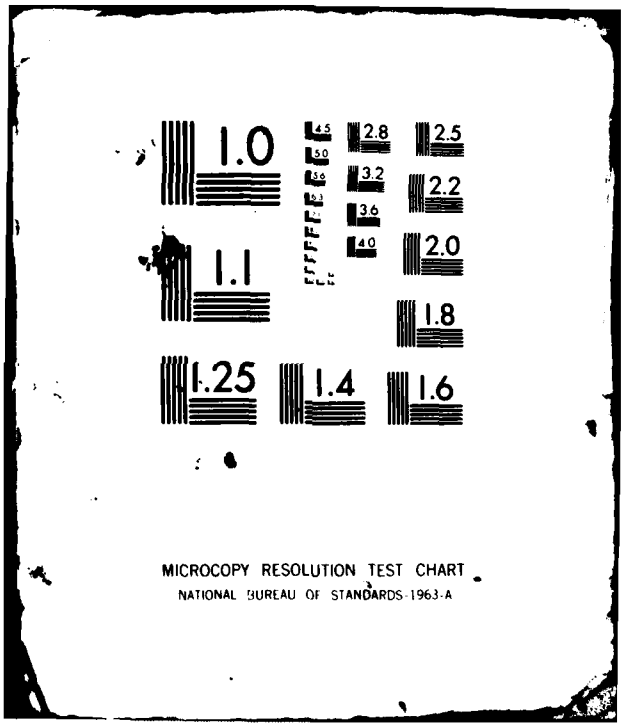
AD-A112 930 NAVY PERSONNEL RESEARCH AND DEVELOPMENT CENTER SAN D--ETC F/G 5/9
EMPIRICAL COMPARISON OF BINARY AND CONTINUOUS PROXIMITY MEASURE--ETC(U)
MAR 82 J J PASS, R E CHATFIELD

UNCLASSIFIED NPRDC-TR-82-36 NL

END
DATE FILMED
4 82
DTIC


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

NPRDC TR 82-36 March 1982

EMPIRICAL COMPARISON OF BINARY AND CONTINUOUS PROXIMITY MEASURES FOR CLUSTERING OCCUPATIONAL TASK DATA

John J. Pass
Robert E. Chatfield

Reviewed by
Martin F. Wiskoff

Released by
James F. Kelly, Jr.
Commanding Officer

Navy Personnel Research and Development Center
San Diego, California 92152

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER NPRDC TR 82-36	2. GOVT ACCESSION NO. AD-A112 930	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) EMPIRICAL COMPARISON OF BINARY AND CONTINUOUS PROXIMITY MEASURES FOR CLUSTERING OCCUPATIONAL TASK DATA		5. TYPE OF REPORT & PERIOD COVERED Final Report 1 Oct 1980-1 Aug 1981
		6. PERFORMING ORG. REPORT NUMBER 12-82-3
7. AUTHOR(s) John J. Pass Robert E. Chatfield		8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS Navy Personnel Research and Development Center San Diego, California 92152		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS Methods for Clustering Tasks ZF000-01-042-04.01.05
11. CONTROLLING OFFICE NAME AND ADDRESS Navy Personnel Research and Development Center San Diego, California 92152		12. REPORT DATE March 1982
		13. NUMBER OF PAGES 23
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Cluster analysis
Comprehensive occupational data analysis programs (CODAP)
Occupational analysis
Proximity measures

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Thirteen binary and three continuous proximity measures were used to cluster-analyze job incumbent profiles of task inventory data. The results were compared (1) to recommend a binary measure for programming into CODAP System 80, a software package used extensively by the military and many other organizations, and (2) to determine to what extent binary measures can produce cluster solutions similar to solutions based on continuous measures. Sixteen 250-by-250 proximity matrices were derived from each of three Navy occupational samples, and the clustering procedure in CODAP was applied to selected matrices. Proximity matrix and cluster solution

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

comparison revealed that (1) there was high variability among binary measures, (2) the Jaccard and Dice measures were the most powerful binary measures, and (3) there was high similarity between the Jaccard and distance measures. The implications of the findings are discussed with reference to the proportion of zero scores in task inventory data. The Jaccard measure is recommended for clustering binary data for tasks and for programming into CODAP System 80.

FOREWORD

The purpose of this research, which was conducted in support of NAVPERSRANDCEN independent laboratory research project ZR000-01-042-04.01.05 (Methods for Clustering Tasks), was to evaluate the use of various proximity measures for the cluster analysis of occupational data. The occupational analysis and design programs of the military services and many other organizations routinely apply cluster analysis techniques to occupational data to develop more effective personnel and training systems.

One decision that can impact on the cluster analysis solution is the selection of a proximity measure. This research empirically evaluated proximity measures for the cluster analysis of occupational task inventory data. The results will be used to select a binary proximity measure to program into the Comprehensive Occupational Data Analysis Programs (CODAP System 80), which are currently being developed by the Department of Defense. The results are further intended for use by federal job analysts in military and civilian agencies.

JAMES F. KELLY, JR.
Commanding Officer

JAMES J. REGAN
Technical Director

SUMMARY

Problem

Cluster analysis techniques are used by the military services to define types of work for several personnel functions. The data sets analyzed are typically job analysts' or job incumbents' scale responses to items within a structured task inventory. An important decision, one that can affect the cluster analysis solution, is the selection of a proximity measure to evaluate similarities among profiles of response scores.

The military services use the clustering procedures available in the Comprehensive Occupational Data Analysis Programs (CODAP). The selection of a binary measure for programming into CODAP System 80, an enhanced IBM version being developed by the Department of Defense, requires an evaluation of proximity measures with the capacity to cluster-analyze occupational task inventory data. While recent research has indicated that binary measures may be able to capture as much profile information as continuous measures, there has been no empirical comparison of cluster solutions produced by the application of various proximity measures to occupational data.

Purpose

The purpose of this effort was to evaluate proximity measures for the CODAP cluster analysis of task inventory data. Specifically, the research was conducted to (1) determine to what extent binary measures can produce cluster solutions of task inventory data similar to solutions based on continuous measures, and (2) recommend a binary proximity measure for programming into CODAP System 80, based on the capability of various binary measures to capture information for cluster analysis.

Approach

Data for analysis, collected by the Navy Occupational Development and Analysis Center (NODAC), consisted of three samples of job incumbent profiles. Each sample was comprised of 250 profiles indicating time spent on various job tasks, based on a 5-point scale. Sixteen proximity matrices were derived for each sample, with each matrix based on the application of one of three continuous or thirteen binary proximity measures. The CODAP clustering procedure, an average linkage procedure, was applied to seven selected proximity matrices from each of the three samples. The evaluation of the binary measures was based on the extent to which the binary matrices and cluster solutions were objectively similar to those based on the continuous measures.

Results

Proximity matrix comparisons revealed that (1) six binary measures obtained high correlations with the continuous measures while seven others yielded low or very low correlations with the continuous measures, (2) the CODAP binary measure (the tau measure) captured less information than did five other binary measures, (3) high variability existed among the binary measures, and (4) the Jaccard binary measure captured more distance measure information than did the Pearson correlation, which is a continuous measure.

Cluster solution comparisons revealed that the relative magnitude of the relationships between measures matched the results obtained for matrix comparisons. However, some differences appeared when the absolute magnitudes of the relationships derived from cluster solutions were compared to those derived from matrix comparisons.

Conclusions

1. The Jaccard and Dice proximity measures are consistently powerful measures, capable of capturing more profile information than many other binary measures.

2. The performance of a proximity measure in cluster analysis can be strongly affected by the proportion of zeros in the data analyzed.

3. The use of selected binary proximity measures will yield cluster solutions highly similar to cluster solutions based on continuous proximity measures.

4. Because the high proportion of zeros in the incumbent profiles analyzed is typical for this type of data set, the findings appear generalizable to data collected from other occupational task inventories.

Recommendations

1. The Jaccard measure should be used to cluster binary data collected from occupational task inventories and should be programmed into CODAP System 80 as the binary proximity measure.

2. Research should be conducted to develop a conceptual model for the interaction between data set response distributions and proximity measure performance.

CONTENTS

LIST OF TABLES

INTRODUCTION

Background

In occupational psychology, cluster analysis methods are used to define types of work by grouping together (1) jobs that are similar on some profile of work-related variables (e.g., McCormick, DeNisi, & Shaw, 1977; Pass & Cunningham, 1978; Dulewicz & Keenay, 1979), or (2) job incumbents with a similar profile of work requirements (e.g., Archer, 1966; Christal & Ward, 1967). In such studies, cluster analysis is typically applied to job analysts' or job incumbents' Likert-type scale ratings for items in a structured work questionnaire or task inventory. By averaging scale rating data within a derived cluster, profiles of tasks, attributes, or other work-related requirements can be derived to define the job cluster. Such cluster profiles are useful for aligning training with actual work performed (DeNisi, 1976) and for streamlining occupational classification systems by combining administratively separate jobs into one job type.

Hundreds of analyses to derive clusters from data on work have already been performed, and the number of analyses will probably continue to increase for at least two reasons. First, there is a growing need for such job analysis methods as part of efforts to validate or develop personnel tests, procedures, and policies that comply with federal employment guidelines (U.S. Equal Employment Opportunity Commission, 1978). For example, researchers concerned with demonstrating synthetic validity have used clustering techniques to derive a family of jobs on which to validate common predictors (McCormick et al., 1977). Second, a powerful computerized data analysis system is being installed at an increasing rate by government agencies around the world. The comprehensive occupational data analysis programs (CODAP), a computerized data analysis and report system originally developed by the U.S. Air Force, is capable of cluster-analyzing from 2000 (for the current IBM version) to 7000 (for the UNIVAC version) job incumbent profiles. Because the CODAP System is used extensively by Department of Defense (DoD) agencies (e.g., the occupational analysis programs of the Navy, Air Force, Army, and Marine Corps), the CODAP clustering procedure is the focus of the present research.

Cluster analysis studies conducted in DoD agencies typically derive job types from occupational task inventory responses. In these studies, job incumbents are clustered on the basis of similar profiles of task requirements. In effect, each job type is a group of positions rated or analyzed by their incumbents. The CODAP hierarchical clustering procedure used in DoD studies and in the present research is based on work by Ward (1961), but it is not the well known minimum variance procedure frequently referred to in the literature (e.g., Borgen & Weiss, 1971; Blashfield, 1976, 1980). Instead, the procedure is an average linkage procedure; that is, the value of the proximity measure determining the clustering is equal to the average of the proximity values for each member of one cluster paired with every member of the other cluster (Archer, 1966). This is an important distinction, because differences have been demonstrated (e.g., Blashfield, 1976) among the properties of solutions based on the average linkage and the minimum variance methods. There are two proximity measures available for clustering in the CODAP package--a distance measure, the time option (hereafter referred to as the overlap between measure) and a binary measure, the task option (hereafter referred to as the task measure).

Problem

In light of recent findings documenting the information-providing potential of task inventory data (Pass & Robertson, 1980), and after examination of the CODAP task formula, it appears that other binary proximity measures might be able to capture more

profile information for clustering than does the current task measure. Several researchers have defined or proposed various measures of profile proximity for a number of general pruposes (e.g., Cronbach & Gleser, 1953; Nunnally, 1967; Cheetham & Hazel, 1969), but few studies (e.g., Hamer & Cunningham, in press) have compared cluster solutions based on different proximity measures for specific data. While the capacity of a proximity measure to reflect more information would enable the derivation of a more valid cluster solution, there has been no research that empirically compares binary measures to continuous measures for the clustering of occupational task data. Such a comparison is required to establish a recommendation of a binary proximity measure for programming into CODAP System 80, an enhanced IBM verison of CODAP currently being developed by DoD.

Purpose

The purpose of this effort was to evaluate proximity measures for the CODAP cluster analysis of task inventory data. Specifically, the research was conducted to (1) determine to what extent binary measures can produce cluster solutions of task inventory data similar to solutions based on continuous measures, and (2) recommend a binary proximity measure for programming into CODAP System 80, based on the capability of various binary measures to capture information for cluster analysis.

APPROACH

Data

The Navy Occupational Development and Analysis Center (NODAC) collected data for analysis from job incumbents in three Navy occupations: the aviation machinist's mate (AD) rating (N = 2538), the electronics technician (ET) rating (N = 2546), and the yeoman (YN) rating (N = 2771). A subsample of 250 was drawn from each total sample by means of a systematic random sampling procedure described by Kish (1965).[1] The subsamples contained job incumbents from eight different pay grades (skill levels).

The data consisted of incumbent profiles of responses to job tasks. There were [illegible] tasks in the inventory for AD, 597 for ET, and 529 for YN. Job incumbents were instructed to estimate the time spent performing each task in an inventory by selecting the appropriate score on the following scale:

Score	Time Spent
1	Very little
2	
3	Average
4	
5	Very much

Instructions were to leave the response item blank for any task that was not performed. Blanks were treated as zeros in subsequent calculations.

---

[1]The YN sample was reduced to 249 because one response profile contained incomplete data.

Table 1 presents the distributions of response scores in percentages for the three samples. The response scores were proportionalized within each incumbent's profile; that is, each score was divided by the sum of the incumbent's scores, thereby yielding a value for each task that summed to 100 percent (1.0 for proportions) for all tasks performed by each incumbent. This standardization, which is automatically calculated by the current version of CODAP, is intended to remove possible sources of rater bias, including mean scale differences among raters.

Table 1

Response Score Percentage Distributions

Sample[a]	Response Score					
	0	1	2	3	4	5
Aviation Machinist's Mate (AD)	86.5	2.1	2.5	6.4	1.5	1.0
Electronics Technician (ET)	86.9	2.3	2.1	5.0	1.5	2.2
Yeoman (YN)	82.8	2.7	3.4	7.5	2.0	1.6

[a] $\underline{N}$ = 250 for samples AD and ET and 249 for sample YN.

Proximity Measures

The three continuous proximity measures analyzed used standardized response values or proportions within profiles, while the binary proximity measures used only zero and one values, the latter representing all nonzero proportions. The continuous measures analyzed were distance, overlap-between, and the Pearson correlation coefficient (see appendix for formulas).

When applied to the proportionalized data, the distance measure has been symbolized by Cronbach and Gleser as D'. D' does not measure profile level (or mean) information, only profile shape and scatter information (Cronbach & Gleser, 1953). The overlap between measure is the only continuous measure available in CODAP. (Additional programming was required to include the values of the other proximity measures in the CODAP clustering algorithm.) This measure is defined as the sum of the proportionalized minimum values for corresponding tasks across the two profiles being compared. Applied to proportions, this index uses profile shape and scatter information. Except where all response values are zero, the overlap between measure is a linear transformation of D' and thus has also been considered a distance measure for this study. Unlike the other two continuous measures, the magnitude of the Pearson correlation coefficient reflects only profile shape information and is affected by pairs of zero scores for corresponding tasks across profiles. These paired zero scores will not add to the magnitude of the other two continuous measures.

The 13 binary measures analyzed were:

1. Jaccard
2. Dice
3. Task (also known as second Kulczynski)
4. Otsuka
5. Correlation ratio
6. Phi*
7. Simple matching*
8. Rogers and Tanimoto*
9. Hamaan*
10. Sokal distance*
11. Number of features of difference
12. First Kulczynski
13. Yule*

(Asterisks indicate measures for which magnitudes will be affected by zero scores on corresponding tasks for the profiles being compared.) The formulas[2] for all of these binary measures have been included in the appendix.

Evaluation of Proximity Measures

Continuous proximity measures are, in general, capable of reflecting more information about profile data than are nominal measures, which, in the binary case, only indicate the presence or absence of some profile variable. For this reason, the criterion employed to evaluate the 13 binary measures was the extent to which they could capture the same information contained in the values of the continuous measures. If the data analyzed contained only the zero and one category of nonzero response scores, a binary measure could pick up as much information as a continuous measure. However, examination of the task response distributions indicates that this is not the type of data set analyzed here. As Table 1 indicates, nonzero responses are distributed throughout the entire 5-point scale in the three samples. Because certain continuous measures, such as the distance measures, can use more profile information than can measures such as the Pearson correlation coefficient, they may be considered appropriate criterion measures against which to judge binary measures, as well as the Pearson measure itself. The amount of information captured by the proximity measures was calculated by analyzing similarity among proximity matrices and making cluster solution comparisons.

Proximity Matrix Comparisons

For each sample, 16 250-by-250 proximity matrices were derived, each based on the application of one proximity measure. A Pearson correlation coefficient was calculated between each possible pair of matrices, calculated on nondiagonal corresponding matrix cell values. Binary proximity measures were evaluated in terms of their capability to capture as much information as the continuous measures.

Cluster Solution Comparisons

The CODAP hierarchical clustering algorithm was applied to seven selected proximity measure matrices for each of the three samples. The selection of the seven measures was based on the researchers' interest in specific measures and on the decision to compare measures that were markedly different in terms of the matrix comparison results. An iterative procedure, the CODAP clustering method, first clusters the two most similar profiles and then groups the next most similar profiles or clusters at

---

[2]Cheetham and Hazel (1969) have done the tedious job of defining numerous binary proximity measures (including these analyzed in this study) and describing the general properties of the indices.

subsequent stages (iterations) until all profiles are contained in one cluster (Archer, 1966). The similarity among the seven resultant solutions for each sample was determined by calculating the Pearson correlation coefficient on the iteration number where the same profiles were first clustered in any two of the seven hierarchical solutions being compared.

RESULTS

Comparison Among Proximity Matrices

Table 2 displays the correlations between each binary measure matrix and each of the three continuous measure matrices. The results are highly consistent across the three samples analyzed. As expected, the distance and the overlap between measures were perfectly correlated, as demonstrated by the identical correlation coefficients obtained for each sample. Six binary measures (numbers 1 through 6 in Table 2) consistently obtained high or very high correlations with the continuous measures. For each of the three samples, the CODAP task measure captured less of the distance information than did five other binary measures.

High variability among binary measures may also be seen in Table 2. In fact, correlation coefficients ranged from 1.00 to about zero (disregarding sign). Perfect correlations were obtained among matrices derived from four out of the six measures whose calculation is affected by zero scores on corresponding tasks (numbers 7 through 10 in Table 2). These measures tended to capture almost none of the continuous measure variance. Of considerable importance, the Jaccard and Dice binary measures each accounted for about 95 percent of the variance for distance and overlap-between measures. Because these two binary measures were nearly perfectly correlated and because the Dice formula is slightly more complex, it was decided that only the Jaccard measure would be further evaluated.

The stability of the findings in Table 2 across the three samples was determined both by intercorrelating the coefficients in the overlap-between columns across samples and by intercorrelating the coefficients in the Pearson r columns across samples. The results, presented in Table 3, document the high stability of the findings.

Proximity matrix comparisons revealed two additional interesting findings, which are displayed in Table 4. First, in each sample, the relationship between the Jaccard binary measure and the distance measure is higher than the relationship between the distance measure and the Pearson correlation coefficient (a continuous measure). Second, the Pearson correlation coefficient appears to be more variable with respect to the sample analyzed than does the Jaccard measure; that is, when applied to one sample, the Pearson correlation captures more distance measure information than when it is applied to another sample. The reason for both of these findings appears to be the large but different mean number of zeros in the profile for each of the three samples; that is, as the mean number of zeros increases, the information-capturing capability of the Pearson correlation coefficient decreases.

Comparison Among Cluster Solutions

The correlation of the iteration numbers between solutions for the same sample revealed that the relative magnitude of relationships between measures matched the results obtained for matrix comparisons. However, the absolute magnitude of the correlations obtained from cluster solutions differed from those obtained by matrix

Table 2

Correlation of Binary Measures with Continuous Measures by Sample

	Binary Measure	Continuous Measures								
		ET (N = 250)			AD (N = 250)			YN (N = 249)		
		Distance	Overlap-between	Pearson r	Distance	Overlap-between	Pearson r	Distance	Overlap-between	Pearson r
1.	Jaccard	-.968	.968	.810	-.979	.979	.893	-.975	.975	.848
2.	Dice	-.969	.969	.815	-.978	.978	.892	-.974	.974	.850
3.	Task	-.720	.720	.790	-.864	.864	.901	-.795	.795	.843
4.	Otsuka	-.916	.916	.843	-.952	.952	.913	-.937	.937	.879
5.	Correlation ratio	-.903	.903	.812	-.933	.933	.892	-.927	.927	.853
6.	Phi[a]	-.823	.823	.931	-.897	.897	.960	-.855	.855	.922
7.	Simple matching[a]	-.060	.060	.199	.051	-.051	.100	.114	-.114	-.002
8.	Rogers and Tanimoto[a]	-.060	.060	.199	.051	-.051	.100	.115	-.115	-.002
9.	Hamaan[a]	-.060	.060	.199	.051	-.051	.100	.114	-.114	-.002
10.	Sokal distance[a]	.059	-.059	-.199	-.051	.051	-.100	-.114	.114	.002
11.	Number of features of difference	.060	-.060	-.199	-.051	.051	-.100	-.114	.114	.002
12.	First Kulczynski	-.309	.309	.181	-.375	.375	.261	-.340	.340	.217
13.	Yule[a]	-.459	.459	.709	-.701	.701	.793	-.517	.517	.677

[a]In these measures, magnitudes are affected by pairs of zero scores on corresponding tasks for profiles being compared.

Table 3

Stability of Relationships Between Binary Proximity Measures and Continuous Proximity Measures

Binary Measures Compared by Sample	Intercorrelations of Continuous Measures	
	Overlap-between	Pearson Correlation Coefficient
AD vs. ET	.974	.977
AD vs. YN	.988	.987
ET vs. YN	.975	.950

Table 4

Proximity Matrix Comparisons of the Pearson Correlation Coefficient and Jaccard Measures with the Distance Measure

Measures Correlated with Distance	Correlations by Sample		
	ET	AD	YN
Pearson correlation coefficient	-.830	-.903	-.871
Jaccard	-.968	-.979	-.975
$\bar{X}$ number of zeros	502	352	460

comparison, especially for the simple matching measure. For example, the Jaccard binary measure correlated at about .90 with the distance measure, but the simple matching measure that obtained near-zero correlations for the matrix comparisons obtained substantial negative correlations for the cluster solution comparisons (see Table 5). Also notable are the expected identical (or, due to rounding error, nearly identical) solutions obtained for the distance and overlap-between measures. The apparently high stability of these findings across the three samples was confirmed by correlating, between samples, the matrices of coefficients presented in Table 5. The obtained rs were: AD versus ET, .983; AD versus YN, .978; and ET versus YN, .992.

Table 5

Correlation of Profile Stage (Iteration) Numbers Between Cluster Solutions for Selected Proximity Measures

Measure	Continuous Measures			Binary Measures			
	Distance[a]	Pearson	Overlap-between	Jaccard	Task	Yule	Simple Matching
ET Sample							
Distance	---	.789	.980	.885	.831	.332	-.322
Pearson		---	.786	.679	.670	.427	.043
Overlap-between			---	.870	.815	.322	-.296
Jaccard				---	.909	.336	-.368
Task					---	.400	-.344
Yule						---	.382
Simple matching							---
AD Sample							
Distance	---	.735	1.00	.906	.762	.411	-.110
Pearson		---	.735	.661	.584	.391	.130
Overlap-between			---	.906	.762	.411	-.110
Jaccard				---	.805	.364	-.215
Task					---	.524	.056
Yule						---	.434
Simple matching							---
YN Sample							
Distance	---	.820	.999	.914	.843	.124	-.545
Pearson		---	.834	.748	.705	.315	-.226
Overlap-between			---	.909	.838	.129	-.538
Jaccard				---	.901	.172	-.568
Task					---	.270	-.517
Yule						---	.464
Simple matching							---

[a]Distance values were inverted before clustering.

DISCUSSION

It is important to comment on the appropriateness of using various binary proximity measures for the cluster analysis of occupational task inventory data. The results clearly demonstrated that the Jaccard and Dice binary measures were capable of capturing more information than were other binary measures, including the CODAP task measure. However, since many other binary measures were not capable of capturing much profile information, they are inappropriate for use as proximity measures in clustering such occupational data.

The high similarity between solutions based on certain continuous measures and solutions based on certain binary measures has possible application as well as theoretical interest. This similarity is apparently accounted for by the large proportions of zeros in the data sets analyzed. The data sets consisted of about 85 percent zeros and 15 percent nonzeros (see Table 1). Thus, the unique variance capturable by the continuous measures resided in the 15 percent of nonzero scores. Given this type of data set to be analyzed, there will be little loss of continuous information if a binary measure such as the Jaccard measure is used. If there is reason to believe that data collected on a binary scale have higher validity than those collected on a continuous scale (as demonstrated by Hartley, Brecht, Pagerey, Weeks, Chapanis, & Hoecker, 1977), then a decision to cluster binary data will not risk any substantial loss of information and might increase the validity of the obtained solution. Further analysis could produce a utility curve displaying the binary versus continuous information differential as the proportion of zero data points varies.

An alternative approach to analyzing such data sets would be to delete from the calculation of proximity values any pairs of zero scores on corresponding variables for any two profiles being compared (as in pair-wise deletion). This procedure would change the proportion of zeros in the data, with subsequent effect on the performance of different proximity measures. This alternative approach needs to be examined for its effect on cluster solution validity.

Further evidence for the interaction between the type of data set and proximity measure performance was found in the relatively poor performance of the measures whose magnitude was affected by pairs of zero scores on common profile tasks (see results for such binary measures in Table 2 and for the Pearson correlation coefficient in Table 4). Measures such as Jaccard, Dice, and distance were not so affected, and the relationships were predictably stronger. The fact that the distance measures in the study use more profile information than does the Pearson correlation coefficient has justified their use as criteria against which to judge the binary measures.

The present findings and the additional research suggested above would not be valuable if the data sets analyzed were dissimilar from most other data sets of task inventory information. To the contrary, the proportion of zeros is usually very high when task inventory data are collected, simply because an incumbent only does (or only responds to) a small proportion of a large number of inventory tasks (e.g., usually more than 400 in Navy task inventories). Thus, it is reasonable to apply the findings to analysis of task inventory data in general.

It is also useful to comment on the selection of comparative analyses for evaluating proximity measures. The relationships among measures, as demonstrated in absolute values of correlation coefficients, differ substantially from cluster solution comparisons to proximity matrix comparisons. Final evaluation should be based on the cluster solutions.

CONCLUSIONS

1. The Jaccard and Dice proximity measures are consistently powerful measures, capable of capturing more profile information than are many other binary measures.

2. The performance of a proximity measure in cluster analysis can be strongly affected by the proportion of zeros in the data analyzed.

3. The use of selected binary proximity measures will yield cluster solutions highly similar to cluster solutions based on continuous proximity measures.

4. Because the high proportion of zeros in the incumbent profiles analyzed is typical for this type of data set, it appears that the findings are generalizable to data collected from other occupational task inventories.

RECOMMENDATIONS

1. The Jaccard measure should be used to cluster binary data collected from occupational task inventories and should be programmed into CODAP System 80 as the binary proximity measure.

2. Research should be conducted to develop a conceptual model for the interaction between data set response distributions and proximity measure performance.

REFERENCES

Archer, W. B. Computation of group job descriptions from occupational survey data (PRL-TR-66-12). Lackland Air Force Base, TX: Personnel Research Laboratory, Aerospace Medical Division, December 1966.

Blashfield, R. K. Mixture model tests of cluster analysis: Accuracy of four agglomerative hierarchical methods. Psychological Bulletin, 1976, 83(3), 377-388.

Blashfield, R. K. The growth of cluster analysis: Tryon, Ward, and Johnson. Multivariate Behavioral Research, 1980, 15, 439-458.

Borgen, F. H., & Weiss, D. J. Cluster analysis and counseling research. Journal of Counseling Psychology, 1971, 18, 583-591.

Cheetham, A. H., & Hazel, J. E. Binary (presence-absence) similarity coefficients. Journal of Paleontology, 1969, 43(5), 1130-1136.

Christal, R. E., & Ward, J. H., Jr. The MAXOF clustering model. In M. Lorr and S. B. Lyerly, Conference on cluster analysis of multivariate data (Final Report, Office of Naval Research, Contract No. N00014-67-C-0175). Washington, DC: Catholic University of America, June 1967.

Cronbach, L. J., & Gleser, G. C. Assessing similarity between profiles. Psychological Bulletin, 1953, 50(6), 456-473.

DeNisi, A. S. The implications of job clustering for training programs. Journal of Occupational Psychology, 1976, 49, 105-113.

Dulewicz, S. V., & Keenay, G. A. A practically oriented and objective method for classifying and assigning senior jobs. Journal of Occupational Psychology, 1979, 52, 155-166.

Hamer, R. M., & Cunningham, J. W. Cluster analyzing data contaminated with individual rating tendencies: A comparison of profile association measures. Applied Psychological Measurement, in press.

Hartley, C., Brecht, M., Pagerey, P., Weeks, G., Chapanis, A., & Hoecker, D. Subjective time estimates of work tasks by office workers. Journal of Occupational Psychology, 1977, 50, 23-56.

Kish, L. Survey sampling. New York: John Wiley and Sons, 1965.

McCormick, E. J., DeNisi, A. S., & Shaw, J. B. The use of the position analysis questionnaire (PAQ) for establishing the job component validity of tests (Tech. Rep. No. 5). West Lafayette, IN: Purdue University, Department of Psychological Sciences, June 1977.

Nunnally, J. C. Psychometric theory. New York: McGraw-Hill, 1967.

Pass, J. J., & Cunningham, J. W. Occupational clusters based on systematically derived work dimensions: Final report. JSAS Catalog of Selected Documents in Psychology, 1978, 8, 192. (Ms. No. 1661).

Pass, J. J., & Robertson, D. W. Methods to evaluate scales and sample size for stable task inventory information (NPRDC Tech. Rep. 80-28). San Diego: Navy Personnel Research and Development Center, May 1980. (AD-A085 600)

U.S. Equal Employment Opportunity Commission. Uniform guidelines on employee selection procedures. Washington, DC: Federal Register, 1978, 43(166), 38295-38309.

Ward, J. H., Jr. Hierarchical grouping to maximize payoff (Tech. Note WADD-TN-61-29). San Antonio, TX: U.S. Air Force Laboratory, Wright Air Development Division, Air Research and Development Command, March 1961.

APPENDIX

PROXIMITY MEASURE FORMULAS

Table A-1

Continuous Proximity Measure Formulas

Distance: $\sum_{k=1}^{n}\left[P_{ik} - P_{jk}\right]$ Overlap-between: $\sum_{k=1}^{n} \min\left(P_{ik} : P_{jk}\right)$

Pearson Correlation Coefficient:

$$\frac{n\sum_{k=1}^{n} P_{ik}\,P_{jk} - \left(\sum_{k=1}^{n} P_{ik}\right)\left(\sum_{k=1}^{n} P_{jk}\right)}{\sqrt{n\sum_{k=1}^{n}\left(P_{ik}\right)^2 - \left(\sum_{k=1}^{n} P_{ik}\right)^2}\sqrt{n\sum_{k=1}^{n}\left(P_{jk}\right)^2 - \sum_{k=1}^{n}\left(P_{jk}\right)}}$$

Where:

- i: Any one profile
- j: Any other profile
- k: Any one of n tasks in profile
- n: N of tasks in profile
- P: Any one of k task scores in profile

Table A-2

Binary Proximity Measure Formulas

Correlation Ratio:	$\frac{C_{ij}^2}{T_i T_j}$	Otsuka:	$\frac{C_{ij}}{\sqrt{T_i T_j}}$
Dice:	$\frac{2C_{ij}}{2C_{ij} + N_i + N_j}$	Phi:	$\frac{C_{ij}A_{ij} - N_i N_j}{\sqrt{T_i T_j \; N_i + A_{ij} \; N_j + A_{ij}}}$
First Kulczynski:	$\left[\frac{C_{ij}}{N_i - N_j}\right]$	Rogers and Tanimoto:	$\frac{C_{ij} + A_{ij}}{G + N_i + N_j}$
Hamaan:	$\frac{C_{ij} + A_{ij} - N_i + N_j}{G}$	Simple Matching:	$\frac{C_{ij} + A_{ij}}{G}$
Jaccard:	$\frac{C_{ij}}{C_{ij} + N_i + N_j}$	Sokal Distance:	$\sqrt{1 - \frac{C_{ij} + A_{ij}}{G}}$
Number of Features of Difference:	$N_i + N_j$	Task:	$\frac{C_{ij}}{2T_i} + \frac{C_{ij}}{2T_j}$
		Yule:	$\frac{C_{ij}A_{ij} - N_i N_j}{C_{ij}A_{ij} + N_i N_j}$

Where:

- i: Any one profile
- j: Any other profile
- A_{ij}: N of common zero-scored tasks in profile$_i$ and profile$_j$
- C_{ij}: N of common nonzero-scored tasks in profile$_i$ and profile$_j$
- G: Grand total of nonzero-scored tasks in all profiles anlayzed
- N_i: N of nonzero-scored tasks present in profile$_i$ and absent in profile$_j$
- N_j: N of nonzero-scored tasks present in profile$_j$ and absent in profile$_i$
- T_i: Total N of nonzero-scored tasks in profile$_i$
- T_j: Total N of nonzero-scored tasks in profile$_j$

Note. All formulas are presented in Cheetham and Hazel (1969). The formula for First Kulczynski has been modified by application of the absolute function.

DISTRIBUTION LIST

Director of Manpower Analysis (ODASN(M))
Chief of Naval Operations (OP-01), (OP-11), (OP-12) (2), (OP-13), (OP-14), (OP-15), (OP-115) (2), (OP-140F2), (OP-987H)
Chief of Naval Material (NMAT 0722), (NMAT 08L)
Deputy Chief of Naval Material (Technology)
Chief of Naval Researcch (Code 200), (Code 440) (3), (Code 442), (Code 448)
Chief of Information (OI-213)
Chief of Naval Education and Training (02), (N-5)
Chief of Naval Technical Training (016)
Commandant of the Marine Coprs (MPI-20)
Commander Naval Data Automation Command (Library)
Commander Fleet Training Group, Pearl Harbor
Commander Naval Military Personnel Command (NMPC-013C)
Commanding Officer, Naval Aerospace Medical Institute (Library Code 12) (2)
Commanding Officer, Naval Education and Training Program Development Center (Technical Library) (2)
Commanding Officer, Naval Education and Training Support Center, Pacific
Commanding Officer, Naval Health Research Center
Commanding Officer, Naval Health Sciences Education and Training Command
Commanding Officer, Naval Regional Medical Center, Portsmouth, VA (ATTN: Medical Library)
Commanding Officer, Naval Training Equipment Center (Technical Library)
Commanding Officer, Office of Naval Research Branch Office, Chicago (Coordinator for Psychological Sciences)
Director, Naval Civilian Personnel Command
Director, Training Analysis and Evaluation Group (TAEG)
Officer in Charge, Naval Occupational Development and Analysis Center
President, Naval War College (Code E114)
Superintendent, Naval Postgraduate School
Commander, Army Researcch Institute for the Behavioral and Social Sciences, Alexandria (PERI-ASL)
Director, U.S. Army TRADOC Systems Analysis Activity, White Sands Missile Range (Library)
Chief, Army Research Institute Field Unit--USAREUR (Library)
Chief, Army Research Institute Field Unit, Fort Harrison
Commander, Air Force Human Resources Laboratory, Brooks Air Froce Base (Scientific and Technical Information Office)
Commander, Air Force Human Resources Laboratory, Lowry Air Force Base (Technical Training Branch)
Commander, Air Force Human Resources Laboratory, Williams Air Force Base (AFHRL/OT)
Commander, Air Force Human Resources Laboratory, Wright-Patterson Air Force Base (AFHRL/LR)
Director, Occupatioal Measurement Center, Randolph Air Force Base
Commandant Coast Guard Headquarters
Commanding Officer, U.S. Coast Guard Institute
Commanding Officer, U.S. Coast Guard Research and Development Center, Avery Point
Superintendent, U.S. Coast Guard Academy
Director, Science and Technology, Library of Congress
Defense Technical Information Center (DDA) (12)