Journal of Classification (2022) 39:487–509
https://doi.org/10.1007/s00357-022-09413-z

Understanding the Adjusted Rand Index and Other Partition Comparison Indices Based on Counting Object Pairs

Matthijs J. Warrens[1] · **Hanneke van der Hoef**[1]

Accepted: 23 March 2022 / Published online: 22 July 2022

Abstract

In unsupervised machine learning, agreement between partitions is commonly assessed with so-called external validity indices. Researchers tend to use and report indices that quantify agreement between two partitions for all clusters simultaneously. Commonly used examples are the Rand index and the adjusted Rand index. Since these overall measures give a general notion of what is going on, their values are usually hard to interpret. The goal of this study is to provide a thorough understanding of the adjusted Rand index as well as many other partition comparison indices based on counting object pairs. It is shown that many overall indices based on the pair-counting approach can be decomposed into indices that reflect the degree of agreement on the level of individual clusters. The decompositions (1) show that the overall indices can be interpreted as summary statistics of the agreement on the cluster level, (2) specify how these overall indices are related to the indices for individual clusters, and (3) show that the overall indices are affected by cluster size imbalance: if cluster sizes are unbalanced these overall measures will primarily reflect the degree of agreement between the partitions on the large clusters, and will provide much less information on the agreement on smaller clusters. Furthermore, the value of Rand-like indices is determined to a large extent by the number of pairs of objects that are not joined in either of the partitions.

Keywords Clustering comparison · External validity indices · Reference standard partition · Trial partition · Wallace indices · Cluster size imbalance

1 Introduction

The problem of measuring agreement between two different partitions of the same finite set of objects reappears continually in many scientific disciplines (Hennig et al., 2015;

✉ Matthijs J. Warrens
m.j.warrens@rug.nl

Hanneke van der Hoef
h.vanderhoef@hotmail.com

[1] Groningen Institute for Educational Research, University of Groningen, Grote Rozenstraat 3, 9712 TG, Groningen, The Netherlands

Hubert, 1977; Pfitzner et al., 2009; Van der Hoef & Warrens, 2019). For example, in unsupervised machine learning, to evaluate the performance of a clustering method, researchers typically assess agreement between a reference standard partition that purports to represent the true cluster structure of the objects (golden standard), and a trial partition produced by the method that is being evaluated (Halkidi & Batiskis, 2002; Jain, 2010; Wallace, 1983). High agreement between the two partitions may indicate good recovery of the true cluster structure.

Agreement between partitions can be assessed with so-called external validity indices (Albatineh et al., 2006; Brun et al., 2007; Pfitzner et al., 1996; Warrens, 2008b, d). External validity indices can be roughly categorized into three approaches, namely 1) counting object pairs (Albatineh et al., 2006; Albatineh & Niewiadomska-Bugaj, 2011a; Warrens, 2008c), 2) information theory (Vinh et al., 2009; 2010; Kvalseth, 1987), and 3) matching sets (Fränti et al., 2014; Steinley et al., 2016). Most external validity indices are of the pair-counting approach, which is based on counting pairs of objects placed in identical and different clusters. Information theoretic indices are based on concepts like the mutual information, Shannon entropy (Shannon, 1948) and joint entropy (Kvalseth, 1987; Pfitzner et al., 2009). These indices assess the difference in information between two partitions. Finally, set-matching indices are based on matching entire clusters, usually using the matched parts of each cluster, while ignoring the unmatched parts (Fränti et al., 2014; Meilă, 2007; 2016).

Commonly used external validity indices are the Rand index (Rand, 1971) and the Hubert-Arabie adjusted Rand index (Hubert & Arabie, 1985; Steinley, 2004; Steinley et al., 2016; Steinley et al., 2015; Warrens, 2008c; Chacón, 2019; Chacón & Rastrojo, 2020). Both these indices are based on counting pairs of objects. The adjusted Rand index corrects the Rand index for agreement due to chance (Albatineh et al., 2006; Warrens, 2008b). Milligan and Cooper (1986), Milligan (1996), and Steinley (2004) proposed to use the adjusted Rand index as a standard tool in cluster validation research. However, the Rand index continues to be a popular validity index, probably because it has a simple, natural interpretation (Anderson et al., 2010).

Researchers tend to use and report validity indices that quantify agreement between two partitions for all clusters simultaneously (Albatineh & Niewiadomska-Bugaj, 2011b; Alok et al., 2014; Kim et al., 2009; Milligan & Cooper, 1986; Yu et al., 2012). Since these overall measures give a general notion of what is going on, it is usually difficult to pinpoint what their values, usually between 0 and 1, actually reflect. Values of overall indices are generally hard to interpret, except for values close to 0 or 1.

The goal of this study is to provide a thorough understanding of the adjusted Rand index as well as many other indices based on counting object pairs. We analyze three different families of indices. We focus on indices based on pair-counting because these are most commonly used (Pfitzner et al., 2009; Van der Hoef & Warrens, 2019). To enhance our understanding of overall indices, we show that various overall indices can be decomposed into indices that reflect the degree of agreement on the level of the individual clusters. More precisely, we show that the overall indices are weighted means (variously defined) of indices that can be used to assess agreement for individual clusters of the partitions. In many cases the weights of these means are quadratic functions of the cluster sizes.

The decompositions show that, if the cluster sizes differ, measures like the Jaccard index (Jaccard, 1912) and the Hubert-Arabie adjusted Rand index tend to mainly reflect the degree of agreement between the partitions on the large clusters. The indices provide little to no information on the smaller clusters. This susceptibility to cluster size imbalance has been observed previously in the literature for some indices (Pfitzner et al., 2009; Van der Hoef & Warrens, 2019). The analyses presented in this paper amplify these previous studies by providing insight into how this phenomenon actually works and to which indices it applies. Furthermore, the values of Rand-like indices are determined to a large extent by the number of pairs of objects that are not joined in either of the partitions.

The paper is organized as follows. The notation is introduced in Section 2. In Sections 3, 4 and 6 we present decompositions of three families of indices. Section 3 focuses on indices that are functions of the two asymmetric Wallace indices (Wallace, 1983). Prototypical examples of this family are the Jaccard index and an index by Fowlkes and Mallows (1983). Decompositions of indices that are adjusted for agreement for chance (Albatineh et al., 2006; Warrens, 2008b) are presented in Section 4. A prototypical example of this family is the Hubert-Arabie adjusted Rand index. In Section 5 we present artificial and a real-world example to illustrate how the indices associated with the families in Sections 3 and 4 are related. In Section 6 we analyze indices that are functions of both the Wallace indices and two indices that focus on pairs of objects that are not joined together in the partitions. A prototypical example of this family is the Rand index (Rand, 1971). In Section 6.2 we consider particular properties of the Rand-like family defined in Section 6.1. Finally, Section 7 contains a discussion, our recommendations and some ideas for future research.

2 Notation

In this section we introduce the notation. Suppose the data are scores of n objects on k variables. Let $U = \{U_1, U_2, \ldots, U_I\}$ and $Z = \{Z_1, Z_2, \ldots, Z_J\}$ denote two partitions of the objects, for example, a reference standard partition and a trial partition that was obtained with a clustering method that is being evaluated. Let $\mathbf{N} = \{n_{ij}\}$ be a matching table of size $I \times J$ where n_{ij} indicates the number of objects placed in cluster U_i of the first partition and in cluster Z_j of the second partition. The cluster sizes in respective partitions are the row and column totals of $\mathbf{N}$, that is,

$$|U_i| = n_{i+} = \sum_{j=1}^{J} n_{ij} \qquad \text{and} \qquad |Z_j| = n_{+j} = \sum_{i=1}^{I} n_{ij}. \tag{1}$$

Table 1 is an example of matching table $\mathbf{N}$. Table 1 is based on a data set that contains information on E. coli sequences (Horton & Nakai, 1996; Lichman, 2013). The data set consists of 336 proteins belonging to 8 classes (reference partition), which are the localization sites: cytoplasmic (cp), inner membrane without signal sequence (im), inner membrane lipoprotein (imL), inner membrane, cleavable signal sequence (imS), inner membrane proteins with an uncleavable signal sequence (imU), outer membrane (om), outer membrane lipoprotein (omL), and periplasmic (pp). For all proteins, 7 features were calculated from amino acid sequences.

Table 1 presents the matching table of the reference partition and a K-means clustering (Huo et al., 2016; Jain, 2010; Steinley, 2006) of the E. coli sequences. All 7 features were used in the K-means clustering, and solutions with 3–10 clusters were estimated. The clustering solution with $K = 4$ clusters had the highest value of the Dunn index (Dunn, 1974). Thus, the trial partition of Table 1 consists of 4 clusters.

Following Fowlkes and Mallows (1983), the information in the matching table $\mathbf{N}$ can be summarized in a fourfold contingency table (like Table 2) by counting several different

Table 1 Matching table of a reference partition and a K-means clustering of E. coli sequences

Reference partition		Trial partition				
Proteins		Z_1	Z_2	Z_3	Z_4	Totals
cp	$= U_1$	5	0	137	1	143
im	$= U_2$	8	0	1	68	77
imL	$= U_3$	0	1	0	1	2
imS	$= U_4$	1	0	0	1	2
imU	$= U_5$	0	0	1	34	35
om	$= U_6$	2	18	0	0	20
omL	$= U_7$	0	5	0	0	5
pp	$= U_8$	46	1	4	1	52
Totals		62	25	143	106	336

types of pairs of objects: $N := n(n-1)/2$ is the total number of pairs of objects,

$$T := \sum_{i=1}^{I} \sum_{j=1}^{J} \binom{n_{ij}}{2} \tag{2}$$

is the number of object pairs that were placed in the same cluster in both partitions,

$$P := \sum_{i=1}^{I} \binom{n_{i+}}{2} \tag{3}$$

is the number of object pairs that were placed in the same cluster in partition U, and

$$Q := \sum_{j=1}^{J} \binom{n_{+j}}{2} \tag{4}$$

is the number of object pairs that were placed in the same cluster in partition Z. The bottom panel of Table 2 gives a representation of the matching table in terms of the counts N, T, P and Q. Furthermore, define $a := T$, $b := P - T$, $c := Q - T$ and $d := N + T - P - Q$. Quantity b (c) is the number of object pairs that were placed in the same cluster in partition

Table 2 Two 2 × 2 contingency table representations of matching table **N**

First partition	Second partition		
	Pair in the same cluster	Pair in different clusters	Totals
Representation 1			
Pair in the same cluster	a	b	$a+b$
Pair in different clusters	c	d	$c+d$
Totals	$a+c$	$b+d$	N
Representation 2			
Pair in the same cluster	T	$P-T$	P
Pair in different clusters	$Q-T$	$N+T-P-Q$	$N-P$
Totals	Q	$N-Q$	N

U (Z) but in different clusters in partition Z (U). The quantity d is the number of object pairs that are not joined in either of the partitions.

The top panel of Table 2 gives a representation of the matching table using the counts a, b, c and d. The latter notational system is commonly used for expressing similarity measures for 2 × 2 tables (Heiser & Warrens 2010; Warrens 2008a, b, d; Warrens 2019).

Several authors have proposed indices specifically for assessing agreement between partitions (Fowlkes & Mallows, 1983; Hubert & Arabie, 1985; Rand, 1971; Wallace, 1983). However, if the agreement between the partitions is summarized as in the top panel of Table 2, one may use any similarity index from the vast literature on 2 × 2 tables (Albatineh et al. 2006; Baulieu 1989; Hubálek 1982; Pfitzner et al. 2009; Warrens 2008a, d, e; Warrens 2019). Moreover, each index that has been specifically proposed for assessing agreement between partitions, has a precursor in the literature on 2 × 2 tables (see Tables 4, 5 and 11 for specific examples).

Table 3 is the fourfold table corresponding to Table 1, and is an example of Table 2. Table 3 summarizes the information in matching Table 1 on E. coli sequences in terms of the four types of pairs of objects.

3 Functions of the Wallace Indices

Wallace (1983) considers the following two asymmetric indices. The first index

$$W = \frac{T}{P} = \frac{a}{a+b} \tag{5}$$

is the proportion of object pairs in the first partition that are also joined in the second partition (Severiano et al., 2011). The second index

$$V = \frac{T}{Q} = \frac{a}{a+c} \tag{6}$$

is the proportion of object pairs in the second partition that are also joined in the first partition. Table 4 presents twelve examples of indices from the literature that are increasing functions of conditional probabilities (5) and (6). Some of these functions, for example, the Dice index (Dice, 1945)

$$D = \frac{2WV}{W+V} = \frac{2T}{P+Q}, \tag{7}$$

which is the harmonic mean of (5) and (6), are rather simple functions of the Wallace indices (e.g., sum, product, geometric mean, arithmetic mean, minimum, maximum), while other functions, for example, the Jaccard coefficient, are more complicated functions of (5) and (6). Table 4 is a list of partition comparison indices that are functions of both W and V.

Table 3 The 2 × 2 contingency table corresponding to Table 1

Reference partition	Trial partition		
	Pair in the same cluster	Pair in different cluster	Totals
Pair in the same cluster	13398	1804	15202
Pair in different clusters	4511	36567	41078
Totals	17909	38371	56280

Table 4 Indices that are increasing functions of Wallace indices (5) and (6)

Source	Formula 1	Formula 2
Jaccard (1912)	$a/(a+b+c)$	$WV/(W+V-2WV)$
Gleason (1920), Dice (1945), Sørenson (1948)	$2a/(2a+b+c)$	$2WV/(W+V)$
Kulczyński (1927), Driver and Kroeber (1932)	$a/2(a+b)+a/2(a+c)$	$(W+V)/2$
Braun-Blanquet (1932)	$a/(a+\max(b,c))$	$\min(W,V)$
Simpson (1943)	$a/(a+\min(b,c))$	$\max(W,V)$
Ochiai (1957), Fowlkes and Mallows (1983)	$a/\sqrt{(a+b)(a+c)}$	$\sqrt{WV}$
Sorgenfrei (1958), Cheetham and Hazel (1969)	$a^2/(a+b)(a+c)$	WV
Sokal and Sneath (1963)	$a/(a+2b+2c)$	$WV/(2(W+V)-3WV)$
McConnaughey (1964)	$(a^2-bc)/(a+b(a+c)$	$W+V-1$
Johnson (1967)	$a/(a+b)+a/(a+c)$	$W+V$
Van der Maarel (1969)	$(2a-b-c)/(2a+b+c)$	$4WV/(W+V)-1$
Legendre and Legendre (1998)	$3a/(3a+b+c)$	$3WV/(W+V+WV)$

The middle column of Table 4 gives the formulas in terms of the regular 2 × 2 tables. The last column of Table 4 gives the formula in terms of W and V. All indices in Table 4 are increasing functions of W and V. Hence, to understand all indices in Table 4, it is instrumental to first understand the values produced by indices (5) and (6).

The Wallace indices can be decomposed into the following indices for the individual clusters of partitions U and Z. Define for $U_i \in U$ the (relative) weights

$$P_i := \binom{n_{i+}}{2} \qquad \text{and} \qquad p_i := \frac{P_i}{P}, \tag{8}$$

which are, respectively, the number and proportion of object pairs in cluster U_i, and the index

$$w_i := \sum_{j=1}^{J} \binom{n_{ij}}{2} \Big/ \binom{n_{i+}}{2}, \tag{9}$$

which is the proportion of object pairs in cluster U_i that are joined in partition Z. For example, for the first row of Table 1 (cluster U_1) we have

$$P_1 = \binom{143}{2} = 10153 \qquad \text{and} \qquad p_1 = \frac{P_1}{P} = \frac{10153}{15202} = 0.67,$$

and

$$w_1 = \frac{\binom{5}{2} + \binom{0}{2} + \binom{137}{2} + \binom{1}{2}}{\binom{143}{2}} = \frac{10 + 9316}{10153} = 0.92.$$

Furthermore, define for $Z_j \in Z$ the (relative) weights

$$Q_j := \binom{n_{+j}}{2} \quad \text{and} \quad q_j := \frac{Q_j}{Q}, \tag{10}$$

which are, respectively, the number and proportion of object pairs in cluster Z_j, and the quantity

$$v_j := \sum_{i=1}^{I} \binom{n_{ij}}{2} \Big/ \binom{n_{+j}}{2}, \tag{11}$$

which is the proportion of object pairs in cluster Z_j that are joined in partition U.

Indices (9) and (11) can be used to assess the agreement between partitions U and Z on the level of the individual clusters. Index (9) (or (11)) has value 1 if all objects in cluster U_i (Z_j) are in precisely one cluster of partition Z (U), and value 0 only if no two objects from cluster U_i (Z_j) are paired together in partition Z (U). Index (9) is a measure of sensitivity (recall, classification rate) (Ting, 2011) that does not require any matching between clusters from partitions U and Z.

We have the following decomposition for the first Wallace index. Index (5) is a weighted average of the indices in (9) using the P_i's (or p_i's) as weights:

$$W = \frac{\sum_{i=1}^{I} w_i P_i}{\sum_{i=1}^{I} P_i} = \sum_{i=1}^{I} w_i p_i. \tag{12}$$

Decomposition (12) shows that the overall W value will in large part be determined by the w_i values of the clusters with high P_i values, that is, the large clusters, since each P_i is a quadratic function of the cluster size. The overall W value will be high if, for each large cluster, its corresponding objects are assigned to the same cluster of partition Z, regardless of the w_i values associated with smaller clusters.

Furthermore, we have the following decomposition for the second Wallace index. Index (6) is a weighted average of the indices in (11) using the Q_j's (or q_j's) as weights:

$$V = \frac{\sum\limits_{j=1}^{J} v_j Q_j}{\sum\limits_{j=1}^{J} Q_j} = \sum_{j=1}^{J} v_j q_j. \tag{13}$$

Similarly, decomposition (13) shows that the overall V value will in large part be determined by the v_j values of the clusters with high Q_j values, that is, the large clusters. The overall V value will be high if, for each large cluster, its corresponding objects are put in the same cluster of partition U.

Decompositions (12) and (13) show that the indices in Table 4 are functions of the w_i's and v_j's of the individual clusters. Their values are largely determined by the w_i values and v_j values associated with the large clusters. For example, the Dice index is simply a weighted average of the w_i's and v_j's, using the P_i's and Q_j's as weights:

$$D = \frac{\sum\limits_{i=1}^{I} w_i P_i + \sum\limits_{j=1}^{J} v_j Q_j}{\sum\limits_{i=1}^{I} P_i + \sum\limits_{j=1}^{J} Q_j}. \tag{14}$$

The decompositions in (12), (13), and (14) are further explored with numerical examples in Section 5.

4 Chance-Corrected Functions

Most indices from the literature have value 1 if there is perfect agreement between the two partitions. However, for many indices it is unclear under which conditions their theoretical lower bound, for example 0, is attained. Therefore, when partitions are compared, it is usually convenient that the index of choice has value 1 if the partitions are completely similar and value 0 if the partitions are statistically independent. In this study, two partitions are considered statistically independent if we have, for all i and j,

$$\binom{n_{ij}}{2} = \frac{1}{N}\binom{n_{i+}}{2}\binom{n_{+j}}{2},$$

that is, the binomial coefficient $\binom{n_{ij}}{2}$ can be factored into a product of binomial coefficients with integers from the row and column totals. For example, the Wallace indices in (5) and (6) have value 1 if the partitions are identical. However, their value is not necessarily 0 under statistical independence of the partitions.

If a similarity measure S does not have value 0 under statistical independence, it can be corrected for agreement due to chance using the formula

$$AS = \frac{S - E(S)}{1 - E(S)}, \tag{15}$$

where expectation $E(S)$ is conditional upon fixed row and column totals of matching table $\mathbf{N}$, and 1 is the maximum value of S regardless of the marginal numbers (Albatineh & Niewiadomska-Bugaj, 2011a; Albatineh et al., 2006; Warrens, 2008b).

Assuming a generalized hypergeometric model for matching table $\mathbf{N}$, we have the expectation (Fowlkes & Mallows, 1983; Hubert & Arabie, 1985)

$$E\binom{n_{ij}}{2} = \frac{1}{N}\binom{n_{i+}}{2}\binom{n_{+j}}{2}. \tag{16}$$

Summing identity (16) over all cells of $\mathbf{N}$ we obtain

$$E\,(T) = \frac{PQ}{N}. \tag{17}$$

Using Wallace index (5) in (15), together with identity (17), yields the adjusted index (Severiano et al., 2011)

$$AW = \frac{NT - PQ}{P(N-Q)} = \frac{ad - bc}{(a+b)(b+d)}. \tag{18}$$

Furthermore, inserting Wallace index (6) into (15) yields

$$AV = \frac{NT - PQ}{Q(N-P)} = \frac{ad - bc}{(a+c)(c+d)}. \tag{19}$$

Table 5 presents five examples of indices from the literature that are increasing functions of adjusted indices (18) and (19). A well-known example is the adjusted Rand index (Cohen, 1960; Hubert & Arabie, 1985; Steinley, 2004; Steinley et al., 2015; Steinley et al., 2016; Warrens, 2008c)

$$AR = \frac{2(NT - PQ)}{N(P+Q) - 2PQ} = \frac{2(ad - bc)}{(a+b)(b+d) + (a+c)(c+d)}. \tag{20}$$

The adjusted Rand index in (20) is the harmonic mean of (18) and (19)(19). If $NT \neq PQ$, we have $AW > AR > AV$ if $P < Q$ and $AW < AR < AV$ if $P > Q$. The adjusted Rand index is what we get if we use the Rand index (defined below) in (28) in correction for chance formula (15). Moreover, the adjusted Rand index is also obtained if we use the Dice index in (7) in (15), that is, $AR = AD$ (Albatineh et al., 2006).

Indices (18) and (19) can be decomposed into the following indices for the individual clusters of partitions U and Z. Using (9) in (15) we obtain

$$Aw_i = \frac{N \sum_{j=1}^{J} \binom{n_{ij}}{2} - \binom{n_{i+}}{2} Q}{\binom{n_{i+}}{2}(N-Q)}. \tag{21}$$

Furthermore, inserting (11) into (15) yields

$$Av_j = \frac{N \sum_{i=1}^{I} \binom{n_{ij}}{2} - \binom{n_{+j}}{2} P}{\binom{n_{+j}}{2}(N-P)}. \tag{22}$$

Similar to indices (9) and (11), indices (21) and (22) can be used to assess the agreement between partitions U and Z on the level of the individual clusters. Index (21) (or (22)) has value 1 if all objects in cluster U_i (Z_j) are in precisely one cluster of Z (U), and value 0 under statistical independence. Index (21) is a measure of sensitivity (recall, classification rate) that does not require any matching between clusters from partitions U and Z.

Table 5 Indices that are increasing functions of (18) and (19)

Source	Formula 1	Formula 2
Doolittle (1885)	$(ad - bc)^2/(a+b)(a+c)(b+d)(c+d)$	$AW \cdot AV$
Yule (1912) (phi coefficient)	$(ad - bc)/\sqrt{(a+b)(a+c)(b+d)(c+d)}$	$\sqrt{AW \cdot AV}$
Loevinger (1947)	$(ad - bc)/\min[(a+b)(b+d), (a+c)(c+d)]$	$\max(AW, AV)$
Cohen (1960), Hubert and Arabie (1985)	$2(ad - bc)/[(a+b)(b+d) + (a+c)(c+d)]$	$(2AW \cdot AV)/(AW + AV)$
Fleiss (1975)	$(ad - bc)/2(a+b)(b+d) + (ad - bc)/2(a+c)(c+d)$	$(AW + AV)/2$

Index (18) is a weighted average of the indices in (21) using the P_i's (or p_i's) as weights:

$$AW = \frac{\sum\limits_{i=1}^{I} Aw_i P_i}{\sum\limits_{i=1}^{I} P_i} = \sum_{i=1}^{I} Aw_i p_i. \tag{23}$$

Decomposition (23) shows that the overall AW value will in large part be determined by the Aw_i values of the clusters with high P_i values, that is, the large clusters, since each P_i is a quadratic function of the cluster size. The overall AW value will be high if, for each large cluster, its corresponding objects are assigned to the same cluster of the second partition, regardless of the Aw_i values associated with smaller clusters.

Furthermore, index (19) is a weighted average of the indices in (22) using the Q_j's (or q_j's) as weights:

$$AV = \frac{\sum\limits_{j=1}^{J} Av_j Q_j}{\sum\limits_{j=1}^{J} Q_j} = \sum_{j=1}^{J} Av_j q_j. \tag{24}$$

Similarly, decomposition (24) shows that the overall AV value will in large part be determined by the Av_j values of the large clusters (that is, clusters with high Q_j values). The overall AV value will be high if objects that are together in a large cluster are also put together in the first partition.

Decompositions (23) and (24) show that the adjusted Rand index is simply a weighted average of the Aw_i's and Av_j's, using the P_i's and Q_j's as weights:

$$AR = \frac{\sum\limits_{i=1}^{I} Aw_i P_i + \sum\limits_{j=1}^{J} Av_j Q_j}{\sum\limits_{i=1}^{I} P_i + \sum\limits_{j=1}^{J} Q_j}. \tag{25}$$

Hence, the value of the adjusted Rand index will in large part be determined by the Aw_i values and Av_j values corresponding to large clusters.

5 Numerical Examples

In this section, we present examples to illustrate how the building blocks in (9) and (11) are related to the Wallace indices in (12) and (13), and how the building blocks in (21) and (22) are related to the adjusted Wallace indices in (18) and (19). We first consider two toy examples. In addition, we consider the data on E. coli sequences in Table 1.

5.1 Toy Example 1

Table 6 is a matching table of two partitions of four clusters each. The two partitions both consist of two relatively large clusters ($n_{1+} = n_{2+} = n_{+1} = n_{+2} = 20$) and two small clusters ($n_{3+} = n_{4+} = n_{+3} = n_{+4} = 8$). Table 7 presents various row, column and overall statistics corresponding to Table 6. Since Table 6 is symmetric, the row and column statistics are identical. First of all, there is perfect agreement between the partitions on the two large clusters, which is reflected in the corresponding (adjusted) cluster indices: w_1, w_2, v_1, v_2,

Table 6 Matching table of two partitions with four clusters each, and perfect agreement on the large clusters

Reference partition	Trial partition				
	Z_1	Z_2	Z_3	Z_4	Totals
U_1	20	0	0	0	20
U_2	0	20	0	0	20
U_3	0	0	4	4	8
U_4	0	0	4	4	8
Totals	20	20	8	8	56

Aw_1, Aw_2, Av_1 and Av_2 are all equal to 1.00. Furthermore, the two small clusters are completely (uniformly) mixed up, which is reflected in the corresponding cluster indices: w_3, w_4, v_3, v_4, Aw_3, Aw_4, Av_3 and Av_4 are all substantially lower than unity.

The overall indices W, V and D are weighted averages of the w_i's and v_j's (see decompositions (12), (13) and (14)), and overall indices AW, AV and AR are weighted averages of the Aw_i's and Av_j's (see decompositions (23), (24) and (25)). The weights used in these weighted averages are the p_i's and q_i's. The weights associated with the two large clusters (0.44) are 7.33 times as high as the weights associated with the two small clusters (0.06). Larger clusters simply have larger weights (see Equations (8) and (10)). The values of the overall indices are therefore much closer to the values of the cluster indices associated with the large clusters (1.00) than the values of the indices corresponding to the small clusters (0.43 or 0.20).

For example, using the values of the w_i's and p_i's in Table 7, decomposition (12) is equal to

$$W = \sum_{i=1}^{4} w_i p_i = (1.00)(0.44) + (1.00)(0.44) + (0.43)(0.06) + (0.43)(0.06) = 0.93.$$

Thus, due to the high weights of the cluster indices associated with the large clusters, the values of the overall indices primarily reflect the degree of agreement on the two large clusters, which happens to be perfect agreement. The values of the overall indices are therefore quite close to unity.

5.2 Toy Example 2

Similar to Table 6, Table 8 is a matching table of two partitions of four clusters each. Again, the two partitions both consist of two relatively large clusters

Table 7 Row, column and overall statistics for the data in Table 6

Row statistics				Column statistics				Overall indices			
i	w_i	Aw_i	p_i	j	v_j	Av_j	q_j				
1	1.00	1.00	0.44	1	1.00	1.00	0.44	W	0.93	AW	0.90
2	1.00	1.00	0.44	2	1.00	1.00	0.44	V	0.93	AV	0.90
3	0.43	0.20	0.06	3	0.43	0.20	0.06	D	0.93	AR	0.90
4	0.43	0.20	0.06	4	0.43	0.20	0.06				

Table 8 Matching table of two partitions with four clusters each, and perfect agreement on the small clusters

Reference partition	Trial partition				
	Z_1	Z_2	Z_3	Z_4	Totals
U_1	10	10	0	0	20
U_2	10	10	0	0	20
U_3	0	0	8	0	8
U_4	0	0	0	8	8
Totals	20	20	8	8	56

($n_{1+} = n_{2+} = n_{+1} = n_{+2} = 20$) and two small clusters ($n_{3+} = n_{4+} = n_{+3} = n_{+4} = 8$). Table 9 presents various row, column and overall statistics corresponding to Table 8. Unlike Table 6, there is perfect agreement between the partitions on the two small clusters in Table 8, which is reflected in the corresponding (adjusted) cluster indices: w_3, w_4, v_3, v_4, Aw_3, Aw_4, Av_3 and Av_4 are all equal to 1.00. Furthermore, the two large clusters are completely (uniformly) mixed up, which is reflected in the corresponding cluster indices: w_1, w_2, v_1, v_2, Aw_1, Aw_2, Av_1 and Av_2 are all substantially lower than unity.

The overall indices W, V and D are weighted averages of the w_i's and v_j's, and overall indices AW, AV and AR are weighted averages of the Aw_i's and Av_j's. As was the case in toy example 1, the weights associated with the two large clusters (0.44) are 7.33 times as high as the weights associated with the two small clusters (0.06). The values of the overall indices are therefore much closer to the values of the cluster indices associated with the large cluster (0.47 or 0.27) than the values of the indices associated with the small clusters (1.00). Thus, the values of the overall indices primarily reflect the degree of agreement on the two large clusters, which happens to be substantially lower than unity.

5.3 E. coli Sequences Example

Table 10 presents various row, column and overall statistics corresponding to Table 1, which is the matching table associated with the E. coli sequences data. Consider the row indices first. Most of the cp proteins are grouped together ($w_1 = 0.92$ and $Aw_1 = 0.88$). Many of the im proteins are grouped together ($w_2 = 0.79$ and $Aw_2 = 0.69$). None of the imL and imS proteins are grouped together ($w_3 = w_4 = 0.00$ and $Aw_3 = Aw_4 = -0.47$). Most of the imU proteins are grouped together ($w_5 = 0.94$ and $Aw_5 = 0.92$). Many of the om proteins are grouped together ($w_6 = 0.81$ and $Aw_6 = 0.72$). All of the omL proteins

Table 9 Row, column and overall statistics for the data in Table 8

Row statistics				Column statistics				Overall indices			
i	w_i	Aw_i	p_i	j	v_j	Av_j	q_j				
1	0.47	0.27	0.44	1	0.47	0.27	0.44	W	0.54	AW	0.36
2	0.47	0.27	0.44	2	0.47	0.27	0.44	V	0.54	AV	0.36
3	1.00	1.00	0.06	3	1.00	1.00	0.06	D	0.54	AR	0.36
4	1.00	1.00	0.06	4	1.00	1.00	0.06				

Table 10 Row, column and overall statistics for the data in Table 1

Row statistics				Column statistics				Overall indices	
i	w_i	Aw_i	p_i	j	v_j	Av_j	q_j		
1	0.92	0.88	0.67	1	0.57	0.41	0.11	W	0.88
2	0.79	0.69	0.19	2	0.54	0.37	0.02	V	0.75
3	0.00	– 0.47	$< .001$	3	0.92	0.89	0.57	D	0.81
4	0.00	– 0.47	$< .001$	4	0.51	0.33	0.31	AW	0.83
5	0.94	0.92	0.04					AV	0.65
6	0.81	0.72	0.01					AR	0.73
7	1.00	1.00	0.001						
8	0.79	0.68	0.09						

are grouped together ($w_7 = Aw_7 = 1.00$). Finally, many of the pp proteins are grouped together ($w_8 = 0.79$ and $Aw_8 = 0.68$).

The overall indices $W = 0.88$ and $AW = 0.83$ reflect that many of the proteins from the same class are grouped together in the K-means clustering. The overall values are weighted averages of the cluster indices associated with the rows of Table 1. The W value and AW value are almost completely determined by the values of the cluster indices associated with the two large classes, the cp and im proteins ($p_1 = 0.67$ and $p_2 = 0.19$). The values of the indices associated with the five smallest classes (imL, imS, imU, om, and omL) are basically immaterial for the calculation of the values of W and AW.

Next, consider the column indices. Since there are 8 classes of proteins and the K-means clustering consists of only 4 clusters, the recovery of the cluster structure as represented in the reference partition cannot be perfect. That is, some of the protein classes will be lumped together in the same cluster. The indices associated with clusters Z_1, Z_2 and Z_4 reflect that the clusters contain more than one type of protein ($v_1 = 0.57$, $v_2 = 0.54$, $v_4 = 0.51$, and $Av_1 = 0.41$, $Av_2 = 0.37$ and $Av_4 = 0.33$). Furthermore, the indices associated with cluster Z_3 tell us that at least one of the protein classes was recovered rather well by the K-means clustering ($v_3 = 0.92$ and $Av_3 = 0.89$).

The overall indices $V = 0.75$ and $AV = 0.65$ reflect that some proteins from different classes have been grouped together in the K-means clustering. Recall that these overall statistics are weighted averages of coefficients of individual clusters associated with the columns of Table 1. The V value and AV value are completely determined by the values of the cluster indices associated with the three large clusters Z_1, Z_3 and Z_4 ($q_1 = 0.11$, $q_3 = 0.57$ and $q_4 = 0.31$). The value of the index associated with the smallest cluster (Z_2) is not relevant for the calculation of the values of V and AV.

Finally, the Dice index $D = 0.81$ and the adjusted Rand index $AR = 0.73$ are harmonic means of, respectively, $W = 0.88$ and $V = 0.75$, and $AW = 0.83$ and $AV = 0.65$. Compared to the ordinary arithmetic mean of two numbers, the harmonic mean puts a bit more emphasis on the smallest of the two numbers. Therefore, the values of D and AR lie between, respectively, the values of W and V, and AW and AV, and just a little bit closer to the overall indices V and AV.

In summary, the three data examples show that the overall indices that belong to the families of indices based on the Wallace indices in (5) and (6) and the adjusted Wallace indices in (18) and (19) are quite susceptible to cluster size imbalance. The overall indices tend to mainly reflect the degree of agreement between the partitions on the large clusters. They provide little to no information on the degree of agreement on the smaller clusters.

6 Rand-Like Indices

6.1 Definitions

In addition to Wallace indices (5) and (6), we may consider the following two asymmetric indices. The first index

$$W^* = \frac{N+T-P-Q}{N-P} = \frac{d}{c+d} \tag{26}$$

is the proportion of object pairs not placed together in partition Z that are also not joined in partition U. The second index

$$V^* = \frac{N+T-P-Q}{N-Q} = \frac{d}{b+d} \tag{27}$$

is the proportion of object pairs not placed together in partition U that are also not joined in partition Z. The quantity $N + T - P - Q$ in the numerator of (26) and (27) is the number of pairs that are not joined in either of the partitions. As an indication of agreement between the partitions, this quantity is rather neutral, counting pairs that are not clearly indicative of agreement (Wallace, 1983).

Table 11 presents eight examples of indices that are increasing functions of the four conditional probabilities (5), (6), (26), and (27). For example, the well-known Rand index (Rand, 1971) is given by

$$R = \frac{N+2T-P-Q}{N} = \frac{a+d}{a+b+c+d}. \tag{28}$$

The Rand index is a weighted average of indices (5), (6), (26), and (27), using the denominators of the indices as weights:

$$R = \frac{WP + VQ + W^*(N-P) + V^*(N-Q)}{P+Q+N-P+N-Q}. \tag{29}$$

Furthermore, combining (29) with (12) and (13) we have the decomposition

$$R = \frac{\sum_{i=1}^{I} w_i P_i + \sum_{j=1}^{J} v_j Q_j + W^*(N-P) + V^*(N-Q)}{\sum_{i=1}^{I} P_i + \sum_{j=1}^{J} Q_j + 2N - P - Q}. \tag{30}$$

The decomposition in (30) shows that the Rand index can also been seen as a weighted average of the w_i's, v_j's and W^* and V^*, using the P_i's, Q_j's and $(N - P)$ and $(N - Q)$ as weights.

6.2 Numerical Examples for Rand-Like Indices

Indices that belong to the families of indices based on the Wallace indices in (5) and (6) and the adjusted Wallace indices in (23) and (24) can be understood in terms of indices for individual clusters (see Sections 3 and 4, respectively). However, this is quite different for the family of indices from Subsection 6.1. These Rand-like indices are increasing functions of the Wallace indices in (5) and (6) as well as the asymmetric indices in (26) and (27). The

Table 11 Indices that are increasing functions of (5), (6), (26) and (27)

Source	Formula 1	Formula 2
Sokal and Michener (1958), Rand (1971)	$(a+d)/(a+b+c+d)$	$R=$ formula (28)
Rogers and Tanimoto (1960)	$(a+d)/(a+2b+2c+d)$	$R/(2-R)$
Hamann (1961)	$(a-b-c+d)/(a+b+c+d)$	$2R-1$
Sokal and Sneath (1963)	$2(a+d)/(2a+b+c+2d)$	$2R/(R+1)$
Sokal and Sneath (1963)	$ad/\sqrt{(a+b)(a+c)(b+d)(c+d)}$	$\sqrt{WVW^*V^*}$
Sokal and Sneath (1963)	$a/4(a+b)+a/4(a+c)+ d/4(b+d)+d/4(c+d)$	$(W+V+W^*+V^*)/4$
Rogot and Goldberg (1966)	$a/(2a+b+c)+d/(b+c+2d)$	$WV/(W+V)+W^*V^*/(W^*+V^*)$
Warrens (2008a)	$4ad/(4ad+(a+d)(b+c))$	$4/(W^{-1}+V^{-1}+W^{*-1}+V^{*-1})$

Rand index, which is the prototypical example of this family, may be interpreted as the ratio of the number of object pairs placed together in a cluster in each of the two partitions and the number of object pairs assigned to different clusters in both partitions, relative to the total number of object pairs. Rand-like indices combine two sources of information, object pairs put together in both partitions, which is reflected in Wallace indices (5) and (6), and object pairs assigned to different clusters in both partitions, which is reflected in indices (26) and (27).

To understand what the values of Rand-like indices may reflect requires knowledge of how the two sources of information on object pairs contribute to the overall values of the indices. The above interpretation suggests that both sources may contribute equally. Results presented in Warrens and Van der Hoef (2020) show that this is not the case. In this paper it is shown how the Rand index (Rand, 1971) is related to the four asymmetric indices (5), (6), (26) and (27). Warrens and Van der Hoef (2020) systematically varied artificial data examples. The results of their study can be summarized as follows. In many situations, including cases of high, medium and low agreement between the partitions, and statistical independence of the partitions, the number of object pairs assigned to different clusters in both partitions is (much) higher than the number of object pairs that are combined in both partitions.

Decomposition (29) shows that the Rand index is a weighted average of the indices W, V, W^* and V^* using, respectively, the quantities P, Q, $(N-P)$ and $(N-Q)$ as weights. The results of Warrens and Van der Hoef (2020) have two consequences: 1) the values of W and V are usually (much) smaller than the values of W^* and V^*; 2) the values of P and Q are usually (much) smaller than the values of $(N-P)$ and $(N-Q)$. The second consequence implies that the value of the Rand index will in many cases in large part be determined by the values of W^* and V^*. Furthermore, together the two consequences imply that the Rand index will usually produce a high value, say between 0.70 and 1.00, because $(N-P)$ and $(N-Q)$, the weights associated with W^* and V^*, will in general also be high. Since all Rand-like indices presented in Table 11 are increasing functions of W, V, W^* and V^*, these indices will generally produce high values as well.

The results in Warrens and Van der Hoef (2020) can be illustrated with the data examples from the previous section. Table 12 gives the values of indices W, V, W^*, V^* and R and relative weights P/N, Q/N, $(N-P)/N$ and $(N-Q)/N$ for the three data examples from

Table 12 Values of indices and weights for the three data examples from Section 5

	Data examples		
Statistic	Toy 1	Toy 2	E. coli
W	0.93	0.54	0.88
V	0.93	0.54	0.75
W^*	0.97	0.82	0.89
V^*	0.97	0.82	0.95
R	0.96	0.74	0.89
P/N	0.28	0.28	0.27
Q/N	0.28	0.28	0.32
$(N-P)/N$	0.72	0.72	0.73
$(N-Q)/N$	0.72	0.72	0.68

Section 5. In all three examples the relative weights P/N and Q/N are much smaller than the relative weights $(N-P)/N$ and $(N-Q)/N$. Thus, in each example the value of the Rand index will be influenced more by the values of W^* and V^* than by the values of the Wallace indices W and V. Furthermore, in each example the values of W^* and V^* are quite high.

In summary, the Rand-like indices tend to reflect how much object pairs have been assigned to different clusters in both partitions. A first consequence is that they will generally produce high values (say between 0.70 and 1.00). A second consequence is that cluster size imbalance is less of an issue for these indices.

7 Discussion

7.1 Conclusions

For assessing agreement between two partitions researchers usually use and report overall measures that quantify agreement for all clusters simultaneously. Since overall indices only give a general notion of what is going on, their values are often hard to interpret. The goal of this study was to provide a more thorough understanding of the adjusted Rand index as well as many other indices based on counting object pairs. We analyzed three families of indices in this paper. We presented decompositions of the overall indices into indices that reflect the degree of agreement on the level of the individual clusters. The decompositions make explicit what the building blocks of the overall indices are and how they are weighted, and thus provide insight into what information the values of overall indices may reflect.

Indices that are based on the Wallace indices, for example, the Jaccard index and an index by Fowlkes and Mallows, or the adjusted Wallace indices, for example, the adjusted Rand index, are quite susceptible to cluster size imbalance. The importance of an (adjusted) cluster index in the computation of these overall indices is a (roughly) quadratic function of the size of the corresponding cluster. For example, an (adjusted) cluster index corresponding to a cluster that is twice as big as a second cluster will receive four times the weight of the cluster index corresponding to the second cluster. Thus, if cluster sizes differ, overall measures based on the (adjusted) Wallace indices will primarily reflect the degree of agreement between the partitions on the large clusters, and will provide much less information on the agreement on smaller clusters. The contribution of small clusters to the overall agreement will in many cases be small or even negligible, depending on how their size compares to the larger clusters.

Susceptibility to cluster size imbalance of various indices has previously been observed in the classification literature (De Souto et al., 2012; Fränti et al., 2014; Van der Hoef & Warrens, 2019). The analyses presented in this paper add some details to these studies by providing insight into how this phenomenon actually works, and to which indices it applies. The various indices are weighted means of cluster indices, and it is this weighting that introduces the susceptibility to cluster size imbalance.

In addition to the Wallace indices and adjusted Wallace indices, a third family of indices consists of Rand-like indices. These indices can be decomposed into a row and a column index that reflect how many object pairs are put together in both partitions, and into a row and a column index that reflect how many object pairs are put in different clusters in both partitions. They tend to reflect how much object pairs have been assigned to different clusters in both partitions. They will generally produce high values (say, between 0.70 and 1.00). Moreover, cluster size imbalance is less of an issue for these indices.

A negative property of the Rand index that has been noted in the classification literature is that its value concentrates in a small interval near the value 1 (Fowlkes & Mallows, 1983; Meilă, 2007). The analyses presented in Warrens and Van der Hoef (2020) and in this paper provide insight into how this property works. Furthermore, the analyses in this paper show that the property also applies to other Rand-like indices, that is, indices that are increasing functions of the same building blocks as the Rand index.

In this paper we focused on indices that are based on counting pairs of objects. This type of index, especially the adjusted Rand, is most commonly used. Some of the ideas presented in this paper can be applied to other types of partition comparison indices. For example, decompositions of various normalizations of the mutual information (Pfitzner et al., 2009) are presented in Van der Hoef and Warrens (2019) and Van der Hoef and Warrens (2020). It turns out that these information theoretic indices are also susceptible to cluster size imbalance, but in a more complicated way than the indices based on counting pairs of objects.

7.2 Practical Recommendations

Based on the findings in the literature and the results of this study, we have the following recommendations for studying agreement between two partitions. Since they provide much more detailed information than a single overall number, we generally recommend researchers to examine and report the adjusted indices for the individual clusters presented in (21) and (22). When there is a large number of clusters, reporting all cluster indices is perhaps not feasible. One solution here is to report the distribution of the values of the cluster indices for each partition. Another solution for this case is to summarize the cluster indices by counting how many cluster indices have a value above a certain threshold that reflects high agreement (say, 0.95) and all values below a certain number that indicates poor agreement (say, 0.50).

However, if a single number provides sufficient granularity to answer the research question(s), researchers can resort to an overall measure for quantifying agreement. When one uses indices that are based on the (adjusted) Wallace indices, for example, the adjusted Rand index, one should keep it mind that these indices are susceptible to cluster size imbalance. These overall indices will only reflect the ‘average’ agreement on all clusters of the partitions if the sizes of all clusters are approximately the same. However, if the cluster sizes differ, these indices will primarily reflect the degree of agreement between the partitions on the large clusters, and will provide much less information on the agreement on smaller clusters. The latter may not be desirable in all situations, for example, when the small clusters are the more interesting clusters.

Since cluster size imbalance is quite common in practice, it may be worthwhile to consider overall measures that are not susceptible to unbalanced cluster sizes (see, for example, Pfitzner et al. (2009) and Fränti et al. (2014)). Furthermore, the (adjusted) Wallace indices are susceptible to cluster size imbalance because their building blocks (cluster indices) are weighted in the computation. We obtain robust overall measures if we consider ordinary averages of cluster indices instead of weighted averages. For example, robust variants of the Wallace indices that are not susceptible to cluster size imbalance can be defined as

$$W^{\dagger} := \frac{1}{I}\sum_{i=1}^{I} w_i \tag{31}$$

and

$$V^{\dagger} := \frac{1}{J}\sum_{j=1}^{J} v_j. \tag{32}$$

The variants of the Wallace indices in (31) and (32) are, respectively, ordinary averages of the I cluster indices corresponding to the clusters of the reference partition (the w_i's) and the J cluster indices corresponding to the clusters of the trial partition (the v_j's). Furthermore, analogous robust variants of the adjusted Wallace indices are given by

$$AW^{\dagger} := \frac{1}{I}\sum_{i=1}^{I} Aw_i \tag{33}$$

and

$$AV^{\dagger} := \frac{1}{J}\sum_{j=1}^{J} Av_j. \tag{34}$$

Using the quantities in (31) and (32), and following the format of (7), a robust variant of the Dice index that is not susceptible to cluster size imbalance is given by

$$D^{\dagger} := \frac{2W^{\dagger}V^{\dagger}}{W^{\dagger}+V^{\dagger}}, \tag{35}$$

which is the harmonic mean of (31) and (32). Moreover, using the quantities in (33) and (34), a robust variant of the adjusted Rand index can be defined as

$$AR^{\dagger} := \frac{2AW^{\dagger}AV^{\dagger}}{AW^{\dagger}+AV^{\dagger}}. \tag{36}$$

The variant of the adjusted Rand index in (36) is the harmonic mean of (33) and (34).

Finally, we generally recommend against the use of the Rand index and the Rand-like indices considered in Section 6. The values of these indices are determined to a large extent by the number of pairs of objects that are not joined in either of the partitions, which is not clearly indicative of agreement (Wallace, 1983).

7.3 Future Research

The indices by Pfitzner et al. (2009) and Fränti et al. (2014), and the indices in (35) and (36), may potentially be better indices than the overall indices that are susceptible to cluster size imbalance. However, it should be noted that, compared to the adjusted Rand index, the indices by Pfitzner et al. (2009) and Fränti et al. (2014), and the indices in (35) and (36), have not been studied comprehensively and may have hidden limitations. A topic for future research could be an extensive study of the properties of these indices.

Whenever cluster sizes differ, the overall indices based on the (adjusted) Wallace indices will give more weight to the degree of agreement on the larger clusters, and less weight to the agreement on the smaller clusters. If the cluster sizes differ not too much the values produced by these overall measures may to some extent still reflect the 'average' agreement on all clusters of the partitions. In other words, there may be cases in which these overall indices, including the adjusted Rand index, are not too bad at reflecting the degree of agreement on small clusters. Identifying these cases could be another topic for future research. Such a study should consider different forms of cluster size imbalance, varying both the number of clusters and the relative sizes of the clusters.

Data Availability All data analyzed in this article are fully presented in the article.

Declarations The authors declare no competing interests. Furthermore, the research study did not involve human participants and/or animals.

References

Albatineh, A. N., & Niewiadomska-Bugaj, M. (2011a). Correcting Jaccard and other similarity indices for chance agreement in cluster analysis. *Advances in Data Analysis and Classification*, *5*(3), 179–200.

Albatineh, A. N., & Niewiadomska-Bugaj, M. (2011b). MCS: A method for finding the number of clusters. *Journal of Classification*, *28*, 184–209.

Albatineh, A. N., Niewiadomska-Bugaj, M., & Mihalko, D. (2006). On similarity indices and correction for chance agreement. *Journal of Classification*, *23*(2), 301–313.

Alok, A. K., Saha, S., & Ekbal, A. (2014). Development of an external cluster validity index using probabilistic approach and min-max distance. *International Journal of Computer Information Systems and Industrial Management Applications*, *6*, 494–504.

Anderson, D. T., Bezdek, J. C., Popescu, M., & Keller, J.M. (2010). Comparing fuzzy, probabilistic, and possibilistic partitions. *IEEE Transactions on Fuzzy Systems*, *18*, 906–917.

Baulieu, F. B. (1989). A classification of presence/absence based dissimilarity coefficients. *Journal of Classification*, *6*(1), 233–246.

Braun-Blanquet, J. (1932). *Plant sociology: The study of plant communities*. New York: Authorized English translation of Panzensoziologie. McGraw-Hill.

Brun, M., Sima, C., Hua, J., Lowey, J., Carroll, B., Suh, E., & Dougherty, E.R. (2007). Model-based evaluation of clustering validation measures. *Pattern Recognition*, *40*, 807–824.

Chacón, J. E. (2019). A close-up comparison of the misclassification error distance and the adjusted Rand index for external clustering evaluation. arXiv:1907.11505.

Chacón, J. E., & Rastrojo, A. I. (2020). Minimum adjusted Rand index for two clusterings of a given size. arXiv:2002.03677.

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

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37–46.

De Souto, M. C. P., Coelho, A. L. V., Faceli, K., Sakata, T. C., Bonadia, V., & Costa, I.G. (2012). A comparison of external clustering evaluation indices in the context of imbalanced data sets. Brazilian Symposium on Neural Networks, pp. 49–54.

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, *26*, 297–302.

Doolittle, M. H. (1885). The verification of predictions. *Bulletin of the Philosophical Society of Washington*, *7*, 122–127.

Driver, H. E., & Kroeber, A. L. (1932). Quantitative expression of cultural relationship. *The University of California Publications in American Archaeology and Ethnology*, *31*, 211–256.

Dunn, J. (1974). Well separated clusters and optimal fuzzy partitions. *Cybernetics*, *4*, 95–104.

Fleiss, J. L. (1975). Measuring agreement between two judges on the presence or absence of a trait. *Biometrics*, *31*, 651–659.

Fowlkes, E. B., & Mallows, C. L. (1983). A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association*, *78*, 553–569.

Fränti, P., Rezaei, M., & Zhao, Q. (2014). Centroid index: Cluster level similarity measure. *Pattern Recognition*, *47*, 3034–3045.

Gleason, H. A. (1920). Some applications of the quadrat method. *Bulletin of the Torrey Botanical Club*, *47*, 21–33.

Halkidi, M., & Batiskis, Y. (2002). Cluster validity methods: Part I. *SIGMOD Record*, *31*, 40–45.

Hamann, U. (1961). Merkmalsbestand und Verwandtschaftsbeziehungen der Farinose. Ein Betrag zum System der Monokotyledonen. *Willdenowia*, *2*, 639–768.

Heiser, W. J., & Warrens, M. J. (2010). Families of relational statistics for 2×2 tables. In H. Kaul, & H. Mulder (Eds.) *Advances in interdisciplinary applied discrete mathematics* (pp. 25–52). Singapore: World Scientific.

Hennig, C., Meilă, M., Murtagh, F., & Rocci, R. (2015). *Handbook of cluster analysis*. New York: Chapman and Hall/CRC.

Horton, P., & Nakai, K. (1996). A probablistic classification system for predicting the cellular localization sites of proteins. Intelligent Systems in Molecular Biology, pp. 109–115.

Hubálek, Z. (1982). Coefficients of association and similarity based on binary (presence absence) data: An evaluation. *Biological Reviews*, *57*, 669–689.

Hubert, L. J. (1977). Nominal scale response agreement as a generalized correlation. *British Journal of Mathematical and Statistical Psychology*, *30*, 98–103.

Hubert, L. J., & Arabie, P. (1985). Comparing partitions. *Journal of Classifications*, *2*(1), 193–218.

Huo, Z., Ding, Y., Liu, S., Oesterreich, S., & Tseng, G. (2016). Meta-analytic framework for sparse K-means to identify disease subtypes in multiple transcriptomic studies. *Journal of the American Statistical Association*, *111*, 27–52.

Jaccard, P. (1912). The distribution of the ora in the Alpine zone. *The New Phytologist*, *11*, 37–50.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, *31*(8), 651–666.

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, *32*, 241–254.

Kim, E.-Y., Kim, S.-Y., Ashlock, D., & Nam, D. (2009). MULTI-K: Accurate classification of microarray subtypes using ensemble k-means clustering. *BMC Bioinformatics*, *10*, 260.

Kulczyński, S. (1927). Die P anzenassociationen der Pienenen. *Bulletin Interna- tional de l'académie Polonaise des Sciences et des Letters, Classe des Sciences Mathematiques et Naturelles, Serie B, Supplément II*, *2*, 57–203.

Kvalseth, T. O. (1987). Entropy and correlation: Some comments. *IEEE Transactions on Systems, Man and Cybernetics*, *17*(3), 519–519.

Legendre, P., & Legendre, L. (1998). *Numerical ecology*. Amsterdam: Elsevier.

Lei, Y., Bezdek, J. C., Chan, J., Vinh, N., Romano, S., & Bailey, J. (2016). Extending information-theoretic validity indices for fuzzy clustering. *IEEE Transactions on Fuzzy Systems*, *25*(4), 1013–1018.

Lichman, M. (2013). UCI Machine Learning Repository. Retrieved from http://archive.ics.uci.edu/ml.

Loevinger, J. A. (1947). A systematic approach to the construction and evaluation of tests of ability. Psychometrika, Monograph No. 4.

McConnaughey, B. H. (1964). The determination and analysis of plankton communities. Marine Research, Special No, Indonesia, pp. 1–40.

Meilă, M. (2007). Comparing clusterings. an information based distance. *Journal of Multivariate Analysis*, *98*(5), 873–895.

Meilă, M. (2016). Criteria for comparing clusterings. In C. Hennig, M. Meilă, F. Murtagh, & R. Rocci (Eds.) *Handbook of cluster analysis* (pp. 619–636). New York: Chapman and Hall/CRC.

Milligan, G. W. (1996). Clustering validation: Results and implications for applied analyses. In P. Arabie, L. Hubert, & G. De Soete (Eds.) (pp. 341–375). River Edge: World Scientific.

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.

Ochiai, A. (1957). Zoogeographic studies on the soleoid fishes found in Japan and its neighboring regions. *Bulletin of the Japanese Society for Fish Science*, *22*, 526–530.

Pfitzner, D., Leibbrandt, R., & Powers, D. (2009). Characterization and evaluation of similarity measures for pairs of clusterings. *Knowledge and Information Systems*, *19*, 361–394.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, *66*(3), 846–850.

Rogers, D. J., & Tanimoto, T. T. (1960). A computer program for classifying plants. *Science*, *132*, 1115–1118.

Rogot, E., & Goldberg, I. D. (1966). A proposed index for measuring agreement in test-retest studies. *Journal of Chronic Disease*, *19*, 991–10.

Severiano, A., Pinto, F. R., Ramirez, M., & Carriço, J.A. (2011). Adjusted Wallace coefficient as a measure of congruence between typing methods. *Journal of Clinical Microbiology*, *49*, 3997–4000.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 623–656.

Simpson, G. G. (1943). Mammals and the nature of continents. *American Journal of Science*, *241*, 1–31.

Sokal, R. R., & Michener, C. D. (1958). A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, *38*, 1409–1438.

Sokal, R. R., & Sneath, P. H. (1963). *Principles of numerical taxonomy*. San Francisco: W. H. Freeman and Company.

Sørenson, T. (1948). A method of stabilizing groups of equivalent amplitude in plant sociology based on the similarity of species content and its application to analyses of the vegetation on Danish commons. *Kongelige Danske Videnskabernes Selskab Biologiske Skrifter*, *5*, 1–34.

Sorgenfrei, T. (1958). Molluscan Assemblages From the Marine Middle Miocene of South Jutland and Their Environments. Copenhagen: Reitzel.

Steinley, D. (2004). Properties of the Hubert-Arabie adjusted Rand index. *Psychological Methods*, *9*(3), 386–396.

Steinley, D. (2006). K-means clustering: A half-century synthesis. *British Journal of Mathematical and Statistical Psychology*, *59*, 1–34.

Steinley, D., Brusco, M. J., & Hubert, L.J. (2016). The variance of the adjusted Rand index. *Psychological Methods*, *21*(2), 261–272.

Steinley, D., Hendrickson, G., & Brusco, M.J. (2015). A note on maximizing the agreement between partitions: A stepwise optimal algorithm and some properties. *Journal of Classification*, *32*, 114–126.

Ting, K. M. (2011). Sensitivity and specificity. In C. Sammut, & G. Webb (Eds.) *Encyclopedia of machine learning*. Boston: Springer.

Van der Hoef, H., & Warrens, M. J. (2019). Understanding information theoretic measures for comparing clusterings. *Behaviormetrika*, *46*, 353–370.

Van der Hoef, H., & Warrens, M. J. (2020). Understanding Malvestuto's normalized mutual information. In T. Imaizumi, A. Okada, S. Miyamoto, F. Sakaori, Y. Yamamoto, & M. Vichi (Eds.) *Advanced Studies in Classification and Data Science* (pp. 289–299). Springer.

Van der Maarel, E. (1969). On the use of ordination models in phytosociology. *Vegetatio*, *19*, 21–46.

Vinh, N. X., Epps, J., & Bailey, J. (2009). Information theoretic measures for clusterings comparison: Is a correction for chance necessary? In *Icml '09 proceedings of the 26th international conference on machine learning* (pp. 1073–1080). New York: ACM.

Vinh, N. X., Epps, J., & Bailey, J. (2010). Information theoretic measures for clustering comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, *11*, 2837–2854.

Wallace, D. (1983). Comment on a method for comparing two hierarchical clusterings. *Journal of the American Statistical Association*, *78*, 569–576.

Warrens, M. J. (2008a). Bounds of resemblance measures for binary (presence/absence) variables. *Journal of Classification*, *25*, 195–208.

Warrens, M. J. (2008b). On similarity coefficients for 2×2 tables and correction for chance. *Psychometrika*, *73*(3), 487–502.

Warrens, M. J. (2008c). On the equivalence of Cohen's kappa and the Hubert-Arabie adjusted Rand index. *Journal of Classification*, *25*(2), 177–183.

Warrens, M. J. (2008d). On the indeterminacy of resemblance measures for binary (presence/absence) data. *Journal of Classification*, *25*, 125–136.

Warrens, M. J. (2008e). *Similarity coefficients for binary data: Properties of coefficients, coefficient matrices multi-way metrics and multivariate coefficients (Unpublished doctoral dissertation)*. Leiden: Leiden University.

Warrens, M. J. (2019). Similarity measures for 2×2 tables. *Journal of Intelligent and Fuzzy Systems*, *36*, 3005–3018.

Warrens, M. J., & Van der Hoef, H. (2020). Understanding the Rand index. In T. Imaizumi, A. Okada, S. Miyamoto, F. Sakaori, Y. Yamamoto, & M. Vichi (Eds.) *Advanced Studies in Classification and Data Science* (pp. 301–313). Springer.

Yu, Z., You, J., Wong, H.-S., & Han, G. (2012). From cluster ensemble to structure ensemble. *Information Sciences*, *198*, 81–99.

Yule, G. U. (1912). On the methods of measuring the association between two attributes. *Journal of the Royal Statistical Society*, *75*, 579–652.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.