifically in making tradeoffs in importance between objectives, and introduced

combine the objectives differently. Making visual comparisons gives more flexibility in evaluating choices. At this point, the income objective "i" might seem less important, since except for WA, the other five 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

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. Perceptual properties supporting one such scale (red-orange-yellow) are discussed below.

Data can be coded for display by graphic variables such as size, texture, color, gray scale, shape, and orientation. Using graphics has the potential to both aid and hinder data analysis. How graphic variation is perceived can highlight 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 fused 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.

These intuitions have been formalized by Kay and McDaniel (1975, 1978), who have traced a detailed line of argument incorporating studies of color term semantics, neurophysiological evidence, and the explicit use of fuzzy set theory to model the continuity of membership in color categories. Studies of color term semantics indicate that all languages share a universal system of basic color categorization. A total of eleven basic color categories are identified. Six of these, blue, green, yellow, red, white, and black are shown to be biologically based on fundamental neural responses, while orange, pink, purple, brown and gray are described by simple functions of fuzzy intersections among the first six.

Some implications for constructing color scales are briefly mentioned:

a) A color scale should be anchored at each end by basic colors, since these are most easily named and remembered (Rosch, 1977). The basic distinction between degree and kind of fit (some fit vs. no fit) is also expressed well by contrast between basic colors. For example, red contrasts with yellow while degrees of both red and yellow can be seen in intermediate colors.

b) The scales red-orange-yellow and green-chartreuse-yellow differ in that orange is basic while chartreuse is not. Orange as an intermediate category aids discrimination of and gives a separate identity to intermediate degrees of fit, while chartreuse is perceived as an approximation to either green or yellow.

c) Witkowski and Brown (1977) note the special salience of red, as the first basic category to be encoded in evolutionary sequence. To "red-flag", meaning to call attention, attests to this salience.

Color has a dimension of brightness (from white to black or light to dark) as well as dimensions of hue and saturation. Vivid, fully saturated, colors are expected to enhance selectivity. Since maximum saturation is achieved for different hues at different brightness levels (Witkowski and Brown, 1977), there is an implicit scale of brightness values corresponding to a scale of fully saturated colors. These scales should be consistent. For example, scales from yellow to red and yellow to green also run from light to dark. A scale from blue to green is unsatisfactory because cyan in the middle is lighter than blue and green at either end.

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, and how perceptual properties of color can be used to selectively focus attention, distinguish or disregard variation, and evaluate overall degree of fit. An interactive computer graphics program has been described in which categories can be formulated and combined using familiar language. Implementation is based on quantitative techniques in fuzzy set theory.

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. I would like to thank Kevin Thompson of the Department of Labor for his support and many practical examples, and Aaron Marcus of Lawrence Berkeley Laboratory for help with graphic design.

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.
- Chambers, J. (1977). Computational Methods for Data Analysis. Wiley, New York. pp. 219-220.
- Glass, A. L., K. J. Holyoak, and J. L. Santa (1979). Cognition. Addison-Wesley, Reading, Mass. pp. 354-390.
- 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.
- Rosch, E. (1977). Human Categorization. In N. Warren (Ed.), Studies in Cross-Cultural Psychology, Vol. 1, Academic Press, New York. pp. 1-49.
- Rosch, E. (1978). Principles of Categorization. In Rosch and Lloyd (Eds.) Cognition and Categorization, Erlbaum, Hillsdale, N.J. pp. 27-48.
- Simon, H. A. (1977). The New Science of Management Decision Making. Prentice-Hall, Englewood Cliffs, N.J.
- Witkowski, S. R., and C. H. Brown (1977). An explanation of color nomenclature universals. American Anthropologist, 79, 50-57.
- Yager, R.R. (1980a). Fuzzy subsets of type II in decisions. J. of Cybernetics, 10, 137-159.
- Yager, R.R. (1980b). A linguistic variable for importance of fuzzy sets. J. of Cybernetics, 10, 249-260.
- 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.
- Zadeh, L. A., K. S. Fu, and M. Shimura (1975). Fuzzy Sets and their Applications to Cognitive and Decision Processes. Academic Press, New York. p. 29.