Title
LOOKING AT MLJLTIVARIATE DATA THROUGH FUZZY SETS

Author
Benson, W.H.

Publication Date
1984-07-01

Lawrence Berkeley Laboratory
UNIVERSITY OF CALIFORNIA

Computing Division

Presented at the First International Conference
on Fuzzy Information Processing, Kauai, HI,
July 22-26, 1984; and to be published in
The Analysis of Fuzzy Information, J. Bezdek, Ed.,
CRC Press, Boca Raton, FL

LOOKING AT MULTIVARIATE DATA THROUGH FUZZY SETS

W.H. Benson

July 1984

Prepared for the U.S. Department of Energy under Contract DE-AC03-76SF00098

DISCLAIMER

LBL-15792

Looking at Multivariate Data Through Fuzzy Sets

William H. Benson

Computer Science Research Department
University of California
Lawrence Berkeley Laboratory
Berkeley, California 94720

July, 1984

This research was supported by the Applied Mathematics Sciences Research Program of the Office of Energy Research, U.S. Department of Energy under contract DE-AC03-76SF00098.

LBL-15792

Looking at Multivariate Data Through Fuzzy Sets

William H. Benson
July 31, 1984
Computer Science Reseach Group
Lawrence Berkeley Laboratory
University of California
Berkeley, CA 94720

ABSTRACT

There is usually a preliminary stage to data analysis, to become familiar with the data and provide a starting point for further analysis. An example from the recent literature illustrating the process of data inspection is reviewed and compared with a more direct approach based on fuzzy sets and computer graphics. A demonstration is provided of the effectiveness and efficiency of direct expression of descriptions and viewpoints, and of direct manipulation of the graphic display.

Acknowledgment:

This work was supported by the Office of Energy Research, U.S. Department of Energy under Contract DE-AC03-76SF00098.

I. INTRODUCTION

This note briefly sketches some graphical techniques for looking at multivariate data using basic notions from fuzzy set theory. Starting from the standard statistical data structure of cases and variables, new variables are defined as fuzzy restrictions on the ranges of the original variables. These new variables are defined most conveniently in terms of linguistic expressions,[1] such as "variable X1 is high", "X2 is low", etc. Evaluating each case for compatibility with a list of linguistic expressions yields a new table consisting of membership values. Graphic operations are then applied to this table to help detect and recognize interesting features.

These graphic operations, like graphical techniques in general, are not intended to provide a statistical analysis, but rather to "help suggest hypotheses, check assumptions, and interpret results."[2] The following section briefly summarizes results from a recent article advocating a descriptive approach to multivariate data. Some graphical techniques based on fuzzy sets are then illustrated in the context of the specific data set examined in the article.

II. DIRECT INSPECTION

The sample data set and analysis are taken from a recent article addressed to statistics teachers in which the author looks at the relation between social factors and disease incidence for the 21 wards (comparable to census tracts) in the city of Hull, England.[3] The sample data set consists of observations on four indicators of social welfare (rates per thousand population of "overcrowding", "no inside toilet", "do not possess a car", and "unskilled workers") and three for infectious diseases (rates of incidence of jaundice, measles and scabies). All variables run in the same direction, with lower values indicating better conditions; higher values poorer conditions.

A preliminary data reduction step evaluates the quartiles for each variable and replaces each numerical value by the ordinal quartile number (1,2,3, or 4) in which it lies. Inspection of the table after this data reduction reveals a structured relationship between the social variables, suggesting certain cases should be grouped together and examined separately. The variables X1, X3, and X4 are very similar (especially for low values), and very different from X2. There are five affluent cases where X1, X3, and X4 are all low together (called group A), and a different five cases where X2 is low (group B). It turns out that the behavior of X2 can be accounted for by the large proportion of public housing in group B wards. The remainder is called group C. Thus high X1, X3, and X4 correspond to low X2 in group B, but to high X2 in group C. Since this multivariate relationship differs between B and C, it is inappropriate to analyze them together.

Looking at the disease variables, one can observe that they are related, since no ward has both a high and a low score for these variables. In general, groups A and B have at least 2 out of 3 low scores while group C has at most 1 out of 3. One exception (group D) is noted which calls for further investigation. These four groups are shown in Table 1, where quartiles have already been evaluated for the variables. The tentative conclusion follows that disease rates are related to social factors, in particular to either decent housing or general affluence.

III. GRAPHIC REPRESENTATION

Although the standard maxim is to let the numbers speak for themselves, this need not

be taken to exclude graphic representations and spatial arrangements designed to encourage a particular set of visual comparisons. A scatter plot, for example, invites a visual judgment of linearity. While not a substitute for regression analysis, scatter plots still allow a useful estimate to be made of the degree of (linear) association between two variables.

Some insight into concepts and comparisons relevant to multivariate data may be gained from reviewing the previous example. The sub-verbal protocols of data inspection are not much reported in the literature, but to the extent this one is representative, then the relevant comparisons involve characterizations such as "low", "high", "almost all low", "at least 2 out of 3", "at most 1 out of 3", etc. These comparisons are made easier by the preliminary data reduction step and by bringing rows and columns to be compared into proximity. The data reduction by quartiles, which effectively reduces the resolution at which the data can be seen, leads to much quicker recognition of low and high values; re-organizing rows and columns also brings together items to be counted.

The method illustrated below is designed to encourage comparisons of relative magnitude and cardinality by giving these concepts explicit graphic representation. For example, a related table can be formed by evaluating each case according to membership functions for "X1 is high", "X1 is low", etc. This new table allows comparisons to be made from several different perspectives (such as high vs. low) without loss of resolution throughout the range. Moreover, within this related table, visual comparison of measures of fuzzy cardinality and extremism can be made by reorganizing the table cells so that the table coordinates temporarily reflect these measures, instead of the original case and variable dimensions. This re-organization can be easily accomplished by selectively ranking columns within rows and rows within columns.

Returning to the original example and data set, the right half of Figure 1 shows fourteen new variables for the original 21 wards. These are membership functions in the fuzzy sets for "low" and "high" for each of the original seven variables. The functions express a continuous range of membership values from 0 (no membership), through intermediate values, to 1 (full membership). The membership functions have been defined in terms of the order statistics for each variable. For example, membership for "high" is defined to be 0 below the median, taking increasingly higher values up to 1.0 at the highest value, excluding outliers. This definition has been used because it is intuitively clear, can be confirmed by visual inspection, and is comparable to the quartile analysis in Table 1. This continuous range is represented graphically by shading from light (0) to dark (1.0). Shading is used to discourage misleading and inappropriate ratio comparisons among membership values and help blur the distinction between nearly equal membership values.[4] The shading does not represent the original data directly, but instead shows where (which cases) and how well (how dark) the data fit terms such as "high" and "low".

IV. VISUAL COMPARISONS

The left half of Figure 1 shows the same values, row for row, as the right half, but ranked in descending order, dark to light from left to right, within the corresponding column group. For example, the first four columns represent one definition of the fuzzy cardinality "how many social variables are low." If "all four social variables are low" are necessary for the concept of affluence, then ward 3 and ward 5 are seen to satisfy this condition strongly, and three others weakly.

A more useful approximation to a concept of affluence would take exceptions into account, such as an expression like "most of the social variables are low." The first four columns in Figure 1 then evaluate four different definitions for "most": "at least one", "at least two", "at least three", and "at least four."

A further re-organization of the left half of Figure 1 presents a basis for judgment whether there are sensible definitions of relevant concepts (affluence, social deprivation, healthy, unhealthy) with sufficient power to discriminate interesting cases. I.e. whether a substantial number of conditions are satisfied to a substantial degree for a substantial number of cases.

In Figure 2, each column is ranked individually, dark to light from top to bottom. The table cells are the same as in the left and right parts of Figure 1, but they can no longer be identified by case and variable. For example, within the first group of four columns, one can visually compare different definitions of "most are low". From left to right (within each column group) the definition becomes more stringent, so that at some point too few cases tend to be discriminated to constitute an interesting study group. Conversely the least stringent definitions tend to discriminate too many.

The different definitions can also be compared along a dimension of extremism. A measure of extremism for the n highest cases with respect to each definition can be seen within the corresponding column at the nth row from the bottom of the display. For example, extremism falls off abruptly after the four highest cases for which "at least three social variables are low" (column 3), and after the five highest cases for which "at least two social variables are low" (column 2). Provided the highest case has full membership, this corresponds to Yager's definition of extremism,[5] Abs(Max F(x) + Min F(x) - 1), since each column has been ranked in ascending order.

Since extremism measures both lack of diversity as well as intensity of membership, this idea can be used to identify a group of cases for closer inspection. The two examples just mentioned are the best candidates for closer study in that they are the largest groups satisfying the "most are low" or "most are high" criteria with sufficient cardinality at a high level of extremism.

In Figure 3, the group consisting of the five highest cases for which "at least two social variables are low" has been identified and isolated at the bottom of the display by ranking column 2 alone in ascending order. These are the same five cases called group A in Table 1. Inspection shows three specific variables all low together with X2 the only exception. Among the remaining cases, attention is focussed where X2 is low by ranking the second column in the right half of the display in ascending order. By a sequence of re-orderings, the five cases in group B where X2 is uniquely low are identified and isolated, and the remaining cases arranged roughly in decreasing social deprivation by ranking according to "at least three social variables are high" (column 7). The association between the disease/health variables and the social variables is then made clear visually, as well as the exception in the first row with poor (high) scores for the social variables, but low disease rates.

V. SUMMARY AND CONCLUSIONS

Data inspection is seen as an active process in which the data is described and compared from different viewpoints. Both description and comparison are facilitated by direct representation of concepts useful for summarizing relations between variables. Concepts such as levels of variables, cardinality, and extremism are expressed directly in terms of fuzzy sets, with membership pictured graphically. The graphics

display is designed to be re-organized according to principles of direct manipulation, in order to encourage systematic comparison between different viewpoints and allow an efficient search for interesting structure.

REFERENCES

1. Zadeh, L.A., Outline of a new approach to the analysis of complex systems and decision processes, *IEEE Trans. Syst., Man & Cybern.*, SMC-3, No. 1, 28, 1973.

2. Hartigan, J.A., and Kleiner,B., A mosaic of television ratings, *The American Statistician*, 38, No. 1, 32, 1984.

3. Goldstein, M., Preliminary inspection of multivariate data, *The American Statistician*, 36, No. 4, 358, 1982.

4. Benson, W.H., An application of fuzzy set theory to data display, in *Fuzzy Set and Possibility Theory*, Yager, R.R., Ed., Pergamon Press, 1982, 429.

5. Yager, R.R., Opposites and measures of extremism in concepts and constructs, *Intl. J. of Man-Machine Studies*, 18, No. 3, 307, 1983.

FIGURE CAPTIONS

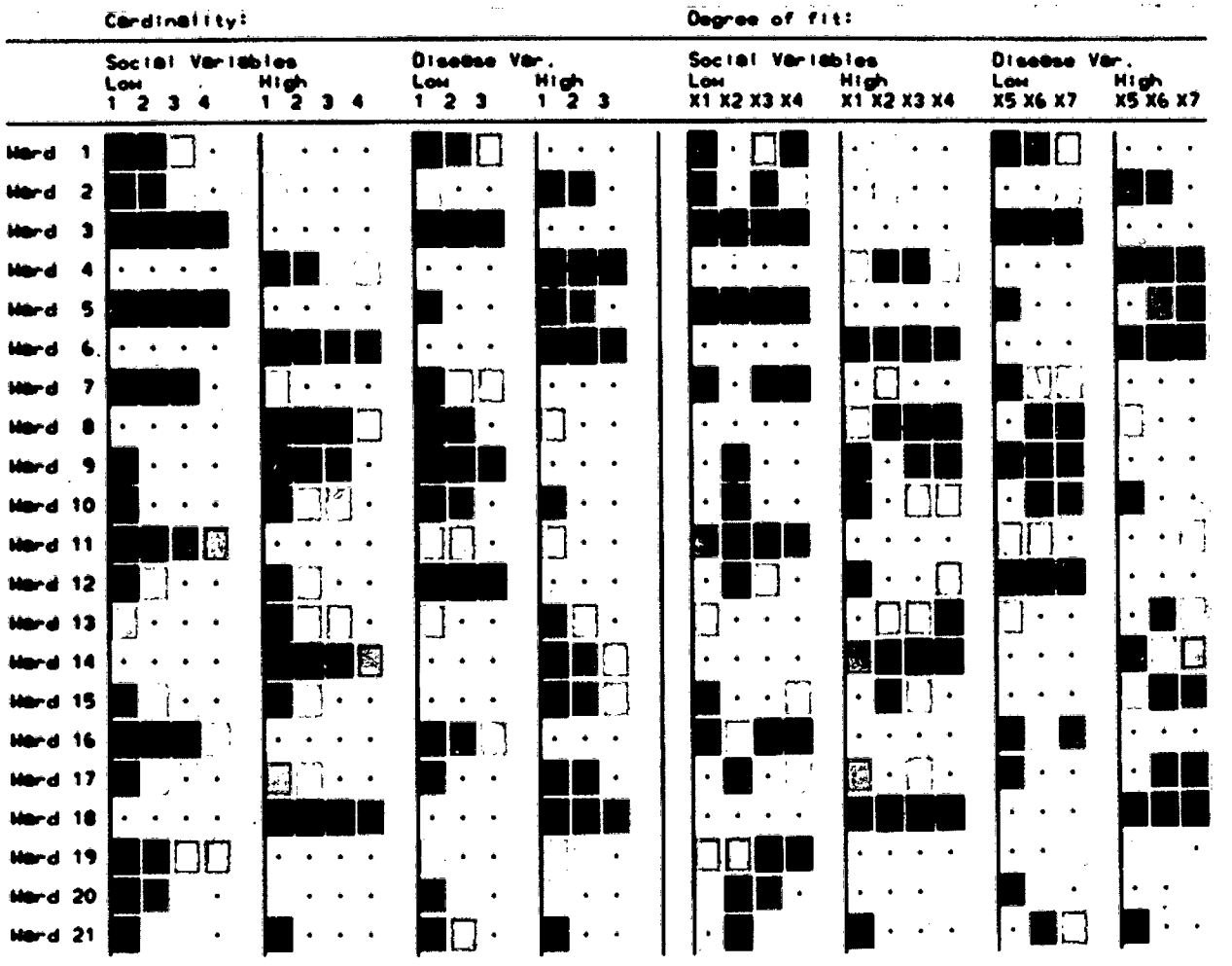
Figure 1. Right side: Cases evaluated for membership functions for "high" and "low". Left side: Right side re-ordered to help make visual estimates of cardinality.

Figure 2. Figure 1 re-ordered, allowing different groups of cases to be compared according to a measure of extremism.

Figure 3. Recovering the structured multivariate relationship shown in Table 1. The two groups at the bottom have mostly low disease rates.

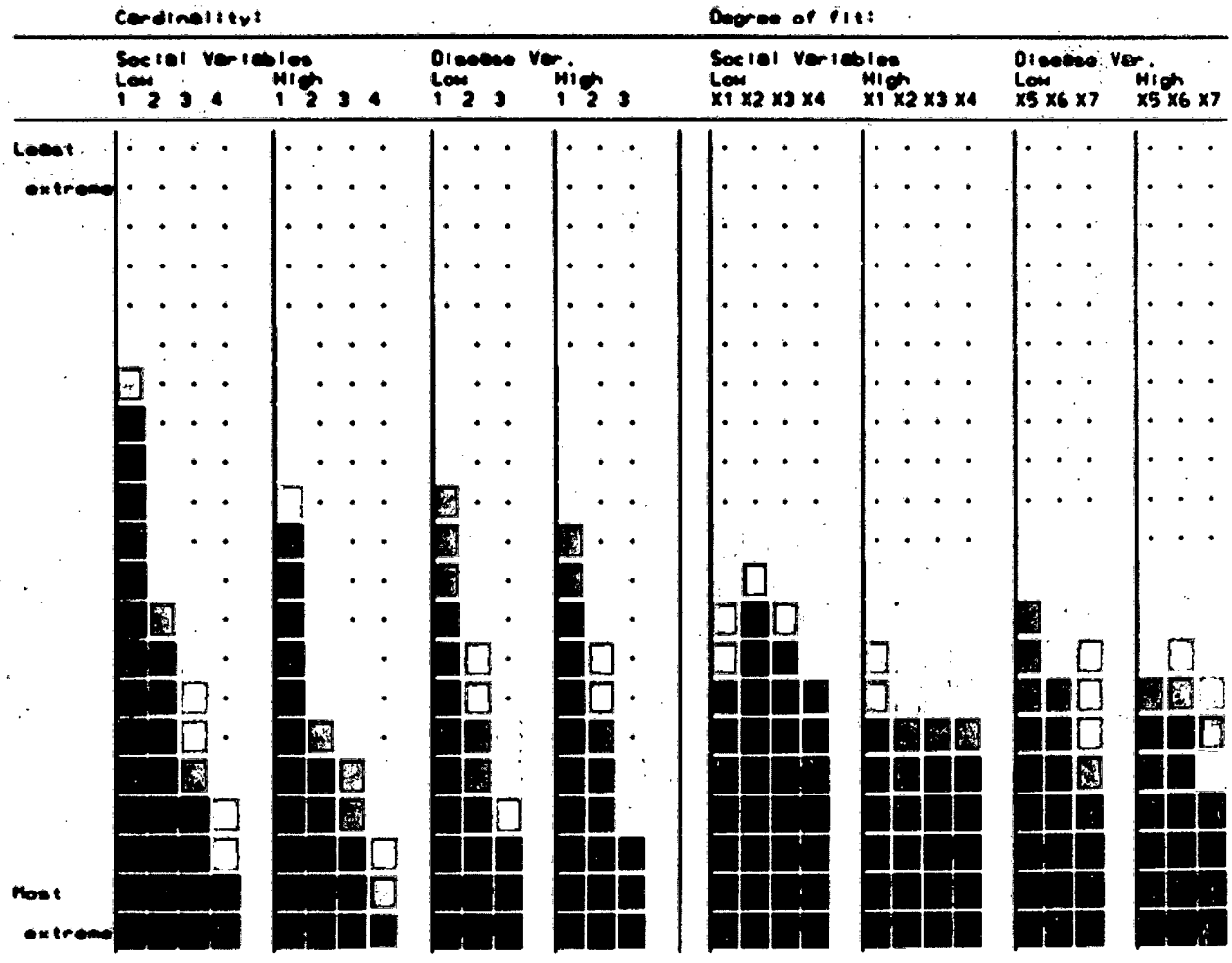
Table 1. Cases are grouped to show a structured multivariate relationship. (from Table 3 in Goldstein[3])

Group	Ward	Social Variables X1	X2	X3	X4	Disease Variables X5	X6	X7
Group D	Ward8	3	4	4	4	3	1	1
Group C	Ward19	2	2	1	2	3	3	3
	Ward11	2	2	2	2	2	3	3
	Ward17	3	2	3	2	2	4	4
	Ward2	2	3	2	2	4	4	2
	Ward13	3	3	3	4	3	3	3
	Ward15	2	4	3	3	3	4	4
	Ward4	3	4	4	3	4	4	4
	Ward14	3	4	4	4	4	3	3
	Ward6	4	4	4	4	4	4	4
	Ward18	4	4	4	4	4	4	4
Group B	Ward9	4	1	4	4	1	1	1
	Ward10	4	1	3	3	3	1	2
	Ward12	4	1	2	3	1	1	1
	Ward20	2	1	2	3	2	2	3
	Ward21	4	1	3	2	4	2	2
Group A	Ward1	1	3	2	1	1	2	1
	Ward3	1	2	1	1	1	1	1
	Ward5	1	2	1	1	2	3	4
	Ward7	1	3	1	1	1	2	2
	Ward16	1	3	1	1	2	2	2



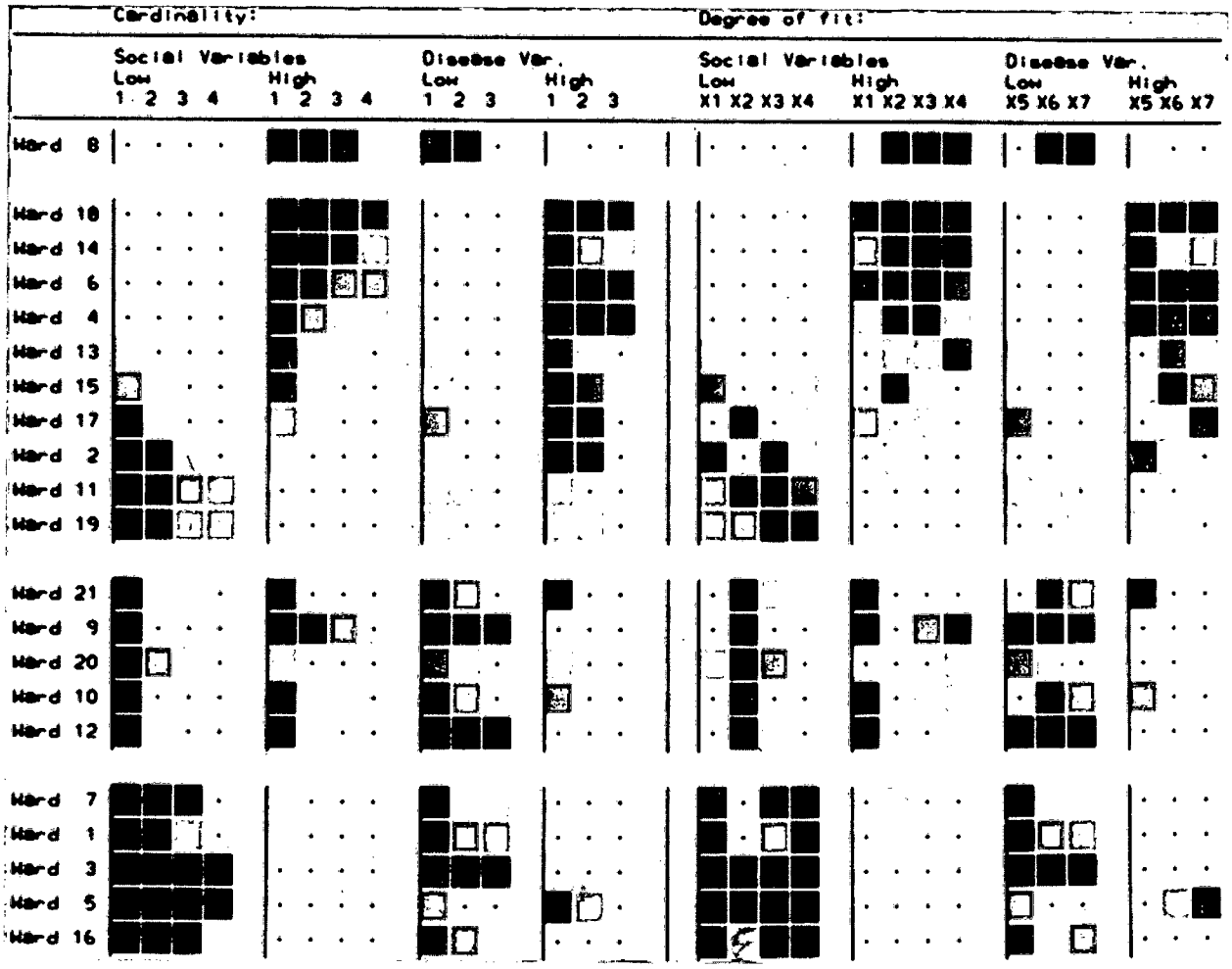
CBB 847-5303

Figure 1



CBB 847-5309

Figure 2



CBB 847-5305

Figure 3

This report was done with support from the Department of Energy. Any conclusions or opinions expressed in this report represent solely those of the author(s) and not necessarily those of The Regents of the University of California, the Lawrence Berkeley Laboratory or the Department of Energy.

Reference to a company or product name does not imply approval or recommendation of the product by the University of California or the U.S. Department of Energy to the exclusion of others that may be suitable.

TECHNICAL INFORMATION DEPARTMENT
LAWRENCE BERKELEY LABORATORY
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720