Information Processing & Management Vol. 26, No. 6, pp. 777–790, 1990
Printed in Great Britain.

0306-4573/90 $3.00 + .00

DIRECT COMPARISON OF BIBLIOMETRIC MODELS

MARK T. KINNUCAN and DIETMAR WOLFRAM
School of Library and Information Science, University of Western Ontario,
Elborn College, London, Ontario, Canada, N6G 1H1

(*Received* 1 *February* 1990; *accepted in final form* 31 *May* 1990)

Abstract—This study describes a technique for statistically comparing bibliometric models, and illustrates its use with three different examples. The technique is based on the idea of comparing full and restricted models as developed in analysis of variance, regression, and log-linear models. In bibliometrics, any two models where one is a special case of the other can be thought of as a full model and a restricted model. One can use the likelihood-ratio chi-square statistic, which has gained acceptance with log-linear models, as a test statistic to directly compare the full model and the restricted model. The first two examples involved Lotka's law. In the first example we investigated the feasibility of applying a single set of global parameter values to eight different author productivity distributions drawn from two different disciplines. In the second example we looked at whether or not a finite maximum productivity level was necessary as an additional parameter in Lotka-type models of author productivity. The final example compared three different forms of a model of library circulation frequencies.

Much of the work in bibliometrics involves the fitting of theoretical probability distributions to empirical data that are in the form of frequency distributions (e.g., Nelson and Tague, 1985; Sichel, 1985). A variety of phenomena have been studied in this way, including book circulation, index term usage, citation age, and author productivity. In the simplest case the distribution of a single variable (a univariate distribution) is modelled. Much statistical modelling in bibliometrics makes use of univariate distributions, and the work to be reported here focuses on these. Theoretical distributions that have been fit include the negative binomial, Poisson, and Zipf distributions, among others.

Often these distributions have one or more parameters that must be estimated from the data. The values of the parameters are frequently chosen to maximize the fit between the theoretical distribution and the observed data. The methods of minimum chi-square, least squares, and maximum likelihood all follow this approach, differing only in how they define the best fit. Once the optimal parameter values have been chosen, the researcher customarily tests the fit of the distribution to the data by means of a statistical test such as the chi-square test or the Kolmogorov–Smirnov test. Cooper and Weekes (1983) provide a good introduction to techniques of fitting and testing statistical models. They cover several of the most frequently used continuous and discrete theoretical distributions.

A common situation in bibliometrics is to have two or more competing distributions being considered as appropriate models for a set of data. The purpose of this article is to propose that wherever possible researchers should use direct comparisons in deciding between competing models. A direct comparison in the sense used here employs a single test statistic that allows the researcher to estimate the probability that one of the models should be rejected *in favor of the other model.* Direct comparisons appear to be employed rarely in bibliometric modelling, although there are situations where they could be used to advantage.

The most straightforward situation in which one can make a direct comparison of models is where one of the models under consideration is a special case of the other model. This often arises when the more specific model can be derived from the more general model be fixing the values of one or more of the parameters of the latter model. For example, the

Requests for reprints should be sent to Mark T. Kinnucan, School of Library and Information Science, University of Western Ontario, Elborn College, London, Ontario, N6G 1H1, Canada.

geometric distribution can be derived from the negative binomial distribution by setting one of the parameters of the negative binomial distribution equal to one. In such a situation, the two models are said to be hierarchically related, with the more specific model nested within the more general model.

This article is concerned specifically with comparisons of nested models. The methods for comparing non-nested models are not as straightforward or well developed as the methods for nested models. We will have a few comments on comparisons of non-nested models at the end of the article.

STATISTICAL INFERENCE AS MODEL COMPARISONS

We first became aware of the idea of comparing nested statistical models in the context of the analysis of variance (ANOVA). In the linear models approach to ANOVA, tests of main effects and interactions are expressed as the comparison of two linear models (Judd and McClelland, 1989; Maxwell and Delaney, 1990). The more elaborate model for a given test is called the full or augmented model, and the less detailed model is called the restricted or compact model.

For example, two-way ANOVA usually includes tests of the main effects of the two factors and a test of the interaction between them. In the test of the interaction, the full model is

$$Y_{ijk} = M + A_i + B_j + AB_{ij} + e_{ijk}, \tag{1}$$

and the restricted model is

$$Y_{ijk} = M + A_i + B_j + e_{ijk}, \tag{2}$$

where Y_{ijk} is an observed data value, M is a parameter for the grand mean of all the observations, A_i is a parameter representing the main effect of factor A, B_j is a parameter representing the main effect of factor B, AB_{ij} is a parameter representing the interaction of A and B, and e_{ijk} represents the error of prediction of the model for that particular observation. The restricted model posits that the interaction is absent, so the parameters representing it (the AB term) are dropped from equation (2).

Because it has more free parameters, the full model always provides a better fit to the data than the restricted model. The philosophy behind model comparisons in ANOVA is that in choosing between the two models, one needs to look at *how much* better the fit of the full model is than the fit of the restricted model. Does the increase in the ability to predict the data afforded by the full model warrant an increase in the complexity of the explanation, as is indicated by the addition of one or more free parameters? The ideal of parsimony dictates that one should adopt the simplest model one can. This means that the restricted model should only be rejected when the weight of evidence suggests it is untenable.

The conventional F-test provides a measure of whether the fit of the full model is sufficiently better than the fit of the restricted model to cause one to reject the latter in favor of the former. The F-test essentially compares the increase in error encountered when one moves from the full model to the restricted model against the error associated with the full model, taking into account the number of additional parameters used by the full model (Judd and McClelland, 1989, pp. 83–87). When the increase in error is large enough, as determined by the table of the F distribution, then the restricted model is rejected in favor of the full model. That is, the interaction is said to be significant.

A major advantage of the linear models approach to ANOVA is that it allows considerable flexibility in the specification of null and alternative hypotheses. One need only keep in mind that the effect being tested is always indicated by the term contained in the full model, but not in the restricted model. Any two linear models that only differ by a single term may serve as the full and restricted models for a test of that term. The implications

of choosing different models to compare when testing for a particular effect in ANOVA are discussed by Appelbaum and Cramer (1974) and Howell and McConaughy (1982).

The view of ANOVA described here illustrates the logic of model comparisons that we propose to apply to bibliometrics. In each case we identify a full model and a restricted model, and we use a formal statistical test to address the question of whether the additional complexity of the full model is necessary. However, as will be explained, we do not compare bibliometric models with F tests, but with chi-square tests.

In detail, our approach is actually more like log-linear models than ANOVA. Log-linear models are an approach to the analysis of categorical data that arose partially as an emulation of the linear models approach to ANOVA (Bishop, Fienberg, and Holland, 1975; Marascuilo and Busk, 1987). In log-linear models, the logarithms of the expected frequencies of the cells of a multiway contingency table are modeled by the sums of parameters that represent the main effects and interactions of the variables that comprise the rows, columns, layers, etc. of the table. For example, the traditional chi-square test of association for a two-way table can be thought of as comparing a log-linear model of independence to a model that includes an interaction term. The log-linear model of independence (the restricted model) can be written as

$$\log(m_{ij}) = u + u1_{(i)} + u2_{(j)}, \tag{3}$$

where m_{ij} is the expected frequency of the ijth cell according to the model, and the u terms are the parameters of the model. The first parameter, u, is based on the total number of observations (it is analogous to the "grand mean" term in ANOVA), $u1_{(i)}$ is a row-effect parameter, and $u2_{(j)}$ is a column-effect parameter.

The full model can be written as

$$\log(m_{ij}) = u + u1_{(i)} + u2_{(j)} + u12_{(ij)}, \tag{4}$$

where u, $u1_{(i)}$, *and* $u2_{(j)}$ are as defined above, and $u12_{(ij)}$ represents the interaction of the row variable and the column variable. With a two-way table, this model has as many parameters as data points, leaving it with no degrees of freedom. It is thus called a "saturated" model, and will always yield expected frequencies that exactly match the observed frequency in each cell. Thus, comparing a restricted model to a saturated model is equivalent to comparing that restricted model to the data.

The full model in a comparison of log-linear models need not be the saturated model. For example, consider the model

$$\log(m_{ij}) = u + u1_{(i)}. \tag{5}$$

This model states that we need to take the row margins into account in our estimate of the logs of the expected frequencies, but not the column margins. In other words, equation (5) says that the observations should occur in the different categories of the column variable with equal frequency. Comparing equation (5) as the restricted model to equation (3) as the full model constitutes a test of whether the different values of the column variable do in fact occur with equal frequency in the population. By analogy to ANOVA, this comparison is sometimes called a test of the main effect of the column variable.

This last illustration shows that a particular model can serve as the restricted model in one comparison and as the full model in another, as equation (3) does above. What equations get used for the full and restricted models in a comparison depends on what question the researcher wants to ask. As with ANOVA, the effect being tested is the one represented by the term that is contained in the full model, but not in the restricted model.

LIKELIHOOD RATIO CHI-SQUARE STATISTIC AND MAXIMUM LIKELIHOOD ESTIMATION

Log-linear models often use an alternative to the traditional Pearson chi-square statistic customarily associated with contingency tables. This alternative is called the likelihood ratio

chi-square statistic, and is typically denoted as G^2 (Bishop *et al.*, 1975). G^2 can be calculated for any log-linear model. It is defined as

$$G^2 = 2 * \sum x_{ij} * \log(x_{ij}/m_{ij}), \tag{6}$$

where x_{ij} is the observed cell frequency and m_{ij} is the expected cell frequency under the model being considered. The logarithms are to the base e. Like Pearson's chi-square, the likelihood ratio chi-square can be used to test a single log-linear model against a set of cross-classified data—degrees of freedom are calculated the same as for Pearson's chi-square, and the value of G^2 is compared to a table of chi-square to determine whether the model fits the data. In fact, G^2 and Pearson's chi-square usually yield very similar values for a given problem. G^2 for a saturated model always equals zero.

The main advantage of G^2 is that in addition to being useful for testing a single model against a set of data, it can also be used to compare any two nested models. The usefulness of G^2 for model comparisons derives from the fact that, unlike Pearson's chi-square, it exhibits the property of *additivity* (Bishop *et al.*, 1975). Additivity can be explained as follows. Suppose model A is nested within model B (i.e., model A is the restricted model and model B is the full model). Additivity holds when the measure of fit obtained when testing model A against the data is equal to the sum of the measure of fit obtained in a comparison of model A to model B, plus the measure of fit obtained when testing model B against the data. If we let $G^2(A)$ stand for the value of G^2 obtained in a test of model A, and we let $G^2(B)$ stand for the value obtained in a test of model B, and we let $G^2(B)(A)$ stand for the improvement in fit afforded by model B over model A, then additivity says that

$$G^2(A) = G^2(B)(A) + G^2(B). \tag{7}$$

Bishop *et al.* (1975) call $G^2(B)(A)$ the conditional measure of fit. Additivity allows a researcher to obtain $G^2(B)(A)$ quickly by subtraction once $G^2(A)$ and $G^2(B)$ have been calculated.

The conditional measure of fit tells the researcher whether the full model is a significantly better description of the data than the restricted model. If the full model is not much of an improvement over the restricted model, then $G^2(B)(A)$ will have a nonsignificant value compared to a chi-square table with degrees of freedom equal to the number of parameters in the full model minus the number of parameters in the restricted model. The conditional measure of fit is what would be used to test the main effect of the column variable in the illustration above.

Equation (7) provides the basis for the approach to the direct comparison of models we are advocating. In the examples below, we describe three different situations in which one might want to compare bibliometric models. Each comparison involves restricted and full models. In each example we calculate G^2 for each of the models under consideration. We then use equation (7) to obtain values for $G^2(B)(A)$ by subtraction, and test the significance of the resulting value as a guide for selecting the most appropriate model. Riefer and Batchelder (1988) have used the same approach for comparisons of different models of cognitive processing.

In addition to its usefulness for model comparisons, the likelihood ratio chi-square statistic has another important property that is significant in both the contexts of log-linear models and bibliometric models. One can find maximum likelihood estimates (MLEs) of the parameters of a model by identifying those values of the parameters that minimize G^2 (Bishop *et al.*, 1975). That is, minimizing G^2 is equivalent to maximizing the likelihood function. This property is useful to note in the present context, because maximum likelihood estimation is becoming increasingly important in bibliometrics (Nicholls, 1986). All the parameter estimates in the examples to follow are maximum likelihood estimates obtained by minimizing G^2.

The rest of this paper consists of three examples of the use of G^2 to make direct comparisons of bibliometric models. All three examples involve the reanalysis of data already

presented. The first two examples use data collected by Nicholls (1986, 1987) on author productivity. The final example uses data from Brownsley and Burrell (1986) on library circulation. Both author productivity and library circulation provide fertile ground for model comparisons because in both areas several different bibliometric models, or at least several different forms of a given bibliometric model, such as Lotka's law, have been proposed.

In the first example, we look at the question of how a researcher might treat the situation in which he or she has several different data sets, all being modelled by the same basic model. Should the researcher assume that it is just the form of the model that is the same across the different data sets, or should the actual parameter values be identical across the data sets? In the example, we use G^2 to compare these possibilities. In the second example, we compare two different forms of Lotka's law, one of which contains an extra parameter that the other one lacks. Again, we use G^2 to guide this comparison. Finally, in the third example we use G^2 to compare three different models of library circulation proposed by Brownsley and Burrell (1986). This example includes the idea of mixtures of distributions, which is absent from the first two examples.

EXAMPLE 1: INDIVIDUAL LOTKA PARAMETERS

For our first example, we chose eight author productivity data sets from the hundred or so collected by Nicholls (1987). We selected four data sets for each of two subject areas: information science and biology. The data sets were chosen for their size and comparability.

The Lotka hypothesis of author productivity states that there exists an inverse relationship between the number of authors and the number of papers each author produces (Pao, 1985; Nicholls, 1986). Numerous studies have investigated this phenomenon. When one examines the literature of a subject area, one typically finds that many authors only contribute one paper to the discipline, whereas very few are prolific contributors. Mathematically, the Lotka hypothesis is typically represented by:

$$g(x) = k/x^b, \tag{8}$$

where x represents the number of papers a given author produces in a discipline, $g(x)$ the proportion of authors contributing x papers, and k and b are parameters to be estimated. Note that k represents the expected proportion of authors contributing one paper, because $g(1) = k/1^b = k$.

If one has several different data sets, it is possible to estimate the parameters of equation (8) separately for each data set. However, it would obviously be more parsimonious to model all of the data sets with a single set of parameters. If, as in our example, some of the data sets are drawn from one discipline and some from another, then an intermediate possibility arises. It could be that one set of parameters would suffice for one discipline, but that a different set of parameters would be required to model the other.

These three possibilities form a set of hierarchically nested models. The fullest model is the one in which each data set has its own b and k parameters. We call this case the individual parameters model. The intermediate model is the one in which there are two sets of b and k parameters, one set for each discipline. We call this case the discipline parameters model. Finally, the most restricted model attempts to explain all eight data sets with a single value of b and a single value of k. We call this the global parameters model.

Theoretical distributions tend to fit observed author productivity distributions poorly for the more productive authors; there are very few prolific authors, and gaps in this portion of the distribution are common. Low expected cell frequencies also tend to be a problem with prolific author data. Therefore, it is a common practice to group together the higher productivity values. We followed this practice, collapsing all authors who had published eight or more papers into a single cell of the table for each data set. This approach enabled us to have the same number of cells for each data set, simplifying degree-of-freedom calculations.

In applications of Lotka's law, the theoretical maximum number of papers that could be written by an author, x_{max}, is sometimes considered to be a parameter of the distribu-

tion (Tague and Nicholls, 1987). In this example we assumed that x_{max} was infinite, and did not attempt to estimate it. We take a closer look at x_{max} in the next example. In practical terms, treating x_{max} as infinite meant that, for each data set, we calculated the expected proportion of authors in all the cells but the last one (i.e., the eighth one) from equation (8), and then we computed $g(8)$ as $1 - [g(1) + g(2) + \ldots + g(7)]$. The proper interpretation of $g(8)$, then, is that it is the expected proportion of authors who wrote eight *or more* papers.

As mentioned above, the k parameter represents the proportion of authors with only one paper in the data set. Pao (1985) has shown how the value of k can be calculated using an approximation procedure once the value of b is known. This approximation assumes that the theoretical x_{max} is infinite. When one uses Pao's procedure, the value of k is entirely dependent on the value of b. This means that k and b are not independent parameters. This fact must be taken into account when calculating the number of degrees of freedom for the significance tests.

Because the technique for finding maximum likelihood estimates outlined above involves minimizing chi-square values, the observed and theoretical distributions must be expressed in terms of frequencies rather than proportions. This is accomplished for the theoretical distribution by multiplying the proportions given by equation (8) separately for each data set by the total number of authors in that data set. Since there are eight data sets, this procedure costs each model eight degrees of freedom on top of those lost in estimating the parameters of that model. Thus, with 64 cells (eight data sets with eight cells each), the global parameters model has 55 degrees of freedom, the discipline parameters model has 54 degrees of freedom, and the individual parameters model has 48 degrees of freedom.

We obtained maximum likelihood estimates of the b parameter by finding the value of b that minimized G^2. We used a FORTRAN function-minimization routine called STEPIT (Chandler, 1965) to accomplish the minimization. We wrote the Pao approximation for estimating k into the driver program for STEPIT.

Three different STEPIT runs were conducted. In each run, we programmed STEPIT to find a single value of G^2 that was minimized over all eight data sets simultaneously. That is, G^2 was found in each case by summing $x_i * \log(x_i/m_i)$ over all 64 cells at once. For the first run, we instructed STEPIT to estimate eight different b parameters in finding the minimum value of G^2, in the second run we instructed it to estimate two b parameters, and in the third run we instructed it to estimate a single b parameter. The three runs corresponded to the individual, discipline, and global parameter models, respectively. Comparing the values of G^2 obtained in the three runs allowed us to test the relative merits of trying to explain author productivity in terms of a single curve for all data sets or in terms of a single curve for each discipline.

The results of the three STEPIT runs are shown in Table 1. Looking first at the individual parameter results, it is apparent that three of the values of b within each discipline are close to each other, but that each discipline has one discrepant value. On the whole, the values of b for information science tend to be higher than the values of b for biology. This latter result is seen more clearly in the second part of Table 1, where the discipline parameters are presented. Finally, not too surprisingly, STEPIT calculated a global value of b that is intermediate between the value for biology and the value for information science.

The G^2 values for all three models are significant, which says that none of the three models adequately describes the data. This result is not too surprising, given that the power of chi-square tests increases with sample size and that the total number of observations across all eight data sets is over 10,000. Riefer and Batchelder (1988, p. 325) referred to this problem as the rejection of "good, but technically incorrect," models. They argued that such models should not be summarily dismissed. Figure 1 gives an example of the expected author productivity frequencies under each of the three models for one of the data sets in biology. The observed frequencies are also shown on the figure. It is obvious from Fig. 1 that the models fit the shape of the data reasonably well.

Now let us consider comparisons of the models. The difference in G^2 between the global parameters model and the discipline parameters model is 297.5, which is quite significant with one degree of freedom. This result says that productivity in biology and in-

Table 1. Parameter values and chi-square for individual, discipline, and global versions of Lotka's Law fitted to data sets in biology and information science

Data set	b	k
Individual Parameters Model		
B1	2.581	0.762
B2	1.787	0.526
B3	2.451	0.735
B4	2.472	0.739
IS1	3.297	0.868
IS2	2.512	0.748
IS3	2.705	0.786
IS4	2.640	0.774
$G^2 = 264.3$, $df = 48$		
Discipline Parameters Model		
Biology	2.295	0.697
Information science	2.913	0.820
$G^2 = 668.0$, $df = 54$		
Global Parameters Model	2.594	0.765
$G^2 = 965.5$, $df = 55$		

formation science cannot both be described by a single equation. Furthermore, the difference in G^2 between the discipline parameters model and the individual parameters model is 403.6. This time the test has six degrees of freedom, but the G^2 difference is still quite significant. This result tells us that we can obtain a significantly better explanation of the data if we allow each data set to have its own parameters for Lotka's law than if we try to explain all of the data sets within a discipline with a single set of parameters.

It is not surprising that such a result appeared. As was mentioned above, one data set in each discipline differed noticeably from the others in the same discipline. This outcome makes it hard to say whether the difference between the disciplines is real or not. The data certainly suggest that biology has lower values of b and k than information science. Lower values of b and k occur when there is a higher proportion of prolific authors, so perhaps prolific authors are rarer in information science. However, with only four data sets for each discipline, it is hard to say for sure. It would be interesting to see if the pattern held up with, say, ten data sets in each discipline.

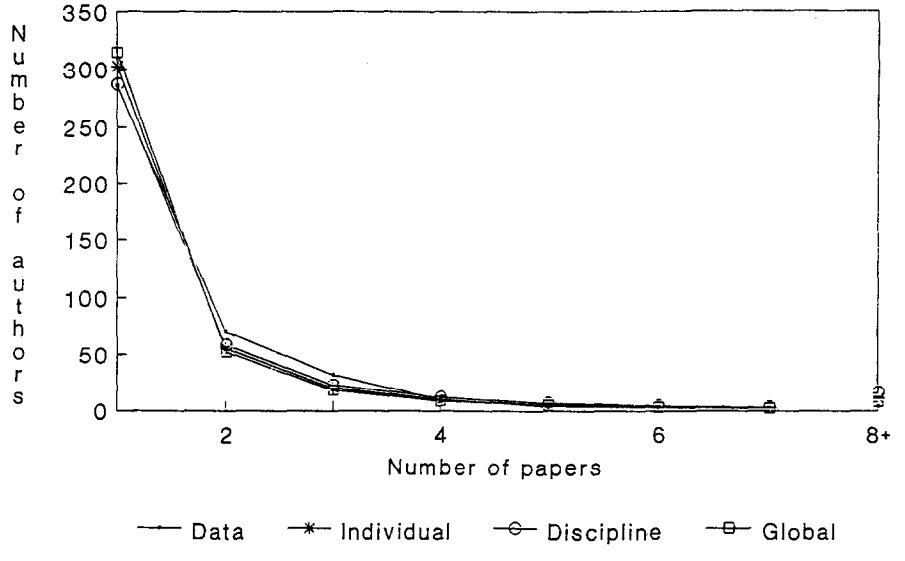

Fig. 1. Example of the fit of three models to one data set of author productivity.

EXAMPLE 2: LOTKA'S LAW WITH FINITE MAXIMUM

In the second example, we applied the technique of comparing models with G^2 to the situation in which one is trying to decide between two different forms of a bibliometric model, where one of the forms involves an additional parameter that the other lacks. Specifically, we looked at the question of whether the theoretical maximum number of papers an author might write should be included in a model of author productivity. We again used Nicholls' (1987) data for this example.

When one counts the number of publications for each author in any actual bibliography or data set, there will, of course, be a finite maximum to the number of papers authored by any one person. Tague and Nicholls (1987) have argued that this fact should be acknowledged in modelling author productivities and other Zipf-distributed data. Indeed, it is possible to treat $x_{\max}$ as a parameter of the model to be estimated from the data. In that case, Lotka's law becomes

$$g(x) = k/x^b, \qquad x = 1,2,\ldots,x_{max}. \tag{13}$$

Is it really necessary to include $x_{\max}$ in models of author productivity? Using the methodology outlined above, we set out to compare a full model that estimates $x_{\max}$ from the data to a restricted model that assumes it to be infinite. We conducted the comparisons for ten data sets drawn from those compiled by Nicholls (1987). Only one of the ten was also used in Example 1 above. The data sets chosen for this illustration had between 5 and 12 cells (i.e., $x'_{\max}$ was between 5 and 12), and had no empty cells (i.e., no gaps in the productivity distribution).

Unlike the first example, each data set was analyzed separately. We collapsed each data set to five cells (rather than eight) by grouping all authors with five or more publications into the fifth cell. With these exceptions, we carried out the estimation of b and k for the restricted model using exactly the same procedures as in the first example. In particular, we used Pao's approximation to calculate k for each data set.

For the full model, we needed to estimate $x_{\max}$ as well as b, and to calculate k in a different way. With ungrouped data, the maximum likelihood estimate of $x_{\max}$ is simply $x'_{\max}$ (Tague and Nicholls, 1987). However, because we had grouped together the higher author productivities, we took a different approach. On the assumption that the maximum likelihood estimate of $x_{\max}$ would be the one that gave the lowest value of G^2, and because $x_{\max}$ can only take on integer values, we conducted separate STEPIT runs for $x_{\max} = 5,6,7$, etc. Each STEPIT run produced the best-fitting value of b for that value of $x_{\max}$ and a value of G^2 for that particular combination of b and $x_{\max}$. We continued these runs until we found the *globally* minimum value of G^2. What are reported in Table 2 are the values of b, k, and $x_{\max}$ associated with the globally minimum G^2 value for each data set.

Pao's approximation is not appropriate for calculating k when $x_{\max}$ is assumed to be finite, although k is still determined by the other parameters. It is calculated as

$$k = t \Big/ \sum_{1}^{x_{\max}} (1/x^b), \tag{9}$$

where t is the total number of authors in the data set (Tague and Nicholls, 1987).

Tables 2 and 3 show the results of these analyses for the full and restricted models, respectively. Table 2 shows that the full model fits these data sets quite well. Only two of the ten G^2 values are significant at the .05 level. The estimated value of $x_{\max}$ was between five and eight for all the data sets except set 4. This was an anomalous data set with a virtually flat likelihood function, which made estimation of $x_{\max}$ difficult. It also had an unusually high estimated value of b.

Table 3 shows the results for the restricted model and the results of the comparison of the full and restricted models. Several observations can be made about this table. First of all, the values of b and k are consistently slightly higher for the restricted model than for the full model. This is because one effect of making $x_{\max}$ finite is to increase the pro-

Table 2. Fitting Lotka's Law with finite maximum productivity (the full model) to ten author productivity data sets

Data set	No. of authors	ML estimates			G^2	df	p
		x_{max}	b	k			
1	210	8	3.68	.903	4.91	2	n.s.
2	1282	7	3.44	.885	3.26	2	n.s.
3	2461	5	3.43	.886	7.44	2	<.05
4	408	35[a]	4.22	.936	.94	2	n.s.
5	736	7	2.38	.743	6.15	2	<.05
6	198	7	3.60	.897	4.60	2	n.s.
7	164	5	2.38	.755	3.85	2	n.s.
8	411	6	2.15	.702	2.98	2	n.s.
9	386	8	2.26	.714	1.80	2	n.s.
10	170	5	2.40	.759	3.70	2	n.s.

[a]The likelihood function is virtually flat for values of x_{max} from x'_{max} to infinity for this data set.

portion of authors who have contributed just one paper. The same effect can be accomplished by increasing the value of b. Secondly, the restricted model provides an adequate fit to the data for only half of the ten data sets. This is in contrast to the full model, which fit eight of the ten. This contrast is borne out when the difference in G^2 is computed. Five of the ten G^2_{diff} values are significant at the .05 level or higher. In other words, assuming x_{max} to be finite and estimating it from the data yields a significantly better model in half of the cases. It appears that making x_{max} a finite parameter sometimes provides a better fit, but not always.

EXAMPLE 3: LIBRARY CIRCULATION

Our final example looks at the number of times different books circulate from a library collection over a given period of time. This example allows us to illustrate how G^2 can be used when a researcher has "doubly nested" models—three models in a single hierarchy. It also allows us to remark on a specific problem that arises when testing a model with mixtures of distributions.

Circulation data are generally tallied in the form of how many books have circulated once, how many twice, etc., and how many not at all. Quentin Burrell (1980, 1982; Brownsley and Burrell, 1986) has studied library circulation data extensively. Brownsley and Burrell (1986) directed their attention specifically to public library circulation, as most of the previous research had dealt with circulation in academic libraries. Brownsley and Burrell were able to make use of machine-readable circulation data for public libraries in the United Kingdom, which became available as a result of a law that mandated payments to authors for the loan of their books from public libraries. Brownsley and Burrell were interested in

Table 3. Lotka's Law with infinite maximum productivity (the restricted model) and model comparisons

Data set	ML estimates					G^2 diff.		
	b	k	G^2	df	p		df	p
1	3.70	.904	4.95	3	n.s.	.04	1	n.s.
2	3.49	.887	4.20	3	n.s.	.94	1	n.s.
3	3.54	.891	19.10	3	<.001	11.66	1	<.001
4	4.22	.936	.94	3	n.s.	.00	1	n.s.
5	2.56	.758	11.97	3	<.01	5.82	1	<.02
6	3.64	.899	4.68	3	n.s.	.08	1	n.s.
7	2.70	.784	12.59	3	<.01	8.74	1	<.01
8	2.44	.731	14.48	3	<.01	11.50	1	<.001
9	2.43	.730	4.64	3	n.s.	2.84	1	n.s.
10	2.71	.787	12.43	3	<.01	8.73	1	<.01

three different models of public library circulation data: the geometric distribution, a single negative binomial distribution, and a mixture of three negative binomial distributions. Each of the latter two distributions is a generalization of the preceding distribution in the list, so they form a good set of candidates for the approach to model comparisons being illustrated here.

Brownsley and Burrell reported the observed circulation frequency data for one of the public libraries (Library 05: Hillhead), as well as the expected frequencies under each of the models. They also reported Pearson chi-square for each model, but not G^2. We decided to reanalyze the data for this same library using the approach we followed in the previous examples.

One problem with modelling circulation data is how to deal with items that do not circulate in the time period under study. Burrell (1982) has argued that such items are of two kinds. Some of them are "dead" items of no interest or which never circulate for other reasons, such as being lost or in the reference collection, and others are active items that just did not happen to circulate during the study period.

Analogous problems arise in other applications of statistical modelling, such as models of the number of visits made to a medical clinic. The number of people who did not visit the clinic includes those not sick, as well as those who should have visited the clinic but did not (Gross and Miller, 1981).

Our approach to this problem is based on the work of Cohen (1966; see also Everitt and Hand, 1981). Cohen proposed that when both dead and active items (to use the language of the present example) are included in the total count, such situations should be modelled with distributions of the form

$$g(x) = \begin{cases} (1-q) + q * f(0) & \text{for } x = 0 \\ q * f(x) & \text{for } x = 1,2,\ldots \end{cases} \tag{10}$$

where $f(x)$ is any discrete distribution and q is the proportion of all the items that are active. Cohen's general approach can be used even if one does not have an accurate count of the total number of items. In this case, q is not estimated, but the number of active items, both circulating and noncirculating, is estimated. It follows from Cohen's approach that the total number of active items, N_A, can be estimated as

$$N_A = N_C/(1 - f(0)), \tag{11}$$

where N_C is the number of items that did circulate and $f(0)$ is the estimate of the proportion of active items that did not circulate provided by the theoretical distribution under consideration (Burrell, 1982; Gross and Miller, 1981). Thus, for all three of the models to be considered we estimated N_A, and multiplied the expected proportions by that value to obtain the expected frequencies, rather than multiplying them by the total number of items that did circulate, N_C.

The form of the geometric distribution used here was

$$f(x) = p * (1-p)^x, \qquad 0 < p < 1, \tag{12}$$

where $x = 0,1,2,\ldots$ is the number of times a book might circulate, $f(x)$ is the proportion of books that circulated that number of times, and p is the parameter of the model. The geometric model produces a straight line when $f(x)$ is plotted on a logarithmic scale and x is plotted on a linear scale.

For the negative binomial distribution, we used a recursive equation based on that given by Cooper and Weekes (1983).

$$f(0) = p^k$$

and

$$f(x) = f(x-1) * (1-p) * (x+k-1) / x, \tag{13}$$

where $0 < p < 1$ and $k > 0$. The terms x and $f(x)$ have the same definition as before, and p and k are the parameters of the distribution. This formulation of the negative binomial distribution allows noninteger values for k. When $k = 1$, the negative binomial distribution reduces to the geometric distribution.

Brownsley and Burrell (1986) considered a mixture of three negative binomial distributions as a third model because the public libraries they studied had three distinct components to their collection: adult fiction, adult nonfiction, and juvenile literature. Thus, their most elaborate model allowed separate p and k parameters for each collection plus two mixing parameters that represented the proportion of the total circulations that could be attributed to adult fiction and to adult nonfiction. General works on fitting and testing mixtures of distributions are provided by Everitt and Hand (1981), Titterington, Smith, and Makov (1985), and McLachlan and Basford (1988). Harris (1983) discusses mixtures of geometric and negative binomial distributions specifically.

By expanding on eqn. 13, the mixture model can be written as follows:

$$f(0)_i = m_i * p_i^{ki}, \qquad \text{for } i = 1,2,3$$

$$f(0) = \sum f(0)_i$$

$$f(x)_i = m_i * f(x-1)_i * (1-p_i) * (x + k_i - 1)/x$$

$$f(x) = \sum f(x)_i$$

and

$$m_3 = 1 - m_1 - m_2. \tag{14}$$

Thus, the mixed negative binomial model has eight parameters: three p's, three k's, and two independent mixing parameters, m_1 and m_2. Equations 14 were written into the driver program for STEPIT, and it was allowed to find the values of the parameters that minimized G^2. Several runs were made with different starting values because of problems with local minima. The results presented are for the run with the lowest G^2 of all the runs. We are still not sure we have found the globally best parameter values but, as will be explained, there is some indication that the values we obtained are reasonable.

Table 4 shows the results of fitting the three models. Note that the best-fitting values of p and k for the negative binomial distribution were fairly close to the values found for the geometric distribution. The value of p went from .150 in the geometric to .166 in the negative binomial, while k went from 1 (by definition) to 1.16. Even though the change in

Table 4. Parameter values and goodness of fit for models of library circulation

Geometric Distribution			
$p = .150$			$G^2 = 491.21$, $df = 24$
Negative Biomial Distribution			
$p = .166$	$k = 1.16$		$G^2 = 382.92$, $df = 23$
Mixed Negative Binomial Distribution			
$p_1 = .285$	$k_1 = 2.22$	$m_1 = .69$	$G^2 = 66.36$, $df = 17$
$p_2 = .090$	$k_2 = .89$	$m_2 = .25$	
$p_3 = .807$	$k_3 = 3.11$	$m_3 = .06$*	

*This value is not a free parameter. It is obtained by subtraction.

the parameter values was slight, it did afford a better fit. G^2 decreased by 108.29 with the addition of a single parameter.

An even more pronounced change occurred with the introduction of the mixed negative binomial distribution. The three values of p obtained with this model are quite different from each other, as are the three values of k, indicating that the three subcollections are behaving differently. Figure 2 shows that the mixed negative binomial distribution fits the data substantially better than the other two models.

Ordinarily we would back up this claim by pointing out that the reduction in G^2 of 316.56 is significant with 6 degrees of freedom. Technically, however, the test is not appropriate in this situation because certain regularity conditions (cf. Bishop *et al.*, 1975, pp. 509–511) are not met when dealing with mixtures of distributions (Everitt and Hand, 1981; McLachlan and Basford, 1988). This means that, in this case, the change in G^2 is not necessarily distributed as chi-square. However, Everitt and Hand argue that the test can still be used informally, and in any case the improvement in fit is so pronounced that it can hardly be dismissed.

Furthermore, we obtained reasonable values of the parameters of the mixed negative binomial distribution. When we fitted this distribution, we did not constrain the solution in any way to associate a specific component distribution with a particular sub-collection. For instance, we didn't know a priori whether m_1 would refer to proportion of circulating titles that were adult fiction, adult nonfiction, or juvenile literature. Nor did we constrain the mixing proportions themselves, except of course to be between 0 and 1 and to add to 1. In fact, the only data we used were the aggregate data. Nonetheless, the mixing proportions obtained as part of our solution seem to match very closely the actual proportions of circulating titles in the three sub-collections. Brownsley and Burrell did not actually report the number of titles circulating in each of the three sub-collections, but they did report the number of "issues," or circulation occurrences, for each sub-collection. According to their figures, adult fiction accounted for 64% of the issues, adult nonfiction for 27%, and juvenile literature for 9%. These percentages match our mixing proportions (.69, .25, and .06) closely, suggesting that the mixed negative binomial distribution is indeed a good model for these data.

This result also suggests that the first set of parameters refers to adult fiction, the second set to adult nonfiction, and the third set to juvenile literature. Further support for this

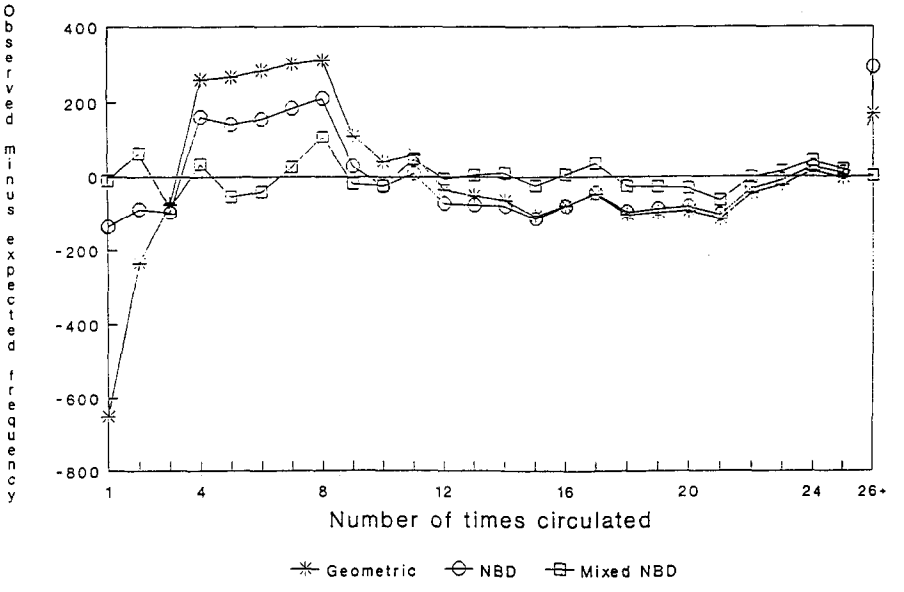

Fig. 2. Lack of fit for three models of library circulation.

interpretation comes from an inspection of the values obtained for k, which Brownsley and Burrell refer to as the index of the negative binomial distribution. According to them, the index should be greater than one for adult fiction and less than one for adult nonfiction. As can be seen from Table 4, labelling the first sub-collection as adult fiction and the second sub-collection as adult nonfiction is consistent with this rule.

It is interesting to note that Brownsley and Burrell tended to prefer the geometric distribution for this particular library, considering it to be generally adequate. The improved fit and ready interpretability of the mixed negative binomial distribution lead us to believe that it is a better model for these circulation data.

CONCLUSION

This paper has shown three different examples of how direct comparisons of bibliometric models can be made, and of the kinds of interpretations that can be given to such comparisons. We find the idea of asking questions in terms of full models versus restricted models, thereby pitting parsimony against accuracy of prediction, to be an extremely useful framework within which to think about bibliometrics, and indeed statistics in general. In addition, we think that the likelihood ratio chi-square statistic, G^2, has merit and we would like to see it gain greater acceptance in bibliometric modelling.

One drawback of the approach we have outlined in this article is that it can only be used for models that are hierarchically related. Sometimes bibliometric researchers want to compare distinct models. For example, Gelman and Sichel (1987) fitted both a beta-binomial distribution and a negative binomial distribution to library circulation data. These distributions do not stand in a hierarchical relationship to one another. Comparisons of models that are distinct in this way tend to be made on a more ad hoc basis. Systematic approaches to such model comparisons are sometimes available, however (Gilchrist, 1984, pp. 157–162). One approach that appears especially promising involves comparing models in terms of the Akaike Information Criterion (*AIC*) (Akaike, 1974; Sclove, 1987). As its name implies, *AIC* is based on concepts from information theory (Bozdogan, 1987).

AIC for an arbitrary model k can be defined as

$$AIC = -2 * \log[\max L(k)] + 2 * m(k) \tag{15}$$

where $\max L(k)$ is the maximum of the likelihood function over the parameters of model k and $m(k)$ is the number of parameters used by model k. The best model is usually thought to be the one that yields the lowest value of *AIC*. A particularly interesting feature of *AIC* is the way it penalizes models for the number of parameters they contain. This emphasizes the importance of simpler models.

AIC has found considerable application in the modelling of time series data and is gaining wider acceptance in factor analysis. Furthermore, Takane (1987) and Sakamoto, Ishiguru, and Kitigawa (1986) discuss the use of *AIC* to guide the selection of models of contingency table data. Since this work also involves categorical data, it is the closest application of *AIC* to bibliometrics that we are aware of. In fact, in his comparisons of models of cross-classified data, Takane used a formula for *AIC* that is based on G^2:

$$AIC = G^2 - 2 * df, \tag{16}$$

where df is the number of degrees of freedom associated with the model under consideration. If one has already calculated G^2 as part of estimating the parameters for a model, eqn. 16 makes the use of *AIC* for model comparisons completely straightforward.

Acknowledgement – This work was supported by Grant A8088 from the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

Akaike, H. (1974). A new look at statistical model identification. *IEEE Transactions on Automatic Control, 19*, 716-723.

Appelbaum, M.I. & Cramer, E.M. (1974). Some problems in the nonorthogonal analysis of variance. *Psychological Bulletin, 81*(6), 335-343.

Bishop, Y.M.M., Fienberg, S.E., & Holland, P.W. (1975). *Discrete multivariate analysis: Theory and practice*. Cambridge, Mass.: MIT Press.

Bozdogan, H. (1987). Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52*(3), 345-370.

Brownsley, K.W.R. & Burrell, Q.L. (1986). Library circulation distributions: Some observations on the PLR sample. *Journal of Documentation, 42*(1), 22-45.

Burrell, Q. (1980). A simple stochastic model for library loans. *Journal of Documentation, 36*, 115-132.

Burrell, Q. (1982). Alternative models for library circulation data. *Journal of Documentation, 38*(1), 1-13.

Chandler, P.J. (1965). *Subroutine STEPIT: An algorithm that finds the values of the parameters which minimize a given continuous function*. Bloomington: Indiana University Quantum Chemistry Program Exchange.

Cohen, A.C. (1966). A note on certain discrete mixture distributions. *Biometrics, 22*, 566-572.

Cooper, R.A. & Weekes, A.J. (1983). *Data, models, and statistical analysis*. Totowa, N.J.: Barnes and Noble.

Everitt, B.S. & Hand, D.J. (1981). *Finite mixture distributions*. London: Chapman & Hall.

Gelman, E. & Sichel, H.S. (1987). Library book circulation and the beta-binomial distribution. *Journal of the American Society for Information Science, 38*(1), 3-12.

Gilchrist, W. (1984). *Statistical modelling*. Chichester, U.K.: Wiley.

Gross, A.J. & Miller, M.C. (1981). Some applications of statistical distribution theory to biology and medicine. In C. Taillie et al. (Eds.), *Statistical distributions in scientific work*. Vol. 6: *Applications in physical, social, and life sciences* (pp. 317-336). Dordrecht, Holland: Reidel.

Harris, C.M. (1983). On finite mixtures of geometric and negative binomial distributions. *Communications in Statistics (A), 12*, 987-1007.

Howell, D.C. & McConaughy, S.H. (1982). Nonorthogonal analysis of variance: Putting the question before the answer. *Educational and Psychological Measurement, 42*(1), 9-24.

Judd, C.M. & McClelland, G.H. (1989). *Data analysis: A model-comparison approach*. San Diego: Harcourt, Brace Jovanovich.

Marascuilo, L.A. & Busk, P.L. (1987). Log-linear models: A way to study main effects and interactions for multidimensional contingency tables with categorical data. *Journal of Counseling Psychology, 34*(4), 443-455.

Maxwell, S.E. & Delaney, H.D. (1990). *Designing experiments and analyzing data: A model comparisons approach*. Belmont, CA: Wadsworth.

McLachlan, G.J. & Basford, K.E. (1988). *Mixture models: Inference and applications to clustering*. New York: Dekker.

Nelson, M.J. & Tague, J.M. (1985). Split size-rank models for the distribution of index terms. *Journal of the American Society for Information Science, 36*(5), 283-296.

Nicholls, P.T. (1986). Empirical validation of Lotka's law. *Information Processing & Management, 22*(5), 417-419.

Nicholls, P.T. (1987). *The Lotka hypothesis and bibliometric methodology*. Ph.D. Thesis, School of Library and Information Science, University of Western Ontario.

Pao, M.L. (1985). Lotka's law: A testing procedure. *Information Processing & Management, 21*(4), 305-320.

Riefer, D.M. & Batchelder, W.H. (1988). Multinomial modeling and the measurement of cognitive processes. *Psychological Review, 95*(3), 318-339.

Sakamoto, Y., Ishiguru, M., & Kitagawa, G. (1986). *Akaike information criterion statistics*. Dordrecht, Holland: Reidel.

Sclove, S.L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika, 52*(3), 333-343.

Sichel, H.S. (1985). A bibliometric distribution which really works. *Journal of the American Society for Information Science, 36*(5), 314-321.

Tague, J.M. & Nicholls, P.T. The maximal value of a Zipf size variable: Sampling properties and relationship to other parameters. *Information Processing & Management, 23*(3), 155-170.

Takane, Y. (1987). Analysis of contingency tables by ideal point discriminant analysis. *Psychometrika, 52*(4), 493-513.

Titterington, D.M., Smith, A.F.M., & Makov, U.E. (1985). Statistical analysis of finite mixture distributions. Chichester, U.K.: Wiley.