Bradford's Law and Fuzzy Sets: Statistical Implications for Library Analyses

Stephen J. Bensman

Stephen J. Bensman is a technical services librarian at Louisiana State University Libraries in Baton Rouge, Louisiana, USA. His main duties have been in cataloging and classification, collection development, and statistical analyses for purposes of collection development and management. He has a master's degree in library science and a doctorate in history, both from the University of Wisconsin in Madison, Wisconsin, USA, where he worked for a while as the foreign law librarian. He has published articles on Soviet legal bibliography, bibliometrics, as well as the socioeconomic structure of the scholarly and scientific journal system. He is owned by a dog named Tramp (shown in the photograph). He may be contacted at: notsjb@lsu.edu.

It was more than sixty years ago that S.C. Bradford published his first paper on the bibliometric law that bears his name.[1] Since that time much has been written on Bradford's Law of Scattering. Yet there is still much to be learned from this law. After considerable practical and theoretical research on the evaluation and utilization of scientific journals, I have come to the conclusion that Bradford's Law is actually a conflation of two concepts of vast importance in the utilization of statistical techniques in library analyses - probability distributions and fuzzy sets. As a matter of fact, this law can be considered as a mathematical description of a probabilistic model for the formation of fuzzy sets. In this paper I will discuss the statistical implications of Bradford's Law as a generator of fuzzy sets.

Bradford's Law

Bradford was Chief Librarian of the Science Museum Library (SML) in South Kensington, London. His main aim in the research, which led to the Law of Scattering, was to improve the coverage of science literature by the indexing and abstracting services. He was particularly disturbed by the gaps in this coverage, estimating that approximately 500,000 of the 750,000 scientific articles published each year were missed by the abstracting and indexing journals. The reason for this oversight was suspected to be the manner in which the literature of a subject was distributed among the periodicals containing it. Bradford summed up the hypothesis that was investigated by the research in the following terms:

> ... An alternative hypothesis...is that, to a considerable extent, the references are scattered throughout all periodicals with a frequency approximately related inversely to the scope. On this hypothesis, the aggregate of periodicals can be divided into classes according to relevance of scope to the subject concerned, but the more remote classes will, in the aggregate, produce as many references as the more related classes. The whole range of periodicals thus acts as a family of successive generations of diminishing kinship, each generation being greater in number than the preceding, and each constituent of a generation producing inversely according to its degree of remoteness.[2]

This hypothesis was tested with two sets of references from the current bibliographies being compiled in the SML. One of the sets pertained to Applied Geophysics; the other set was constructed from references in Lubrication. In this paper I will restrict the analysis to the Applied Geophysics set, mentioning only that the results in Lubrication were basically the same. The Applied Geophysics set encompassed references for the four years 1928-1931 inclusive, and it contained 1,332 references to 326 journals. When ranked in descending order, the journals ranged from one journal receiving 93 total references to 109 journals receiving one reference each. In his report Bradford described the results of grouping the journals in each set into the following three classes:

(a) those producing (on average) more than 4 references a year;
(b) those producing more than 1 reference and not more than 4 a year; and
(c) those producing 1 reference or less a year.

Table 1 below presents the results of Bradford's description of the grouping of the Applied Geophysics journals into classes. It is evident that these results validated the hypothesis being tested. Thus, class (a) accounted for only 2.8 percent of the journals but 32.2 percent of the references; class (b) contained 18.1 percent of the journals but 37.5 percent of the references; and class (c) had 79.1 percent of the journals but merely 30.3 percent of the references.

As Table 1 clearly shows, whereas the number of journals in each class rises exponentially, the number of references accounted for by the journals in each class remains approximately the same. The picture that emerged from the data caused Bradford to give the following verbal formulation to the Law of Scattering:

> ...if scientific journals are arranged in order of decreasing productivity of articles on a given subject, they may be divided into a nucleus of periodicals more particularly devoted to the subject and several groups or zones containing the same number of articles as the nucleus, when the number of periodicals in the nucleus and succeeding zones will be as $1 : n : n^2$... [2]

Here it should be pointed out that Bradford never related the number of articles on a given topic in various journals to the sizes of those journals. This was a question that did not interest him, because his main aim was to coordinate the indexing and abstracting agencies' handling of titles in such a way as to ensure full coverage on any topic. His focus was therefore on the articles on a topic and how they were distributed over various titles. We are dealing, therefore, in terms of percentages of articles. However, to classify the journals in the standard way, it would be necessary to investigate the matter in terms of the journals themselves - not the articles - and to verify how big a percentage of articles in a given journal was dedicated to a given topic. It could be that the title with the most articles on a topic - particularly, a narrow topic - would be a large multidisciplinary journal with only a small portion of its articles dedicated to the topic. Here I have followed Bradford's method of classifying journals by percentage of articles for the entire topic and not by percentage articles of a given journal. Nevertheless, the assumption underlying Bradford's law as revealed by the law's phrase "a nucleus of periodicals more particularly devoted to the subject" is that of a core of journals with most of their articles devoted to a topic, and it is this assumption on which I am basing this paper.

In terms of fuzzy set theory, the important aspect of Bradford's Law of Scattering is that it demonstrated the truth of the final point in the initial hypothesis that for any given subject set "[the] whole range of periodicals...acts as a family of successive generations of diminishing kinship, each generation being greater in number than the preceding, and each constituent of a generation producing inversely according to its degree of remoteness."

Class	Journals		References	
	No.	%	No.	%
(a) 4+ References per Year [93 to 17 Total References]	9	2.8	429	32.2
(b) 2-4 References per Year [16 to 5 Total References]	59	18.1	499	37.5
(c) 1-0.1 References per Year [4-1 Total References]	258	79.1	404	30.3
TOTALS	**326**	**100**	**1,332**	**100**

Table 1: Bradford Journal Classes in Applied Geophysics, 1928-1931 (inclusive).

Fuzzy Sets - Zadeh

Classical set theory is based on the idea that we can make clear, exact distinctions between groups. According to this theory, we should always be able to tell exactly whether an individual is definitely in a group or definitely outside a group. In his book on fuzzy logic Kosko traces this concept back to Aristotle, whose ideas in this respect he summed up in the following manner:

> Aristotle's binary logic came down to one law: A or not-A. Either this or not this. The sky is blue or not blue. It can't be both blue and not blue. It can't be A and not-A.[3]

"A or not-A" is a simple statement of the law of the excluded middle. In classical set operations A is assigned the number 1, whereas not-A is assigned the number 0.

Bradford certainly thought in terms of classical set theory. This is evident in a paper which he delivered in 1944 before the British Society for International Bibliography. In this paper Bradford explored the bases of the Universal Decimal Classification, defining a class as a "set of beings or things, having something in common."[4] He utilized the system of symbols and logic presented in George Boole's book *The Laws of Thought* of 1854 to demonstrate how the human mind logically classifies things into such sets. He derived an equation, which he considered as forming the basis of Boole's calculus of logic, coming to the following conclusion:

> We have, therefore, the law that "It is impossible that the same

> quality should both belong to and not belong to the same thing," which is Aristotle's Principle of Contradiction, which [Boole] regarded as the most certain of all principles.
>
> The fact that this equation is of the second degree, with two roots, 0 and 1, indicates the we perform the process of classification, by separation into pairs of opposites, e.g., men and not men, and we notice that only values 0 and 1, apply to whatever class we designate by any symbol x.4

Nevertheless, Bradford was well aware that "the mutual exclusiveness of classes is not always practicable."4

Classical sets are called 'crisp' sets in the literature in order to distinguish them from 'fuzzy' sets. The latter concept was first developed in a paper published in 1965 by Lotfi Zadeh.[5] In this paper Zadeh described fuzzy sets and their importance thus:

> More often than not, the classes of objects encountered in the real physical world do not have precisely defined criteria of membership. For example, the class of animals clearly includes dogs, horses, birds, etc. as its members, and clearly excludes such objects as rocks, fluids, plants, etc. However, such objects as starfish, bacteria, etc. have an ambiguous status with respect to the class of animals. The same kind of ambiguity arises in the case of a number such as 10 in relation to the "class" of all real numbers which are much greater than 1.
>
> Clearly, the "class of all real numbers which are much greater than 1," or "the class of beautiful women," or "the class of tall men," do not constitute classes or sets in the usual mathematical sense of these terms. Yet, the fact remains that such imprecisely defined "classes" play an important role in human thinking, particularly in the domains of pattern recognition, communication of information, and abstraction.

Zadeh defined a 'fuzzy set' as "a class of objects with a continuum of grades of membership," and he stated, "Such a set is characterized by a membership (characteristic) function which assigns to each object a grade of membership ranging between zero and one" (p. 338). In set theory, a Set A ("the class of tall men," to use one of Zadeh's examples) is delimited as a subset of some Universe of Discourse X ("the class of all men"). According to Zadeh's concept, a membership function assigns to each member x of the Universe of Discourse X a 'grade of membership' in A that ranges from 0 (not-A) to 1 (A). He pointed out that his fuzzy sets differed from ordinary crisp ones in that with the latter the membership function could only take on two values, 0 and 1, according to as x does or does not belong to A.

In classic set operations with two sets, for example, an observation is assumed to have a membership of one in both sets, i.e. a human can be simultaneously a girl and an undergraduate and therefore is in both these overlapping sets. However, with fuzzy sets, the membership is not clear, and this can introduce all sorts of exogenous variables. Taking the example above, the girl is actually bisexual. From this one can see how fuzzy sets can complicate the relationships.

In their book on measurement in information science Boyce, Meadow, and Kraft state, "The major measurement issue associated with fuzzy sets is the assignment of the value representing the degree of set membership."[6] According to these authors, this is often a more or less arbitrary process, and in respect to the indexing of documents they suggest two basic methods: subjectively by human indexers; or empirically by computer software on the basis of the frequency of word counts in the documents.

Bradford himself provided in his report on the Law of Scattering an empirical basis for deriving a membership function applicable to this law.[7] The empirical basis is inherent in the method used to establish classes (a), (b), and (c), which are shown in Table 1 above. Reviewing this process, the method was to divide the total number of references for each journal by the number of years encompassed by the sample - four years in the case of Applied Geophysics - to arrive at the average number of references per year. The classes were then defined according to the following criteria: (a) those producing more than 4 references a year; (b) those producing more than 1 reference and not more than 4 a year; and (c) those producing 1 reference or less a year. Replication of this technique resulted in quotients with no more than two decimal places, so that it was possible to establish the following class boundaries: between (a) and (b) at 4.01; between (b) and (c) at 1.01; and between (c) and the zero class, which I added and named (d), at 0.01. If one considers class (a) - which accounted for 2.8 percent of the journals but 32.2 percent of the references in Applied Geophysics - as "a nucleus of periodicals more particularly devoted to the subject," then it is possible to derive the following membership function for Bradford's sets:

> If the number of references per year to a journal is greater than 4, then the membership grade of this journal equals 1; but if the number of references per year to a journal equals or is less than 4, then the membership grade of this journal equals the number of references to it per year divided by 4.01.

The number 4.01 was selected as the divisor in the second part of the membership function, because this number marked the lowest limit of the nuclei. Applying this membership function to Bradford's data yielded the results shown below in Table 2 for Applied Geophysics. In this table Bradford's classes plus the additional zero class (d) have been named in accordance with the following fuzzy set principles that show descending set membership: (a) = A; (b) = A and not-A; (c) = not-A and A; and (d) = not-A. Inspection of this table reveals that below the nucleus or class (a) the membership grade of the journals skews rapidly downward as the number of these journals skews rapidly upward until the vast bulk of the journals can be considered to be only marginally in the Applied Geophysics set. The number of journals in the zero class (d) has been left deliberately open, as this is a complex question, which Bradford himself never successfully answered.

Classes	No. References per Year	Journals Producing References	Membership Grade (*)
(a) Applied Geophysics	23.25	1	1.000
	21.50	1	1.000
	14.00	1	1.000
	12.00	1	1.000
	11.50	1	1.000
	8.75	1	1.000
	7.00	1	1.000
	5.00	1	1.000
	4.25	1	1.000
Classes a/b Boundary	4.01		1.000
(b) Applied Geophysics/ Not Applied Geophysics	4.00	4	0.998
	3.75	1	0.935
	3.50	5	0.873
	3.00	1	0.748
	2.75	2	0.686
	2.50	5	0.623
	2.25	3	0.561
	2.00	8	0.499
	1.75	7	0.436
	1.50	11	0.374
	1.25	12	0.312
Classes b/c Boundary	1.01		0.252
(c) Not Applied Geophysics / Geophysics	1.00	17	0.249
	0.75	23	0.187
	0.50	49	0.125
	0.25	169	0.062
Classes c/d Boundary	0.01		0.002
(d) Not Applied	0.00	?	0.000

Table 2. Bradford's Law in Terms of Fuzzy Set Theory: Applied Geophysics, 1928-1931, inclusive.

Statistical Implications - Pearson

The inherent fuzziness of Bradford sets has major implications for the utilization of statistical techniques in library analyses. These implications derive from the principle of the unity of science, which Bradford placed at the basis of his law. "According to this principle," Bradford stated, "every scientific subject is related, more or less remotely, to every other scientific subject."[8] Due to this principle, as the membership grade of the documents or journals in a given Bradford set diminishes, the way is opened for materials from other scientific disciplines and therefore for influences exogenous to this set. The result is the inhomogeneity - sometimes the extreme inhomogeneity - of Bradford sets.

This process of inhomogenization complicates what Karl Pearson once described as "the fundamental problem of science."[9] Pearson described the essence of this problem in the 1911 edition of his *The Grammar of Science* in the chapter in which he introduced to the broader public the new concepts of contingency and correlation. In his approach to science, Pearson started out from the principle of classification, stating, "The classification of facts and the formation of absolute judgments upon the basis of this classification - judgments independent of the idiosyncrasies of the individual mind - essentially sum up the *aim and method of modern science.*"[10] According to his view variability is an essential characteristic of reality, and it plays an important role in establishing the conditions under which science operates. Thus, he wrote:

> ...The conclusions of the physicist and the chemist are based on *average* experiences, no two of which exactly agree; at best they are routines of perception which have a certain variability. This variability they may attribute to errors of observations, to impurities in their specimens, to the physical factors of the environment, but it none the less exists and, when it is removed by a process of averaging, we pass at once from the perceptual to the conceptual, and construct a model universe, not the real universe.[11]

On the basis of the variability of phenomena Pearson developed a new theory, by which he replaced the traditional idea of causation with the concept of category of association. He explained the new theory in the following passage:

> If we realize individuality at the basis of all existence, and sameness as a relative term depending on the fineness of classification, then we see that cause and effect ...only connote a degree of likeness, not an absolute repetition. The law of causation is a conceptual figment extracted from phenomena, it is not of their very essence. The actual problem before mankind is a far wider one than that of "causation," and may be summed up as follows: If the "causes" have such and such a degree of likeness, how like will be the "effects" be? Here in the broadest sense anything is a cause which antedates or accompanies a phenomenon, and we ask if we vary that cause to what degree we vary or change the

> phenomenon. If we say that variation of the cause produces no effect on the phenomenon we have absolute independence; if we found variation of this cause absolutely and alone varied the phenomenon we should say that there was absolute dependence. Such absolute dependence of a phenomenon on a single measurable cause is certainly the exception.... It would correspond to a true case of the conceptual limit--of whose actual existence we have our grave doubts. But between these two limits of absolute independence and absolute dependence all grades of association may occur.[12]

Pearson placed "the fundamental problem of science" within this context of classification and variation, writing:

> The universe is made up of innumerable entities, each probably individual, each probably non-permanent; all man can achieve is to classify by measurement or observation of characteristics these entities into classes of *like* individuals. Within these classes variation can be noted, and the fundamental problem of science is to discover how the variation in one class is correlated with or contingent on the variation in a second class.[13]

Pearson illustrated the new theory with a scatter diagram plotting variable A against variable B. The points on the diagram were scattered in the general shape of a curve. According to him, a physicist would handle the diagram by photographing it from 50 yards off or looking at it through an inverted telescope. By such methods the scattered points reduce to a smooth curve, and actual experience is replaced by mathematical function. Pearson then utilized the scatter diagram to sum up the relationship of causation to correlation thus:

> Take any two measurable classes of things in the universe of perceptions, physical, organic, social or economic, and it is such a dot or scatter diagram.... In some cases the dots are scattered all over the paper, there is no association of A and B; in other cases there is a broad belt, there is only moderate relationship; then the dots narrow down to a "comet's tail," and we have close association. Yet the whole series of diagrams is continuous; nowhere can you draw a distinction and say here correlation ceases and causation begins. Causation is solely the conceptual limit to correlation when the band gets so attenuated, that it looks like a curve.[14]

From the perspective of his new theory, Pearson criticized the old view of cause and effect in the following manner:

> ... Any variation within the existences in one class is found to be associated with a corresponding variation among the existences in a second class. Science has to measure the degree of stringency, or of looseness in these concomittant [sic] variations. Absolute independence is the conceptual limit at one end to the looseness of the link, absolute dependence is the conceptual limit at the other end to the stringency of the link. The old view of cause and effect tried to subsume the universe under these two conceptual limits to experience--and it could only fail; things are not in our experience either independent or causative.[15]

Since both his contingency coefficient and his correlation ratio designated absolute independence with 0 and absolute dependence with 1, it is evident that Pearson also was interested in the fuzzy area between 0 and 1, i.e., the fuzzy area of the excluded middle.

Outliers - Barnett and Lewis

Pearson's process of statistical inference through the measurement of the effect of one set upon another is complicated by the fuzziness of Bradford sets through the mechanism of outliers. Beckman and Cook describe an outlier as "a subjective, post-data concept,"[16] and this assessment is shared by Barnett and Lewis in a book that can be considered the standard treatment of the topic. In this book Barnett and Lewis define an outlier in a set of data to be "*an observation (or subset of observations) which appear to be inconsistent with the remainder of that set of data.*"[17] They then set forth the critical issue involved in outliers thus:

> The phrase 'appears to be inconsistent' is crucial. It is a matter of subjective judgement on the part of the observer whether or not some observation (or set of observations) is picked out for scrutiny. What really matters is whether or not some observations are *genuine members* of the main population. If they are not, but are **contaminants** (arising from some other distribution), they may frustrate attempts to draw inferences about the original (basic) population.[18]

Barnett and Lewis closely connect the problem of outliers with assumptions about the probability distribution underlying the population. An observation, which may appear to be an outlier under the assumption of a normal distribution, would not arouse any special concern if the observer were expecting a highly skewed distribution of the type to which biological, social and information data usually conform. Therefore they set as the null or working hypothesis of any discordancy test for outliers some basic probability model for the generation of all the data with no contemplation of outliers. If significant evidence is found for the rejection of the working hypothesis, Barnett and Lewis indicate a number of "contamination" or "outlier-generating" models that may serve as alternative hypotheses.

Of these alternative hypotheses two have the most relevance for this paper. The first is what they call the "deterministic alternative," which covers the case of outliers resulting from gross human errors of measurement, recording, etc. The second Barnett and Lewis term the "mixture alternative," where it is posited that the sample under investigation reflects contamination from a population other than that represented by the basic model and that such "foreign" sample members, or contaminants, are showing themselves as outliers.[19] Given the rapidly diminishing membership grade of members of a Bradford set and its concomitant rapid opening to such contaminants, the "mixture alternative" is of the utmost import for statistical analyses of library data.

Practical Demonstration

I now give a demonstration of the above concepts by utilizing data resulting from a project to restructure the serials holdings of Louisiana State University (LSU).[20] As part of the preparations for this project the faculty of the LSU Department of Chemistry were surveyed in April 1993 on their serials needs. Here it is necessary to emphasize that only the faculty of the Department of Chemistry were surveyed; the Departments of Biochemistry and Chemical Engineering were not included in this survey. The LSU Chemistry faculty were asked to identify those serials important to them for research and teaching purposes from the entire serials universe, without restricting themselves to the ones on subscription at LSU. Their selections were classified according to the subject categories assigned them in the 1993 *Science Citation Index Journal Citation Reports (SCI JCR)* published by the Institute for Scientific Information (ISI).[21]

In conformance with Bradford's Law of Scattering the LSU Chemistry faculty's journal selections ranged over numerous ISI subject categories, among which were the following: Engineering, Electrical and Electronic; Environmental Sciences; Geosciences; Materials Science, Ceramic; Nutrition and Dietetics; Physics; and Radiology and Nuclear Medicine. Due to this, it was decided to restrict the sample only to those journals that were classed by ISI in the various branches of Chemistry, including Chemical Engineering as well as Crystallography. As an exception, the ISI subject category Spectroscopy was also included due to the emphasis of the LSU Department of Chemistry on it, even though this discipline is generally considered part of Optics within Physics. The final result was a sample of 154 journals.

Three quantitative variables were employed to measure the scientific value of these 154 journals: LSU Faculty Score; Total SCI Citations in 1993; and 1993 SCI Impact Factor. Of these measures, only the first two were found to be valid.[22] LSU Faculty Score was considered to be the key measure of scientific value, because the logic of the journal set had been defined by a survey of the Department of Chemistry as well as for philosophical reasons. It was derived in the following manner. The Chemistry professors had been requested to name ten titles, state whether these titles had to be on campus or could be accessed through remote document delivery, and then to rank the titles in descending order from 10 to 1. A title was scored in the following manner: 10 points each time it was selected by a professor; another 10 points if the professor stated that it had to be on campus; and the points from 10 to 1, depending on the rank the professor assigned it. Twenty-five Chemistry professors responded to the survey, and the 154 journals ranged in LSU Faculty Score from 10 to 755 for the *Journal of the American Chemical Society.*

To validate the LSU Faculty Score, I correlated it with Total SCI Citations to determine how well LSU faculty ratings corresponded to the opinion of the publishing segment of the scientific community. It was in performing this operation that I came across a severe outlier problem. Table 3 below gives the distribution of both variables over classes defined by quartiles. This table clearly shows that we are dealing with a distribution of the Bradford type, as the upper end of the distributions account for most of the value. Thus, in Bradford's Applied Geophysics set 2.8 percent of the journals accounted for 32.2 percent of the references, and here the upper quartile class accounted for 62.5 percent of the 154 journals' combined faculty score and 80.2 percent of their total citations.

The frequency distributions of the faculty score and total citations are graphically shown below in Figures 1 and 2, which clearly manifest evidence of the presence of contaminants from a population other than the one being modeled by LSU Faculty Score. The contaminants are not revealed by the extreme observations on the right, which are a usual occurrence in distributions of the Bradford type, but in the relative positions of the *Journal of the American Chemical Society* (Faculty Score - 755; Total SCI Citations - 148,900) and the *Journal of Biological Chemistry* (Faculty Score - 197; Total *SCI* Citations - 231,324). In Figure 1, which shows the frequency distribution of the 154 journals by LSU Faculty Score, the position of the *Journal of the American Chemical Society* on the extreme right fits the logic of the set and is not surprising. However, in Figure 2, which depicts the distribution of these journals by Total *SCI* Citations, the position of the *Journal of*

	LSU FACULTY SCORE		TOTAL SCI CITATIONS	
	Quartile Class Range	% Faculty Score for All	Quartile Class Range	% Total Citations for All Journals
Upper	755 to 111	62.5	231,324 to 11,685	80.2
Upper Middle	110 to 50	20.6	11,586 to 3,303	13.4
Lower Middle	50 to 33	11.1	3,285 to 1,533	4.6
Lower	32 to 10	5.8	1,526 to 255	1.8

Table 3. Distribution of 154 Chemistry Journals in Descending Order by LSU Faculty Score and Total SCI Citations over Classes Defined by Quartiles.

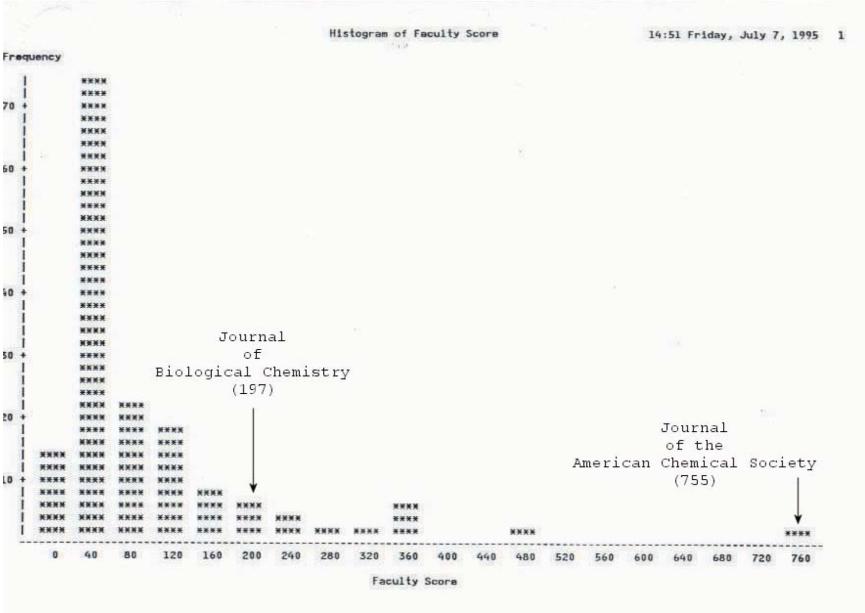

Figure 1. Frequency Distribution of 154 Chemistry Journals by LSU Faculty Score.

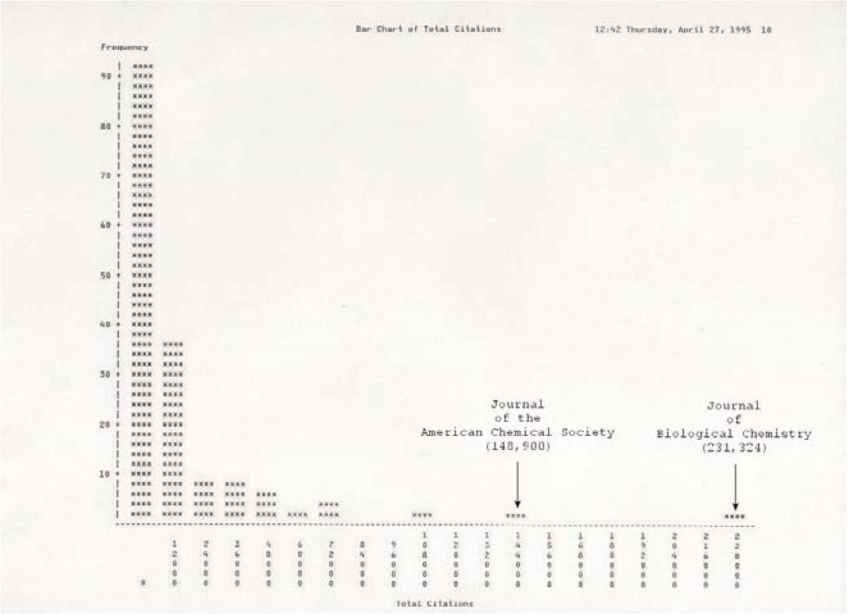

Figure 2. Frequency Distribution of 154 Chemistry Journals by Total SCI Citations.

Biological Chemistry to the right of the *Journal of the American Chemical Society* is surprising and discordant, as it does not fit the logic of the set.

The *Journal of Biological Chemistry* has the definite appearance of being an outlier, and this suspicion is confirmed in Figure 3, which is a scatter diagram plotting LSU Faculty Score against Total SCI Citations. In Figure 3 a hypothetical regression line drawn from the *Journal of the American Chemical Society* to the origin goes directly through the middle of the points, whereas a line drawn from the *Journal of Biological Chemistry* to the origin is below and to the right of all the points. When I constructed this set, I considered Biochemistry to be a branch of Chemistry. My decision in this respect was influenced by the treatment of Biochemistry by the Library of Congress classification schedules as a subset of Organic Chemistry within Chemistry. However, the position of the *Journal of Biological Chemistry* made me suspect otherwise. Subsequent research confirmed this suspicion. Unlike the Library of Congress schedules, the Dewey Decimal Classification has Biochemistry not as a subset of Chemistry but of Biology and Life Sciences. Moreover, not only does LSU have separate departments for Chemistry and Biochemistry, but, in the most recent ratings of US research-doctorate programs by the National Research Council, Chemistry was classified under the rubric of Physical Sciences and Mathematics, whereas Biochemistry was combined with Molecular Biology and placed in the Biological Sciences.[23] Thus, the *Journal of Biological Chemistry* together with a number of other biochemical journals was in my set as a result of the fuzziness of Bradford sets. The *Journal of Biochemistry* was both A and not-A, both Chemistry and Biochemistry, as well as who knows what else.

Methods for Handling Outliers

Barnett and Lewis group the methods for handling outliers into four general categories.[24] There are no hard and fast rules for determining which category of methods should be utilized, because everything depends upon how the outliers arose and the purpose one is trying to accomplish. Barnett and Lewis term one of their categories **rejection.** By this they mean that one discards the outliers, if these cannot be corrected, and then subjects the remaining sample to analysis. This is in effect what I did, when I found five outliers in performing the Pearson product-moment correlation between LSU Faculty Score and Total *SCI* Citations.[25] Analysis of the residuals revealed five outliers, of which four had a low faculty score in respect to their total citations. Two of the latter outliers had been classed by ISI in Biochemistry and Molecular Biology. The initial correlation coefficient was 0.66, and removal of the outliers from the

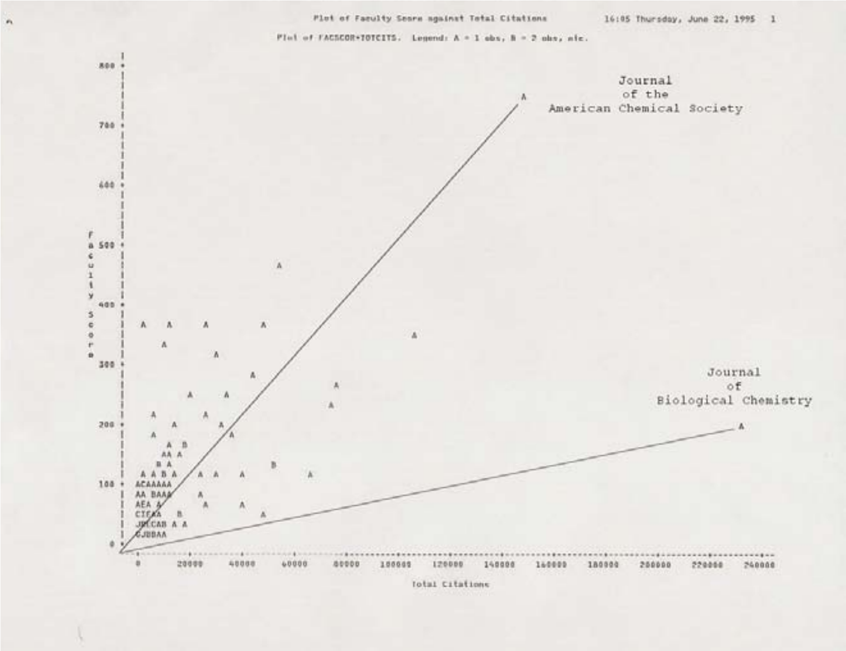

Figure 2. Frequency Distribution of 154 Chemistry Journals by Total SCI Citations.

sample raised this coefficient to 0.72.

A second category of methods for handling outliers is called by Barnett and Lewis **identification**. By this they mean that one should study the discordant outliers as a sign of some unsuspected factors at work in the population under analysis. I also did this, coming to new conclusions on the relationship of Biochemistry to Chemistry. Whereas I first thought of Biochemistry as a branch or subset of Chemistry, I now came to regard it as a separate discipline or set with its own statistical patterns.

Another Barnett and Lewis category of outlier procedures is **incorporation**. The aim of this type of procedures is to replace one homogeneous model with another homogeneous model for the entire sample (incorporating the outliers), in relation to which no observations appear discordant.

The fourth and final category of Barnett and Lewis for handling outliers is **accommodation**. This category is divided by them into two components. The first component contains procedures that are 'robust' or retain reasonable validity in the face of outliers. An example of a robust procedure would be to utilize the chi-square test of independence instead of correlation techniques to investigate the relationship of LSU Faculty Score to Total *SCI* Citations. The chi-square test of independence was pioneered by Karl Pearson on the basis of contingency. Whereas correlation techniques entail the precision of a mathematical function in that they measure either the fit of the data points to a regression line, in the case of the Pearson product-moment correlation, or the relationship of one specific rank to another specific rank, in that of the Spearman rank-order correlation, it is possible to test the correspondence of variables to each other within broad categories with the use of the chi-square test of independence. This possibility is evident in Table 3 above, where it can be seen that the *Journal of the American Chemical Society* and the *Journal of Biological Chemistry* fall in the Upper Quartile Class on both measures of scientific quality.

The other component of **accommodation** encompasses those methods that protect against outliers by placing less importance on extreme values than on other sample members. One such method could be Winsorization, whereby an extreme observation is replaced by its nearest neighbor. By this technique the *Journal of Biological Chemistry* would be assigned the same number of total citations as the *Journal of the American Chemical Society.* However, perhaps a better method of the latter component of **accommodation** would be to apply fuzzy set theory to the handling of outliers. The application of this theory would be empirical in nature and depend on the logic of the set under analysis as well as the purpose of the research. Its main aim would be to adjust the outliers to be proportionate to their membership in the set. In terms of the example being used, one way to accomplish this would be to empirically derive a membership function off LSU Faculty Score and to use the resulting membership grade to adjust the Total *SCI* Citations of the outliers. Another method could be to analyze the ISI database and restrict the citations only to those that pertain to the logic of the set. From this it can be seen that probability theory and fuzzy set theory are not antagonistic but complementary forms of analysis.

References

1. Bradford, S.C. "Sources of Information on Specific Subjects". *Engineering* 137: 85-86 (1934).
2. *Ibid.* A full treatment of the various formulations of Bradford's Law is outside the scope of this paper.
3. Kosko, B. *Fuzzy Thinking: The New Science of Fuzzy Logic.* London: Flamingo, 1994, p. 6.
4. Bradford, S.C. "Some General Principles of a Bibliographical Classification Scheme, with Application to the Universal Decimal Classification". *Proceedings of the British Society for International Bibliography* 6 (3): 57-69 (1944).
5. Zadeh, L.A. "Fuzzy Sets". *Information and Control* 8 (3): 338-353 (1965).
6. Boyce, B.R., C.T. Meadow, and D.H.Kraft. *Measurement in Information Science.* San Diego: Academic Press, 1994, p. 95.
7. Bradford (1934), *op cit.*
8. Bradford, S.C. (1948). "Complete Documentation" in *The Royal Society*

Empire Scientific Conference, June-July 1946: Report. London: The Royal Society, 1948, Vol. 1, pp. 729-748.

[9] Pearson, K. *The Grammar of Science.* 3rd ed., rev. and enl. London: A. and C. Black, 1911. Pt. 1, p. 165.

[10] *Ibid.*, p. 6.

[11] *Ibid.*, p. 154.

[12] *Ibid.*, pp. 156-157.

[13] *Ibid.*, p. 165.

[14] *Ibid.*, p. 170.

[15] *Ibid.*, p. 166.

[16] Beckman, R.J. and R.D. Cook. "Outliers". *Technometrics* 25 (2): 119-149 (1983).

[17] Barnett, V. and T. Lewis. *Outliers in Statistical Data.* 3rd ed. Chichester: J. Wiley, 1995, p. 7. A short, general review of outliers and the procedures for handling them is: Barnett, V. "The Study of Outliers: Purpose and Model". *Applied Statistics* 27 (3): 242-250 (1978).

[18] Barnett and Lewis (1995), *op cit.*, p. 7.

[19] *Ibid.*, pp. 43-49.

[20] Bensman, S.J. "The Structure of the Library Market for Scientific Journals: The Case of Chemistry". *Library Resources & Technical Services* 40 (2): 145-170 (1996); Bensman, S.J. and S.J. Wilder "Scientific and Technical Serials Optimization in an Inefficient Market: A LSU Serials Redesign Project Exercise". *Library Resources & Technical Services* 42 (3): 147-242.

[21] *Science Citation Index Journal Citation Reports: A Bibliometric Analysis of Science Journals in the ISI Database: 1993.* Philadelphia: Institute for Scientific Information, 1994.

[22] Bensman (1996), *op cit.*

[23] Goldberger, M.L., B.A. Maher, and P.E. Flattau, eds. *Research-Doctorate Programs in the United States: Continuity and Change.* Washington, D.C.: National Academy Press, 1995.

[24] Barnett and Lewis (1995), *op cit.*, pp. 27-43. See also Barnett (1978), *op cit.*

[25] Bensman (1996), *op cit.*