Fifth IFSA World Congress (1993), 1362-1365

APPLYING FUZZY MATHEMATICS TO QUANTIFYING HUMAN RESPONSES

R.C. Steinlage*, T.E. Gantner*, P.Y.W. Lim**

*Department of Mathematics, University of Dayton, Dayton, Ohio
**Boise Cascade R&D, Portland, Oregon (Now at the Fine Paper Division of Union Camp Corporation, Franklin, Virginia)

Abstract. Fuzzy mathematics is used to elicit and evaluate human psychophysical responses in panel tests. The fundamental instrument used is a bar graph whose data is then converted to a paired comparison matrix. From this matrix we use the theory of Perron and Froebenius to obtain an eigenvalue and eigenvector which indicates not only the panelist's comparitive responses but also the consistency of the responses from that panelist. Tests were done to evaluate the procedure.

Introduction. Human perception is a complex phenomenon which is difficult to quantify with instruments. For this reason, panels of several or many people are often used to elicit and aggregate subjective judgments. Print quality, taste, smell, sound quality of a stereo system, softness, and grading Olympic divers and skaters are some examples of situations where subjective measurements or judgments are paramount. For lack of better methods to quantify subjective judgments, it is customary to set up a numerical scale such as 1, 2, 3, 4, 5 or 1, 2, 3, ... , 9, 10 for characterizing human responses and subjective judgments with no valid justification except that these scales are easy to understand and convenient to use. But human responses and subjective judgments are psychophysical phenomena that are fuzzy entities and therefore difficult to handle by conventional mathematics and probability theory. The fuzzy mathematical approach provides a more realistic insight into understanding and quantifying human responses.

Approach. The method used to code responses obtained from panelists is especially important when one wishes to make decisions concerning properties or events which are not objectively quantifiable but which must be evaluated subjectively. The problem of coding such responses has been addressed from many directions. In this paper we propose a technique, based in fuzzy mathematics, for quantifying and evaluating subjective responses and then we test our technique in situations where the properties are also objectively measurable. By testing our technique in objective situations, we hope to lend credibility to its use in purely subjective situations. The technique we describe is a refinement of techniques originally proposed by Saaty.

Saaty proposes using five adjectives as "response words" in subjective panel tests. These words indicate that two samples are indistinguishable (1) with respect to a given property or that the difference between them is slight (3), moderate (5), significant (7), or extreme (9). Of course, panelists are permitted to hedge their bets and cast their ballots between two such judgments. (2, 4, 6, or 8). Thus Saaty is proposing a 9 point scale for linguistic or subjective judgments. As Saaty states, this is a good scale in that it provides enough shades of meaning without expecting a panelist to be scrupulous.

After obtaining panel data, the next problem is the analysis of this data. Aside from the usual statistical analysis, a technique that has been shown to be successful in fuzzy or subjective situations is to find the dominant eigenvalue and associated eigenvector for the reciprocal matrix of paired comparisons. This analysis is based on the work of Perron and Froebenius. If n objects A_1, ... , A_n are being compared, these are listed horizontally and vertically to indicate the rows and columns of a matrix M. If A_i is judged to be significantly greater than A_j, then a 7 is placed in row i, column j and 1/7 is placed in row j, column i .

If our objective is to determine the respective weights of n objects, then the resulting eigenvector should indicate the relative weights. If we have perfect information (no judgments are necessary and responses are not restricted to integers and their reciprocals) we could simply fill in the matrix using the ratio of the respective weights: $m_{ij} = w_i/w_j$. We then obtain a reciprocal matrix: $m_{ji} = w_j/w_i = 1/m_{ij}$. It can be shown that $\lambda = n$ is the only non-zero eigenvalue for M and that $W = (w_1, \ldots, w_n)$ is its associated eigenvector; the correct weight determination is indeed obtained as the eigenvector. This eigenvector is unique up to a scalar multiple.

If the experiment was needed, however, perfect information is not available at the outset. But if the responses are a reasonable approximation to the reality of the situation, then the responses will approximate those which would have been placed in the "perfect information" matrix. Hence the eigenvalue should approximate n (the number of samples) and the associated eigenvector should approximate the actual distribution of the property (weight, etc.) among the samples. Thus the eigenvector not only ranks the samples ordinally (indicates smallest, largest, etc.) but also gives a cardinal ranking (indicates relative strengths or weights, etc.). In actuality $\lambda \geq n$, the associated eigenvector $V = (v_1, \ldots, v_n)$ is unique up to a multiplicative constant, and when normalized so that $v_1 + \ldots + v_n = 1$, v_i indicates the percentage of the total (weight) possessed by object i. The eigenvalue λ is a measure of the consistency of the responses given by the panelist. A good rule of thumb is that if $\lambda > n + 2$, the panelist has contradicted himself or herself so many times and/or so egregiously that his or her responses should be ignored. On the other hand, if λ is very close to n, the panelist was very consistent (although not necessarily accurate or correct). In short, the eigenvalue is a good flag to indicate errors in recording data; e.g., a number and its reciprocal may be interchanged.

If a scale much larger than 9 is used, the differences in reciprocals become negligible and some discrimination between samples in the resulting eigenvector will be lost. A collection of objects in which the samples may be too widely diverse should be subjected to a hierarchical analysis. However, our experience indicates that while the above 1 - 9 scale may be appropriate for eliciting and coding human responses, it is not always the proper scale to be used in the ensuing matrix analysis. In fact, the scale used will be reflected in the results. The largest number used is in essence the ratio between the strongest and weakest (or heaviest and lightest, etc.) objects in the resulting eigenvector. Thus an inappropriate numerical scale will lead to undesirable end effects concerning the extremes of the objects being compared. This end effect is extremely volatile when computing percent error on the low end. Our experience indicates that a linear rescaling of the 1 - 9 linguistic scale to a scale determined by the accepted or perceived ratio of the two extreme objects in the given group significantly reduces this end-point effect.

An Example: We considered 6 weights and used the weight ratios to form a "perfect information" matrix M_6. In this case, a weight ratio of 2 indicates that the numerator weight is twice that of the denominator, which is quite different from the linguistic use of the number 2 in the above 1 – 9 scale. The linguistic 2 says two samples are almost indistinguishable. We linearly rescaled the integer entries in M_6 to a 1 – 9 scale to get a reciprocal matrix M_9, as well as to a 1 – 3 scale to get a reciprocal matrix M_3. In all cases the eigenvalue λ was less than 6.005. These low eigenvalues merely indicate consistency, not agreement with experimental measurements. In M_9, the 1 – 9 scale exceeded the actual maximum weight ratio of 6. As a result the eigenvector scale overestimated the heavier weights and underestimated the lighter weights. In M_3, the 1 – 3 scale fell short of the maximum weight ratio of 6. As a result the eigenvector scale underestimated the heavier weights and overestimated the lighter weights. The spread between the two extremes is too large with a scale of 1 – 9 and too small with a scale of 1 – 3. However, all three scales provided the proper ordinal ranking of the weights.

As a practical test of our theory we duplicated Saaty's weight test on five dissimilar objects of various sizes, shapes, and weights. Pairwise comparisons of these objects were made using the linguistic 1 to 9 scale and the corresponding reciprocal matrix was generated. The results were distorted significantly from the actual weight distribution. Nevertheless the eigenvalue $\lambda = 5.30$ was rather low. Again, this low eigenvalue indicates consistency of the responses – not necessarily accuracy of the predictions. On the other hand we observed that the maximum weight ratio was about 3 and we rescaled the original observations linearly to 1 – 3 from 1 – 9. This changed the results considerably. Using the "correct" scale reduced the maximum

relative error from 70% to 12%; the error at the volatile low end was reduced from 70% to 2%. This was accomplished simply by rescaling the original $1-9$ responses to a $1-3$ scale. The experiment was not redone. Had the original experiment been redone with a $1-3$ limitation on responses, much of the "fine tuning" of the responses would have been lost; i.e., not enough linguistic variation would have been permitted.

An alternative to $1-9$ responses (or any numerical response for that matter) is to use a bar graph in which the center position represents equality of the samples and the ends represent extreme dominance of one sample over the other. Using such a bar graph, responses can be interpreted on any numerical scale desired. We used bar graphs of this type in an experiment designed to test the ability of panelists to ascertain small differences in samples when the total magnitude was also small. In short, we wanted to test the applicability of this process to situations in which minor differences must be determined; can the process be "fine tuned" to indicate detailed differences as well as general relationships? Again, we tested the process in a situation where the property in question could also be objectively measured. Without such tests, the process would have little credibility in purely subjective situations. The experiment was to determine subjectively the relative thicknesses of paper, called "caliper" by the paper industry. Caliper is usually determined under laboratory conditions using instruments capable of accuracy to within 10^{-5} inches. Because of the non-uniformity of any given piece of paper, the caliper is usually measured in several spots and an average of these is used as the caliper of that sample. Thus caliper is an imprecise (even fuzzy) measurement made on a given sample of paper. Since fuzzy sets provide a framework in which one can study subjective judgments, we attempted to determine how closely the determination of caliper of paper, made by subjective decisions of panelists, compares with the instrumental measurements of the same samples under laboratory conditions.

Interpreting the panel responses on the bar graphs in the traditional $1-9$ linguistic scale yielded large errors with the largest errors occurring at the low and high ends. The sizes of these errors would seem to severely limit the applicability of the process to situations in which such delicate differences occur and are to be detected. On the other hand, reinterpreting the original panel data on a $1-3.14$ scale, where 3.14 represented the maximum ratio in average measured calipers, improved the results significantly. The distortion at the low and high ends was removed and in several samples, the error was no larger than the variation inherent in the sample itself. In no case was the error more than double the variation in the sample. We think this kind of accuracy obtained from subjective non-quantified judgments is astounding.

We remind the reader again that we are not proposing that this process be used to measure objectively quantifiable properties. Rather, we are testing our theories on objectively quantifiable properties, such as weight and caliper, so as to lend credibility to the process when it is used in situations which are primarily subjective.

Another Experiment. The determination of print quality has been a subject matter for which many sophisticated instrumental approaches have been developed, but human perception is still used as an integral part of the final evaluation. We studied non-impact printer image qualities using paired comparisons elicited from panels. Linguistic expressions and graphic responses were both used for transcribing the panel responses. The responses were analyzed using the techniques described in this paper. The purpose of the experiment was to test the applicability of the process outlined in this paper to situations in which minor differences must be determined. The resulting eigenvalue indicated that the panelists were able to give consistent responses in making the paired comparisons; panel fatigue and/or confusion was not a problem. The eigenvector analysis of the panel responses, when averaged, gave results consistent with what one would expect from viewing the samples. Samples which were ranked as being near in quality required higher levels of magnification before significant differences were observed than did samples which were ranked as being far apart in quality. While instrumental measurements made on greatly magnified images can result in overly stringent purchasing requirements, panel testing brings the determination of print quality closer to the practical marketing situation.

Summary and Conclusions. Many applications in commerce and industry demand far greater relative accuracy than is sometimes evidenced in the use of the paired comparison technique to elicit fuzzy properties. Early indications obtained in a fuzzy analysis should in many cases be refined. Many arguments can be given to justify the utility of a $1-9$ linguistic scale for

subjective responses. However, we have found that it is not necessarily advisable to continue to use the 1 – 9 scale in the computational process which follows the subjective evaluations. In fact, the choice of computational scale greatly influences the results at the extremes - the low and high ends - with related distortions in between. In short, the choice of computational scale dictates (approximately) the ratio between the extremes of the measured property in the given samples as generated by the eigenvector. Too large a scale results in the extremes being separated too far; too small a scale brings the extremes too close together. Taking greater care in making the comparisons cannot correct for this distortion if an inappropriate scale is chosen. This distortion is inherent in the computational process.

Thus in any panel test involving paired comparisons there are two distinct problems :

1. Using an acceptable scale for the panel responses. Whether this scale be linguistic, continuous or otherwise, it would seem that 9 levels on a scale of 1 – 9 is perfectly acceptable. An exception to this rule is that if the samples are too diverse, a hierarchical analysis would be in order.

2. The choice of computational scale should be treated as being independent of the scale used by the panelists. If the results are to be realistic and if the accuracy is to be "fine tuned", the computational scale must be close to the actual ratio between the properties in the extreme samples. This is not a trivial problem, however. If a fuzzy analysis is indeed necessary, then presumably this ratio cannot he obtained objectively. Nevertheless, since some kind of yardstick is probably desired, industry or marketing experts could indicate that one or several scales may be "appropriate". An advantage of the bar graph approach is that it readily lends itself to arbitrary computational scales. Thus results could quickly be processed for several scales and comparisons made. Then other scales could be checked - all without requiring further input from panels.

BIBLIOGRAPHY

[1] R.E. Bellman, Introduction to Matrix Analysis, second edition, McGraw – Hill, New York, 1970.

[2] P. Lim, R.C. Steinlage, and T.E. Gantner, Application of fuzzy set theory for evaluating non-impact printing image qualities, Proceedings of the 1990 International Printing and Graphic Arts Conference, Vancouver BC, Canada, November 1990, pp. 79 - 83.

[3] G.A. Miller, The magical number 7 ± 2: some limits on our capacity for processing information, Psychological Review, vol. 63, 1956, pp. 81 - 97.

[4] T.L. Saaty, A scaling method for priorities in hierarchical structures, Journal of Mathematical Psychology, vol. 15, 1977, pp. 234-281.

[5] T.L. Saaty, Exploring the interface between hierarchies, multiple objectives, and fuzzy sets, Fuzzy Sets and Systems, vol. 1, no. 1, 1978, pp. 57 - 68.

[6] T.L. Saaty, The Analytic Hierarchy Process: Planning, Priority Setting, Resource Allocation, McGraw – Hill, New York, London, 1980.

[7] T.L. Saaty, Decision Making For Leaders: The Analytical Hierarchy Process For Decisions in A Complex World, Lifetime Learning Publications, Belmont, CA, 1982.

[8] T.L. Saaty and L.G. Vargas, The Logic of Priorities: Applications in Business, Energy, Health, and Transportation, Kluwer – Nijhoff, Boston, 1982.