Multivariate Behavioral Research

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/hmbr20

Omega: A General Formulation of the Rand Index of Cluster Recovery Suitable for Non-disjoint Solutions

Linda M. Collins & Clyde W. Dent

Published online: 10 Jun 2010.

To cite this article: Linda M. Collins & Clyde W. Dent (1988) Omega: A General Formulation of the Rand Index of Cluster Recovery Suitable for Non-disjoint Solutions, Multivariate Behavioral Research, 23:2, 231-242, DOI: 10.1207/s15327906mbr2302_6

To link to this article: http://dx.doi.org/10.1207/s15327906mbr2302_6

Multivariate Behavioral Research, 1988, 23, 231–242

Omega: A General Formulation of the Rand Index of Cluster Recovery Suitable for Non-disjoint Solutions

Linda M. Collins
University of Southern California

Clyde W. Dent
Institute for Health Promotion and Disease Prevention Research, University of Southern California

Cluster recovery indices are more important than ever, because of the necessity for comparing the large number of clustering procedures available today. Of the cluster recovery indices prominent in contemporary literature, the Hubert and Arabie (1985) adjustment to the Rand index (1971) has been demonstrated to have the most desirable properties (Milligan & Cooper, 1986). However, use of the Hubert and Arabie adjustment to the Rand index is limited to cluster solutions involving non-overlapping, or disjoint, clusters. The present paper introduces a generalization of the Hubert and Arabie adjusted Rand index. This generalization, called the Omega index, can be applied to situations where both, one, or neither of the solutions being compared is non-disjoint. In the special case where both solutions are disjoint, the Omega index is equivalent to the Hubert and Arabie adjusted Rand index.

In the last decade there has been a geometric increase in the number and variety of clustering procedures, presenting a bewildering array of choices to the consumer. For this reason both comparisons among various procedures and Monte Carlo evaluations of them are important for the user who must decide which approach to take for a particular clustering problem. Central to most evaluations of clustering procedures is the cluster recovery index, sometimes called an external criterion statistic (Milligan & Cooper, 1986). The cluster recovery index serves two purposes. It can give the researcher an idea of the degree of correspondence between two cluster solutions, and it can be a gauge of the degree to which an obtained solution agrees with a criterion solution.

Milligan and Cooper (1986) noted that five cluster recovery indices have gained prominence in the clustering literature: the Rand index (Rand, 1971); the Jaccard index (Downton & Brennan, 1980); the

This research was supported by grant BNS8403126 from the National Science Foundation and by grant R01DA3673 from the National Institute on Drug Abuse.

Requests for reprints should be sent to Linda M. Collins, Department of Psychology, University of Southern California, Los Angeles, CA 90089-1061.

Fowlkes and Mallows (1983) index; the Morey and Agresti (1984) adjusted Rand index; and the Hubert and Arabie (1985) adjusted Rand index. The purpose of the Morey and Agresti and Hubert and Arabie adjustments to the Rand index is to correct for chance levels of agreement, thereby avoiding spuriously large obtained values of the index. Based on Monte Carlo work done before Hubert and Arabie was published, Milligan and associates (Milligan & Schilling, 1985; Milligan, Soon, & Sokol, 1983) recommended use of the Morey and Agresti adjusted Rand index and the Jaccard index. Subsequently Hubert and Arabie pointed out that the Morey and Agresti adjustment does not completely eliminate bias due to chance, and suggested an alternative adjustment. A recent Monte Carlo study by Milligan and Cooper showed that the Hubert and Arabie adjustment to the Rand index outperformed both the Morey and Agresti adjustment and the Jaccard index. Thus, based on current evidence it seems that the Hubert and Arabie adjusted Rand index is the cluster recovery index of choice.

However, there is an important limitation to all five of the cluster recovery indices mentioned above, including the Hubert and Arabie adjusted Rand index. These cluster recovery indices are suitable only for situations involving disjoint (non-overlapping) cluster solutions. If either of the two solutions of interest places an object in more than one cluster, the solutions cannot be compared using the cluster recovery indices listed above. This means that the researcher who wishes to compare a non-disjoint clustering procedure, such as MAPCLUS (Arabie & Carroll, 1980) or BINCLUS (Cliff, McCormick, Zatkin, Cudeck, & Collins, 1986), to a criterion or to another cluster solution must use an alternative approach. This presents an obvious difficulty. The effectiveness of disjoint and non-disjoint clustering procedures cannot be directly compared if they must be evaluated using different cluster recovery indices. Direct comparisons can be made only by using a cluster recovery index suitable for both disjoint and non-disjoint solutions.

The purpose of this paper is to present Omega, a more general formulation of the Hubert and Arabie adjusted Rand index. This general formulation can be applied to situations where both, one, or neither of the solutions being compared is non-disjoint. In the special case where neither solution involves overlapping clusters, this formulation reduces to the Hubert and Arabie adjusted Rand index. Where there are overlapping cluster solutions involved, Omega returns an evaluation of cluster recovery that is in the same metric as the Hubert and Arabie adjusted Rand index.

Table 1

Table Forming the Basis of Several Cluster Recovery Indices

	Solution V		
	Pair of items in same cluster	Pair of items in different clusters	Marginal
Solution U			
Pair of items in same cluster	a	b	a + b
Pair of items in different clusters	c	d	c + d
Marginal	a + c	b + d	N

The Rand Index

Computation of the Rand Index

The Rand index focuses on pairs of objects. A disjoint clustering solution classifies pairs of objects in one of two ways: as belonging together (in the same cluster) or apart (in different clusters, or not in any cluster). The Rand index is a measure of the degree to which each pair of objects is classified the same by the two cluster solutions being compared. Essentially the Rand index calls for the construction of Table 1. Cells a and d in Table 1 represent agreement between the two solutions, and cells b and c represent disagreement. For example, if a pair of objects is classified as being in the same cluster by both solutions, this contributes to cell a in Table 1; if the two objects in the pair are classified as together in Solution U and apart in Solution V, this contributes to cell b; and so forth. Let n represent the number of objects being clustered, and N represent the number of pairs of objects; $a + b + c + d = N = n(n - 1)/2$. Then the Rand index is:

[1] $$(a + d)/N.$$

Conceptually, Equation 1 can be represented as:

[2] pairs classifed in agreement/total number of pairs,

that is, the proportion of pairs classified in agreement by the two solutions.

Examination of Table 1 illustrates why the Rand index cannot be applied to non-disjoint solutions. In an overlapping cluster solution, it may happen that a pair of items is together in more than one cluster. Suppose in Solution U there are overlapping clusters, so that a pair of objects is placed together in two clusters. However, in Solution V the pair of objects is placed together in only one cluster. The two solutions agree in one sense, that is, they both place the objects together, but they disagree in another sense, because in one solution the objects are together in more than one cluster and in the other solution they are together once only. There is no place for this information to be incorporated in the computation of the Rand index.

Hubert and Arabie's Adjustment

In this section Hubert and Arabie's (1985) adjustment to the Rand index will be discussed. In order to facilitate presentation of the Omega index, it is necessary to translate the adjustment into a notation slightly different from the combinatorial notation used by Hubert and Arabie.

Conceptually, the Hubert and Arabie adjustment is of the following form:

$$[3] \qquad \frac{\text{observed index—expected index}}{\text{maximum index—expected index}} \; .$$

That is, the numerator of Equation 3 represents the observed improvement over chance, and the denominator represents the maximum possible improvement over chance. The "maximum index" in the denominator of Equation 3 is always unity.

The adjustment centers on the computation of an expected index. The starting point for Hubert and Arabie's adjustment is the contingency table of Solution U cluster membership by Solution V cluster membership. An example of such a contingency table appears in Table 2, which shows clustering data first presented by Rand (1971) and also reported in Morey and Agresti (1984) and Hubert and Arabie (1985). Let i index the row elements, j index the column elements, and n equal the total number of objects being clustered. Then the marginals of Table 1 can be expressed as follows:

Table 2

Hypothetical Clustering Data from Rand (1971)

		Partition V			
		v_1	v_2	v_3	Marginal
Partition U	u_1	2	1	0	3
	u_2	0	2	1	3
	Marginal	2	3	1	6

[4] $$(a + b) = \sum_i \binom{n_{i\cdot}}{2}$$

[5] $$(c + d) = \binom{n}{2} - \sum_i \binom{n_{i\cdot}}{2}$$

[6] $$(a + c) = \sum_j \binom{n_{\cdot j}}{2}$$

[7] $$(b + d) = \binom{n}{2} - \sum_j \binom{n_{\cdot j}}{2} .$$

These marginals are used to compute cell expectations in the usual manner. Thus

[8] $$\varepsilon(a) = (a + b)(a + c)/N$$

$$= \sum_i \binom{n_{i\cdot}}{2} \sum_j \binom{n_{\cdot j}}{2} \Big/ \binom{n}{2}$$

and

[9] $$\varepsilon(d) = (c + d)(b + d)/N$$

$$= \left[\binom{n}{2} - \sum_i \binom{n_{i\cdot}}{2}\right]\left[\binom{n}{2} - \sum_j \binom{n_{\cdot j}}{2}\right] \Big/ \binom{n}{2}$$

$$= \binom{n}{2} - \sum_i \binom{n_{i\cdot}}{2} - \sum_j \binom{n_{\cdot j}}{2} + \sum_i \binom{n_{i\cdot}}{2} \sum_j \binom{n_{\cdot j}}{2} \Big/ \binom{n}{2} .$$

Then the expected Rand index is

Table 3

Reformatted Hypothetical Clustering Data from Rand (1971)

	Solution V		
	Pair of items in same cluster	Pair of items in different clusters	Marginal
Solution U			
Pair of items in same cluster	2	4	6
Pair of items in different clusters	2	7	9
Marginal	4	11	15

[10]
$$\varepsilon(a + d)/N$$
$$= [(a + b)(a + c) + (c + d)(b + d)]/N^2$$

or alternatively

[11]
$$\varepsilon(a + d)/N$$
$$= 1 + 2\sum_i \binom{n_{i\cdot}}{2} \sum_j \binom{n_{\cdot j}}{2} \Big/ \binom{n}{2}^2 - \left[\sum_i \binom{n_{i\cdot}}{2} + \sum_j \binom{n_{\cdot j}}{2}\right] \Big/ \binom{n}{2} .$$

Equations 10 and 11 are equivalent expressions of the Hubert and Arabie expected Rand index. Equation 10 is the expected Rand index expressed in terms of Table 1, and Equation 11 is the expected Rand index expressed in Hubert and Arabie's (1985) notation. Table 3 contains the data in Table 2 rearranged to fit the Table 1 format, for the convenience of readers who may wish to verify Equations 4 through 11. Applying Equation 10 yields the following result:

$$(24 + 99)/225 = 8.2/15 = .54667$$

which agrees with the expected Rand index for these data computed by Hubert and Arabie (1985) using Equation 11.

Following the general form of Equation 3 and expressing it in terms of the quantities in Table 1, the Hubert and Arabie adjusted Rand index is:

$$[12] \qquad \frac{N(a + d) - [(a + b)(a + c) + (c + d)(b + d)]}{N^2 - [(a + b)(a + c) + (c + d)(b + d)]} \, .$$

The Omega Index

In a disjoint cluster solution each object can appear in at most one cluster, so agreement between two disjoint solutions can be expressed completely in terms of agreement about how many pairs of objects are placed together and how many are placed apart. In contrast, in an overlapping cluster solution it is possible for a pair of objects to be together more than once. In fact, where a solution yields K overlapping clusters it is even possible for a pair of objects to be together in all K clusters. This means that when two non-disjoint cluster solutions are being compared, simply noting agreement about which pairs of objects are placed together and which are placed apart may not be sufficient to describe the degree of agreement between the two solutions. For example, if a cluster solution places a pair of objects together in two clusters and another cluster solution places the pair together in four clusters, these two solutions do not show perfect agreement. Yet an index incorporating information only about whether a pair is placed together or apart would indicate perfect agreement between these two solutions.

In order to assess the degree of agreement between two cluster solutions adequately, information must be included concerning how many pairs of objects belong together in no clusters, how many belong together in exactly one cluster, how many belong together in exactly two clusters, and so forth. The Omega index accomplishes this by expanding Table 1 to dimensions $(J + 1)$-by-$(K + 1)$, where J represents the maximum number of clusters in which a pair of objects appears in Solution 1, and K represents the maximum number of clusters in which a pair of objects appears in Solution 2. The expanded table is illustrated in Table 4. If a pair of objects is placed together in the same number of clusters by each solution, this represents agreement between the solutions; if a pair of objects is placed together in different numbers of clusters by the two solutions, this represents disagreement between the solutions. It is important to note that the total number of pairs, N, remains equal to $n(n - 1)/2$. Just as in Table

Table 4

Expanded Contingency Table for Computation of Omega Index

		Solution 2					
		Number of clusters in which pair is together					
		0	1	2..	k..	K	Marginal
Solution 1	0	A_0					$N_{0.}$
Number of clusters in which	1		A_1				$N_{1.}$
pair is together	2 ⋮			A_2			$N_{2.}$
	j ⋮				A_j		$N_{j.}$
	J					A_J	$N_{J.}$
	Marginal	$N_{.0}$	$N_{.1}$	$N_{.2}$	$N_{.k}$	$N_{.K}$	N

1, each pair appears in one and only one cell of the contingency table. In the special case of a disjoint solution, where a pair of items can be together in at most one cluster, Table 4 reduces to Table 1.

Omega is then computed using the same general form as Equation 3; that is, Omega is a combination of an unadjusted index and an expected index.

An Unadjusted Cluster Recovery Index for the Extended Table

In the $(J + 1)$-by-$(K + 1)$ table shown in Table 4, pairs appearing in the $j = k$ diagonal are those pairs showing agreement between the two solutions, just as the diagonal cells a and d in Table 1 show agreement. These cells showing agreement between the two solutions are labelled A in Table 4. Thus, a simple extension of the unadjusted Rand index to this situation is:

[13]
$$\sum_{j=0}^{\min(J,K)} A_j / N$$

that is, Equation 13 is the proportion of pairs of items classified in the same way by Solutions 1 and 2.

An Expected Omega Index Based on the Extended Table

The expected index based on the extended table is computed following Hubert and Arabie's (1985) recommendations for computing the expected Rand index. Expectations for each A_j are computed using a version of Equation 8, that is, by multiplying the appropriate marginals and dividing by N:

$$[14] \qquad \varepsilon(A_j) = N_{j\cdot}\, N_{\cdot j} \,/\, N$$

so the expected index is:

$$[15] \qquad \varepsilon(\text{Omega}) = \sum_{j=0}^{\min(J,K)} N_{j\cdot}\, N_{\cdot j} \,/\, N^2.$$

When $J = K = 1$, that is, when both clustering solutions involved are disjoint, Equation 15 is equivalent to Equation 10 and Equation 11.

Computation of the Omega Index

Following the general form of Equation 3, the numerator is the difference between the observed index and the expected index, and the denominator is the difference between the maximum index and the expected index:

$$[\sum_{j=0}^{\min(J,K)} A_j \,/\, N - (\sum_{j=0}^{\min(J,K)} N_{j\cdot}\, N_{\cdot j} / N^2)] \,/\, [1 - (\sum_{j=0}^{\min(J,K)} N_{j\cdot}\, N_{\cdot j} \,/\, N^2)]$$

This reduces to the computationally simpler

$$[16] \qquad (N \sum_{j=0}^{\min(J,K)} A_j - \sum_{j=0}^{\min(J,K)} N_{j\cdot}\, N_{\cdot j}) \;/\; (N^2 - \sum_{j=0}^{\min(J,K)} N_{j\cdot}\, N_{\cdot j})$$

When $J = K = 1$, Equation 16 is equivalent to Equation 12.

A Hypothetical Example

Table 5 shows two overlapping cluster solutions on the same hypothetical data set. In Solution 1 there are pairs of objects that are never clustered together, such as *A,F,* and there are pairs of objects that are clustered together once only, such as *A,B*. Although there is overlap between clusters 1 and 2 in this solution, there are no pairs of objects that appear together in more than one cluster. In Solution 2 there is one pair of objects that appears in two clusters, that is, pair *C,D*.

Table 6 contains a contingency table like the one in Table 4,

Table 5

Two Hypothetical Overlapping Cluster Solutions

	Solution 1*				Solution 2		
	Cluster				Cluster		
Object	1	2	3	4	1	2	3
A	X				X		
B	X				X		
C	X				X	X	
D	X	X			X	X	
E		X				X	
F		X				X	
G		X				X	
H			X				X
I			X				X
J				X			X

*An "X" indicates that the row object is a member of the column cluster.

classifying each pair of objects according to the number of times it is clustered together in each solution. That is, the *j,k*th cell in Table 6 contains the number of times a pair of objects is placed together *j* times by Solution 1 and *k* times by Solution 2. For example, there were 12 pairs of objects placed together once by both solutions; there were five pairs of objects placed together once by Solution 1 and never by Solution 2, and so forth.

From the information in Table 6 Omega can be computed using Equation 16:

$$\frac{(45)(39) - [(27)(32) + (17)(13)]}{45^2 - [(27)(32) + (17)(13)]} = .71.$$

An Omega value of .71 can be interpreted as meaning that the two clustering solutions show a 71 percent agreement, over and above any agreement that would be expected due to chance.

Table 6

Cross-Tabulation of Number of Times Each Pair of Objects in Table 5 is Clustered Together by Each Solution

		Solution 2			
		Number of clusters in which pair is together			
		0	1	2	Marginal
Solution 1					
Number of clusters	0	27	5	0	32
in which pair is together	1	0	12	1	13
	Marginal	27	17	1	45

Summary

The cluster recovery indices that have received the most attention in recent literature are suitable only for disjoint solutions. This limitation makes it difficult to compare the effectiveness of disjoint and non-disjoint procedures because they must be evaluated using different criteria. The present paper has introduced a generalization of the Hubert and Arabie adjusted Rand index for assessing cluster recovery. This generalization, called Omega, is suitable for situations where both of the two solutions being compared are disjoint, for situations where one solution is disjoint and one is overlapping, and for situations where both solutions are overlapping. In the special case where both solutions are disjoint, the Omega index is equivalent to the Hubert and Arabie adjusted Rand index.

References

Arabie, P., & Carroll, J. D. (1980). MAPCLUS: A mathematical programming approach to fitting the ADCLUS model. *Psychometrika, 45,* 211–235.

Cliff, N., McCormick, D. J., Zatkin, J. L., Cudeck, R. A., & Collins, L. M. (1986). BINCLUS: Nonhierarchical clustering of binary data. *Multivariate Behavioral Research, 21,* 201–227.

Downton, M., & Brennan, T. (1980, June). *Comparing classifications: An evaluation of several coefficients of partition agreement.* Paper presented at the meeting of the Classification Society, Boulder, CO.

Fowlkes, E. B., & Mallows, C. L. (1983). A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association, 78,* 553–584.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification, 2,* 193–218.

Milligan, G. W., & Cooper, M. C. (1986). A study of the comparability of external criteria for hierarchical cluster analysis. *Multivariate Behavioral Research, 21,* 441–458.

Milligan, G. W., & Schilling, D. A. (1985). Asymptotic and finite sample characteristics of four external criterion measures. *Multivariate Behavioral Research, 20,* 97–109.

Milligan, G. W., Soon, T., & Sokol, L. (1983). The effect of cluster size, dimensionality, and the number of clusters on recovery of true cluster structure. *IEEE Transactions on Panel Analysis and Machine Intelligence, 5,* 40–47.

Morey, L. C., & Agresti, A. (1984). The measurement of classification agreement: An adjustment to the Rand Statistic for chance agreement. *Educational and Psychological Measurement, 44,* 33–37.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association, 66,* 846–850.