INFORMETRICS 87/88
L. Egghe and R. Rousseau (Editors)
© Elsevier Science Publishers B.V., 1988

PARAMETERS FOR CLUSTER ANALYSIS OF LIBRARY OVERLAP

William E. McGRATH

School of Information and Library Studies, State University of New York at Buffalo, Buffalo, New York 14260, USA

Abstract

The relationship between two common approaches to library overlap is discussed : 1) the practice of tabulating frequencies of 0,1,2 ..., n copies of titles in a network, and 2) reporting of the number of co-occurring titles in a symmetric resemblance matrix. It is contended that an understanding of the first approach is important in the practice of the second approach-- e.g., in constructing a resemblance coefficient between any two libraries and in establishing sample size when the matrix is used as input for cluster analysis. The Poisson and negative binomial distributions were tested for fit with several samples of data. The negative binomial distribution provides a much better fit than the Poisson and generates frequencies of 2, the essential ingredient in symmetric overlap matrices, which are very close to frequencies of 2 in the samples. Since the Poisson assumes equal propabilities of success, while the negative binomial assumes variable probability, the better fit may be explained by assuming that each book has its own probability of being acquired.

1. INTRODUCTION

Approaches to Overlap-Distributions and Clustering

"Overlap" is a general term for describing the extent to which libraries have similar collections. There are two common approaches for describing overlap. The first is to tabulate the distribution of titles held by 0,1,2, ..., n libraries, and to compute the mean from that distribution. The second is to report the number of titles held in common by any two libraries in a symmetric matrix as in Table 1, where Library A, for example, has 12 books in common with library B. The two approaches are rarely seen as related, but they obviously are--i.e., the matrix is constructed from the frequencies of 2 in the distribution f(x), x = 0,1,2,...,n. This paper shows how the first approach

Table 1 : Symmetric matrix--with typical overlap data, number of titles held in common by any 2 libraries.

		Library			
		A	B	C	D
	A	--	12	5	10
Library	B	12	--	8	9
	C	5	8	--	14
	D	10	9	14	--

may have a bearing on the second approach, by attempting to fit f(x) to a statistical distribution and by showing why that distribution may be important in the construction of an overlap matrix. The literature of overlap is reviewed by Potter and will not be further reviewed here [1].

In few overlap studies has analysis gone beyond reporting of the mean of the distribution in the first approach, or of the matrix itself (either counts or percentages) in the second approach, or of the sampling procedures for either approach--e.g., how titles were selected from card catalogs.

For a small number of libraries, the relationships represented by these counts are comprehensible. In Table 1, it is clear enough that Libraries C and D are more similar to each other than are Libraries C and A. For a large number of libraries, however, the number of relationships becomes very large, and more difficult to comprehend. For example, what does it mean to compare two libraries in a group of 100 libraries when the number of pair-wise comparisons is 4950 ? Or to compare any three libraries ? Or any four ?

Hierarchical clustering is a useful way to reduce these comparisons to something visualizable and comprehensible (2,3,4,5). Clustering methods usually require input from matrices such as that in Table 1 in which the data represent resemblance, or similarity between two objects--in this case, between any two libraries.

Two problems arise in the application of clustering methods : 1) selection of the appropriate resemblance coefficient and 2) determining the minimum sample size.

Resemblance Coefficients

"Resemblance" is a general term used in this paper to include the concepts of correlation, co-occurrence, similarity, dissimilarity, proximity, distance and so on. Coefficients may be either quantitative or qualitative.

Depending on the circumstances, many of the familiar resemblance coefficients may not be appropriate. For example, a simple count of the number of titles held in common is inappropriate if it does not take into account the total number of titles held by each library. The Pearson correlation coefficient may not meet the assumptions of normality and homogeneity of variance. The Euclidean distance measure--an intuitively clear measure based on the sum or average by hypotenuses in the Pythagorean theorem--is oft-used by cluster analysts. It is appropriate where the data are aggregated--for example, in subject categories and when information about individual titles is not available. But it is inappropriate if the cluster analyst wishes to cluster libraries based on titles instead of subjects. These and other measures, and their appropriate application, are reviewed in Romesburg [6]. Several of these coefficients are now routine options in the Cluster program SPSS-X statistical analysis package [7].

Jaccard Resemblance Coefficient

An appropriate measure for library overlap is the Jaccard coefficient [6]. It is simply constructed and intuitively clear. It is based on the presence or absence of attributes, in this case, books. The Jaccard coefficient is a qualitative measure constructed entirely from the sum of binary 1's and 0's for two libraries as in Table 2.

The Jaccard coefficient is computed from the formula

$$J_{AB} = \frac{a}{a + b + c},$$

and has a positive range from 0.0 to 1.0. Using the data from Table 2(b),

$$J_{AB} = \frac{2}{2 + 1 + 2} = 0.4$$

The Jaccard coefficient is widely used for cluster analysis in numerical taxonomy (7,8). The ownership of a particular book is analogous to the presence or absence of particular attributes in plant or animal specimens, such as the presence or absence of diaphanous wings. By analogy, the library is the specimen and the book is the attribute. For example, owning the Gutenberg

Table 2 : Construction of the Jaccard coefficient, for two libraries; 1 = title owned; 0 = not owned; (a) data; (b) summary table; (c) frequency distribution

Titles owned	Library	
	A	B
Hamlet	1	0
MacBeth	0	0
Iliad	0	0
Odyssey	1	1
Republic	0	0
The Bible	1	1
Oedipus	0	1
Decameron	0	0
Euclid	0	0
Faust	0	1
Total	3	4

(a)

		B	
		1	0
A	1	a = 2	b = 1
	0	c = 2	d = 5

(b)

x	f(x)
0	5
1	3
2	2

(c)

Bible is an attribute of The Library of Congress, but not of the State University of New York at Buffalo. The Jaccard has the advantage of being self-standardizing, so that the effect of one library being larger than another is automatically neutralized. Notice that "d", joint lack of ownership, is not used in the formula. According to most taxonomists, joint lack of attributes should not contribute to similarity between specimens, though some qualitative measures do include it.

Since the Jaccard is computed from binary data, the distribution of its components is obviously a binomial function as in Table 2(c), from which the coefficient

$$J_{AB} = \frac{f(2)}{f(1) + f(2)} = \frac{2}{3 + 2} = 0.4$$

may also be computed. Thus, the coefficient is simply the ratio of the number of books commonly owned to the total number owned.

Problem of Minimum Sample Size

A sufficiently large sample is necessary to establish confidence in the resemblance coefficients as well as in whatever clusters emerge from the cluster analysis. Each coefficient is an estimate of the true resemblance, while the clusters are hypothetical approximations of true clusters.

Confidence in these estimates, of course, depends on the underlying distribution, f(x), x = 0,1,2, ..., n, of the total number of titles in the sample. Since the data are binary, it is assumed that the distribution is one of the binomial family-- e.g., the Poisson or negative binomial, or a mixture. If a sample can be shown to be one of these, than it may be assumed that the sample is large enough.

It should be noted that the distribution in question (f(x), x = 0,1,2,3, ..., n) is for the total sample from which the resemblance matrix is constructed. As pointed out above, each coefficient has its own distribution (f(x), x = 0,1,2) which is a subdistribution of the total sample. Although it is assumed that the subdistribution is some form of binomial, the question is not further addressed in this paper. The question remains to be addressed, however.

Confidence in the clusters is a less familiar problem. In a large network of libraries, the number of coefficients in the resemblance matrix is very large. For three of the samples in this study, the number of libraries was 22, and the number of coefficients was therefore (22*21)/2 = 231. The distribution of these coefficients is necessarily highly skewed, particularly if the number of holdings for each library is highly skewed. If the sample is too small, there will be a large number of zero coefficients, suggesting a zero relationship between two libraries--an unlikely situation, since virtually all libraries will have at least a few books in common.

In either case, confidence in the coefficients, or confidence in the clusters, the problem is the same. There must be some minimum number of titles in the entire sample to establish a standard confidence level. Though no attempt is made in this paper to determine that minimum sample, it is contended that when the sample is large enough to fit a particular distribution the confidence level may be at least minimal.

It is commonly supposed that overlap data must be Poisson distributed, but the Poisson simply does not fit data very well, as will be shown here. The reason may be that the Poisson rests on the assumption of equal probability for each success. For libraries, the assumption would be that each title (book, periodical, etc.) has an equal chance of being acquired. But obviously, this cannot be the case. Books are published in unequal copies. Authors are not equally known. Publishers are not equally regarded. No two books receive equal publicity. Subjects are not equally popular. Books receive unequal reviews, or none at all. Furthermore, the fact that one library has acquired a book may influence another to acquire it. The number of citations a book receives may also influence acquisitions. Therefore, it is reasonable to assume that Poisson trials are not independent and that each book has an unequal chance of being acquired. These circumstances suggests the negative binomial distribution, in which the number of trials varies where the number of successes is constant. For overlap, this translates to a variable number of libraries for a fixed number of copies.

Burrell and Cane have extensively analyzed the negative binomial in application to library circulation data. They assume that each circulating title carries its own "desirability" or probability of circulation [9,10]. Tague argues that "success-breeds-success" phenomena, such as citation frequencies discussed by Price [11], lead to the negative binomial distribution [12]. Nielson and Tague test it and other distributions for fit with frequency of index terms [13]. Further references to the negative binomial are given in foregoing citations.

2. METHOD

The remainder of this paper reports on attempts to fit the Poisson and negative binomial distributions to several samples using the chi-square goodness of fit test. Samples of overlap data were obtained from various sources. These sources are documented in the Appendix. All analysis was done with the Lotus spreadsheet program in a personal computer. The standard formula for the Poisson,

$$P(x) = \frac{m^x e^{-m}}{x!} ,$$

was entered into Lotus. For the negative binomial, the method of moments, was used where the mean is given by kq/p, the variance by kq/p^2, and where p is estimated by m/Var, and k is given by mp/(1-p) [14]. The terms for x = 0,1,2, ..., n, are given by

$$p^k, \; kq \; p^k,$$

and in general, the (xth + 1) term is obtained from the xth term by multiplying by q(x + k - 1)/x.

3. RESULTS

Fitting Distributions to Overlap Data

For all samples tested, the negative binomial distribution provides a much better fit than the Poisson (Tables 3 to 12). For most samples the fit is very good. For example, in all but Tables 5 and 11, no significant difference was found between the data and the negative binomial distribution, whereas the data were significantly different from the Poisson in every case. Where the negative binomial was significantly different (Tables 5 and 11), the fit was still much better than the Poisson. In several cases where the negative binomial was significantly different, it was possible to achieve a better fit by combining categories, though this practice is not necessarily recommended. A more desirable approach may be to test for mixtures of negative binomials or other mixtures. No attempt to do this has been made in this study.

Table 3 : Distribution of a Sample of Titles in the Niagara University-Nioga Network.

x	f(x)	Poisson (a)	Negative Binomial (b)
0	91	72	90
1	23	43	22
2	7	13	10
3	5	3	5
4	3	1	2
5	1	0	1
6	1	0	1
7	1	0	0
Total	132		

(a) Sig. Chi-square = 100, df = 2, 0.05.
(b) n.s. Chi-square = 0.93, df = 5, 0.05.
note : mean = 0.62; variance = 1.55

Table 4 : Distribution of a Sample of Titles in the Niagara Community College Network.

x	f(x)	Poisson (a)	Negative Binomial (b)
0	70	6	71
1	29	14	27
2	15	16	14
3	5	12	8
4	4	7	5
5	4	3	3
6	1	1	2
7	1	0	1
8	0	0	1
9	1	0	0
10	1	0	0
Total	131		

(a) Sig. Chi-square = 687, df = 3, 0.05.
(b) n.s. Chi-square = 1.65, df = 8, 0.05.
note : mean = 1.08; variance = 3.15.

Table 5 : Distribution of a Sample of Titles in the Nioga Network.

x	f(x)	Poisson (a)	Negative Binomial (b)
0	75	37	71
1	40	62	48
2	32	53	31
3	29	30	19
4	4	13	12
5	7	4	7
6 to 10	13	1	11
Total	200		

(a) Sig. Chi-square = 100, df = 5, 0.05.
(b) Sig. Chi-square = 12.47, df = 4, 0.05.
note : mean = 1.7; variance = 4.19.

Table 6 : Distribution of a Sample of Titles in the Niagara University-Nioga, Niagara Community College Network. (Aggregate of Tables 1,2,3.)

x	f(x)	Poisson (a)	Negative Binomial (b)
0	236	129	225
1	92	167	104
2	54	109	58
3	39	47	34
4	12	15	20
5	14	4	12
6	11	1	7
7	6	0	5
8 to 10	8	0	6
Total	472		

(a) Sig. Chi-square = 189.9, df = 5, 0.05.
(b) n.s. Chi-square = 8.89, df = 6, 0.05.
note : mean = 1.3; variance = 3.64.

Table 7 : Distribution of a Sample of Titles in the Louisiana Numerical Register.

x	f(x)	Poisson (a)	Negative Binomial (b)
0	39844	38229	39830
1	5669	8233	5697
2	1369	887	1364
3	386	64	370
4	105	3	107
5	28	0	32
6	6	0	10
7 to 8	7	0	4

(a) Sig. Chi-square = 5773, df = 3, 0.05.
(b) n.s. Chi-square = 5.11, df = 5, 0.05.
note : mean = 0.22; variance = 0.32.

Table 8 : Distribution of Titles, Dawson Data.

x	f(x)	Poisson (a)	Negative Binomial (b)
0	4747	4721	4745
1	155	204	159
2	23	78	20
3 to 4	4	4	4
Total	4929		

(a) Sig. Chi-square = 123, df = 2, 0.05.
(b) n.s. Chi-square = 0.47, df = 1, 0.05.
note : mean = 0.04; variance = 0.06.

Table 9 : Distribution of a Sample of Titles from the North Carolina Data.

x	f(x)	Poisson (a)	Negative Binomial (b)
0	1036	875	1022
1	119	327	142
2	60	61	54
3	26	8	25
4	13	1	13
5	9	0	7
6	5	0	4
7 to 9	3	0	4
Total	1271		

(a) Sig. Chi-square = 418, df = 3, 0.05.
(b) n.s. Chi-square = 5.46, df = 5, 0.05.
note : mean = 0.37; variance = 1.01.

Table 10 : Distribution of Titles from Altman's Study of Overlap in Secondary Schools.

x	f(x)	Poisson (a)	Negative Binomial (b)	Negative Binomial Aggregated (c)
0	1453	543	1410	1410
1	581	948	597	597
2	322	828	353	353
3	230	481	228	228
4	133	210	153	153
5	102	73	106	106
6	76	21	74	74
7	66	5	53	53
8	41	1	38	38
9	35		27	27
10	21		20	20
11	12		14	14
12	7		10	31
13	9		8	
14	4		6	
15	6		4	
16 +	14		3	
Total	3112			

(a) Sig. Chi-square = 4345, df = 7, 0.05.
(b) Sig. Chi-square = 56.03, df = 14, 0.05.
(c) n.s. Chi-square = 16.39, df = 10, 0.05,
frequencies for 12+ combined.
Note 1 : mean = 1.75; variance = 7.19.
Note 2 : Altman's data is reinterpreted to reflect "every other library having a title", thus the 1 copy class becomes the zero class.

Table 11 : Distribution of a Sample of Titles in the Louisiana Union Catalog, 1959-62.

x	f(x)	Poisson (a)	Negative Binomial (b)
0	418	277	427
1	83	200	61
2	25	72	29
3	14	17	17
4	2	3	11
5	6	0	7
6	3	0	5
7	5	0	4
8 to 13	14	0	9
Total	570		

(a) Sig. Chi-square = 686, df = 3, 0.05.
(b) n.s. Chi-square = 20.6, df = 6, 0.05.
note : mean = 0.72; variance = 3.65.

Table 12 : Distribution of a Sample of Titles in the Louisiana Union Catalog 1963-67.

x	f(x)	Poisson (a)	Negative Binomial (b)
0	1102	812	1102
1	115	404	103
2	41	101	47
3	22	17	27
4	14	2	17
5	9		11
6	4		8
7	7		6
8	8		4
9 to 14	14		10
Total	1336		

(a) Sig. Chi-square = 1600, df = 3, 0.05.
(b) n.s. Chi-square = 12.1, df = 7, 0.05.
note : mean = 0.5; variance = 2.66.

Since frequencies of 2 are of essential interest in the construction of overlap matrices, they and their Poisson and negative binomial expected frequencies are summarized in Table 13. The table shows in nearly all samples that the negative binomial predicts frequencies very close to actual data for both large and small samples.

Tabel 13 : Summary of expected Number of Co-occurring Titles, f(2), in Ten Samples.

Co-occuring titles								
Source	Sample Size	Actual	Poisson	Negative Binomial	mean	var	p	k
Niagara Un.	132	7	13	10	0.62	1.55	0.40	0.42
Niagara CC.	131	15	16	14	1.08	3.15	0.34	0.57
Nioga Netw.	200	32	53	31	1.70	4.19	0.40	1.15
NU/NCC/Nio.	472	54	109	58	1.30	3.64	0.36	0.72
Louis. NR	47414	1369	887	1364	0.22	0.32	0.66	0.43
Dawson	4929	23	78	20	0.04	0.06	0.78	0.15
No. Carol.	1271	60	61	54	0.37	1.01	0.37	0.22
Altman Study	5000	322	828	353	1.75	7.19	0.24	0.56
LUC, 1959-62	570	25	72	29	0.72	3.65	0.20	0.18
LUC, 1963-67	1336	41	101	47	0.50	2.66	0.19	0.11

Mean and Variance

One important property of the negative binomial is that the variance exceeds the mean, unlike the Poisson for which the mean and variance are equal. According to Williamson and Bretherton,

> Any ... tendency for one event to increase the probability of another event, will increase the variance of the distribution relative to the mean, and a negative binomial distribution will invariably better fit the data [14] .

For every sample tested in this study, the variance exceeded the mean providing further evidence in favor of the negative binomial.

4. CONCLUSION

Given the close fit to the negative binomial for most samples tested, and the high ratio of variance to the mean, it may be concluded that the negative binomial is a likely distribution for overlap data. This conclusion makes good sense when one considers the circumstances of library acquisitions.

From the discussion above, it appears that acquisition by one library can readily influence the probability of acquisition by another library. One library may purchase a book because another library thought it important enough to acquire. On the other hand, cooperating libraries will often not acquire a book if another in the network already owns it. In either case, each book carries its own probability and the variance is higher than it would be if each book had equal probability, an essential distinction between the Poisson and negative binomial processes.

Finally it is shown that the negative binomial generates frequencies of 2 which are very close to the frequencies of 2 in actual data. Since frequencies of 2 are the essential ingredient in constructing of resemblance matrices for cluster analysis of library overlap, it is suggested that overlap data be routinely tested for the negative binomial distribution when cluster analysis is the research objective.

Further Research

The following questions remain to be answered. What is the exact distribution of the titles owned by any two libraries ? When the Poisson or negative binomial do not fit well, would some mixture of these provide a better fit ? Given a certain number of libraries in a network, what is the minimum sample of titles required to establish confidence in the clusters ? How can statistical confidence in the cluster be tested ? A distinction needs to be made between sampling methods and the nature of the distribution resulting from each. For example, what is the effect of selecting a sample of fixed size from the universe of titles and then checking holdings against each library compared to selecting separate samples of holdings from each library and then checking the holdings of each sample against every other library in the network ?

REFERENCES

[1] Potter, W.G., Studies of Collection Overlap; a literature review, Library Research 14 (1982) pp. 3-21.

[2] McGrath, W.E., Multidimensional map of library similarities, In : Communicating Information : Proceedings of the 43rd Annual Meeting of the American Society for Information Science, Anaheim, CA (1980) pp. 298-300 (White Plains, NY, Knowledge Industry Publications, 1980).

[3] McGrath, W.E., Morphology and the structure of libraries-- a fresh look at descriptive methods for management, Science & Technology Libraries (984) 4, pp. 117-132.

[4] McGrath, W.E., Collection evaluation; theory and the search for structure, Library Trends (1985) 22(3) pp. 241-266

[5] McGrath, W.E. and Hickey, T.B., Research Report prepared for OCLC on multidimensional mapping of libraries based on shared holdings in the OCLC Online Union Catalog (Dublin; Ohio, OCLC, Office of Research, 1983) (OCLC/OPR/RR-83/5).

[6] Romesburg, H.C., Cluster analysis for Researchers, (Belmont, Calif., Lifetime Learning Publications, 1984).

[7] Norusis, M.J., SPSS-X User's Guide, (New York, etc., McGraw-Hill Book Company, 1985).

[8] Sneath, P.H.A. and Sokal, R.R., Numerical Taxonomy (San Francisco, W.H. Freeman, 1973).

[9] Burrel, Q.L., A note on ageing in library circulation model, Journal of Documentation (1985) 41(2) pp. 100-115.

[10] Burrel, Q.L. and Cane, V.R., The analysis of library data, Journal of the Royal Statitistical society (1982) 145(4) pp. 439-471.

[11] Price, D., A general theory of bibliometric and other cumulative advantage processes, Journal of the American Society for Information Science (1976) 27(5) pp. 292-306.

[12] Tague, J., The success-breeds-success phenomenon and bibliometric processes, Journal of the American Society for Information Science (1981) 36(4) pp. 280-286.

[13] Nielson, M.J. and Tague, J.M., Split size-rank models for the distribution of index terms, Journal of the American Society for Information Science (1985) 36(5) pp. 283-296.

[14] Williamson, E. and Bretherton M.H., Tables of the Negative Binomial Probability Distribution (London, New York, John Wiley & Sons, 1963).

APPENDIX--SOURCES OF OVERLAP DATA

The reliability of the analyses in this study may depend partly on the sampling methods of the following sources.

Tables 3-6. Data collected by Frances Wilson, SUNY Buffalo, 1985.

Table 7. Data collected by William McGrath from the Louisiana Numerical Register, ca. 1971.

Table 8. Data published in J.M. Dawson. 1957. The acquisition and cataloging of research in libraries : a study of the possibilities for centralized processing. Library Quarterly 27 pp. 1-22.

Table 9. Published data from Eugene T. Neely. 1971. The Norht Carolina Catalog; an examination and evaluation. Champaign, University of Illinois Graduate School of Library Science (Occasional Paper, no. 99).

Table 10. Published data from Ellen Altman. 1972. Implications of title diversity and collection overlap for interlibrary loan among secondary schools. Library Quarterly 42(2) pp. 177-194.

Table 11-12. Data collected by William McGrath from the Louisiana Union Catalog.

Other sources tested :

Published data from John A. Urquhart and J.L. Schofield. 1972. Overlap of acquisitions in the University of London Libraries; a study and a methodology. J. Librarianship 4(1) pp. 32-47. Results : Poisson significant; negative binomial not significant when data treated as "every other library having a copy."

Published data from W.Y. Arms. 1973. Duplication in union catalogues. J. Doc. 29 (4) pp. 373-379. Results : Poisson significant; negative binomial significant.

Published data from Debora Shaw. Overlap of monographs in public and academic libraries in Indiana. Libr. & Info. Sci. res. 7(3) pp. 275-298. Results : Poisson significant; negative binomial significant.