INFORMETRICS 87/88
L. Egghe and R. Rousseau (Editors)

COMMENTS ON THE SCOPE OF BIBLIOMETRICS

B.C. BROOKES

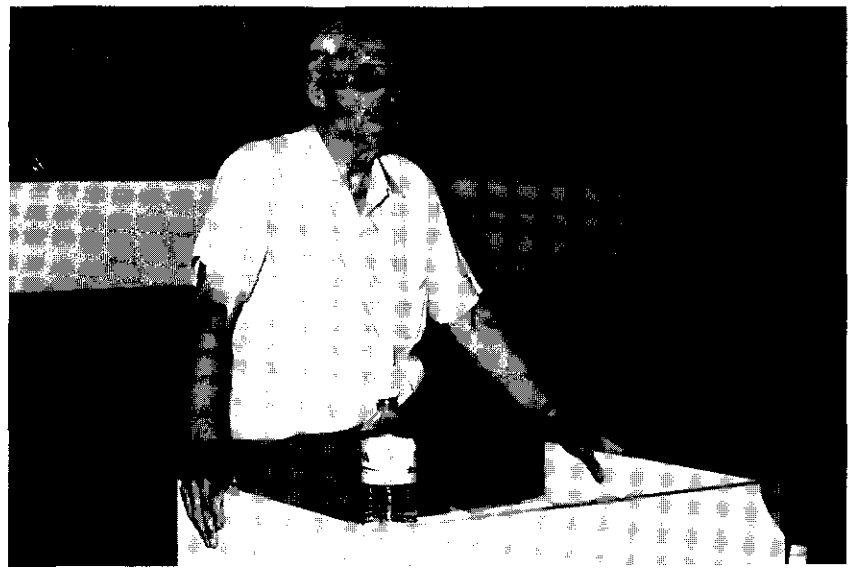
The City University, London, U.K.

Abstract

Bibliometrics emerged as a distinguishable study in 1969 at a time when, in the U.K. (as elsewhere too) university librarians were forced to abandon any 'Alexandrian' aims to be wholly selfsufficient. Close interaction with the new British Library and its Lending Division called for techniques of selection and integration that had hitherto not been seriously needed. The concurrent applications to library work of computers and developments in telecommunications helped to speed up these basic processes but also brought new problems and an abundance of data which invited exploration.
Recent developments in Statistics - notably the analysis and clarification of the 'long-tailed Zipfian' distributions - seem to suggest that bibliometrics is wholly reducible to applied statistics. But critical discussion of one of these frequency distributions - Sichel's Inverse-Gaussian/Poisson - suggests that there remain some aspects of bibliometrics beyond the reach of techniques dependent on the analysis of frequency distributions. There therefore remains a theoretical gap yet to be bridged.

1. TERMINOLOGY

The term bibliometrics dates from 1969. It was proposed by Alan Pritchard (1) to replace the ambiguous term statistical bibliography and other terms variously used up to that time. It was intended to embrace all the quantitative techniques then used to help organize library and information services more effectively. In the West, it was adopted forthwith, with no dissentient voice whatever. And I note that it is now being used retrospectively to reach back to the 19th century.
At about the same time, in the countries of the Eastern bloc, the term scientometrics was adopted to cover the techniques applied to the quantification and analysis of scientific activities including the publication and organizing of books and journals. The pedant may find some differences of connotation between these two terms but, even if he does, the two lines of study seem to be convergent.
A third term informetrics has recently been adopted by F.I.D. but, as far as I can see, it is being used to cover both sciento- and biblio- metrics impartially. It has produced no distinctively new ideas of its own but as it implicitly covers both documentary and electronic forms of information, it may have a future.
In Britain, Pritchard's proposal was timely. The mere coining of an appropriate neologism helped to focus attention on a problem that had arisen in an acute form. Great changes were suddenly being imposed on British library services by the founding of the British Library, especially by the setting up of the British Library's Lending Division at Boston Spa.
Up to that time, all university librarians had aimed to make their own campus wholly self-sufficient in documentation. When they had filled their shelves to bursting point, they had hitherto confidently demanded of the funding agency a new building to house their future acquisitions. But, suddenly, librarians were told



B.C. Brookes

that when they had filled their present shelves they would have to make any new shelf-space they needed by weeding out their least-used material and sending it to Boston Spa. There again it would be filtered: some would be discarded and the rest would be reorganized for wider national use - by direct access, by loan or by photocopy.

In a relatively small compact country like Britain such integration of the national library services was feasible. The integration was aided by the beginnings of computer applications to library and information services. Similar integrating processes were going on elsewhere, though not exactly parallel to those in Britain, notably in the U.S.A.. Yet, even today, there are many countries inadequately served by their documentary resources and which cannot therefore fully exploit the global networks of information systems that telecommunications now bring. It is in those countries that I note the keenest and most basic applications of bibliometrics.

So, suddenly, the simple techniques that bibliometrics could then offer were widely demanded and exploited. At the same time, new information services such as those offered by the indexed documentary databases became more accessible and Cyril Cleverdon in England and Gerard Salton in the U.S.A. began to show us how to use them more effectively. The founding of the I.S.I. by Eugene Garfield initiated a whole new industry of citation analysis on a world-wide scale. And since that time, developments in computers and telecommunications have proliferated. The field of information science abounds in countable entities - in journals, papers, authors, users, citations - all distributed over time, over many countries, over many languages, over many subjects.

This whole new world of endless countabilities invites detailed exploration. And that exploration can be done by just sitting still - as long as you have command of a computer terminal. Bibliometricians find it very intriguing.

In this situation one must expect to see what I can only call "some splashing about" - it is the natural way of trying to find out what bibliometrics can do.

But I am also pleased to discern the beginnings of a more critical approach - serious attempts to push bibliometrics to its limits to discover what those limits are. We are at present exploring only the surface of something - human knowledge and communication - which has as yet unknown depths.
Observing this scene I find myself asking: What is the scope of bibliometrics in this limited sense ? As I scan its current literature, I note ingenious but ad hoc applications of techniques that are basically those of applied statistics. So can bibliometrics be distinguished from applied statistics ? Or does the name bibliometrics merely mark off a particular area to which statistics is applied ? I believe it has something it can call its own. What is it ?

2. APPLIED STATISTICS AND THE LONG-TAILED DISTRIBUTIONS OF BIBLIOMETRICS

To avoid the risk of bias, I consulted a compact encyclopaedia (2) for its description of Statistics. Its densely written compact account mentioned, inter alia, the following:
Statistics: ... the manipulation of numerical data.
Descriptive statistics: ... the classification and presentation of data.
Sampling fundamental to statistics.
Random sampling; impractical to test every element of a population.
Mathematical model of [frequency] distributions.
Goodness of fit.
Probability.
All these ideas and concepts abound in the literature of bibliometrics. There is no doubt that bibliometrics is very heavily indebted to statistics. But I have underlined three items which bear closer examination in our context.
One of these is the concept of the frequency distribution. If we can cast the data we are concerned with into the form of a frequency distribution, then the whole powerful corpus of applied statistics can be brought to bear on its analysis.
However, the earliest frequency distribution of interest to bibliometrics is Lotka's law of 1927 which is usually expressed as the discrete inverse square law:

$$p(x) = k \sum 1/x^2 \quad , \quad 1 < x < \infty .$$

This infinite series is known to converge to the limit of $\pi^2/6$, so $k = 6/\pi^2$. But the mean of this distribution is

$$\frac{6}{\pi^2} \sum 1/x$$

which is infinite.
And so all its higher moments are infinite also.
Statistical sampling theory, however, requires the frequency distributions it was created to analyse to have at least some finite moments - the more the better. A statistician would say of the Lotka distribution that "it has no moments".
So the very first frequency distribution of bibliometrics lies outside the scope of orthodox statistical theory. In the standard text-books it receives no mention whatever.
The second distribution in the history of bibliometrics is that of Bradford (1934) (3). This was originally presented as a rank distribution by Bradford but when such data are cast into the form of a frequency distribution Egghe has shown that this distribution too is closely related to Lotka's.
Though in bibliometrics we find exemplifications of all the frequency distributions arising in everyday applied statistics - the normal or Gaussian, the binomial, the exponential, the negative exponential, the negative binomial, the beta, the gamma and many others - we also find this family of 'long-tailed' distributions which appear to be unique to bibliometrics.

Together with Zipf's laws (1934), derived from the statistics of vocabulary and to which Lotka's and Bradford's laws are also related, these 'long-tailed' laws were long regarded as statistical oddities which inhabited regions 'beyond the pale', beyond the reach of statistical theory.
Happily, the long-tailed denizens of this remote region have recently begun to receive expert attention. The first paper I saw on this topic was by S.D. Haitun, a Russian contributor to scientometrics (4). He called the distributions of orthodox statistics 'Gaussian', because they can all be derived from that distribution, and called the long-tailed distributions of scientometrics Zipfian. He then set to work exploring systematically the statistical properties of Zipfian distributions which he regarded as wholly distinct from the Gaussian. I was intrigued to note that he related the occurrence of these peculiar distributions to the steadily increasing entropy or running down of the physical universe. I took that idea, however, to signify his desperation, as a good scientometrician, to find a physical basis for these Zipfian phenomena.
At this point I came to surmise that whereas Gaussian forms were based on physical effects, the Zipfian forms in some way reflected human judgments as, for example, in deciding whether a paper is relevant to one's interests or not.
This surmise of mine was, however, shattered when the statistician H.S. Sichel revived interest in a distribution known as the Inverse Gaussian and, by compounding it with a Poisson distribution, showed that, in its general form, it captured the whole family of long-tailed Zipfian distributions (5). The most general form of the Inverse Gaussian/Poisson distribution (IGPD) has to be expressed in terms of Bessel functions and looks very intractable. But when its three parameters are assigned certain simple values, the Bessel functions collapse into familiar 'Gaussian' distributions well-known in bibliometrics, such as the negative binomial. More recently, Sichel has used the IGPD to capture the vocabulary of the English language - expressed as a frequency distribution - and has thus achieved with impressive comprehensiveness and precision one of the major objectives Zipf had set himself in the 1930's.
The Inverse Gaussian distribution first arose from a study of the Brownian motion first observed by the botanist Robert Brown in 1827. When examining a suspension of pollen grains in water under a microscope, he noticed that all the grains were in erratic motion, staggering about in all directions. Were they not living particles ? This first hypothesis was quashed when it was seen that particles of finely-ground rock wriggled in just the same way. So the Brownian motion was a purely physical phenomenon.
The physicists Einstein and Schrödinger studied this random motion. One of the measurements taken was to time the particles as they staggered from the centre to the circumference of a circle inscribed on the microscope slide. Schrödinger showed that the distribution of these times conformed with what is now known as the Inverse Gaussian distribution.
The Zipfian distributions can thus be derived from purely physical effects and so are not necessarily related to human judgments at all. But I still find it difficult to swallow the fact that the Bradford law, used in bibliometrics in a highly deterministic way, is derivable from a distribution describing the highly erratic Brownian motion when that distribution is further randomized by compounding it with a Poisson !
What then remains of bibliometrics that has not yet been captured by applied statistics ? There are still a few residuals.

3. RANKS

In bibliometrics we sometimes use ranks but the use of ranks sometimes gives rise to doubts, especially among those mathematicians and statisticians who retain from their formal upbringing any sensitivity to Number Theory. The word rank also has other meanings in the English language, one of which is exemplified in a line from Shakespeare's Hamlet: "O my crime is rank, it smells to heaven !!". This is one of those quotes which my colleagues brought up in England may well have now buried deep in their sub-conscience minds but which still disturbs their

attitudes to the word rank.
As the IGPD has now captured all Zipfian distributions, why not convert bibliometric ranked data to the corresponding frequency distribution and so make full use of all manifold richness of the statistical techniques now available ? Is anything gained by retaining ranks ?
The ranks 1st, 2nd, 3rd, and so on are ordinal numbers - they simply order, by some criterion, the entities of a set to which they are applied. But the numbers 1, 2, 3,.. and so on which we have to use in calculations are cardinal numbers. So, if we use ranks in calculation, we need some device for converting ordinals to cardinals. How can this be done ?
Since M.G. Kendall (6) developed the statistics of rank correlation, ranks have been more widely used in statistics. So it is worth examining how Kendall justifies his conversion. To say that an object of a set has been ranked 5th (he argued) implies that 4 other objects of the set have been given a preference over that 5th object. Similarly, to say that an object of the set has been ranked 7th implies that 6 objects of the set have been given priority over that 7th object.
It is obviously meaningsless to state that 5th + 7th = 12th, or that 5th times 7th is equal to 35th. But we can say that between the 5th and the 7th objects there is a difference of two preferences and as two is a cardinal number, we can calculate !
That is Kendall's argument. He did not explain, however why he should equate 1st to 1, which is what he does. His argument about differences of ranks would be equally valid (or otherwise) if he were to put 1st equal to any arbitrary number, say q, so that 1st = q, 2nd = q + 1, 3rd = q + 2 and so on. Though obviously it helps his calculations to put 1st = 1.
Ranks were introduced into bibliometrics by Bradford and into vocabulary studies by Zipf in the same year - 1934. Both Bradford - a numerate physical chemist - and Zipf introduced ranks into their work because they needed them. Ranks were also adopted by Leimkuhler (7) in his formulation of the Bradford law which is now widely applied. It is now usually expressed as:

$$G(r) = k \ln(1 + br), \quad r = 1, 2, 3, ..$$

or as

$$G(r) = k \ln(a + r) - k \ln a \tag{3.1}$$

where k and a = 1/b are parameters evaluated from the data and G(r) is the cumulated total of the frequencies.
Here implicitly we are putting 1st = 1+b, rth = 1+rb or 1st = a+1, rth = a+r. In the first case, the difference between successive ranks is b and in the second case it is 1.
So, in effect, bibliometrics has adopted Kendall's rank conversion plus a modification for the conversion of 1st. We can therefore claim the authority that rank correlation, now highly developed, confers on this process. Nevertheless, there is an arbitrariness here that needs a more formal basis to make the use of ranks statistically respectable.
For many years, from the late 19th C up to the 1920's mathematicians regarded statistics with grave doubts. The doubts related to probability. What kind of number is this probability ? For more than 50 years, at least in England, mathematicians felt that statistical theory was built on sand. Statisticians, however, did not worry overmuch about this issue; they simply got on with their job of developing their theories and applying them. They solved quantitative problems that pure maths could not reach. The evident successes of statistics helped greatly in allaying the doubts but the basic issues were finally cleared up by Kolmogorov and others in the 1920's.
I sense that statisticians properly brought up on frequency distributions and sampling theory regard ranks in much the same way as mathematicians used to regard probability. So, if bibliometricians wish to retain their rank techniques, they must at least demonstrate that by using ranks they can solve problems

which orthodox techniques cannot reach.
As Sichel (7) has recently applied the IGPD to some of Bradford's own data, we have a chance to compare the two approaches.

4. BRADFORD'S BIBLIOGRAPHICAL PROBLEM

Bradford himself did not apply what I have called 'the usual rank analysis' - that emerged only after a long period of disputation (which still continues). So before attempting any calculation on the Applied Geophysics (AG) data we should remind ourselves about the problem Bradford was concerned with. At that time - the 1930's - the Science Museum Library of which he was Director had established a high reputation for the excellence of their bibliographies on diverse scientific topics.
Bradford estimated that there were about 15,000 periodicals contributing scientific papers at that time; his Library has as good a collection as any in Britain. His staff would have access to Chemical Abstracts and Physical Abstracts and doubtless to other internal lists and files because they were continuously engaged in preparing bibliographies as a free public service to any bona fide requester. The 1930's were in a period well before photocopying, microforms or computers were available in libraries. Books and journals were carried around the world by train and boat; all documentary information was transported by the mail service. It took about a year before one could claim to have searched the previous year's literature.
Bradford's AG data relate to a period of 4 successive years 1929 tot 1932 - though Bradford has to admit that the data for 1932 were not complete when he published his paper. What he noted was that in each successive year new AG sources arose from journals whose average productivity was less than 1 per year. He divided all the productive sources into two main types: those that produced at least 1 paper per year (and therefore at least 4 papers in 4 years) and those whose average annual yield was less than 1 per year. On examining the total AG data retrospectively he found:

Year	Total sources	AV < 1	Annual diff.
1929	145	86	-
1930	218	155	73
1931	281	230	68
1932	326*	258*	45*
* Data incomplete			

What disturbed Bradford was the initially large and evergrowing proportion of sources found in the intermittently productive group. This group, he estimated, of about 1,000 potential sources could not be identified without scrutinizing a much larger number - perhaps 3,000 or so - over a long period. Because this kind of relatively unrewarding work was also going on in other bibliographical agencies around the world, he proposed that all scientific journals be completely indexed by a 'clearing house' in their country of origin and that such clearing houses should inform all others and any other agencies with specific interests of their findings. Only in this way, he felt, could papers of value to science not be missed. A good idea !
Reverting to the AG data, he suggested that the total numbers of sources identifiable year by year could be represented by the formula

$$S_n = a + b\,(1 + r + r^2 + r^2 + r^3 + \ldots r^{n-1})$$
$$= a + b\,(1 - r^n)/(1 - r) \qquad (4.2)$$

where n is the no. of years from the first count, and a = 60, b = 85 and r = 0.9 for the particular case of the AG data. He thus estimated that 'in the limit', S = 910 sources.
What Bradford clearly assumes in his paper is that, apart from the incompleteness of the 4th year, all the data for the previous 3 years were complete. I have no reason to doubt this assumption.
But Bradford's analysis has no clear relation to the usual application of his law.

5. APPLICATION OF THE BRADFORD LAW TO BRADFORD'S AG DATA

Implicit in Bradford's analysis of the AG data is the idea that the corpus of periodicals contributing to a specified topic can be divided into two groups. The first group consists of sources which regularly yield at least one relevant paper per annum and therefore at least 4 in the 4 years of observation. These are the 'nuclear' journals which are easily recognised and cause the bibliographer no problems. The second group includes all those potential sources, identified or otherwise, which may contribute, on average, less than 1 paper per annum. Bradford further assumed that it was reasonable to regard this irregular group of sources as a set of S sources each of which was subject to the same Poisson mean.
If we denote this mean by u and the unknown total of this group by S, we would expect to see:

$$S(1 - e^{-u}),\ S\,(1 - e^{-2u})\ ,\ S(1 - e^{-3u}) \text{ and } S(1 - e^{-4u})$$

sources reveal themselves in the successive years of the 4-year period. The number of new sources identified each year after the first would be:

$$Se^{-u}(1 - e^{-u}),\ Se^{-2u}(1 - e^{-u}) \text{ and } Se^{-3u}(1 - e^{-u}).$$

Comparing these results with Bradford's parameters in (4.2) we see that his $r = e^{-u}$ and the expected numbers of sources in this group to be revealed in year 1 to be $S(1 - e^{-u}) = b$.
Though this analysis looks plausible, it is not reasonable to assign equal means to all the potential sources a priori. Nor is it easy to see how this set relates to the Bradford law to which the nuclear sources conform. It would, however, conform with the law if we regarded this S group to have means ranging from k/r = 1 to k/r = 0 under the graph of y = k/r.
The no. of identifiable sources arising from this mixed Poisson group in year 1 would then be expected to be:

$$I = \int_k^\infty \frac{k}{r}\ e^{-k/r}(e^{k/r} - 1)\ dr,\ k \leqslant r < \infty.$$

On putting k/r = u, with $dr = -k/u^2\ du$, we have

$$I = k \int_0^1 (1 - e^{-u})/u\ du,$$

$$= k \int_0^1 (u - u^2/2! + u^3/3! - \ldots u^r/r! \ldots)/u\ du\ .$$

This function can be integrated term by term as far as needed.
To four significant figures, we thus find I = 0.7966k.
The corresponding number of papers is:

$$p = \int_0^1 k\ (u.e^{-u}.e^{u}.u)/u^2\ du = k.$$

As the average yield of the identified sources would be 1/0.7966 or 1.2553 papers, some of them must yield more than 1.
Does this set of sources and papers fit cleanly on to the upper end of the Bradford log-linearity set up by the nuclear sources ? And, if so, what is the end-point of the linearity ?
Assume that the nuclear group is ranked from 1 to n and the mixed Poisson group from n+1 to m. This fitting imposes the conditions:

$$k\ \ln\ (a + m) - k\ \ln\ (a + n) = k \qquad (5.1)$$

which implies that:

$$(a + m) = e(a + n)$$

and that

$$(a + m) = (a + n) + 1 \qquad (5.2)$$

We thus find that $a + m = Ie/(e-1) = 1.260k$
and that $a + n = I/(e-1) = 0.464k.$

Before these results can be applied to the AG data (Table I) we need to evaluate k and a from those data. Here I find the equation of the line joining G(1) = 93 and G(68) = 928 with the ranks marked off along a log scale.
The reason for choosing G(68) is that it lies close to the end of the nuclear group: G(68) = 928 marks the end of those sources which contribute at least 5 papers over the 4-year period whereas any higher point may reflect the known incompleteness.
So we solve the equations for b and k:

$$G(1) = k\ \ln\ (1 + b) = 93$$
$$G(68) = k\ \ln\ (1 + 68b) = 928.$$

By iteration we find that k = 279 with b = 0.396.
On this basis the totals to be expected in the complete AG bibliography would therefore be:

No of sources $= 1.26k = 352$
and no of papers $= 279\ \ln(1 + 352b) = 1379$.

Bradford's own estimate for S_4 is $60 + 850\ (1 - 0.9^4) = 352$ also, but he did not estimate the total of papers: here I find a total of 1379 as against the 1332 already identified in his data. It is not implausibly high.
I explore these two models in more detail to see how they compare in Appendix B.
Though several disputable issues arise in my calculations, it is important to note that the AG bibliography Bradford reports is the result of a 4- or 5-year patient search by skilled bibliographers and that the Bradford-law techniques are only rarely applied to bibliographic data sets as nearly complete as is the AG set.
Though the interface between the nuclear and the intermittent groups is unlikely to be as clean - as 'unfuzzy' - as I have assumed, it looks as though the totals calculated on the assumption that the log-linearity continues to its end-point provides reasonable estimates of both totals.
The AG analysis suggests that the ratio of nuclear to peripheral intermittent sources is as 1:e but I am not yet satisfied that this ration holds generally - the A.G. may be a special case.
However, a Poisson model of the intermittent sources does at least offer the advantage of making error estimates.

6. APPLICATION OF THE IGPD TO BRADFORD'S AG DATA

Sichel (8) has applied the IGPD to the AG data. We have to remember that his objective was to demonstrate that the IGPD can, with suitable parametric values, be fitted to bibliographic data when the ranked form is converted to the corresponding frequency distribution. This he has indisputably achieved as seen in Table 1. Sichel was wise to choose a data set as nearly complete as that of the AG. But, even so, there are some weaknesses in the fit.

Table I : Bradford's data on Applied Geophysics

Bradford's ranked d.			Sichel's frequency dist.		
				Journals	
r	g	Eg	Papers	Obs.	Calc.
1	93	93	1	169	169.0
2	86	197	2	49	49.0
3	56	235	3	23	25.3
4	48	283	4	17	15.9
5	46	329	5	12	11.1
6	35	364	6	11	8.2
7	28	392	7	7	6.4
8	20	412	8	8	5.1
9	17	429	9-10	8	7.6
13	16	493	11-12	3	5.4
14	15	508	13-15	6	5.7
19	14	578	16-20	6	5.9
20	12	590	21	7	11.4
22	11	612		326	326.0
27	10	662			
30	9	689			
38	8	753			
45	7	802			
56	6	868			
68	5	928			
85	4	996			
108	3	1065			
157	2	1163			
326	1	1332			
			χ^2 = 5.823, 10 d.f.		
			$p\ (\chi^2 \mid 10) = 0.830$		

The most dubious datum in the AG frequency distribution is the value of f(1) = 169. As the missing data were likely to be drawn from the sources of lowest productivity, it is this datum which is most at risk. Yet it is this datum, expressed as a fraction of the total (also incomplete) which Sichel adopts for one of his 3 parameters. The second parameter is the mean of the data which, of course, can be only that of the incomplete data also. The third parameter is put equal to zero a priori.
There is therefore no recognition of the fact that the data are not complete and no awareness of Bradford's problem.

Further doubts arise from the grouping of the ranked data. The most serious doubt arises from the grouping of the first 7 ranked sources, which together yielded 392 of the 1332 or 29% of the total. These, with productivities ranging from 93 to 28 are grouped as 7 items of yield $\geqslant$ 21. But Sichel is much more concerned with sources than with papers.

When the calculated values of the frequency groups are compared with the data, it suggests that the frequency of this group should be 11.4 rather than the 7 of the data. But the implication of this difference is that Bradford's expert bibliographers must have missed four nuclear sources with a mean productivity of about 50 papers - a total of the order of 200. After their long search Bradford's staff would find this idea difficult to accept, I guess. And so do I.

So though Sichel has demonstrated that the IGPD can be fitted to Bradford's data converted to the frequency form, one is left asking the questions: So what? What follows?

His approach also implies that he regards the data as a random sample from the population of the AG literature. It is certainly not that. It constitutes more than 90% of that population and insofar as it can be considered to be a sample at all, it is very highly biassed indeed towards the nuclear sources.

What I most deplore is, however, the discarding of hard-won empirical information implicit in the conversion of the ranked data to the compacted frequency distribution. It is evident, by looking at the two tables, that the ranked data could not be re-constituted from the frequency distribution. So some information has been lost. Applying the Brillouin measure to the two distributions, I find that the frequency distribution has discarded almost 70% of that inherent in the original data. No statistical analysis, however sophisticated, can compensate for the discarding of empirical information on that scale.

7. THE SCOPE OF BIBLIOMETRICS

Though I have been highly critical of Sichel's fitting of the IGPD to Bradford's AG data, what my criticisms amount to is that his work does not help me directly in solving the problem I had set myself - that of trying to estimate the complete from the incomplete AG data. But Sichel's problem was the very different one of using the data as though it were a single datum - a well-known classical set of bibliographic data on which he could test how well the new distribution fitted. In fact, I applaud his efforts to do just that and I confidently expect further results to follow as he continues to exploit and refine the new analytical techniques applicable to Zipfian distributions he has initiated.

But, at the same time, I was also stressing that though Sichel and others may be primarily concerned to extend the reach of applied statistics into bibliometrics, I and many others, including the users who depend on searching the databases for their individual purposes, are still dependent on the ranked distributions of Bradford and Zipf, crude as they may sometimes appear to be. We have learned to live with them and to make ad hoc adjustments to any particular data set we may be concerned with.

As we have seen, the frequency distribution of the AG data reverses the order of the ranked distribution and gives priority to those sources in the ranked tail while relegating to the tail of the frequency distribution the nuclear sources which are most readily accessible and usually the most keenly sought. For someone interested in retrieving papers on a specific topic, the ranked form offers advantages which the frequency form does not. For example, as soon as the user of a database has identified the major nuclear sources of his topic, he can begin to apply the Bradford law, to estimate the two parameters and so estimate his expected totality; if the search is continued and new sources are identified, the first provisional estimate can easily be revised from time to time, especially if the process is computerized.

Yet I have to admit that the Ranked Bradford-law technique is often applied without the discipline that, for example, Bradford displayed in his paper. For current use one of the aspects of the law that needs attention is the introduction of a time parameter, i.e. to repeat Bradford's work on the current databases.

We also need to understand rather better than we do at present the processes that underlie the generation of a literature on a specific topic. Some years ago, a research assistant of mine, Elizabeth Wilkinson, attempted to simulate the generation of such a literature by a computer program which set up a 'success-breeding-success' mechanism - a stochastic process (9). Her first model began with just 1 paper published in 1 journal and the process was set to 'publish' 1,000 papers in as many journals as might be demanded.

The initial results were chaotic: sometimes a single journal would emerge very quickly and thereafter run away with the bulk of the papers. Sometimes a surprisingly flat distribution would arise. There was no consistency whatever.

So we consulted an expert on stochastic processes, David Bartholomew, who suggested that the next model be given an initiating kernel of, say, 20 papers already set up in Bradford form. With this initiating kernel, the resulting distributions were more plausibly Bradfordian though not consistently similar. Elizabeth Wilkinson then restricted the ranges of the two key parameters which automatically set up the probabilities anew after each paper had found its journal. So, at last, the model began to generate fairly consistent distributions, though not quite the Bradford distributions we had expected. It was noticeable that their defects in this respect were similar to those of the initiating kernel, which had been set up perhaps a little too casually. It would have been valuable to have been able to test modifications of the kernel but time and funds had run out.

I mention these matters to suggest that bibliometrics offers some intriguing problems in at least one area of study whic, as far as I can see at present, cannot be reached by the continuing refinement of the IGPD and other Zipfian frequency distributions.

For solving the general problems of bibliometrics, as also of course for many other problems of information science, we have to rely on the methods of applied statistics. The statistician's aim is to establish theoretical models which are generally valid and with which he hopes to capture the interactions involved in the writing, publishing and retrieving of books and papers and of organizing access to them through libraries and databases. The formulations which capture all these effects, summarizing a conflation and compounding of random processes, would be very helpful indeed.

However, the distributions which describe these complex processes will inevitably have very large variances as compared with those of Gaussian distributions. So when the individual data-base user applies them to his particular search, or when the hard-pressed librarian tries to decide how best to spend his limited funds to meet as well as he can the specific interests of his users, they may well find them unhelpful in such specific cases.

So I see a theoretical gap between applied statistics (as we know their techniques at present) and the ranked distributions. There also remains a gap - a practical gap - between theory and the application of both frequency and rank techniques to particular case. And it seems to me that we need to start bridging these gaps from both ends.

So I very much hope that Leo Egghe's initiative in calling together bibliometricians for the first time will be rewarded fairly soon by discernible efforts to integrate bibliometrics into coherence.

REFERENCES

[1] Pritchard, A., Statistical bibliography or bibliometrics?, J. Docum. (1969) 25 (4) pp. 348-349.

[2] Yule, J.D., (Ed), Phaidon concise encyclopaedia of science and technology. (Elsevier, Oxford, 1978).

[3] Bradford, S.C., Sources of information on specific subjects. First published Engineering (1934) Reprinted J. of Inf. Sci. (1985) 10 pp. 176-180.

[4] Haitun, S.D., Stationary scientometric distributions. Scientometrics (1987) 4 pp. 1-25, 89-104, 181-194.

[5] Sichel, H.S., A bibliometric distribution which really works. J. Amer. Soc. Inf. Sci. (1985) 36 (5) pp. 314-321.
[6] Kendall, M.G., Rank correlation methods (Griffin, London 1962).
[7] Leimkuhler, F.F., The Bradford distribution. J. Docum. (1967) 23 (3) pp. 197-207.
[8] Sichel, H.S., The GIGP distribution models with applications to physics literature, Czech. J. Physics B. (1986) 36 (1) B, pp. 133-137.
[9] Wilkinson, E., The Bradford-Zipf distribution, a simulation study. OSTI Report 5172 (London, 1972).

APPENDIX

The simple and the mixed Poisson models applied to Bradford's AG data.

A. The no. of potential sources in Bradford's Poisson group S requires

$$S(1 - e^{-u}) = 0.7966k$$

$$S.u = k .$$

These equations are satisfied by $u = 0.4736$, $e^{-u} = 0.6228$ with $S = 2.112k = 589$ when $k = 279$.

B. For the mixed Poisson model:

$$f(n) = k \int_0^1 e^{-u} \frac{e^u}{n!} \frac{1}{u} \, du$$

$$= \frac{k}{n!} e^{-1} + \frac{(n - 1)}{n} f(n-1) . f(1) = k(1 - e^{-1}) .$$

Comparing the two models we have:

	Simple		Mixed	
	Sources	**Papers**	**Sources**	**Papers**
f(1)	174	174	183	183
f(2)	41	82	38	76
f(3)	6	18	8	24
f(4)	1	4	1	4
	222	278	234	286

The mixed Poisson model thus suggests slightly higher totals of 3 - 4% of both sources and papers than the simple model. The two Poisson means are in the ratio $u/\ln(1 - e^{-1}) = 1.033$.