INFORMETRICS 89/90
L. Egghe and R. Rousseau (Editors)
© Elsevier Science Publishers B.V., 1990

CLUSTER ANALYSIS OF TITLE OVERLAP IN TWENTY-ONE LIBRARY COLLECTIONS IN WESTERN NEW YORK

Frances L. WILSON

Director of Prospect Research
University at Buffalo Foundation

William E. McGRATH

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

Abstract

Title overlap of 21 libraries in western New York, was described by the Jaccard similarity coefficient. The coefficient has two features appropriate for describing overlap : it is self-normalizing and avoids mutual lack of ownership as a basis for similarity. The matrix of 210 coefficients was submitted to 4 methods of cluster analysis : *single linkage*, *complete linkage*, *average linkage and Ward's*. Cophenetic correlation coefficients for single, complete, and average linkage, 0.55, 0.69, and 0.74, respectively, suggested that the average linkage was best. Ward's was judged poorest on the basis of visual inspection. Two clusters from the average linkage method, with 8 and 12 libraries each, were prominent. A third contained only 1 library. The one-dimensional solution to multidimensional scaling (MDS) tended to confirm the clusters. Two variables were hypothesized to account for the two large clusters : *circulation* and *total volumes*. Both were significantly associated with the clusters, but analysis of covariance showed that when *total volumes* were treated as the covariate, circulation was not significant. Hence neither variable was considered explanatory. Principle contribution to overlap analysis was (1) use of the Jaccard coefficient, (2) the attempt to validate overlap clusters using MDS, and (3) an effort to explain clusters by other variables.

STATEMENT OF THE PROBLEM

INTRODUCTION

The purpose of this study was to describe the overall similarity, or overlap, of collections of member libraries of the Nioga Library System (NIOGA), *Niagara University (NU), and Niagara County Community* College (NCCC) and to look for factors which might be associated with that similarity.* Description of these libraries is in the Appendix.

* Though the study was not officially endorsed by these libraries, some were interested in potential applications for further cooperative activities.

BACKGROUND

Overlap of library collections has been of prime importance to librarians interested in potential cooperative activities since the founding of library networks. Though overlap was known to exist, librarians were not sure whether it was high or low nor what they should do about it in either case.

Early studies were usually limited to frequency counts of titles held by two or more libraries. Some studies found that the number of titles shared by any pair of libraries varied greatly. Some pairs shared more titles than other pairs. Other studies found that a small number of titles were held by a large number of libraries or, conversely, that a large number of titles were held by few libraries. Nothing much was made of these findings other than that there was high or low overlap. High overlap was used to justify cooperative collection development and cataloging. Low overlap was used to justify sharing of collections. Either finding was used to justify the formation of networks.

Mathematically, these studies demonstrated simple binary relationships, but did not convey an overall picture. They did not try to characterize the overlap in any general way, while few looked for explanations.

SPECIFIC PURPOSE

This study looks beyond the question of "how much" overlap by describing its general structure and by attempting to explain that structure in terms of other variables.

Since collections are necessarily similar or dissimilar, and since the number of similarities and dissimilarities between libraries is very large - even in a small network - how can one acquire an overall understanding of the network's overall resources without bogging down in local relationships? In other words, how or where can one stand to see the forest? Does the forest contain distinctive stands of oaks, pines and maples, or do the trees intermingle without any species appearing dominant? Do libraries cluster according to concentrations of subjects, or do they all stand together, one indistinguishable from another? And, whether they cluster or not, what are the implications?

Specifically, a high degree of similarity (many titles in common, or intense clustering) would have important implications for centralized acquisitions, mutual collection development, on-line catalogs, and cost benefits, particularly for libraries which are not members of the Nioga System--Niagara University and Niagara County Community College, for example.

A low degree of similarity (few titles in common, or little clustering) would suggest that adding the titles of one or both of the academic libraries to the Nioga System would benefit all the members by providing access through a common library card and the on-line circulation system.

Otherwise, the contribution of the study is in its application of particular methods, namely, the Jaccard coefficient to express overlap between two libraries, cluster analysis computed from those coefficients, multidimensional scaling (MDS) in conjunction with cluster analysis, and in its attempt to validate clusters with other variables.

REVIEW OF LITERATURE

McGrath and Hickey (1980) sampled 57,000 titles from the Online Union Catalog of OCLC and introduced multidimensional scaling (MDS) to the analysis of overlap. MDS, a technique commonly used in numerical taxonomy, psychology, sociology, communications and other disciplines, graphically maps relationships between objects, in this case libraries. MDS is described in detail later in this paper.

Potter (1982) reviewed the literature on collection overlap. He noted great diversity in objectives, methodologies, and reporting which made comparison and generalization difficult. He listed four applications : cataloging, central processing, cooperative collection development, and research studies whose concern was "to achieve a better understanding of the phenomenon of collection overlap".

Shaw (1983) examined overlap among public and academic libraries to assess possible predictors of overlap, concluding that overlap correlates with library type, collection size, publication date, and subject.

Rogers (1984), in a review of selected networking studies, suggested that practitioners need practical studies which can be replicated in libraries and research tools to minimize intuitive decisions.

In a series of papers, McGrath applied cluster analysis and multidimensional scaling to the analysis of both subject overlap within a library and to collection overlap between libraries : circulation by subject (1983); the difference between hypothesis testing and hypothesis generating in the context of collections and networks (1984); the theory of collection evaluation using various collection matrices (1985); centralization and decentralization of academic library collections (1986); use of the Jaccard coefficient for overlap similarities, the relationship between cluster analysis and binary data and fitting the negative binomial distribution to overlap data (1988). McGrath, Geraci and Romney discussed four ways of interpreting a data matrix used in cluster analysis and MDS of library circulation (1986).

METHODOLOGY

CLUSTER ANALYSIS

In early overlap studies, *percentage* was the primary descriptive statistic : percentage of overlap overall, and percentage between two libraries. Two shortcomings of this approach were in the limitations of a pairwise statistic when the number of libraries is very large and its use of percentages as the comparative statistic.

If the number of libraries was small (fewer than ten), percentages could be meaningfully interpretated. For a large network, the number of percentages increased geometrically making interpretation more difficult.

This can be shown from the formula for combinations :

$$C(n,m) = \frac{n!}{m!(n-m)!} \quad ,$$

where C(n,m) is the number of combinations of n libraries in the network taken m at a time. To illustrate m = 2, or pairs : 3 libraries have 3 pairs, 21 have 210, 45 have 990, and 90 libraries have 4005 pairs. Clearly, pairwise comparison of libraries in a large network is uninterpretable without further analysis.

As alternatives to these shortcomings, this study used cluster analysis and the Jaccard similarity coefficient. Cluster analysis is an appropriate method for coping with a large number of relationships. It does so by revealing structure in the overlap data and by providing a basis for generating hypotheses about the meaning of the structure (McGrath, 1984).

The Jaccard coefficient, often used with cluster analysis, has a number of advantages over raw overlap counts, percentages and other coefficients, among which are : it is simple, intuitive, is self-normalizing, and avoids the problem of computing similarity based on lack of ownership.

SAMPLING PROCEDURE AND DATA COLLECTION

Sampling procedure for overlap studies differs with the purpose of the study, availability of data, and the preferences of the investigator. Shaw (1983) investigated overlap of public and academic library collections using titles from the collections and from the OCLC database. McGrath (1980) sampled academic library collections from the OCLC database exclusively. Knightly (1973) examined the relationships between title overlap and subject areas using a sample of titles from the American Book Publishing Record checked against the holdings of each library.

O'Neill (1972) examined title overlap in a study of 18 libraries in Western New York. Using cluster sampling, he randomly selected fifty Library of Congress classifications, and tabulated overlap of titles by examining the shared holdings in each class.

Sampling procedure in these studies were quite different, so that results in each are not necessarily comparable. In this study, titles were randomly sampled directly from the 21 collections. The question was whether or not membership of Niagara University and Niagara County Community College in the NIOGA network would be mutually beneficial.

Two samples of 132 titles each were taken from drawers (also chosen randomly) of the author/title catalogue of NU and NCCC and one of 200 from the NIOGA union list. Serials, children's literature, government documents and audio visual materials were excluded. Differences in imprint were disregarded. Duplicate titles and those common to two or more samples were excluded. The three samples were aggregated for a final sample of four hundred titles.

Bookstein (1983) has discussed sources of bias when sampling from catalogs, for example : inclusion of different formats (government documents, microfiche etc.); variable number of cards per drawer; titles in multiple file categories; thickness of cards. Though these and other sources have probably affected our samples, our intent was not to describe each collection by representative sampling, but to determine the overlap between collections, and in the application of novel methods heretofore not applied to overlap studies.

RAW DATA MATRIX

Data were tabulated in a single 21-library x 400-title data matrix of the form shown in figure 1a. The data are binary, i.e., either a library has the title (1) or does not (0).

CONSTRUCTION OF THE JACCARD RESEMBLANCE COEFFICIENT

The raw overlap data were used to compute the Jaccard similarity coefficient for every pair of libraries. Figure 1 shows the construction of the Jaccard coefficient for two libraries. The coefficient is a function of the variables a, b, c in which a is the number of titles held in common by the two libraries, b is the number held only by one library and c is the number held only by the other library. It is important to note that the number of titles held by neither library - that is, 0 and 0, designating lack of similarity - is not used in the coefficient.

	Libraries	
Titles	A	B
Title 1	1	0
Title 2	0	0
Title 3	1	1
Title 4	0	1
Title 5	1	0
Title 6	1	0
Title 7	0	0
Title 8	1	1
Title 9	0	1
Title 10	1	0
Total	6	4

Figure 1a :
Raw Data Matrix

1 = Owned
0 = Not owned

		Library B	
		1	0
Library A	1	a=2	b=4
	0	c=2	d=2

Figure 1b :
Elements of
Jaccard coefficient

a = 1 1,	both A & B have the title
b = 1 0,	A has title, B does not
c = 0 1,	B has title, A does not
d = 0 0,	Neither A nor B has title

Fig.1 : Construction of the Jaccard Coefficient using 10 "titles" for Libraries "A" and "B". Actual raw overlap data consisted of 400 titles and 21 libraries.

The Jaccard coefficient is the proportion of the number of titles held in common to the total number held,

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

It has a positive range from 0.0 to 1.0. Using the data from figure 1b :

$$J_{AB} = \frac{2}{2+4+2} = 0.25 \ .$$

The Jaccard is popular among numerical taxonomists who support leaving "d" out of the formula, arguing that joint lack of features between specimens should not contribute to their similarity (Romesburg, 1984). Reference to

Jaccard is also found in Sneath and Sokal (1973). It is a similarity measure - i.e., the larger the value, the more similarity between the two objects being measured. The concepts of similarity and distance measure are discussed by Aldenderfer and Blashfield (1984).

RESEMBLANCE MATRIX

Data are input to cluster analysis in the form of a symmetric resemblance matrix, here constructed directly from the raw data matrix. The matrix need not be standardized since binary data are qualitative and dimensionless. Each datum has equal weight in determining similarity (Romesburg, 1984, 143).

CLUSTERING ALGORITHMS

Four common algorithms for computing clusters are : single linkage (SLINK), complete linkage (CLINK), unweighted pair-group method using arithmetic averages (UPGMA), and Ward's minimum variance method. In SLINK, CLINK and UPGMA, the two most similar clusters are merged at each clustering step. Clusters in SLINK are joined at each stage by the single shortest or strongest link between them, tending to produce compact trees. Clusters in CLINK are linked at some maximum distance or minimum similarity. Clusters in UPGMA are linked at each stage according to the largest average similarity between any two groups, producing more interpretable trees.

Ward's method links clusters one step at a time based on the smallest value of an index E, or variance. Aldenderfer and Blashfield (1984) and Romesburg (1984) provide an extensive description of these and other methods. UPGMA is a compromise between the two extremes of SLINK and CLINK (Aldenderfer and Blashfield, 1984). UPGMA produces less distortion in transforming the similarities between objects into a tree (Romesburg, 1984). Calculations for all four methods were performed using the SPSS-X statistical package.

COPHENETIC CORRELATION COEFFICIENT

The cophenetic correlation coefficient is a measure of the goodness of fit between cluster output and the original data matrix. It can be used to test the relative goodness of each of the four cluster algorithms. It is extensively discussed by Romesburg (1984, p.24) and Farris (1969). It is simply the ordinary product moment correlation between the elements of the derived similarities of cluster output and the corresponding original similarities (Sokal and Rohlf, 1962). The coefficient is computed after listing the cophenetic values and the original resemblance coefficients side by side (Table 1). In practice, cophenetic correlation fails to meet certain axiomatic assumptions, for instance, the tree structure may well distort the underlying relationship between the two variables (Rohlf, 1974), so that it is not an entirely satisfactory measure. It should therefore be used with considerable caution.

VALIDATION

Most cluster analysis programs inevitably depict "clusters", but mere depiction does not necessarily mean they are valid or really exist. In this study two methods of confirmation were applied : multidimensional scaling and a test for differences between each cluster.

Table 1 : Cophenetic Matrix and Resemblance Matrix formatted at two Variables

Library Pairs	Cophenetic Matrix	Resemblance Matrix
NT/NIO	1	0.39
NT/NFL	6	0.33
NT/LOC	10	0.28
NT/BTV	18	0.14
NT/MED	18	0.12
NT/NU	22	0.08
NT/NCC	22	0.09
NT/BAK	24	0.05
NT/HOL	24	0.08
NT/LER	24	0.01
......	..	....
......	..	....

MULTIDIMENSIONAL SCALING

A method related to cluster analysis is multidimensional scaling (MDS). ALSCAL, an MDS program now available in SPSS and other packages, was applied to the resemblance matrix.

Cluster analysis and MDS are complementary ways of graphing proximity data. They use the same input data, but differ in the way the data representing objects (libraries, in this study) are graphed. Cluster analysis represents the objects in dimensionless trees, whereas MDS represents them in spatial maps. Cluster analysis converts the similarities to hierarchical relationships, while MDS plots coordinates of objects on a line, a two-dimensional map, or in higher dimensions. It is up to the researcher to determine whether the clusters and higher dimensions are valid. For MDS, Kruskal's STRESS statistic, or the more familiar R-square, are used to evaluate the best of one-, two- and three-dimensional solutions (Kruskal and Wish, 1978).

COMPARISON OF MEAN DIFFERENCES BETWEEN CLUSTERS

One way to test clusters is to treat them as fixed factors and to compare means on some variable associated with each cluster. Significant differences between the means would lend support to their validity.

For all libraries in the study, data for two variables were available which could be tested, *volumes* (total volumes in the library) and *circulation*. Previous studies have found evidence, though not conclusive, that size of collection has an affect on overlap. Also, it is hypothesized that, since different collections have different books, circulation will differ. Two tests were used, the t-test for independent samples in which *volumes* and *circulation* were tested separately, and analysis of covariance in which the variable *volumes* was treated as the covariate, or control, since it is assumed that libraries with more volumes will have greater circulation.

Table 2 : Jaccard Resemblance Coefficients for Library Overlap

	NU	NFL	BTV	LOC	MED	NT	ALB	BAK	HOL	LEW	LER	LYN	MDT	NEW	OAK	RAN	SAN	WIL	YOU	NCC
NFL	0.20																			
BTV	0.06	0.18																		
LOC	0.09	0.28	0.19																	
MED	0.04	0.13	0.22	0.14																
NT	0.08	0.33	0.14	0.28	0.12															
ALB	0.02	0.11	0.14	0.15	0.18	0.15														
BAK	0.02	0.09	0.13	0.10	0.07	0.05	0.13													
HOL	0.03	0.05	0.13	0.13	0.07	0.08	0.14	0.25												
LEW	0.05	0.10	0.08	0.17	0.17	0.18	0.13	0.08	0.16											
LER	0.02	0.06	0.10	0.12	0.07	0.10	0.17	0.16	0.18	0.11										
LYN	0.03	0.06	0.13	0.10	0.11	0.08	0.18	0.12	0.04	0.13	0.08									
MDT	0.01	0.03	0.16	0.06	0.19	0.06	0.23	0.11	0.20	0.19	0.11	0.18								
NEW	0.02	0.05	0.09	0.09	0.13	0.06	0.19	0.11	0.13	0.16	0.18	0.24	0.20							
OAK	0.02	0.04	0.06	0.07	0.07	0.05	0.13	0.10	0.11	0.15	0.05	0.08	0.25	0.05						
RAN	0.02	0.04	0.08	0.09	0.07	0.06	0.03	0.00	0.00	0.15	0.05	0.12	0.11	0.18	0.16					
SAN	0.02	0.08	0.09	0.14	0.11	0.09	0.14	0.08	0.24	0.22	0.12	0.06	0.13	0.13	0.12	0.22				
WIL	0.01	0.01	0.02	0.03	0.03	0.03	0.04	0.00	0.08	0.07	0.07	0.05	0.08	0.18	0.00	0.07	0.11			
YOU	0.02	0.04	0.13	0.04	0.06	0.03	0.09	0.15	0.05	0.14	0.10	0.12	0.10	0.05	0.10	0.00	0.07	0.07		
NCC	0.21	0.15	0.08	0.11	0.05	0.09	0.05	0.02	0.03	0.07	0.01	0.02	0.03	0.02	0.03	0.01	0.02	0.01	0.04	
NIO	0.08	0.30	0.16	0.24	0.16	0.39	0.13	0.06	0.03	0.16	0.09	0.08	0.05	0.06	0.06	0.06	0.09	0.02	0.06	0.08

Mean = 0.10 Min = 0.00 Max = 0.39

RESULTS

RESEMBLANCE MATRIX

Table 2 contains the Jaccard coefficients computed from the raw overlap data, with values ranging from 0 to 0.395 (theoretical range is 0 to 1). The distribution is shown in Table 3.

Table 3 : Distribution of Jaccard Coefficients

Range	Frequency
0.00 - 0.09	110
0.10 - 0.19	81
0.20 - 0.29	16
0.30 - 0.39	3
0.40 - 0.49	0
0.50 - 0.99	0
Total	210
Mean = 0.1	

The distribution is highly skewed, with a relatively low mean. Since overlap distributions tend to be highly skewed (McGrath, 1988), their means are necessarily low. Here, the mean may be typical, and may not have a serious effect on cluster relationships. The impact of this skewed distribution was not otherwise investigated in this study.

CLUSTER ANALYSIS

Figures 2 - 5 show the tree diagrams, or dendrograms, for the SLINK, CLINK, UPGMA and Ward's methods. Each tree has one branch for each library. Branches cannot overlap and the similarity between any two libraries is represented by the height of the lowest node connecting them.

Cophenetic correlation coefficients were computed for SLINK, CLINK and UPGMA (Table 4). The cophenetic coefficient for Ward's was not computed because of the limitation it places on the resemblance coefficient. The coefficients suggest that the UPGMA method yields a better fit.

Visual inspection of the four solutions also suggested that the clusters generated from the average linkage method were more interpretable. Symbols for libraries in each cluster are listed in Table 5.

```
                   5        10        15        20        25
LIBRARY  +---------+---------+---------+---------+---------+

NT       -+------------+
NIO      -+            +---------+
NFL      --------------+         +---------------+
LOC      ------------------------+               +-+
NU       ----------------------------------------+ |
NCC      ----------------------------------------+ |
BTV      --------------------------------------+---+
MED      --------------------------------------+   |
BAK      ------------------------------+---+       +-+
HOL      ------------------------------+   +---+   | |
SAN      ----------------------------------+       | |
RAN      --------------------------------------+-+ | |
LEW      --------------------------------------+ +-+ |
LYN      ----------------------------------+-----+   +-+
NEW      ----------------------------------+     |   | |
MDT      ------------------------------+---+     |   | |
OAK      ------------------------------+   +-----+   | +---+
ALB      ----------------------------------+         | |   |
WIL      --------------------------------------------+ |   |
LER      ----------------------------------------------+   |
YOU      --------------------------------------------------+
```

Fig.2 : Single Linkage Dendrogram

```
                   5        10        15        20        25
LIBRARY  +---------+---------+---------+---------+---------+

NT       -+-----------+
NIO      -+           +------+
NFL      -------------+      +---------------+
LOC      --------------------+               +---------+
BTV      ----------------------+-------------+         +---+
MED      ----------------------+                       |   |
NU       ------------------------+---------------------+   |
NCC      ------------------------+                         |
BAK      ----------------------+----------+                |
HOL      ----------------------+          +----------------+
LER      ---------------------------------+                |
WIL      -----------------------------------------+--------+
YOU      -----------------------------------------+        |
LYN      --------------------+-------+                     |
NEW      --------------------+       +---------------+     |
ALB      ----------------------------+               +-----+
MDT      --------------------+---------------+       |
OAK      --------------------+               +-------+
RAN      ------------------------+--------+
SAN      ------------------------+        +---+
LEW      ---------------------------------+
```

Fig.3 : Complete Linkage Dendrogram

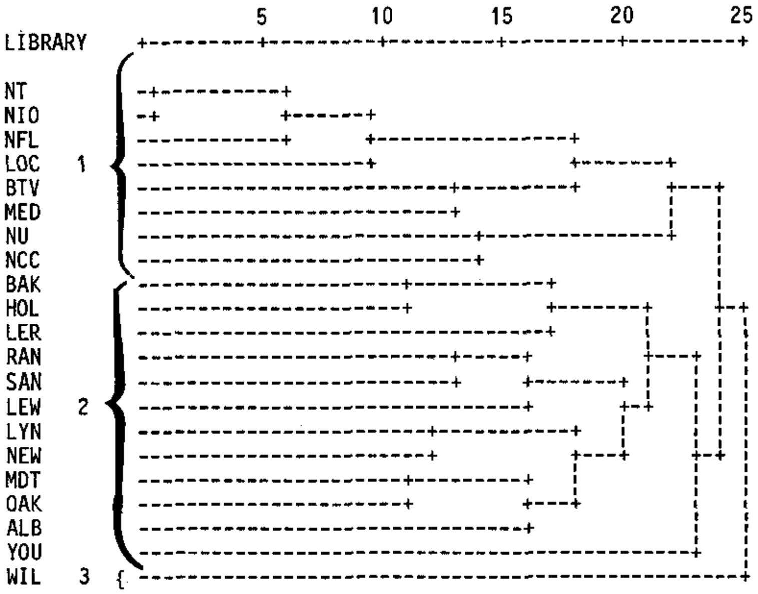

Fig.4 : Average Linkage Dendrogram (between groups) interpreted as three clusters with 8, 12 and 1 libraries each

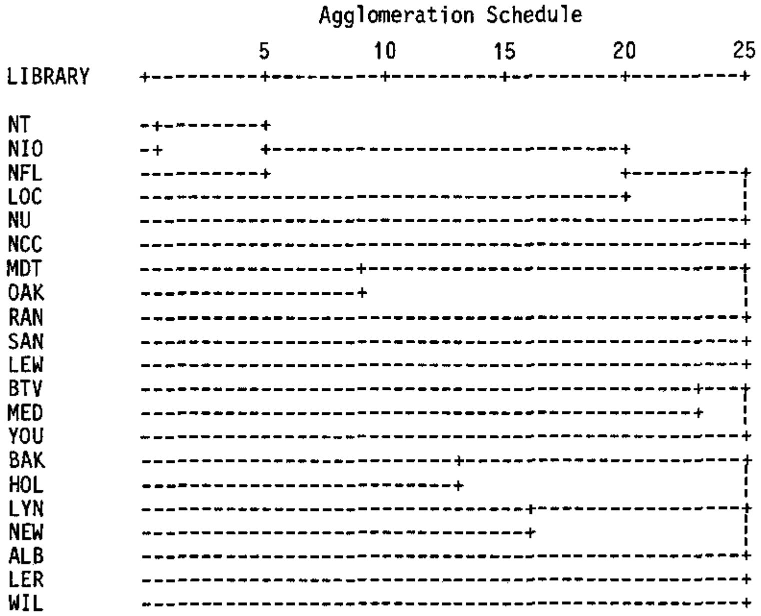

Fig.5 : Dendrogram using Ward's Method

Table 4 : Cophenetic Correlation Coefficients for the Single, Complete and Average Methods of Cluster Analysis

Single Linkage	0.55
Complete Linkage	0.69
Average Linkage	0.74

Table 5 : Clusters generated by the UPGMA method

Cluster 1	Cluster 2	Cluster 3
NT	BAK	WIL
NIO	HOL	
NFL	LER	
LOC	RAN	
BTV	SAN	
MED	LEW	
NU	LYN	
NCC	NEW	
	MDT	
	OAK	
	ALB	
	YOU	

MULTIDIMENSIONAL SCALING

The one-dimensional solution to MDS analysis (Figure 6) shows the same allignment of libraries as in the UPGMA clusters. The same clusters can readily be discerned. The two-dimensional solution was judged to be weak, since its STRESS and R-square were little improvement over one dimension (Table 6). The three- and four-dimensional solutions were uninterpretable and undoubtedly not meaningful.

These results suggest that higher dimensional interpretation can not be justified, at least with this sample, and that the one-dimensional solution is acceptable.

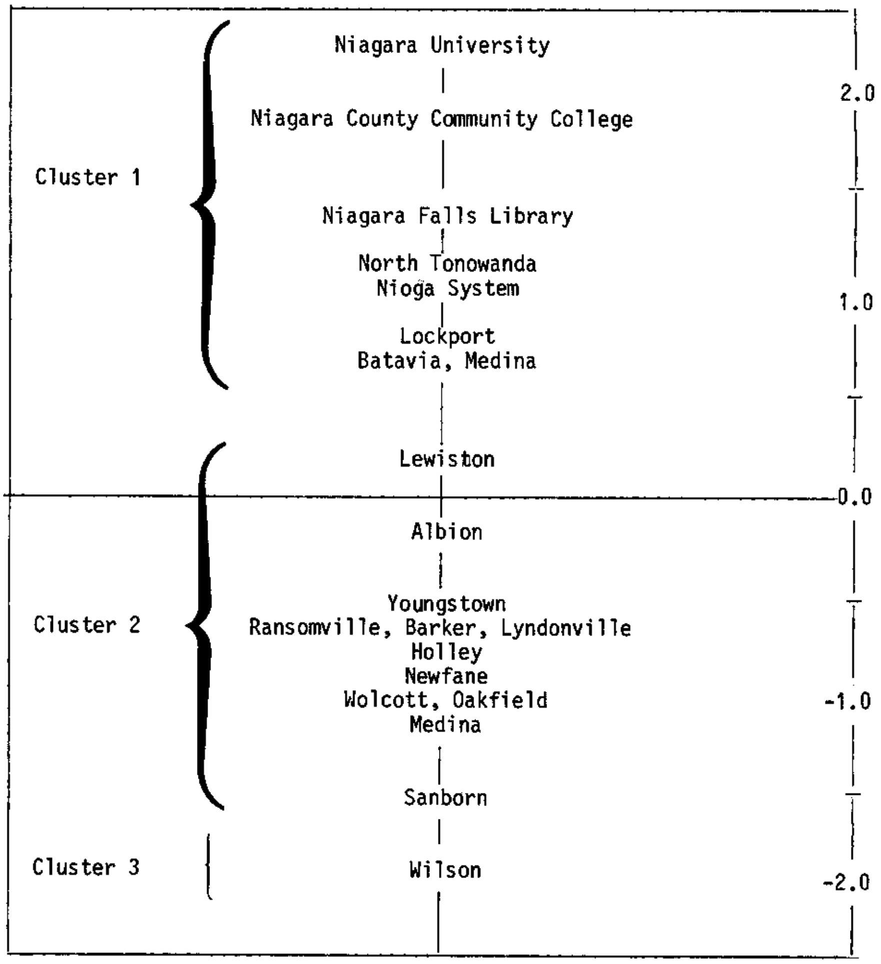

Fig.6 : One-dimensional solution to Multi-dimensional scaling analysis.

Table 6 : STRESS* and R-Square** Values for four Multidimensional Solutions

Dimension	STRESS	R-Square	R-Square Improvement
Four	0.126	0.854	0.047
Three	0.171	0.807	0.067
Two	0.235	0.740	0.161
One	0.389	0.579	0.579 best
None	---	---	

* STRESS values are Kruskal's Stress Formula 1 (Kruskal and Wish, 1978).

** R-Square values for between scaled data (disparities) and their corresponding Jaccard coefficients.

COMPARISON OF MEANS

Since the average linkage method yielded the best clusters - in terms of both interpretability and the cophenetic correlation - Clusters 1 and 2 were treated as fixed factors for the t-test and analysis of covariance. Cluster 3 was not tested since it contained only one library and lacked variance needed for the test.

Table 7 : Results of independent samples t-test for circulation and volumes on clusters 1 and 2 (UPGMA method)

	Circulation		Volumes	
	Cluster 1 (N = 8)	Cluster 2 (N = 12)	Cluster 1 (N = 8)	Cluster 2 (N = 12)
Mean*	4.9258	4.5094	4.9604	4.2764
SD	0.493	0.198	0.286	0.212
Stan.error	0.174	0.057	0.101	0.061
t-value	2.65		6.15	
df	18		18	
2-tail prob.	0.016		0.000	

* Data were log transformed

Results of the t-test suggest Clusters 1 and 2 differ significantly on both *circulation* and *volumes* (Table 7). This would suggest that clustering on the basis of similarity of collections is associated with total size and that the clusters somehow have an effect on circulation. That would be too easy a conclusion, however, since other variables and other effects might also be involved. For example, it is known that *circulation* is highly correlated with *total volumes*, so that if one wants to test on *circulation*, the effect of *total volumes* should be controlled.

This was done using analysis of covariance. Table 8 shows that the covariate, *volumes*, is a significant contributor to the variance of *circulation*. When its effects are removed, the difference between the means of Clusters 1 and 2 is not significant.

Table 8 : Analysis of Covariance on the Difference in *Circulation* in Cluster 1 and Cluster 2, with *Total Volumes* as the covariate

Source of Variance	Sum of Squares	df	MS	F	Sig. of F
Covariate (Volumes)	0.942	1	0.942	8.090	Sig.© = .05
Main Effects (Clusters)	0.040	1	0.040	0.342	not sig.
Explained	0.982	2	0.149		
Residual	1.980	17	0.116		
Total	2.962	19			

* © = 0.05

CONCLUSIONS AND SUGGESTIONS FOR FURTHER RESEARCH

Interpretation of clustering is more subjective than it should be. While three clusters are apparent, interpreting the dendrogram as more than three is harder to defend, given the uncertainties of the sample, the underlying distribution of overlap data, and the confidence of averages in the merging process.

We known that sampling procedures may affect results. Real world data are composed of complex mixtures of multivariate sampling distributions (Aldenderfer and Blashfield, 1984), while bias is always present. Systematic studies of library overlap sampling distributions are scarce (McGrath, 1988).

The distribution of coefficients in overlap matrices is likely to be highly skewed with few large values and many small values. This may be all the more so when holdings themselves are highly skewed, for example, when very large libraries are paired with very small ones. Also, if samples are too small, the matrix will contain many zero values, suggesting that some pairs of libraries have no books in common, an unlikely situation.

The literature of overlap provides little insight into the establishment of minimum sample sizes needed to avoid the problem of skewed distributions of cell entries with too many zeroes. Hence the sample sizes used in this study are not necessarily the correct minimums. McGrath (1988) has suggested that knowledge of the underlying distribution of overlap data would help to determine minimum sample sizes. Certainly, sample sizes should be larger than those required when the distribution is normal. The sample size in our study was probably too small to establish satisfactory variance in the resemblance coefficient, one reason why we were reluctant to interpret the dendrograms as having more than 2 or 3 clusters.

No method is offered for testing averages in the hierarchical merging process, nor for computing confidence intervals, so that final clusters are based on no more than ordinal relationships, i.e., a cluster is placed in the hierarchy according to the absolute value of the averaged coefficients.

Since the distribution of Jaccard coefficients is positively skewed, it contains many more low than high coefficients. Low Jaccard coefficients indicates low similarity between any two libraries. Thus, though the dendrogram may suggest that libraries NT and NIO, for example, are very similar, they are similar only in relation to other libraries. Observed in isolation, or relative to other libraries in a negatively skewed distribution, one might conclude the opposite, that they are not very similar.

A procedure for determining the number of clusters, i.e., where to cut the tree, is not available according to Everitt (1979), who identifies two important reasons why little progress has been made : the lack of a suitable null hypothesis and the complex nature of multivariate sampling distributions. These problems are also addressed by Aldenderfer and Blashfield (1984) who observe that the clusters are generally determined by subjective inspection.

It does appear that the collections in Cluster 1 may be more similar to each other than to those in Cluster 2, but the procedures used in an attempt to validate them offer only minimum support for them.

Significant differences between Clusters 1 and 2 on *circulation* and *total volumes* can not be regarded with much confidence, since both of these variables are not independent of raw overlap sample data. That is, overlap data are necessarily samples of the population of holdings, while *circulation* is also a sample, albeit a biased one, of holdings. In order to test validity of these clusters, test variables must be completely independent of them.

One possibility is a feature common to libraries in Cluster 1. With the exception of the two academic libraries, they are all members of the NIOGA Library System's automated library system or ALMS, while those in Cluster 2 are not, suggesting that such a system either contributes to or is somehow a result of overlap. It also suggests that Niagara University and Nigara County Community College could benefit from membership in the ALMS network.

In summary, further research using more careful sampling procedures should provide a better basis for inference concerning overlap, while further investigation into the underlying distribution should provide greater understanding of appropriate sample size (McGrath, 1988).

A study using the same libraries but titles randomly selected from an external source, with a specified time span might provide a useful comparison to this study. Further attempts to statistically validate the number of clusters selected using preselected multiple variables would provide additional confidence in clusters.

Lastly, the use of cluster analysis of collections or any method for practical solution to the formation of networks may be academic these days, since networks are now well-established. Nevertheless, techniques used in this study offer some insight into overlap and for the enhancement of existing network activities.

REFERENCES

Aldenderfer, Mark S. and Roger K. Blashfield, 1984, *Cluster Analysis.* Sage University Paper series on Quantitative Applications in the Social Sciences, 07-001. Beverly Hills and London : Sage Publications.

Anderberg, Michael R., 1973, *Cluster Analysis for Application.* New York and London : Academic Press.

Bookstein, Abraham, 1983, "Sampling from Card Files". Library Quarterly, 53(3) : 307-312.

Davison, Mark L., 1983, *Multidimensional Scaling.* New York : John Wiley & Sons.

Everitt, B., 1980, *Cluster Analysis.* New York : Halsted.

Everitt, B., 1973, "Unresolved Problems in Cluster Analysis". *Biometrics,* 35 : 169-181.

Farris, J.S., 1969, "On the Cophenetic Correlation Coefficient". *Systematic Zoology,* 18 : 279-285.

Knightly, John Joseph, 1973, *Cooperative Collection Development in Academic Libraries : The Relationship of Book Collections to Curricula of Cooperating Institutions.* Ph.D. diss. The University of Texas at Austin.

Kruskal, Joseph B. and Myron Wish, 1978, *Multidimensional Scaling.* Sage University Paper series on Quantitative Applications in the Social Sciences, 07-001. Beverly Hills and London : Sage Publications.

McGrath, William E., 1980, "Multidimensional Map of Library Similarities". *Proceedings of the American Society for Information Science,* 18 : 298-300.

McGrath, William E., 1983, "Multidimensional Mapping of Book Circulation in a University Library". *College & Research Libraries,* 44(2) : 103-115, March.

McGrath, William E., 1984, "Morphology and the Structure of Libraries - A Fresh Look at Descriptive Methods for Management". *Science & Technology Libraries,* 4 : 117-132.

McGrath, William E., 1985, Collection Evaluation : Theory and the Search for Structure. *Library Trends,* 22(3) : 241-266.

McGrath, William E., 1986, "Circulation Cluster - An Empirical Approach to Decentralization of Academic Libraries". *The Journal of Academic Librarianship,* 12 : 221-26.

McGrath, William E., 1988, "Parameters for Cluster Analysis of Library Overlap". In : *Informetrics 87/88; Select Proceedings of the First International Conference on Bibliometrics and Theoretical Aspects of Information Retrieval, Diepenbeek, Belgium, 25-28 August 1987.* Ed. by Leo Egghe and Ronald Rousseau, Amsterdam, etc., Elsevier Science Publishers, 1988.

McGrath, William E., Diane Geraci and A. Kimball Romney, 1988, "Matrix Models and the Search for Structure in Subject-Majors Circulation ; a Methodological Study". *Library and Information Science Research,* 1091 : 77-94, Jan.-March.

McGrath, William E. and Thomas B. Hickey, 1983, *Multi-dimensional Mapping of Libraries Based on Shared Holdings in the OCLC Online Union Catalogs. Research report prepared for OCLC.* Report number OCLC/OPR/RR-83/5. Dublin, OH : OCLC Online Computer Library Center, Inc.

O'Neill, Edward T., 1972, *A Survey of Library Resources in Western New York.* Prepared in collaboration with Mary Lynn Seanor, Buffalo, NY : Western New York Library Resources Council.

Potter, William Gray, 1982, "Studies of Collection Overlap : A Literature Review". *Library Research,* 14 : 3-21.

Rogers JoAnn V., 1984, "Networking : Selected Research Studies, 1979-1983". *Library & Information Science Research*, 6 : 111-132.

Rohlf, F.J., 1974, "Methods of Comparing Classifications". In : R.F. Johnson, P.W. Frank and C.D. Michener (Eds.), *Annual Review of Ecology and Systematics*, Vol.5. Palo Alto, CA : Annual Reviews Inc.

Romesburg, H. Charles, 1984, *Cluster Analysis for Researchers*. Belmont, CA : Lifetime Learning Publications.

Roscoe, John T., 1975, *Fundamental Research Statistics for the Behavioral Sciences*. New York : Holt, Rinehart and Winston, Inc.

Shaw, Debra, 1983, *Overlap of Monographs in Public and Academic Libraries in Indiana*". Ph.D. Diss., Indiana University.

Sneath, Peter H.A. and Robert R. Sokal, 1973, *Numerical Taxonomy : The Principles and Practice of Numerical Classification*. San Francisco : W.H. Freeman and Company.

Sokal, R.R. and F.J. Rohlf, 1962, "The Comparison of Dendrograms by Objective Methods". *Taxon*, 33 : 40.

APPENDIX

LIBRARIES IN THE STUDY

The Nioga Library System is a cooperative library network of twenty-one participating libraries from Niagara, Genesee and Orleans counties. The central headquarters maintains a union author/title catalog containing its independent holdings and those of its participating libraries. It purchases and processes materials for the members and provides reference service and in-service training. The participating libraries, however, are autonomous with regard to administration, budget, and collection development.

Within the Nioga Library System, five of the largest libraries (North Tonawanda, Niagara Falls, Batavia, Medina and Lockport) have a joint automated library management system (ALMS) which maintains a common database of holdings and an online circulation system. The central processing unit is located and maintained at the NIOGA Headquarters.

Niagara University is a private Catholic university offering undergraduate degrees and a limited number of graduate programs. Niagara County Community College is a two-year public community college.

Symbols	Libraries
NU	Niagara University, Niagara Falls, NY
NCC	Niagara County Community College, Sanborn, NY

Nioga Library System :

NT	North Tonawanda Public Library, North Tonawanda, NY
NIO	NIOGA Library System, Lockport, NY
MFL	Earl W. Brydges Public Library, Niagara Falls, NY
LOC	Lockport Public Library, Lockport, NY
BTV	Richmond Memorial Library, Batavia, NY
MED	Lee-Whedon Memorial Library, Medina, NY
BAK	Barker Free Library, Barker, NY
HOL	Library Community Free, Holley, NY
LER	Woodward Memorial Library, Wolcott, NY
RAN	Ransomville Free Library, Ransomville, NY
SAN	Sanborn-Pekin Free Library, Sanborn, NY
LEW	Lewiston Public Library, Lewiston, NY
LYN	Yates Community Library, Lyndonville, NY
NEW	Newfane Free Library, Newfane, NY
MDT	Middleport Free Library, Middleport, NY
OAK	Haxton Memorial Library, Oakfield, NY
ALB	Swan Library, Albion, NY
YOU	Library of Youngstown, Youngstown, NY
WIL	Wilson Free Library, Wilson, NY