Characterizing and comparing external measures for the assessment of cluster analysis and community detection

Nejat Arinik, Vincent Labatut, Rosa Figueiredo

Characterizing and comparing external measures for the assessment of cluster analysis and community detection

Nejat Arınık, Vincent Labatut & Rosa Figueiredo

January 29, 2021

Abstract

In the context of cluster analysis and graph partitioning, many external evaluation measures have been proposed in the literature to compare two partitions of the same set. This makes the task of selecting the most appropriate measure for a given situation a challenge for the end user. However, this issue is overlooked in the literature. Researchers tend to follow tradition and use the standard measures of their field, although they often became standard only because previous researchers started consistently using them. In this work, we propose a new empirical evaluation framework to solve this issue, and help the end user selecting an appropriate measure for their application. For a collection of candidate measures, it first consists in describing their behavior by computing them for a generated dataset of partitions, obtained by applying a set of predefined parametric partition transformations. Second, our framework performs a regression analysis to characterize the measures in terms of how they are affected by these parameters and transformations. This allows both describing and comparing the measures. Our approach is not tied to any specific measure or application, so it can be applied to any situation. We illustrate its relevance by applying it to a selection of standard measures, and show how it can be put in practice through two concrete use cases.

Keywords: Cluster analysis, Community detection, External evaluation measures, Regression.

Cite as: N. Arinik, V. Labatut & R. Figueiredo. Characterizing and comparing external measures for the assessment of cluster analysis and community detection, in *IEEE Access*. DOI: 10.1109/ACCESS.2021.3054621

1 Introduction

The problem of comparing two partitions of the same set occurs in a number of situations, the most widespread being probably the assessment of clustering (or cluster analysis) and community detection (or graph partitioning) results. In this context, one has computed the clusters of a dataset, or the community structure of a network. This result takes the form of a partition of the set of data points or of set of nodes, respectively. One then wants to compare this estimation with some ground-truth also taking the form of a partition. Alternatively, one has computed several such estimations, and wants to compare them to each other.

This comparison is traditionally performed through some measure able to quantify the similarity between two such partitions. In the context of cluster analysis, these are called *external* measures, as they allow comparing the output of the clustering method to an independent solution (generally the ground truth). In the rest of this article, we will simply call them *measures*, as there is no possible confusion in our context. Examples of such measures include *Adjusted Rand Index* (ARI) [24], Normalized Mutual Information [56] and so on. There are many ways to formalize what one means by "similar", resulting in the proposition of a very large number of measures over the years [39, 60]. In turn, this situation inevitably leads to the publication of a number of surveys aiming at reviewing and comparing all these measures [59].

In the literature, authors proposing new external measures follow a relatively standard workflow. First, they list some mathematical properties which they deem desirable in such measures, e.g. not being sensitive to the number of clusters k [15, 45, 49]. They then show that existing measures do not possess these properties. Finally, they solve this issue by proposing a new measure having these properties, or modifying an existing one to this end.

There are mainly two ways to check whether a measure has a given property. The most robust approach is to proceed analytically, through a mathematical proof (e.g. [38]). However, this task requires

certain skills, and can be difficult or even impossible depending on the considered measure and property. Moreover, the proof is generally not transposable to other measures and properties, which makes it a one-shot effort. This is why the second approach, which is empirical, is much more frequent in the literature (e.g. [46, 49]). It consists in applying some predefined transformations to certain partitions, both designed in a way that is related to the property of interest, and to study how the measure reacts to these perturbations by using it to compare those partitions. For instance, to assess the sensitivity to k, one could increase the number of clusters in the transformed partition, and check how this affects the measure values.

Each application case is likely to bring its own constraints and requirements, so there is no such thing as a "best" measure that would fit *all* situations. One trait considered as positive in one case could very well be perceived as a drawback in another. However, due to the profusion of available measures, selecting the most appropriate one for a given situation is a challenge for the end user. As mentioned before, some survey articles try to compare them, but they focus only a small number of measures [59] and/or properties [2]. More importantly, the comparisons they perform are specific to these measures and properties [59], preventing the end user from including additional measures or properties in the comparison. In practice, the problem of selecting an appropriate measure to compare partitions is generally overlooked, and researchers tend to follow tradition and use the measures frequently appearing in the literature of their field.

In this work, we propose a new framework to solve this issue. It is based on the empirical approach mentioned above, and consequently relies on a set of predefined partitions and parametric partition transformations. We study the effect of each parameter on the measure through multiple linear regression, in order to produce results that the end user can interpret. Our framework is not tied to any specific measure, property, or transformation, so it can be applied to any situation. We illustrate its relevance by applying it to a selection of popular measures, and show how it can be put in practice through two concrete use cases. In addition to these contributions, we review the literature for desirable properties and the partition transformations used to test their presence, and propose a typology of the latter.

The rest of the article is organized as follows. First, in Section 2, we review the literature on external measures, focusing on desired properties, partition transformations, and property assessment methods. Next, in Section 3, we introduce our own framework designed to study and compare measures and their properties. We put it into practice on a selection of widespread measures in Section 4 and discuss its results in Section 5. Moreover, we consider two use cases in Section 6 to further illustrate its relevance. Finally, we review our main findings in Section 7, and identify some perspectives for our work.

2 Literature Survey

In this section, we perform a review of the literature, focusing on three aspects directly related to our work. We first discuss the desirable properties used to characterize measures (Section 2.1). We then survey the partition transformations proposed to empirically show the absence or presence of these properties (Section 2.2). Finally, we give an overview of the evaluation methods used for assessing and comparing the measures based on these transformations (Section 2.3).

2.1 Desirable Properties

As mentioned in the introduction, measures can be characterized in terms of a number of distinct desirable properties. There are many of them, sometimes with minor differences, and the same property is likely to appear under different names and forms in the literature: this makes it difficult to list them exhaustively and compare them. Here, we focus on the most frequently used, and propose a typology to ease their comparison. These properties are listed in Table 1, with a short description, as well as examples of popular measures known to possess them. When the bibliographic sources explicitly name the property, we use the same name in the table. Otherwise, we propose a name based on its description. In the following, we distinguish three main categories, depending on whether the property is related to the *measure interpretation*, to the way it handles *random partitions*, and to its *sensitivity* to certain characteristics of the partitions.

2.1.1 Interpretation-Related Properties

The first category of properties is related to the interpretability of a measure, i.e. how easily its values can be understood by a human operator. This concerns the interpretation of a single value, i.e. what its magnitude means, but also the comparison of several values, and the interpretation of their difference.

Understandability [12, 39, 44, 50] means that the measure has a straightforward interpretation. For instance, the Rand index [47] (RI) is the proportion of element pairs for which both partitions agree. Other measures have less direct interpretations, for example the Standardized Mutual Information [50] (SMI) is a normalized version of the mutual information corresponding to the number of standard deviations the mutual information is away from the mean value, for a specific null distribution. At the other end of the spectrum, composite measures such as the F-measure [6] do not have a straightforward interpretation, as they combine other measures. This property is generally obtained by construction.

The *Fixed Range* property [59, 60, 62] means that the measure is designed so that its values are restricted to a predefined interval, which is often $[0, 1]$. This property eases the comparison of scores obtained on different partitions.

It is also the case of the *Value Validity* property [30]. Let m_1, m_2, m_3 and m_4 represent four numerical values obtained with some external measure for several pairs of partitions. In addition to order (size) comparisons such as $m_1 > m_2$, when a measure possesses the Value Validity property, the *difference* between several pairs of partitions, such as $m_1 - m_2 > m_3 - m_4$ or $m_1 - m_2 > k(m_3 - m_4)$ (for constant k), can also be interpreted.

Convex Additivity [37, 38] concerns the case where one partition is a refinement of another partition (i.e. there is a hierarchical relationship between them). With a measure possessing this property, the difference in overall score can be expressed as a weighted sum of the score differences between individual clusters.

2.1.2 Handling of Independent Partitions

The second category of properties focuses on how two independent partitions should be treated by the measure.

The *Constant Baseline* [1, 46, 50, 59, 60] property deals with statistical independence, i.e. the case where one compares two partitions sampled independently at random. This property specifies that in this situation, the measure should return a constant value. In practice, this constant value is very often zero, in particular when the maximal value is 1 (cf. also the *Fixed Range* property), see for instance the Adjusted Rand Index [24] (ARI).

The traditional approach to bring this property to an existing measure is to apply a so-called *correction for chance*. It consists in subtracting to the measure the score estimated for two independent partitions, and possibly in normalizing the resulting expression, in order to get a fixed range. This is how Hubert & Arabie derived their Adjusted Rand Index [24] (ARI) from the original Rand Index [47], but the method had been used before in other contexts [8, 19]. Note that there is a number of ways to define the null model used to estimate the measure score under the assumption of independent partitions [17], with no consensus emerging regarding which of these models is the most appropriate.

Certain authors consider two independent partitions as the worst possible case [45], meaning that the resulting score should correspond to the measure minimal value. On the contrary, others make a distinction between independence and worst case [18, 65], a property that is called *Baseline-Minimum Distinction*. They generally place the constant baseline midway between the respective scores of the worst and base cases. This is for instance the case of the ARI, which ranges from -1 to $+1$, 0 being the constant baseline. In practice though, cases with scores lower than the constant baseline are rare, and have not been studied much in the literature [65].

2.1.3 Sensitivity to Partition Characteristics

The last category of properties concerns the sensitivity of the measures to certain characteristics of the compared partitions. The main such characteristics are the number of clusters, the number of elements, the size of the clusters, and various descriptors allowing to express how similar the partitions are. These characteristics are often considered separately, and sometimes several at once.

In this category, the most frequent property is probably k*-invariance*. Certain measures such as the Normalized Mutual Information tend to favor partitions depending on the number of clusters they contain when compared with a reference partition [59], a bias that a number of authors want to avoid [4, 43, 51, 58, 60, 64]. For example, suppose that one compares a ground truth partition to two estimated partitions differing only in their number of clusters. A biased measure will reach a noticeably higher value for one of these partitions due to this single difference.

By comparison, the *Discriminativeness* property relies on the difference in number of clusters between the compared partitions [4, 23, 46]. It states that the measure score should decrease when this difference increases. Put differently, the score should be larger when both partitions contain similar numbers of clusters than when they differ on this point.

The n-*invariance* property is analogous to the k-invariance, except it is defined relative to the number of elements in the dataset [38, 60, 62], instead of the number of clusters. It allows comparing measure scores computed on datasets of different sizes, as n-invariant measures are not affected by such changes.

Authors do not agree on whether a measure should be sensitive or not to cluster size. This disagreement concerns partitions constituted of clusters which are imbalanced in terms of size, i.e. containing large and small clusters. Certain authors want the measure to focus mainly on the larger clusters, as they consider smaller ones as negligible [22, 40, 62]. Others adopt the *Insensitivity to Cluster Size* property and assume that all clusters are equally important regardless of their size, and that the measure should not be sensitive to cluster size imbalance [15, 45, 49].

Finally, some properties focus on how the measure should quantify the differences between pairs of partitions. Suppose we compare one primary partition to two different secondary partitions, resulting in two scores. The *Monotonicity* property states that the score of the most similar pair of partitions should be consistently higher or smaller (depending on whether the measure expresses similarity or dissimilarity) [18, 49, 63]. In addition, the *Proportionality* property states that the difference between these scores should be proportional to how close the secondary partitions are [34]. On the contrary, certain authors expect the measure score to rapidly change in presence of even the smallest differences, which corresponds to a non-linear behavior [45]. More generally, some authors want the measure to be *sensitive to small differences* [14, 38], whereas some others, on the contrary, want the measure to ignore what are considered as marginal differences [15]. It is important to stress that these are very generic properties, as the notion of proximity between two partitions can be understood in a number of ways.

2.1.4 Discussion

As explained in the introduction, and as summarized in Table 1, certain of the properties described in this section are obtained by construction, or verified through an analytical proof, whereas others are shown empirically, by applying specific transformations to a set of partitions. This is generally the case when the mathematical proof is impractical or too difficult to make.

In this article, we adopt an empirical approach, therefore we focus only on the latter type of properties. This includes the properties of our second (Comparison with Random Partitions) and third (Sensitivity to Partition Characteristics) categories. The framework that we propose does not necessarily handles the properties exactly as they are described here: we sometimes had to reformulate them to ease experiments and make the framework more generic. It relies on a set of variables similar to those used in the literature to define these properties (number of clusters, number of elements, cluster size distribution, etc.). Our framework is able to handle properties on which authors disagree, such as the sensitivity to cluster size distribution or to small differences.

2.2 Partition Transformations

Like for the desired properties, the literature exhibits a large number of different partition transformations, which are not always named, and when they are, not always similarly. This makes it difficult to identify and compare them. Here, we focus on the most frequent ones and use their most consensual names. Table 1 indicates the transformations used in the literature to assess the presence of each listed property. One can distinguish two types of partition transformations: random vs. deterministic.

2.2.1 Random Transformations

Random transformations consist in randomly distributing all the elements of the reference partition over a number of clusters to form the new partition. These transformations mainly differ in the probability distributions they rely upon. Such processes can be seen more as shuffling than transformations, as the original partition has no effect on the result. In essence, the goal is to obtain a partition as independent as possible from the original one. They are mainly used to check the existence of the Handling of Independent Partitions category of properties [4, 17, 18, 49–51, 58]. But several works leverage random transformations to look for other desirable properties, too. Certain authors force the shuffled partition to have various numbers of clusters and imbalanced cluster sizes, in order to check the k-invariance [4, 18, 51, 58, 64] and Insensitivity to Cluster Size [49, 61] properties, respectively. Others shuffle the original partition with an increasing level of randomness in order to test for the Monotonicity [18] and Proportionality [34] properties.

Table 1: Overview of the main desirable properties appearing in the literature, with examples of measures possessing them, and transformations used for their assessment.

Category	Desirable Property	Example measures	Related Transformations
Interpretation-Related	Fixed Range [59, 60, 62]	NMI [56], NVI [59]	▪ None (proof)
	Convex Additivity [38]	RI [47], Mirkin [41], VI [38], χ^2 distance [24]	▪ Splitting into unequal parts [38] (proof)
	Understandability [12, 38, 44, 50]	RI [39], JI [39], SMI [50], Split-Join [12]	▪ None (proof)
	Value Validity [30]	MI_c [30]	▪ None (proof)
Handling of Independent Partitions	Constant Baseline [1, 46, 50, 59, 60]	ARI [24], AMI [58], rNMI [64], RMI [43], FNMI [4], cNMI [32]	▪ Fragmenting every cluster [45] ▪ Random shuffling [4, 17, 18, 49–51, 58]
	Baseline-Minimum Distinction [18, 65]	ARI [24], SMI [50]	▪ Random shuffling [18]
Sensitivity to Partition Characteristics	k-invariance [4, 43, 51, 58, 60, 64]	ARI [24], VI [38]	▪ Random shuffling [4, 18, 51, 58, 64] ▪ Splitting into singleton clusters [43] ▪ Swap with single cluster [49]
	n-invariance [38, 60, 62]	VI [38], FMI [14], NMI [56], ARI [24]	▪ None (proof)
	Discriminativeness [4, 23, 46]	ARI [24], GNMI [4], FNMI [4]	▪ Merging whole clusters [4, 46] & Splitting into unequal parts [4, 46]
	Sensitivity to Small Differences [14, 15, 38, 45]	VI [38],FMI [14]	▪ Swap with all clusters [45]
	Insensitivity to Cluster Size [15, 45, 49]	PSI [49]	▪ Swap with single cluster & remove [49] ▪ Fragmenting a single cluster [49] ▪ Random shuffling [22, 61]
	Monotonicity [18, 49, 63]	PSI [49], Element-centric [18]	▪ Merging a whole cluster with a part of other cluster [49] ▪ Merging parts of different clusters [11, 52] ▪ Merging whole clusters & splitting into equal parts [63] ▪ Swap with single cluster [49] ▪ Swap with all clusters [49] ▪ Random shuffling [18]
	Proportionality [34, 38] vs. Non-linearity [45]	Kappa index [34]	▪ Random shuffling [34]

2.2.2 Deterministic Transformations

Deterministic transformations are used more frequently in the literature, probably because they offer a better control of the changes applied to the original partition. We distinguish five categories of such transformations. We call the first one *Remove*, and it consists in deleting some elements from a cluster without erasing it completely. Although it is used to check the Insensitivity to Cluster Size property in the literature [49], it has the drawback of affecting simultaneously two aspects of the partition: cluster size, and number of elements n. For this reason, it is not frequently used.

The second transformation category is *Split*, which consists in dividing a cluster into multiple smaller parts. Two variants mainly appear in the literature: splitting into *equal* [38, 39] vs. *unequal* parts [4, 38, 39, 46]. There is also a specific case of the first variant, consisting in splitting a cluster into only singleton clusters [43, 47, 48]. This transformation category is used in the literature to test several distinct properties. Hierarchical splits (i.e. refinements of a partition) constitute an important part of

the small experiments proposed by Meilă [38, 39], and allow to check the Convex Additivity property. Reichart and Rappoport [48] compare a reference partition to two estimated partitions differing mainly in their number of clusters: slightly perturbed reference vs. singleton clusters. They expect that singleton clusters are less similar to the reference, and a measure should not favor singleton clusters in such a case (cf. k-invariance property). Rabbany et al [46] apply repeated split operations onto the ground truth of several real-world networks and then compare them to check the Discriminativeness property.

Transformation *Merge* is the reciprocal of Split, as it gathers nodes belonging to different clusters into the same cluster. It also appears under three forms: merging a whole cluster with a *whole* other cluster [4, 38, 46, 63] vs. a *part* of another cluster [49], and merging parts of different clusters [52]. Note that the last two transformations are not *pure*, in the sense that a Split is performed before the Merge. Regarding the desirable properties, since Merge is the reciprocal of Split, all the properties tested through Split can be also be tested by using Merge. On top of that, some authors leverage Merge to test for *Monotonicity*, in two different ways: Rezaei and Fränti [49] enlarge incrementally a specific cluster by moving elements from the other clusters, whereas Rosenberg and Hirschberg [52] merge same-sized parts of each cluster to create new clusters, which they consider as *noise*.

The next two transformations can be viewed as combinations of Split and Merge, and they are also frequently used in the literature. *Swap* consists in interchanging a number of elements between pairs of (generally equal-sized) clusters. In practice, this operation is usually repeated for each cluster, using one of two different forms: each cluster swaps elements with *only one* different cluster [38, 49] vs. *all* other clusters [38, 45, 49]. In the literature, the first form is mainly used with a range of the number of clusters to check the k-invariance property. In the experiments of Rezaei and Fränti [49], the authors keep the cluster sizes fixed, independently from the number of clusters. However, this increases the number of total elements, which arguably introduces a side effect in their experiments. The second form induces more perturbation of the original partition compared to first one, and the experiments in the literature mainly focus on the desirable properties related to this aspect, which are Monotonicity [49] and Sensitivity to Small Differences [45].

Finally, the idea behind the *Fragment* transformation is that elements belonging to the same cluster in the original partition are placed in different clusters in the transformed partition, as much as possible. Two variants mainly appear in the literature: fragmenting a *single* cluster vs. *all* of them. The former [49] is only used to change marginally the underlying partition structure, whereas the aim of the latter [22, 45, 47] is to obtain two maximally different partitions. In the literature, these variants are used to check the Insensitivity to Cluster Size [22, 49] and Constant Baseline [45] properties, respectively.

2.2.3 Discussion

Besides these categories, the literature also contains transformations which can be expressed as combinations of some of these categories [46, 49, 63]. It is important to stress that transformations are typically defined *ad hoc*, specifically to test for a particular property of interest, and on some predefined partitions. For this reason, each author adopts a different angle, and it is hard to find two articles with the exact same methodology, targeting the exact same desired properties. In turn, this makes it difficult to compare transformations and measures from one paper to the other. To solve this issue, there is clearly a need for a unified view.

Another important limitation of the existing work is the lack of control over the original partition and its transformation. Some authors use a single parameter, for example the number of clusters in the transformed partition [11, 50]. However, there are other aspects likely to affect the outcome, such as the number and size of the clusters in the original partition, or the intensity of the transformation, and they are not taken into account simultaneously in the literature. This results in a relatively incomplete assessment of the measure properties.

In Section 3.1.2, we try to solve both these issues, by proposing a unified set of transformations designed to cover most of the literature, and by defining a set of parameters to get the appropriate level of control.

2.3 Assessment Methods

After having described the properties that authors want to find in partition comparison measures and the related partition transformations, we now turn to the methods used in the literature to check the presence or absence of these desired properties based on these transformations. We distinguish two families of approaches: visual inspection vs. statistical methods, more specifically correlation and regression.

2.3.1 Visual Inspection

Visual inspection is perhaps the most intuitive way to characterize the behavior of a measure. Typically, one plots the value of the measure as a function of some parameter used to control the partition transformation, e.g. the number of clusters produced. Authors usually expect a monotonic trend, e.g. proportional increase or decrease in [18]. Some are more specific and look for a specific pattern, e.g. the so-called *knee shape* used in [46] for a parameter controlling the number of clusters in the transformed partition. It requires the function to reach its maximum when the numbers of clusters in the original and transformed partitions match, and to decrease when there are too few or too many clusters in the transformed partition.

There are mainly two limitations to visual inspection. First, it is not an objective method, so limit cases can be difficult to judge. Second, it can handle only a very limited number of distinct parameters at once, especially if one wants to compare several measures and consider several properties, or assess how parameters interact. Statistical methods allow to solve the first issue, by providing an objective score. There are mainly two types of statistical tools used in the literature to assess measure properties: correlation and regression.

2.3.2 Correlation

A *correlation coefficient* quantifies the dependence between two random variables. In our context, and like with visual comparison, these variables are on the one hand the score computed with the measure of interest, and on the other hand a parameter controlling the partition definition or transformation. Many authors [23, 65] use the popular Pearson's product-moment correlation coefficient, which measures the linear dependence between the variables. Others use a rank-order correlation coefficient such as Kendall's (e.g. [62]) or Spearman's (e.g. [46]), which relies on the rank of the values rather than on the values themselves. Compared to Pearson's, such coefficients are able to detect a non-linear dependence, and can thus lead to different conclusions [46].

Besides objectivity, another advantage of correlation coefficients over visual inspection is that they summarize the dependence through a single value, which allows representing a number of pairwise relationships in a single table. However, this approach too becomes cumbersome when one wants to consider simultaneously a certain number of parameters and/or measures [11]. Moreover, multiple pairwise correlation values are not able to capture the potential interactions between the parameters (i.e. changing one parameter value may affect the partition or transformation feature controlled by another parameter).

2.3.3 Regression

Regression analysis does not suffer from this limitation, though. In its simplest form, it consists in describing the functional relation between a dependent variable and an independent variable [3]. In our context, those are the considered measure and a parameter of interest, respectively. However, multiple regression allows considering several independent variables at once, i.e. several parameters in our case. Another advantage over correlation is that the regression model can be used not only for interpretation, but also for prediction purposes [35].

To the best of our knowledge, the work of Saxena & Navaneetham [54] is the only one that uses multiple regression analysis to assess the similarity of external evaluation measures. The authors study the effects of three input parameters (cluster size, number of dimensions and number of clusters) on a single measure (the ARI). On top of the regression, they also assess the significance of these effects, and compare the relative importance of the parameters through their associated regression coefficients.

As we will see in Section 3.2.2, our method goes in the same direction as Saxena & Navaneetham [54], but with a more complex model, for the following reasons. First, the set of transformations that we propose in Section 3.1.2 requires to handle more parameters, and therefore to include more independent variables in the model. Second, not only do we study the direct effect of each parameter on the measure, but also their interactions. Third, we consider several distinct measures, and we want to assess and compare the relative importance of the effects that the parameters have on them, which requires a specific processing.

3 Proposed Framework

In this section, we describe the framework that we propose to analyze the behavior of a set of considered measures. It is independent from these measures, so we describe it in a generic way, for any selection

of measures.

Our framework is constituted of two parts. The first one consists in characterizing the considered measures through the partition transformation-based principle mentioned in the Introduction (Section 3.1). The second part is to perform an appropriate regression analysis in order to interpret these characteristics and compare the measures (Section 3.2).

3.1 Characterization of the Measures

Our objective is to quantify how similar two partitions are through several external measures, under different scenarios, and then to assess how the resulting values are affected when one of the partitions undergoes systematic and controlled changes. Unlike the common approach taken in the literature, we generate the necessary data in a fully parametric way in order to get a greater control. For the same reason, our approach is deterministic.

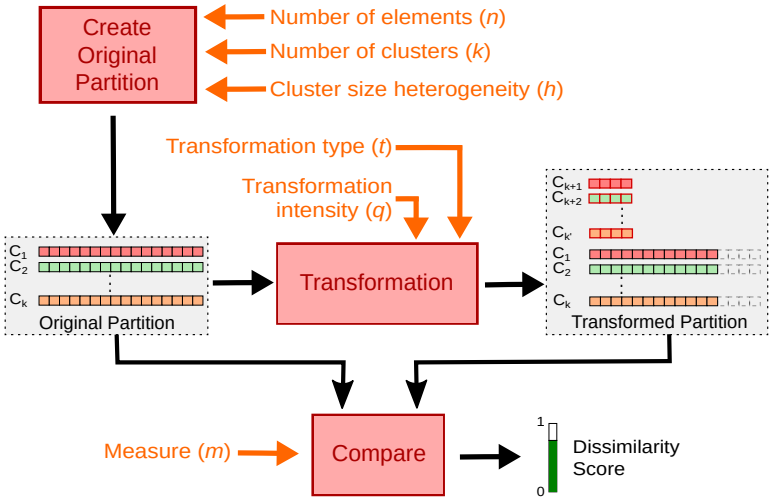

Figure 1: General Framework, with parameters represented in orange. For illustration purposes, the *k New Clusters* transformation is used to produce the output partition with k new clusters. Figure available at 10.6084/m9.figshare.13109813 under CC-BY license.

Our three-step method is summarized in Figure 1, and detailed in the rest of this section. The first step is to create a base partition, called *original partition*, and controlled by three parameters (Section 3.1.1). The second step consists in applying to this partition a transformation controlled by two other parameters (Section 3.1.2). This leads to a second partition, which we call *transformed partition*. Finally, the third step is to compute the selected external measures in order to assess how similar the original and transformed partitions are (Section 3.1.3). The whole process is repeated with an adequate number of different parameter values, in order to cover the parameter space.

3.1.1 Creating the Reference Partition

We control the generation of the reference partition through three parameters: the number of elements n, the number of clusters k and the heterogeneity of the cluster sizes h. The first two parameters allow to control the most basic aspects of the partition. These are frequently targeted in the literature, albeit not always through explicit parameters.

The last one is much more uncommon, and lets us control how much cluster sizes vary in the same partition, and therefore to get more realistic cluster sizes. Similar concepts appear in the literature, for example when dealing with balanced vs. imbalanced cluster sizes, but not under the form of such a convenient parameter, to the best of our knowledge. It ranges from 0 to 1. When $h = 0$, all clusters have the same size (i.e. so-called *balanced* cluster sizes), whereas they get imbalanced when $h > 0$, and the differences between their sizes increase when h gets closer to 1. More formally, the smaller cluster has a size of $s_1 = \alpha$ and the i^{th} smallest cluster has a size of $s_i = s_{i-1} + \beta$, whereas α and β depend on k, n and h. In particular, β is proportional to h. See Appendix C for details. This choice is a form of compromise allowing to obtain very heterogeneous cluster sizes even for a small n and/or a large k.

3.1.2 Applying the Parametric Transformations

After having generated the original partition at the previous step, we now want to change it in order to get the transformed partition. Based on our review of the existing work (Section 2.2), we propose a set of five parametric transformations aiming at fulfilling several constraints. We want to cover most of the transformations used in the literature, in order to deal with as many desired properties as possible, while keeping our transformations as simple (and thus interpretable) as possible and avoiding overlap between them. We discard *Remove*, as it changes n, which in our context is a parameter of the first step of our process. As mentioned before, all our transformations are deterministic in order to offer better control.

We note t the *nature* of the transformation, and use it later as a categorical variable during the regression analysis. We define a parameter q to specify the *intensity* of the transformation, i.e. the proportion of elements it involves. It ranges from 0, meaning no transformation at all, to 1, in which case the transformation involves all elements. We want to give the same importance to all clusters when applying the transformation, which means that it affects all of them. However clusters may have different sizes, depending on the heterogeneity of cluster sizes h. To deal with this situation, we make the number of elements concerned by the transformation in each cluster proportional to the cluster size.

The five transformations that we propose are illustrated in Figure 2, on two example reference partitions (Subfigure 2a). Both contain $n = 72$ elements, represented as numbered squares in the figure, and distributed over $k = 3$ clusters, represented by colors. However, the top partition is balanced ($h = 0$) whereas the bottom one is moderately imbalanced ($h = 0.5$). Each other subfigure shows the partitions resulting from a specific transformation with intensity $q = 1/6$. Note that all these transformations allow to test by construction whether or not a measure is sensitive to some framework parameters. On top of that, they can be used to test certain desirable properties from Section 2.1, as explained in the rest of this section and summarized in Table 2.

3.1.2.1 k New Clusters, t_{knc}

This transformation takes a predefined proportion of each cluster from the original partition, and creates a new cluster with these elements, resulting in k additional clusters (Subfigure 2b). The effect of this proportion on the transformed partition is mirrored in 0.5. For instance, transforming 40% and 60% of the elements give the same transformed partition. For this reason, we scale q so that it corresponds to twice this proportion, which allows us to keep the same $[0; 1]$ range as for the other transformations.

It is worth noting that the transformed partition is a subpartition of the original one, in the sense that each one of its clusters is included in one original cluster. Parameters k and h therefore affect the way the created subclusters relate to the original clusters. This transformation consequently allows testing for the Convex Additivity property, which states that a measure should not be affected when comparing refinements of the same partition. Concretely, we conclude that a measure has this property if it is not affected by k and h when applying this transformation.

3.1.2.2 Singleton Clusters, t_{sc}

All the elements affected by this transformation become singletons, i.e. single-element clusters (Subfigure 2c). This can be viewed as an extreme form of partition refinement, in the sense that each such singleton cluster is fully part of one of the original clusters. Therefore, like k New Clusters, but to a lesser extent, this transformation allows testing the Convex Additivity property through parameters k and h. Moreover, it allows checking the Sensitivity to Small Differences by considering the effect of parameter q. To be consistent with the nature of this property, it is necessary to focus on relatively small values of q (i.e. a limited transformation magnitude), though.

Parameter q can also be used to assess the Discriminativeness property, as increasing q largely increases the number of clusters in the transformed partition. Therefore, a measure which is affected by an increasing q is likely to discriminate more between transformed partitions whose number of clusters is closer to k (and hence to possess this property [46]). This is particularly true when the measure scores cover the whole $[0; 1]$ range. Parameter k can also be used, indirectly, to check the k-invariance property. Indeed, the number of clusters created by this transformation does not depend on k, and is generally much larger than k. So increasing k changes noticeably the number of clusters in the original partition, but not in the transformed one. Consequently, a measure which is marginally or never affected by changes in k when undergoing this transformation can be considered as k-invariant.

3.1.2.3 1 New Cluster, t_{onc}

Like the previous transformation, this one takes a proportion of each original cluster, but it gathers these elements to create a *single* cluster instead of k distinct ones (cf. Subfigure 2d). If we switch the original and transformed partitions, this transformation can alternatively be seen as the removal of a same-sized cluster, i.e. distributing proportionally the elements of a single cluster over the others. This is similar to the transformations used in [49] to test for the Insensitivity to Cluster Size property. In our case, if increasing k results in a substantial change in the measure score (all other things remaining equal), then this indicates that the measure is likely to treat the clusters equally, i.e. that it holds the property [49].

3.1.2.4 Neighbor Cluster Swaps, t_{ncs}

This transformation moves a proportion of each cluster into its neighbor cluster. Each cluster swaps elements with exactly one different cluster (cf. Subfigure 2e). Like for k New Clusters, the effect of this proportion on the transformed partition is mirrored in 0.5 for certain values of h. We therefore rescale it in the same way as before, in order to obtain a parameter q ranging from 0 to 1. This transformation allows to test for the Insensitivity to Cluster Size property through parameter h. By design, the number of clusters in the original and transformed partitions are the same. Hence, this transformation does not interfere with k and h. If increasing h has a substantial effect on the measure score, then this indicates that the measure is not likely to treat the clusters equally, i.e. it does not hold the property.

3.1.2.5 Orthogonal Clusters, t_{oc}

This transformation uses a proportion of each cluster to create new clusters, in such a way that all of their elements come from different original clusters (cf. Subfigure 2f). The resulting clusters are orthogonal to the original ones, in the sense that each original cluster is represented equally in the new clusters.

Applying this transformation with different values of k has an effect on the number of clusters in the transformed partitions, such that the difference in number of clusters between the original and compared partitions substantially decreases, when k increases. This is similar to the transformations used in [18]. The main difference is that the authors shuffle completely the transformed partitions, whereas this randomization process is tuned with the parameter q in our case. Therefore, like in [18], this transformation can be used, to a lesser extent, to test for the k-invariance property. If a measure is marginally or never affected by changes in k when undergoing this transformation can be considered as k-invariant.

Moreover, this transformation can also be used to check the Proportionality property with parameter q, as in [34]. If the scores of a distance measure increase linearly with increasing values of q, then we say that the measure validates this property. Finally, like in *Singleton Clusters*, the Sensitivity to Small Differences property can be also checked through small values of q, i.e. a limited transformation magnitude [45].

3.1.3 Computing and Normalizing the Measures

The third step is very straightforward and simply consists in computing the measures for each pair of partitions generated, in order to compare the reference partition with each transformed partition. Note that during the regression analysis, the measure of interest is considered as a categorical variable noted m.

In order for these values to be comparable, one has to make sure they respect two constraints, though. First, some measures of the literature quantify the similarity between two partitions, whereas others assess their *dis*similarity. For comparison purposes, all measures compared within our framework should express the same concept, be it similarity or dissimilarity. Without loss of generality, we assume in the rest of our framework that all considered measures are dissimilarity measures (possibly after having undergone an appropriate transformation).

Second, all measures are not necessarily defined on the same range, which means that some of them must be normalized in order to allow comparison. Many measures are defined on $[0; 1]$, so this seems like a consensual choice.

3.2 Regression Analysis

The second part of our framework consists in analyzing all the dissimilarity values obtained during the first part. In the following, we first introduce our proposed regression model (Section 3.2.1). We then

Property	Transformation & Parameter	Description
k-invariance	t_{sc} & k [43]	The measure is marginally affected by changes in k when undergoing this transformation.
	t_{oc} & k [18]	The measure is marginally affected by changes in k when undergoing this transformation.
Discriminativeness	t_{sc} & q [46]	Increasing q results in a substantial change in the measure score for this transformation.
Insensitivity to Cluster Size	t_{onc} & k [49]	Increasing k results in a substantial change in the measure score this transformation.
	t_{ncs} & h [49]	The measure is marginally or never affected by this transformation, for increasing h.
Convex Additivity	t_{sc} & k, h [38]	The measure is not affected by k or h for this transformation.
	t_{knc} & k, h [38]	The measure is not affected by k or h for this transformation.
Proportionality	t_{oc} & q [34]	The measure score increases proportionally with q.
Sensitivity to Small Differences	t_{oc} & q [45]	Even small values of q have a substantial effect on the measure score.
	t_{sc} & q	Even small values of q have a substantial effect on the measure score.

Table 2: The six desirable properties selected from Section 2.1, together with the framework parameters and transformations that allow testing them. The bibliographic references indicate matching situations from the literature, when available.

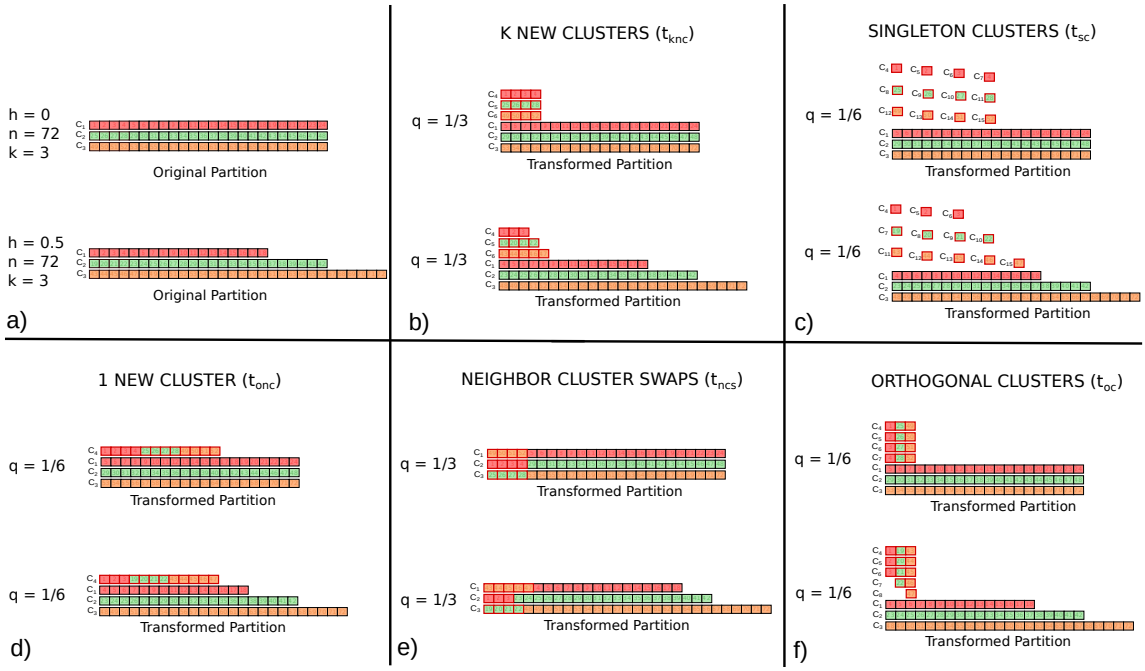

Figure 2: Parametric partition transformations used in our framework. Subfigure a) shows two reference partitions containing both $n = 72$ elements and $k = 3$ clusters, but differing on the heterogeneity of the cluster sizes: balanced ($h = 0$) vs. moderately imbalanced ($h = 0.5$). The 5 transformations, illustrated in Subfigures b)–f), are applied to these two original partitions to produce corresponding transformed partitions. Transformation intensity is $q = 1/6$, or equivalently $1/3$ for both transformations concerned with rescaling. Figure available at 10.6084/m9.figshare.13109813 under CC-BY license.

turn to relative importance analysis (Section 3.2.2), which aims at determining how much the framework parameters affect the measures depending on the applied transformations.

3.2.1 Model Design

In our context, the dependent variable is a dissimilarity score in $[0, 1]$, which we note y, whereas the independent variables correspond to the five parameters of the framework (n, k, h, q, t) and the nature of the measure used to compute the score (m). Four of them are therefore quantitative (n, k, h and q), and two are categorical (t and m).

We study the relation between these variables through a multiple linear regression model. Note that in this type of model, the linearity constraint concerns the regression coefficients, and not the independent variables. This means that independent variables can appear as polynomial terms in the model, and that the model can contain interaction terms corresponding to products of independent variables. There exist more complex types of regression models (e.g. polynomial regression), which could better fit our data. We chose to use a linear regression nevertheless, because it is much more interpretable [9], a property which is particularly important in our case.

The presence of *categorical* independent variables makes it necessary to adopt a specific approach, by comparison to a straightforward model including only numeric dependent variables, and there are several methods to do so [9]. Among them, we decide to use so-called *dummy variables*, as they allow to avoid splitting the model in several parts, which in turns makes it easier to compare the estimated regression coefficients [21].

Our multiple linear regression model is as follows

$$\begin{aligned} y = \sum_i \sum_j \Big(& \beta_{0ij} t_i m_j \\ & + \beta_{1ij} n t_i m_j + \beta_{2ij} k t_i m_j + \beta_{3ij} p t_i m_j + \beta_{4ij} h t_i m_j \\ & + \beta_{5ij} n k t_i m_j + \beta_{6ij} n h t_i m_j + \beta_{7ij} n p t_i m_j \\ & + \beta_{8ij} k h t_i m_j + \beta_{9ij} k p t_i m_j + \beta_{10ij} h p t_i m_j \Big) \\ & + \epsilon, \end{aligned} \tag{1}$$

where the $\beta_{.ij}$ are the regression coefficients, t_i and m_j are the dummy variables, and ϵ is the common error, which is assumed independent and normally distributed with mean 0 and standard deviation σ. Each dummy variable is binary, and represents one specific value of a categorical variable: transformations for t_i ($1 \leq i \leq T$) and measures for m_j ($1 \leq j \leq M$), where T (resp. M) is the number of transformations (resp. measures). The model focuses on various types of interactions between the independent variables. The second line contains terms describing interactions between the categorical variables and each *single* numeric variable. The third line deals, in addition, with interactions between *pairs* of quantitative variables. These terms are likely to introduce some amount of collinearity with the corresponding terms from the previous line. In order to solve this issue, we center all the quantitative independent variables [28]. In order to keep the model interpretable, we do not include any higher order term.

3.2.2 Relative Importance Analysis of Independent Variables

As this stage, we have a multiple linear regression model able to represent the relations between our framework parameters and the scores of the measures. Next, we want to assess the relative importance (also called relative strength [20] or effect size [57]) of the terms constituting our model.

In our context, *relative importance* refers to the contribution of an independent variable, by itself and in combination with other independent variables, to the prediction or the explanation of the dependent variable [26]. Such notion can be formalized in a number of ways, therefore several methods have been proposed [26], originating from different research fields. Nevertheless, they are designed with a common goal in mind, which is to handle both problems frequently occurring in multiple regression analysis and making this task challenging: 1) multi-collinearity between independent continuous variables; and 2) non-linearity of regression models. Since our independent variables are perfectly uncorrelated by design, and since we consider a purely linear model, all of these methods are relatively equivalent in our case. Therefore, we select the most straightforward approach, consisting in using squared standardized regression coefficients (SRC), or *squared β weights* [26, 42], to assess the relative importance.

When the regression terms are by design perfectly uncorrelated, zero-order correlations and β weights are equivalent [26]. Thus, squared β weights sum to the explained variance of the dependent variable [26], generally noted R^2. This implies that squared β weights can be used as a means of decomposing R^2 according to the terms of the model [42]. That is, a squared β weight close to zero makes a regression term less important, from which we can deduce that it does not play a key role in explaining the observed variance for the dependent variable y.

Having a similar beta weight is not sufficient to conclude that two terms have the same importance: the significance of their difference must be statistically tested [21]. In the presence of such significance we can confirm the superiority of the same variable in one transformation type (similarly, for one measure) over the others. The importance analysis framework includes this test for all pairs of β weights.

4 Experimental Setup

In order to illustrate how to use our framework and interpret its results, we now apply it to a selection of popular external measures. In this section, we define our experimental setup. We first describe briefly these measures (Section 4.1), before turning to the dataset and the regression assumptions (Section 4.2). The results are presented afterwards, in Section 5.

4.1 Selected Measures

In the literature, external measures are divided into three main categories based on the basic principle they rely upon [39, 60]: 1) Pair-counting, 2) Set-matching (or set overlaps) and 3) Information-theory. Among them, the pair-counting measures are the most studied ones. In line with this, for our experimental setup we select 6 widely used measures covering all three categories, with a prevalence of pair-counting measures. The formal description is given in the Appendix (Section B): in this section, we focus on the principle underlying these measures, as well as their similarities and differences.

A pair of elements can be handled in only two different ways in a given partition: either they belong to the same cluster or to two different clusters. *Pair-counting* measures are based on the idea of comparing how two partitions of the same dataset handle each pair of elements. For a given pair, there is *positive agreement* between the partitions if its elements belong to the same cluster in both partitions; *negative agreement* if they belong to different clusters in both partitions; and *disagreement* otherwise. The *Rand Index* (RI) [47] is the proportion of agreement relative to the total number of element pairs. Hubert and Arabie's *Adjusted Rand Index* (ARI) [24] is based on the RI, but additionally includes a *correction for chance*. The *Jaccard Index* (JI) was originally defined to compare sets [25], but it is also used as an external measure [7]. It completely ignores negative disagreements, as it corresponds to the proportion of positive agreements relative to the number of disagreements and positive agreements. The *Fowlkes-Mallows Index* (FMI) [14] also ignores negative agreements, as it is based on a score corresponding to the proportion of positive agreements relative to the number of pairs belonging to the same cluster *in one partition*. This score is computed separately for each one of the two compared partitions, and the Fowlkes-Mallows Index is the geometric mean of the resulting values.

To represent the category of *set-matching* measures, we select the F-measure (F). Note that this name is sometimes used in the literature as a synonym of *harmonic mean*, and therefore covers several distinct measures (e.g. [17, 46]). We use the definition of Artiles *et al.* [6], according to which the F-measure is the harmonic mean of two quantities called *Purity* and *Inverse Purity*. In order to compute the Purity of a cluster from the first considered partition, one needs first to identify the cluster from the second partition with which it has the largest intersection. The Purity then corresponds to the proportion of the first cluster which belongs to this intersection. The Purity of the first partition is the total purity of its clusters. The Inverse Purity is simply the Purity of the second partition relative to the first. Finally, the F-measure is the harmonic mean of the Purity and Inverse Purity.

Information-theoretical measures are generally based on the notion of *Mutual Information* [10]. The principle behind these measures is to consider each partition as a categorical random variable, whose possible values are the clusters. The mutual dependence between these variables can then be interpreted as the similarity between the partitions. There are a number of variants of the notion of mutual information, in particular several normalizations have been proposed (see for instance [59]). In this work, we focus on the sum normalization as defined in [56], which is very widespread, and results in the so-called *Normalized Mutual Information* (NMI).

As mentioned in Section 3.1.3, our framework expects that all measures express the same concept, either dissimilarity or similarity, and that they are all defined on the same fixed range. Regarding the latter point, all the selected measures are originally ranging from 0 to 1 except ARI, which can output negative values in theory. However, in practice it is very rare to get negative values for ARI. In the context of our experiments, it is always positive, so we decided not to perform any additional change. Regarding the former point, we adjust our selected measures through a simple subtraction, so that they all quantify the *dissimilarity* between partitions. We note the resulting measures as follows: D_{RI} (Rand Index), D_{ARI} (Adjusted Rand Index), D_{FMI} (Fowlkes-Mallows Index), D_{JI} (Jaccard Index), D_F (F-measure) and D_{NMI} (Normalized Mutual Information).

4.2 Dataset and regression assumptions

We generate our data through the process presented in Section 3.1, using the following parameter values. For the number of elements n, we choose values arithmetically compatible with the desired numbers of clusters, ranging from $3,240$ to $12,960$ with increments of $1,080$. The number of clusters k ranges from 2 to 11. The heterogeneity of the clusters size h ranges from 0 to 0.9 with increments of 0.1. Regarding the transformations, their intensity q ranges from 0.1 to 1, also by increments of 0.1, and the nature t of the transformation itself is one among t_{sc} (*Singleton Clusters*), t_{onc} (*1 New Cluster*), t_{knc} (k *New Clusters*), t_{ncs} (*Neighbor Cluster Swaps*), t_{oc} (*Orthogonal Clusters*), as defined in Section 3.1.2. In the end, the different combinations of our parameter values produce a total of $50,000$ pairs of partitions.

There are several standard assumptions to check before performing a linear regression [9, 20, 28]: 1) sufficient sample size, 2) linear relationships, 3) no or little multicollinearity, 4) multivariate normality, and 5) homoscedasticity. We review them here for our dataset and framework. First, our sample size of $50,000$ observations is large enough for getting reliable estimates of the regression. Second, after a visual inspection we observe that the relation between the dependent variable and the independent variables appear to be linear, except for k and q in which case it looks rather curvilinear. We stick to the linear model for the sake of readability and understandability, though. Third, by design of our dataset, the observations are independent and there is no collinearity between the independent variables. Fourth, the large size of our dataset makes the possible presence of outliers unlikely to affect our results [13]. For the same reason, the central limit theorem guarantees that the residuals will be approximately normally distributed. Fifth, a visual inspection reveals that the variance of y increases with parameters q (intensity of the transformation) and k (number of clusters), which means the data are not completely homoscedastic. The standard way of solving this issue is to introduce non-linear terms in the model [9, 20, 28], but again we want to keep it simple, and moreover the observed level of heteroscedasticity does not prevent us from interpreting the regression coefficients [20].

5 Results and Discussion

We now assess, compare and discuss the performance of the considered measures when applied to the generated dataset. We first show the relevance of our method through visual inspection (Section 5.1), then present our results in further detail (Section 5.2).

5.1 Visual inspection

To show the relevance of our method we highlight two aspects of our analysis through the visual inspection of Figure 3: 1) slope coefficients and 2) interaction between parameters. As we will see in Section 5.2, those aspects allow our method to identify similarities and differences between the considered measures, and therefore to discriminate between them.

Plot 3a shows how the measure scores evolve as functions of q, for the *Singleton Clusters* transformation, while the other parameters are fixed to arbitrary values. One can observe that all the scores increase with q, albeit in different ways. Overall, D_{RI} has the smallest slope coefficient, followed by D_{NMI}, and they are therefore the least sensitive to this transformation. We observe that D_{JI}, D_{ARI}, D_{FMI} and D_F get similar scores for extreme q values, but are relatively different when q gets closer to 0.5. Plot 3b is built upon the same principle, except it focuses on k instead of q. As before, all measures differ in terms of absolute score values, but this time one can detect similar certain trends. In particular, D_{JI}, D_{FMI} and D_F remain unchanged, whereas D_{RI}, D_{ARI} and D_{NMI} decrease with k. These two plots show that our framework is able to produce situations for which the measures behave differently. Moreover, they also show that the slope coefficients, which constitute the basis of our analysis, are able of capturing these differences.

As mentioned in Section 2.3.1, the common way to assess the performances of the measures is trough visual inspection, which requires fixing many parameters, as we did just now, as such plots are able to handle only a limited number of parameters at once. Plot 3c illustrates the limitation of this approach by showing the evolution of the D_{RI} score for the *1 New Cluster* transformation, as a function of both k and q. When considering only k, the D_{RI} score is always monotonic. However the nature and slope of the trend depend on q: increasing for $q \geq 0.7$ vs. decreasing for $q < 0.7$. This means that there is an interaction between both parameters. This type of joint effect between parameters is hard to detect when using only plots, as it requires considering all possible combinations of parameters. However, it is captured by the interaction terms present in our regression model, as we will see in Section 5.2.

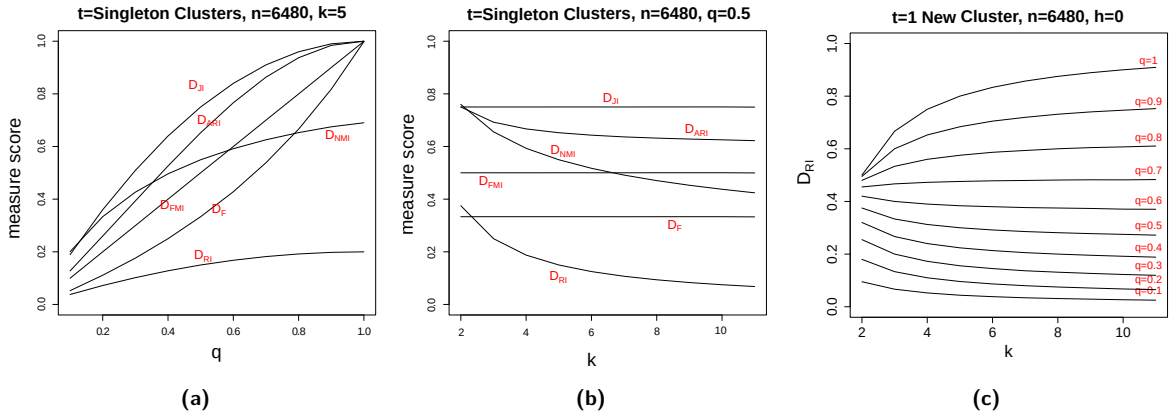

Figure 3: (a) Score of each measure as a function of p, for the *Singleton Clusters* transformation. (b) Score of each measure as a function of k, for the *Singleton Clusters* transformation. (c) D_{RI} score as a function of k, for the *1 New Cluster* transformation, and for several values of q. Figures available at 10.6084/m9.figshare.13109813 under CC-BY license.

5.2 Relative importance analysis

We first discuss the effect of the framework parameters on each measure (Section 5.2.1), and compare them. Along with our discussion, we identify the desirable properties possessed by each measure, as well. We then show how this analysis can be leveraged to derive a typology of the measures (Section 5.2.2).

5.2.1 Effect of the Parameters

We show all the results from our relative importance analysis in Figure 4, using stacked barplots. We describe these plots globally here, for matters of convenience, before interpreting them in the rest of this section. The figure contains 6 barplots (i.e. subfigures), each one corresponding to a specific dissimilarity measure. Each barplot is constituted of 5 stacked bars, each one corresponding to a different transformation. The segments constituting these stacked bars represent the regression terms from (1). Their colors and order match the legend, and their height corresponds to the associated regression coefficient β in (1). More precisely, the segment heights are proportional to the square root of the squared β coefficients.

The larger the segment height, the more important the regression term for the measure and transformation represented by the considered stacked bar. The values they represent are unitless, and we perform no upper bound normalization in order to ease comparisons between transformations and measures. Differences between segment heights are not always statistically significant, though. The exhaustive list of significant differences at p-value ≤ 0.05 is given in Appendix (Figures 5 and 6) for the sake of completeness. However, we find it difficult for the reader to cross-check them systematically with Figure 4. It is more intuitive to use the following rule of thumb: if one can visually detect a difference between two bars of Figure 4, then it is statistically significant.

Finally, there is a last bit of information in Figure 4, under the form of triangles placed over certain segments and representing monotonic behaviors. Upward (resp. downward) triangles indicate that the measure score consistently increases (resp. decreases) when the concerned parameter increases, independently from the other parameters. This information can be seen as complementary to the relative importance analysis. Suppose that a given parameter is similarly important for several measures, i.e. it affects them to roughly the same extend. The triangles allow distinguishing the measures qualitatively, based on the nature of this effect (see Section 6 for a practical example).

Overall, we can observe that all measures are strongly affected by q, and to a lesser extent by k and h. On the contrary, n has close to no effect on the measures. This effect of q on all measures also appears under a different form in Figure 3a. As shown by the triangles in Figure 4, the measure score increases with q in all cases. This general behavior is intuitively sound, as q controls the intensity of the transformation. There are differences, as illustrated in Figure 3a, in the way the measures are affected by q and the other parameters, though, and we can also see some punctual effects due to interactions between parameters. In the following, we consider each measure and discuss the results displayed in Figure 4.

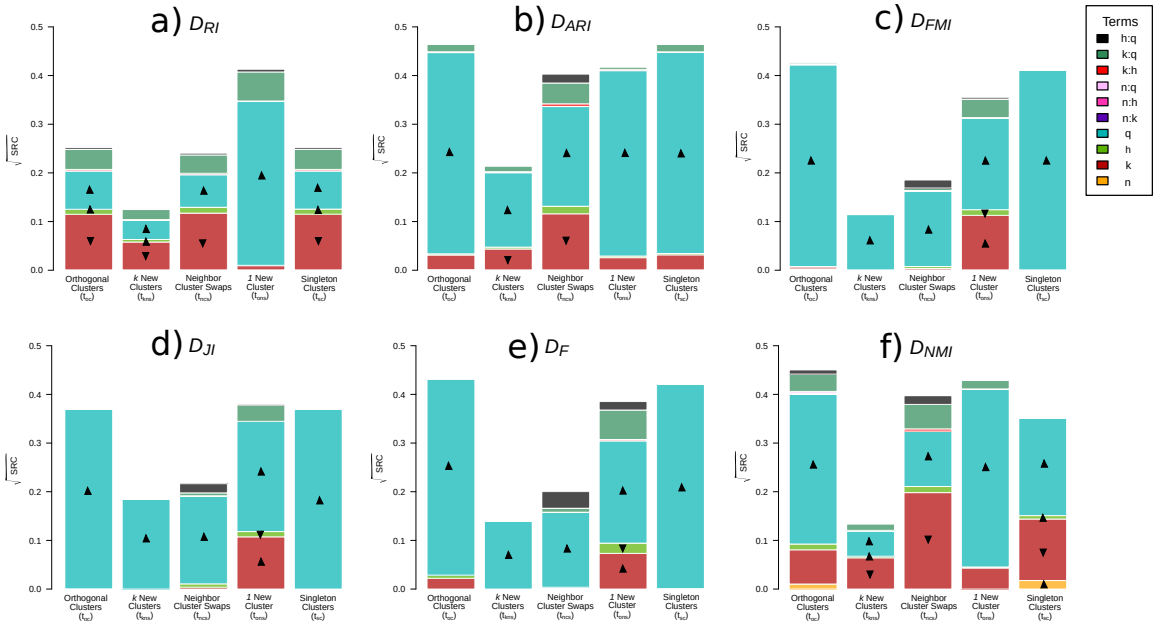

Figure 4: Results of the relative importance analysis, for measures a) D_{RI}, b) D_{ARI}, c) D_{FMI}, d) D_{JI}, e) D_F and f) D_{NMI}. The order of the terms in each bar is shown in the legend. The relative importance scores represented on the y-axis are square-roots. Upper (resp. lower) triangles indicate an increasing (resp. decreasing) trend of measure scores, when the corresponding parameter increases, independently of the values of the other parameters. Figure available at 10.6084/m9.figshare.13109813 under CC-BY license.

5.2.1.1 RI

We can distinguish roughly two categories of transformations regarding D_{RI}, depending on how the measure is affected by the parameters. The first category contains only *1 New Cluster*, for which we observe a sensitivity almost twice as important as for the other transformations, which is unique among the considered measures. Also, this transformation exhibits a very strong effect of q, and to a lesser extent of the interaction between q and k, as also illustrated in Figure 3c from a different perspective. The second category contains the rest of the transformations, for which parameter importance is more balanced between q, k, and their interaction.

As mentioned in Section 4.1, D_{RI} considers positive and negative agreements equally. When applying a transformation of the second category, an increase in q causes the number of positive agreements to decreases, whereas the negative agreements are largely preserved. This prevents D_{RI} from using its whole nominal range $[0, 1]$, as also pointed out by Meilă [39] and Vinh *et al.* [59]. This in turns explains the observed smaller effect of q. As explained in Section 3.1.2.2, we can infer from a small q effect for *Singleton Clusters* that D_{RI} does not possess the Discriminativeness property, a conclusion that confirms the results of Rabbany *et al.* [46].

On the contrary, there is a relatively substantial effect of k for the transformations of the second category, which is due to them largely preserving negative agreements, as already noticed for q. As explained in Section 3.1.2, the large effect of k for the *Singleton Clusters* and *Orthogonal Clusters* transformations indicates that the measure is not k-invariant. Similarly, the large effect of k for the *Singleton Clusters* and k *New Clusters* transformations indicates that it does not have the Convex Additivity property. These findings are in line with the results of Rabbany *et al.* [46] and Amelio & Pizzuti [4] regarding k-invariance, and Meilă [39] regarding Convex Additivity.

The relatively small effect of k for *1 New Cluster* shows that D_{RI} is sensitive to variations in the cluster sizes (cf. no Insensitivity to Cluster Size), as already pointed out by Rezaei & Fränti [49] for pair-counting measures. The absence of any significant effect of h in the results of *Neighbor Cluster Swaps* corroborates this finding. Regarding the remaining parameters, *1 New Cluster* is also the only transformation which seems not to be affected by h. Finally, n does not seem to affect D_{RI} at all.

5.2.1.2 ARI

Overall, q has a much stronger effect on D_{ARI} when compared to D_{RI}, which results in a total sensitivity approximately twice as large for all transformations except *1 New Cluster* (which is already large in D_{RI}). Based on the effects observed for *Singleton Clusters*, D_{ARI} seems to validate the Discriminativeness

property much more than D_{RI}.

The effect of k is much lower than in D_{RI}, for all transformations except *Neighbor Cluster Swaps*. According to certain results obtained by Meilă [39] for a similar transformation, the correction term in D_{ARI} can be sensitive to variations in the cluster number and sizes, which may explain our observation. Regarding *1 New Cluster*, the effect of k in D_{RI} is already small, and the correction present in D_{ARI} only slightly increases it. Therefore, D_{ARI} is sensitive to the variations of the cluster sizes, too (cf. Section 3.1.2.3).

The effect of k observed for *Singleton Clusters* and *Orthogonal Clusters* is much smaller than in D_{RI}, which indicates that the measure is k-invariant (Sections 3.1.2.2 and 3.1.2.5). This is consistent with the fact that D_{ARI} was designed specifically to make D_{RI} k-invariant, a property already verified empirically by Rabbany *et al.* [46]. However, this effect is still noticeable, which shows that the measure is not completely k-independent. Similarly, the effect of k for k *New Clusters* is smaller than in D_{RI} but still considerable. Based on these two observations, we can conclude that D_{ARI} does not possess the Convex Additivity property (Section 3.1.2.1).

The introduction of chance correction has a side-effect on h, as it has a much smaller effect on D_{ARI} compared to D_{RI}, for all transformations. This is consistent with a similar observation pointed out by Romano *et al.* [51]. Interaction-wise, the effect of k:q is much weaker than in D_{RI}, probably due to the lower overall effect of k, except for *1 New Cluster*. Finally, n does not seem to affect D_{ARI} at all.

5.2.1.3 FMI & JI

We jointly discuss both other pair-counting measures, D_{FMI} and D_{JI}, because their results are very similar and differ only on the magnitude of the effect of q. The main difference with the other measures is that q is the only perceptible effect for three transformations: *Orthogonal Clusters*, *k New clusters* and *Singleton clusters*. Consequently, both measures differ from the two previous ones regarding certain desirable properties. First, like D_{ARI} but unlike D_{RI}, both measures possess the Discriminativeness property. Second, unlike D_{RI} and D_{ARI}, they seem to validate the Convex Additivity property. It is worth stressing that, in theory, D_{FMI} and D_{JI} are not supposed to possess this last property [38], *strictly* speaking. However, our results show that in practice they behave as if they do, at least *to some extent*, and under some conditions (here: when the number of elements n is large enough).

The effect of k is negligible for all transformations but *1 New Cluster*, i.e. the second category of transformations previously identified for D_{RI}. These transformations affect only marginally negative agreement, which explains why the effect of k is so small here, compared to D_{RI}. This effect is small for *Singleton Clusters* and *Orthogonal Clusters*, so we can conclude that both measures appear to validate the *k-invariance property* (Sections 3.1.2.2 and 3.1.2.5). The strong effect of k for *1 New Cluster* indicates that these measures possess the Insensitivity to Cluster Size property (Section 3.1.2.3).

Regarding the other effects, one can observe that unlike D_{RI} and D_{ARI}, h has a small effect only for *1 New Cluster*. Furthermore, not only do k and q have a strong effect for this transformation, but their interaction does too. Finally, overall, n has no significant effect on both measures.

5.2.1.4 F-measure

Unlike the previous measures, which rely on pair-counting, D_F is based on set-matching. Nevertheless, the observed effects are very similar to those of D_{FMI} and D_{JI}. We observe essentially two differences. The first is that k and h have a relatively noticeable effect for *Orthogonal Clusters*. The second is that the effect of interaction h:q is stronger for *Neighbor Cluster Swaps* and *1 New Cluster*. D_F still validates the same properties as D_{FMI} and D_{JI} do, despite these small differences.

5.2.1.5 NMI

The results obtained for the information-theoretical measure D_{NMI} are very similar to those of D_{RI}, qualitatively speaking, and to those of D_{ARI}, in terms of magnitude of the effect observed for each transformation. Like D_{RI}, D_{NMI} behaves in the same way for all the four desirable properties, and this is consistent with the observations from the literature. For instance, Meilă [38] proves that the rescaling performed by some measures for normalization purposes, such as NMI, have the effect of breaking the Convex Additivity property. Moreover, Newman *et al.* [43], like others [4, 46, 59], show that NMI tend to favor partitions with more clusters when compared with a reference partition (cf. *no* k-invariance), and that this behavior can be smoothed by correcting *NMI* for chance.

A clear difference between D_{NMI} and all the other measures is that n has a very visible effect for *Orthogonal Clusters* and *Singleton Clusters*. This seems to be an artefact of the normalization for

these transformations, which would match the observation made by Amelio & Pizzuti [4], rather than a violation of the n-invariance property. Indeed, the information-theoretic measures are n-invariant by construction [38].

5.2.1.6 General Observations

For the sake of clarity, we roughly summarize in Table 3 the discussion that takes place throughout the current section regarding the presence or absence of desirable properties within the considered measures. We observe that three measures validate all 4 properties (D_F, D_{JI}, D_{FMI}), whereas two measures have none of them (D_{RI}, D_{NMI}). The last one, D_{ARI}, holds an intermediary position, as it possesses the k-invariance and Discriminativeness properties like D_F, D_{JI} and D_{FMI}, whereas it shares the same behavior with D_{RI} and D_{NMI} regarding Insensitivity to Cluster Size and Convex Additivity.

Let us now conclude this section by highlighting the main observations we could draw from the relative importance analysis. First, it is important to stress that the results produced by our framework are consistent with those published in the literature, including both theoretical and empirical works. This is summarized in Table 3. Second, the systematic nature of our approach helps uncovering properties not already described in the literature. For instance, Rezaei & Fränti [49] state that set matching measures are more suitable regarding the Insensitivity to Cluster Size property. Nevertheless, we find out that the pair-counting measures D_{JI} and D_{FMI} also possess this property. Third, our framework allows us to state that some measures possess certain properties at least partially, or under certain conditions. Indeed, our framework does not predict the presence of a property in a Boolean way, but rather on some continuous spectrum, through regression. Put differently, instead of predicting whether a measure has a property or not, we can estimate *how much* it possesses this property, and assess how this can change depending on the parameter values. For instance, as mentioned above, we can say that D_{ARI} validates the Discriminativeness property much more than D_{RI}, based on the effect of q for *Singleton Clusters*.

	k-invariance (t_{sc} and t_{oc} with k)	Discriminativeness (t_{sc} with q)	Insensitivity to Cluster Size (t_{onc} with k, t_{ncs} with h)	Convex Additivity (t_{sc} and t_{knc} with k and h)
D_{RI}	✗ [4, 39, 46]	✗ [46]	✗ [49, 55]	✗ [39]
D_{ARI}	✓ [46]	✓ [46]	✗ [49, 55]	✗ [39]
D_{FMI}	✓ [18]	✓	✓	✓
D_{JI}	✓ [18]	✓ [46]	✓	✓
D_F	✓	✓	✓ [55]	✓
D_{NMI}	✗ [4, 18, 43, 46, 59]	✗ [4, 46]	✗ [49, 55]	✗ [39]

Table 3: Relations between four desirable properties and the considered measures, based on our results presented in Figure 4. The method used to check whether a measure has a property is summarized between parenthesis in the first line, and additional details can be found in Section 3.1.2. The bibliographic references show matching observations found in the literature, when available.

5.2.2 Typology of Measures

We now show how a typology of the measures can be built based on the results shown in Figure 4, through a cluster analysis. First, we compute a distance matrix comparing all pairs of stacked bars constituting the plots from this figure. For this purpose, we represent each stacked bar by a vector of proportions, each value corresponding to a term of the regression model (i.e. a segment of the stacked bar). We use the Hellinger distance [33], which was designed to compare pairs of discrete probability distributions. Second, we perform the cluster analysis by applying the k-medoids method [27] to our distance matrix. This method requires us to specify the desired number of clusters, though. To find the most appropriate number, we apply the standard approach consisting in performing the clustering using all possible values, and then selecting the most appropriate one. For this purpose, we use the Silhouette measure, a well-known internal criterion [53], but we also take into account a more subjective constraint of parsimony (i.e. we want a small number of clusters).

The analysis results in 5 clusters of stacked bars, for a Silhouette of 0.55. Table 4 shows the distribution of the bars from Figure 4 over these clusters, each one being represented as a specific color. The blue cluster corresponds to bars in which there is a relatively balanced main effect of k and q, and a minor effect of h and k:q. In the brown cluster, the situation is quite similar but q supersedes k. In the red cluster, q even more prevalent, and both minor effects are even smaller. The orange cluster contains bars in which all effects are negligible compared to q. Finally, bars from the green cluster are dominated by q and exhibit a minor effect of h:q.

	Orthogonal Clusters	k New Clusters	Neighbor Cluster Swaps	1 New Cluster	Singleton Clusters
D_{RI}					
D_{ARI}					
D_{FMI}					
D_{JI}					
D_F					
D_{NMI}					

Table 4: Comparison of the measures based on the characterization provided by our framework and shown in Figure 4. We use the Hellinger distance and k-medoids to identify groups of similar behaviors, each one being represented by a color in the table.

Table 4 shows that each transformation produces a different vertical pattern, which indicates that the transformations we selected in our framework are not redundant in the way they allow characterizing the measures. The measures can be compared using the horizontal patterns present in the table. Roughly speaking, there is a first group constituted of D_{FMI}, D_{JI}, D_F; a second containing D_{RI} and D_{NMI}; and D_{ARI} is apart. We see that this characterization is consistent with the results in Table 3. The fact that these groups of measures, which are automatically obtained, match the ones identified manually based on our knowledge of the desired properties, indicates that this clustering-based method could be useful when the user is not able to (or does not want to) express their desired properties *a priori*. Indeed, for a given collection of available measures, this method allows identifying clusters of measures possessing a similar behavior: these clusters can then be characterized *a posteriori*, and the user can select a measure from the cluster considered as the most appropriate to the considered application.

To sum up, not only does our analysis allows distinguishing the effects of the framework parameters over transformation types and measures, but it also makes it possible to categorize the measures based on their empirical behavior. Our results confirm the findings of Pfitzner *et al.* [45], which indicate that the categorization of the measures based on their sole definitions (cf. Section 4.1) does not necessarily hold when it comes to comparing them through experiments.

6 Practical cases

In practice, an external evaluation measure is usually needed in two situations frequently occurring in the context of cluster analysis or community detection. In the first, one wants to compare an estimated partition to a partition of reference. This typically happens when one has applied some algorithm in order to estimate a partition of their data, and wants to quantify how similar it is to some available ground truth partition. In this context, the measure is used to assess the performance of the partitioning method. In the second situation, there is no reference partition involved: one wants to compare two estimated partitions. For instance, one has access to several partitions and wants to assess them in the absence of any ground truth. These partitions could either result from the application of several distinct partitioning methods to the same data, or from the application of single method able to output several solutions for the same input data. In this context, one would use a measure to assess how similar these partitions are, in order to check whether the methods reach a relative consensus.

The external measure has a central role in both situations, as different measures are likely to result in very different outcomes. The choice of an appropriate measure depends on a number of factors, including the broad situation, but also the nature of the application at hand and other contextual aspects such as the behavior expected by the user. In particular, it is worth stressing that not all transformations and parameters are relevant in all cases.

In the following, we illustrate all these aspects through two use cases, each one corresponding to one of the two broad situations described above. First, we treat the partitioning of the well-known cluster analysis method k-means, in a case where the ground truth is known (Section 6.1). Second, we turn to the cluster analysis of a set of estimated partitions, in the context of a study aiming at analyzing votes at the European Parliament (Section 6.2).

6.1 Comparing Estimated Partitions with Ground-Truth

In order to illustrate the comparison of some estimated partitions with the ground truth, we leverage the work of Fränti & Sieranoja. In [16], they study the behavior of k-means, and more precisely how this clustering method is affected by certain properties of the considered data. To this end, they propose a benchmark constituted of various artificially generated datasets together with their associated ground truth. Certain properties of these data are controlled through a set of parameters. Among these, some

are only related to the k-means algorithm (e.g. number dimensions of the data, spatial overlap) and not to the problem of partition comparison, so we ignore them in the rest of our discussion. Fränti & Sieranoja want to assess how changes in the properties controlled by the parameters affect the algorithm performance. For this purpose, they use the ARI and the Centroid Index (CI). The latter is a clustering comparison measure defined by Fränti *et al.* in a previous article [15]. It focuses on *global* partition differences concerning the number of clusters, by opposition to what the authors call *point-level differences*, i.e. *local* differences concerning the cluster borders. However, it was designed specifically to handle centroid-based clustering methods, which is why it is not part of the selection of measures we study in Section 4.1.

Let us now suppose that one wants to use a measure selected among the one discussed in Section 4.1. We can leverage the objectives of Fränti & Sieranoja as described in [16], as well as their methodology, to infer which measure behavior is desirable in terms of our own framework. First, the main parameters used to *directly* control the ground truth partitions in [16] are the number of objects, which corresponds to our parameter n, and the number of clusters, which is the same as our k. Although it is not controlled by a specific parameter, they also consider datasets with various levels of cluster size imbalance, a feature related to our h. The authors compare scores produced on data obtained by using different values of these parameters. For these comparisons to be relevant, it is necessary that these parameters affect the measure as little as possible, in order for it to reflect only changes in algorithm performance.

It appears clearly in the article that, for the authors, incorrectly estimating the number of clusters is the most serious error that k-means can make, by opposition to so-called point-level errors which concern only cluster borders. The transformations of our framework which are the most relevant to this situation are therefore those that change the number of clusters, i.e. all of them but *Neighbor Cluster Swaps*. Moreover, due to the nature of the considered data and clustering method, it is very unlikely to see singleton clusters appear in the considered partition (this would require the presence of very eccentric outliers). Therefore, transformation *Singleton clusters* is not relevant in this situation. This leaves us with *1 New Cluster*, k *New Clusters* and *Orthogonal Clusters*.

Let us now assess the relevance of the measures studied in Section 4.1 with respect to the criteria we identified. Based on our results from Section 5.2, we can identify two categories of measures in this situation. First, D_{RI}, D_{ARI} and D_{NMI} are sensitive to k, especially for k *New Clusters* and *Orthogonal Clusters*, and to a lesser extent, to h. They differ on *1 New Cluster*, as D_{RI} is much less sensitive to these parameters when considering this transformation. The second category contains D_{FMI}, D_{JI} and D_F, which exhibit sensitiveness to k and h only for *1 New Cluster*, and *Orthogonal Clusters* in the case of D_F. In conclusion, the second category is more appropriate to the situation described in [16], with a preference for D_{JI} which, overall, is less sensitive to the parameters of interest than the D_{ARI} used in the original study.

6.2 Comparing Estimated Partitions

We now illustrate the case of comparing several estimated partitions with each other. In [5], Arinik *et al.* study voting data from the European Parliament (EP), in order to identify voting patterns, i.e. how Members of the EP (MEP) are split in various factions depending on the topic of the considered legislative texts. Formally, they model the MEPs' voting behavior as a multiplex signed graph, and perform community detection on each layer to identify so-called voting patterns, i.e. partitions of the set of MEPs. This specific application brings specific constraints on the partitions, which can contain at most three communities: 1) a single community in case of unanimity (all MEPs vote either *For* or *Against* the legislative text); 2) two communities when there is either an antagonistic situation (i.e. some MEPs support the concerned document and the rest oppose it), or a unanimous community with an additional community of abstentionists; 3) two antagonistic communities with an additional community of abstentionists.

Arinik *et al.* identify the partition of MEPs (voting pattern) associated to each legislative text in their corpus. They then use an external measure to compute the dissimilarity between each pair of partitions, and perform a cluster analysis in order to identify groups of similar patterns, which they finally discuss relatively to the application context. In order to select the most appropriate measure, they adopt a qualitative approach consisting in identifying some partitions of interest and comparing how different measures behave when comparing them. Among D_{RI}, D_F, D_{ARI} and D_{NMI}, they conclude that D_F and D_{RI} are the most appropriate for their situation, with a slight advantage to D_F.

We propose to use our results from Section 5.2, and in particular from Figure 4, to solve the same measure selection problem, but based on the method presented in this article. Note that, in the following, we use the term *cluster* instead of *community*, for the sake of consistency with the rest of the article. It is important to stress that some parameters and transformations are not relevant here, due to the application context. First, the case of $k = 1$ is not applicable for some transformations.

Therefore, we apply all our transformations if there is more than one cluster in the original partition. For *k New Clusters* and *Singleton Clusters*, this means that we get at least four clusters in the transformed partition. This is incompatible with the fact that all the compared partitions of this application contain at most three clusters, so we exclude both transformations. Second, in this context, the *Orthogonal Clusters* transformation can be applied only when there are two clusters in the original partition, and only one element in each cluster is affected by the transformation. In this case, this transformation results in the same transformed partition as with *1 New Cluster*, therefore we also exclude *Orthogonal Clusters*.

This leaves us with two transformations. The first is *1 New Cluster*, which we apply only when the original partition has two clusters, in which case the transformation produces an additional cluster in the transformed partition. The second is *Neighbor Cluster Swaps*, which we apply only when the original partition has either two or three clusters, but not a single one. For both these transformations, there is no constraint on parameters h and q. However, as explained above, our analysis must focus only on certain values of k. Finally, in this context all the considered partitions contains the same number of elements, which means n is fixed and can therefore be ignored in our discussion.

Next, based on the description given by Arinik *et al.* of what they consider to be an appropriate measure for their application needs, we express the desired behavior of the measure with respect to the remaining parameters and transformations. The measure must be sensitive to the *1 New Cluster* transformation as, in this context, detecting an extra cluster or missing one is an important error, since there are only a few possible clusters of MEPs. When k increases, so does the diversity of the cluster created by this transformation, in the sense that its elements come from more distinct original clusters. In this context, this is an important difference with the original partition, so we want the score of the measure to increase with k. By comparison, it is desirable that the dissimilarity score decreases when h increases, as this means most elements of the extra cluster come from the same original cluster, an error which is less serious. For the same reason, the effect of k should be stronger than that of h.Transformation *Neighbor Cluster Swaps* consisting in mixing the original clusters to get the transformed partition without changing the number of clusters makes the clusters more different, when q increases. In this application context, it is important that this type of difference between partitions is taken into account, so the measure must be sensitive to it. Changes in k and h do not affect the mixing much, so the measure score is expected to be largely independent from these parameters.

Let us now study which measures studied in Section 4.1 fit the constraints described above. Regarding transformation *1 New Cluster*, it appears that only D_{FMI}, D_{JI} and D_F behave appropriately. When considering transformation *Neighbor Cluster Swaps*, we can see that, even if it is a small one, k as an effect on D_{FMI} and D_{JI}. In conclusion, based on these observations, we would select D_F in this context, a choice that incidentally matches the one made through a more qualitative and heuristic method in [5].

7 Conclusion

In this article, we have presented a new evaluation framework to address the problem of selecting an appropriate measure to compare partitions. We want not only to compare measures, but also to produce results that the end user can easily interpret. For this purpose, based on our review of the literature, we designed a set of predefined partitions and parametric partition transformations in order to generate a benchmark dataset. Our two-step framework first computes the considered measures for these partitions, then conducts a regression and relative importance analysis to determine how the measures are affected by the transformations. We illustrated its relevance by applying it to a selection of standard measures. We showed that our framework allows identifying the desirable properties possessed by each measure. For some of them, our results confirm empirical and theoretical findings already published in the literature. For others, the systematic nature of our approach even uncovers properties not mentioned before in the literature. Furthermore, we propose a typology of the considered measures based on their characteristics. Overall, our results confirm the findings of Pfitzner *et al.* [45], which indicate that categorizing measures based on their mathematical definitions does not necessarily match experimental comparison. Finally, we demonstrated how our framework can be put in practice through two concrete use cases: comparing an estimated partition to a partition of reference, and comparing several estimated partitions with each other.

Our work could be extended in several ways. First, our method can be applied systematically to other external measures, for the sake of completeness. It is particularly important to include the recently proposed measures for an up-to-date comparison, which would prevent from following the tradition of using only well-established measures without regard for their relevance. Second, similar to the previous point, some new parametric transformations can be proposed to closely investigate the performance of

the measures on a specific subject. For instance, there is an important number of measures aiming at correcting Mutual Information for chance in the literature. Including some specific transformations could enable to concentrate more on the aspect related to the number of clusters. Finally, by proposing relevant parameters and transformations, our general method could be adapted to handle objects similar to partitions, such as covers, to compare overlapping clusters (e.g. [18, 23]), or edge-aware community similarity measures, to compare community structures while taking graph topology into account (e.g. [31, 46]).

Acknowledgment

The authors thank Thomas Opitz, Etienne Klein from INRAE PACA and Pierre-Michel Bousquet from LIA for their feedback and guidance on certain statistical points.

References

[1] A. N. Albatineh and M. Niewiadomska-Bugaj. "Correcting Jaccard and other similarity indices for chance agreement in cluster analysis". In: *Advances in Data Analysis and Classification* 5.3 (2011), pp. 179–200. DOI: 10.1007/s11634-011-0090-y.

[2] A. N. Albatineh, M. Niewiadomska-Bugaj, and D. Mihalko. "On Similarity Indices and Correction for Chance Agreement". In: *Journal of Classification* 23.2 (Sept. 2006), pp. 301–313. DOI: 10.1007/s00357-006-0017-z.

[3] E. C. Alexopoulos. "Introduction to multivariate regression analysis". In: *Hippokratia* 14.Suppl 1 (2010), pp. 23–28.

[4] A. Amelio and C. Pizzuti. "Correction for Closeness: Adjusting Normalized Mutual Information Measure for Clustering Comparison". In: *Computational Intelligence* 33.3 (2016), pp. 579–601. DOI: 10.1111/coin.12100.

[5] N. Arinik, R. Figueiredo, and V. Labatut. "Multiple partitioning of multiplex signed networks". In: *Social Networks* 60 (2020), pp. 83–102.

[6] Javier Artiles, Julio Gonzalo, and Satoshi Sekine. "The SemEval-2007 WePS Evaluation: Establishing a Benchmark for the Web People Search Task". In: *Proceedings of the 4th International Workshop on Semantic Evaluations*. SemEval '07. Prague, Czech Republic: Association for Computational Linguistics, 2007, pp. 64–69. URL: http://dl.acm.org/citation.cfm?id=1621474.1621486.

[7] Asa Ben-Hur, Andre Elisseeff, and Isabelle Guyon. "A stability based method for discovering structure in clustered data". In: *Pacific Symposium on Biocomputing 2002*. Ed. by R. B. Altman, A. K. Dunker, L. Hunter, and T. E. Klein. World Scientific, 2001, pp. 6–17. DOI: 10.1142/9789812799623_0002.

[8] J. Cohen. "A Coefficient of Agreement for Nominal Scales". In: *Educational and Psychological Measurement* 20 (1960), pp. 37–46. DOI: 10.1177/001316446002000104.

[9] J. Cohen, P. Cohen, S. G. West, and L. S. Aiken. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. 3rd Edition. Routledge, 2002. ISBN: 9780203774441. DOI: 10.4324/9780203774441.

[10] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. 2nd Edition. Wiley-Interscience, 2006. ISBN: 10 0-471-24195-4.

[11] B. E. Dom. "An Information-theoretic External Cluster-validity Measure". In: *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence*. UAI'02. Morgan Kaufmann Publishers Inc., 2002, pp. 137–145. ISBN: 1-55860-897-4. URL: http://dl.acm.org/citation.cfm?id=2073876.2073893.

[12] S. Dongen. *Performance Criteria for Graph Clustering and Markov Cluster Experiments*. Tech. rep. 4. Amsterdam, The Netherlands, The Netherlands: National Research Institute For Mathematics and Computer Science, 2000. DOI: 10.5445/IR/1000011477.

[13] J. J. Faraway. "Practical regression and ANOVA using R". Accessed on 07/2020, https://cran.r-project.org/doc/contrib/Faraway-PRA.pdf. 2002. URL: https://cran.r-project.org/doc/contrib/Faraway-PRA.pdf.

[14] E. B. Fowlkes and C. L. Mallows. "A Method for Comparing Two Hierarchical Clusterings". In: *Journal of the American Statistical Association* 78.383 (1983), pp. 553–569. DOI: 10.1080/01621459.1983.10478008.

[15] P. Fränti, M. Rezaei, and Q. Zhao. "Centroid index: Cluster level similarity measure". In: *Pattern Recognition* 47.9 (2014), pp. 3034–3045. DOI: 10.1016/j.patcog.2014.03.017.

[16] P. Fränti and S. Sieranoja. "K-means properties on six clustering benchmark datasets". In: *Applied Intelligence* 48.12 (2018), pp. 4743–4759. DOI: 10.1007/s10489-018-1238-7.

[17] Alexander J. Gates and Yong-Yeol Ahn. "The Impact of Random Models on Clustering Similarity". In: *Journal of Machine Learning Research* 18.1 (2017), pp. 3049–3076. ISSN: 1532-4435. URL: http://dl.acm.org/citation.cfm?id=3122009.3176831.

[18] Alexander J Gates, Ian B Wood, William P Hetrick, and Yong-Yeol Ahn. "Element-centric framework unifies overlaps and hierarchy". In: *Scientific Reports* 9.1 (2017). ISSN: 2045-2322. DOI: 10.1038/s41598-019-44892-y.

[19] L. A. Goodman and W. H. Kruskal. "Measures of Association for Cross Classification". In: *Journal of the American Statistical Association* 49.268 (1954), pp. 732–64. DOI: 10.2307/2281536.

[20] D. N. Gujarati. *Basic Econometrics*. Ed. by A. Bright. McGraw-Hill, 2003.

[21] M. Hardy. *Regression with Dummy Variables*. Quantitative Applications in the Social Sciences. SAGE Publications, Inc., 1993. ISBN: 9780803951280. DOI: 10.4135/9781412985628.

[22] H. van der Hoef and M. J. Warrens. "Understanding information theoretic measures for comparing clusterings". In: *Behaviormetrika* 46.2 (2019), pp. 353–370. DOI: 10.1007/s41237-018-0075-7.

[23] D. Horta and R. J. G. B. Campello. "Comparing Hard and Overlapping Clusterings". In: *Journal of Machine Learning Research* 16.93 (2015). Editor: Marina Meila, pp. 2949–2997. URL: http://jmlr.org/papers/v16/horta15a.html.

[24] L. Hubert and P. Arabie. "Comparing partitions". In: *Journal of Classification* 2.1 (1985), pp. 193–218. DOI: 10.1007/bf01908075.

[25] P. Jaccard. "Étude comparative de la distribution florale dans une portion des Alpes et des Jura". In: *Bulletin de la Société Vaudoise des Sciences Naturelles* 37.142 (1901), pp. 547–579. DOI: 10.5169/seals-266450.

[26] J. W. Johnson and J. M. Lebreton. "History and Use of Relative Importance Indices in Organizational Research". In: *Organizational Research Methods* 7.3 (2004), pp. 238–257. DOI: 10.1177/1094428104266510.

[27] L. Kaufman and P. J. Rousseeuw. "Partitioning Around Medoids". In: *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, 2009. DOI: 10.1002/9780470316801.ch2.

[28] M. H. Kutner, C. J. Nachtsheim, J. Neter, and W. Li. *Applied linear statistical models*. Ed. by B. Gordon. 5th Edition. McGraw-Hill Irwin, 2005. ISBN: 0-07-238688-6.

[29] O. T. Kvålseth. "Entropy and Correlation: Some Comments". In: *IEEE Transactions on Systems, Man, and Cybernetics* 17.3 (1987), pp. 517–519. DOI: 10.1109/tsmc.1987.4309069.

[30] O. T. Kvålseth. "On Normalized Mutual Information: Measure Derivations and Properties". In: *Entropy* 19.11 (2017), pp. 631–645. DOI: 10.3390/e19110631.

[31] V. Labatut. "Generalised measures for the evaluation of community detection methods". In: *International Journal of Social Network Mining* 2.1 (2015), pp. 44–63. DOI: 10.1504/ijsnm.2015.069776.

[32] D. Lai and C. Nardini. "A corrected normalized mutual information for performance evaluation of community detection". In: *Journal of Statistical Mechanics: Theory and Experiment* 2016.9 (Sept. 2016), p. 093403. DOI: 10.1088/1742-5468/2016/09/093403.

[33] L. M. Le Cam. *Asymptotic Methods in Statistical Decision Theory*. Springer-Verlag New York, 1986. ISBN: 978-1-4612-4946-7. DOI: 10.1007/978-1-4612-4946-7.

[34] X. Liu, H.-M. Cheng, and Z.-Y. Zhang. "Evaluation of Community Detection Methods". In: *IEEE Transactions on Knowledge and Data Engineering* 32.9 (2019), pp. 1736–1746. DOI: 10.1109/tkde.2019.2911943.

[35] P. Luo, H. Xiong, G. Zhan, J. Wu, and Z. Shi. "Information-Theoretic Distance Measures for Clustering Validation: Generalization and Normalization". In: *IEEE Transactions on Knowledge and Data Engineering* 21.9 (2009), pp. 1249–1262. DOI: 10.1109/tkde.2008.200.

[36] E. Marczewski and H. Steinhaus. "On a certain distance of sets and the corresponding distance of functions". In: *Colloquium Mathematicum* 6.1 (1958), pp. 319–327.

[37] M. Meilă. "Comparing Clusterings by the Variation of Information". In: *Learning Theory and Kernel Machines*. Ed. by Bernhard Schölkopf. Springer Berlin Heidelberg, 2003, pp. 173–187. ISBN: 978-3-540-45167-9. DOI: 10.1007/978-3-540-45167-9_14.

[38] Marina Meilă. "Comparing clusterings—an information based distance". In: *Journal of Multivariate Analysis* 98.5 (2007), pp. 873–895. DOI: 10.1016/j.jmva.2006.11.013.

[39] Marina Meilă. "Criteria for comparing clusterings". In: *Handbook of cluster analysis*. Ed. by C. Hennig, M. Meila, F. Murtagh, and R. Rocci. 1st Edition. Chapman and Hall/CRC, 2015. Chap. 27, pp. 619–635. ISBN: 9780367570408.

[40] G. W. Milligan and M. C. Cooper. "A Study of the Comparability of External Criteria for Hierarchical Cluster Analysis". In: *Multivariate Behavioral Research* 21.4 (1986), pp. 441–458. DOI: 10.1207/s15327906mbr2104_5.

[41] B. Mirkin. "Mathematical Classification and Clustering: From How to What and Why". In: *Classification, Data Analysis, and Data Highways*. Ed. by I. Balderjahn. Springer Berlin Heidelberg, 1998, pp. 172–181. ISBN: 978-3-642-72087-1. DOI: 10.1007/978-3-642-72087-1_20.

[42] L. L. Nathans, F. L. Oswald, and K. Nimon. "Interpreting multiple linear regression: A guidebook of variable importance". In: *Practical Assessment, Research and Evaluation* 17.9 (2012), pp. 1–19.

[43] M. E. J. Newman, G. T. Cantwell, and J. G. Young. "Improved mutual information measure for classification and community detection". In: *Phys. Rev. E* 101.4 (2020), p. 042304. DOI: https://doi.org/10.1103/PhysRevE.101.042304.

[44] J. J. O'Brien, M. T. Lawson, D. K. Schweppe, and B. F. Qaqish. "Suboptimal Comparison of Partitions". In: *Journal of Classification* 37.2 (2019), pp. 435–461. DOI: 10.1007/s00357-019-09329-1.

[45] D. Pfitzner, R. Leibbrandt, and D. Powers. "Characterization and evaluation of similarity measures for pairs of clusterings". In: *Knowledge and Information Systems* 19.3 (2008), pp. 361–394. DOI: 10.1007/s10115-008-0150-6.

[46] R. Rabbany, M. Takaffoli, J. Fagnan, O. R. Zaïane, and R. J. G. B. Campello. "Communities validity: methodical evaluation of community mining algorithms". In: *Social Network Analysis and Mining* 3.4 (2013), pp. 1039–1062. DOI: 10.1007/s13278-013-0132-x.

[47] W. M. Rand. "Objective Criteria for the Evaluation of Clustering Methods". In: *Journal of the American Statistical Association* 66.336 (1971), pp. 846–850. DOI: 10.1080/01621459.1971.10482356.

[48] R. Reichart and A. Rappoport. "The NVI Clustering Evaluation Measure". In: *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*. CoNLL '09. Boulder, Colorado: Association for Computational Linguistics, 2009, pp. 165–173. ISBN: 978-1-932432-29-9. URL: http://dl.acm.org/citation.cfm?id=1596374.1596401.

[49] M. Rezaei and P. Fränti. "Set Matching Measures for External Cluster Validity". In: *IEEE Transactions on Knowledge and Data Engineering* 28.8 (2016), pp. 2173–2186. DOI: 10.1109/tkde.2016.2551240.

[50] S. Romano, J. Bailey, N. X. Vinh, and K. Verspoor. "Standardized Mutual Information for Clustering Comparisons: One Step Further in Adjustment for Chance". In: *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*. ICML'14. Beijing, China: JMLR.org, 2014, pp. II-1143–II-1151. URL: http://dl.acm.org/citation.cfm?id=3044805.3045020.

[51] S. Romano, N. X. Vinh, J. Bailey, and K. Verspoorn. "Adjusting for Chance Clustering Comparison Measures". In: *Journal of Machine Learning Research* 17.134 (2016), pp. 1–32. URL: http://jmlr.org/papers/v17/15-627.html.

[52] Julia Bell Rosenberg Andrewand Hirschberg. "V-Measure: A conditional entropy-based external cluster evaluation". In: *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)* (2007), pp. 410–420. DOI: 10.7916/d80v8n84.

[53] P. J. Rousseeuw. "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis". In: *Journal of Computational and Applied Mathematics* 20 (Nov. 1987), pp. 53–65. DOI: 10.1016/0377-0427(87)90125-7.

[54] P. C. Saxena and K. Navaneetham. "The Effect of Cluster Size, Dimensionality, and Number of Clusters on Recovery of True Cluster Structure Through Chernoff-Type Faces". In: *Journal of the Royal Statistical Society (the Statistician)* 40.4 (1991), pp. 415–425. DOI: 10.2307/2348731.

[55] M. C.P. de Souto, A. L.V. Coelho, K. Faceli, T. C. Sakata, V. Bonadia, and I. G. Costa. "A Comparison of External Clustering Evaluation Indices in the Context of Imbalanced Data Sets". In: *2012 Brazilian Symposium on Neural Networks*. Ed. by IEEE Computer Society Press. IEEE, Oct. 2012, pp. 49–54. DOI: 10.1109/sbrn.2012.25.

[56] A. Strehl and J. Ghosh. "Cluster Ensembles - A Knowledge Reuse Framework for Combining Multiple Partitions". In: *Journal of Machine Learning Research* 3 (2002). Editor: Claire Cardie, pp. 583–617. URL: https://www.jmlr.org/papers/v3/strehl02a.html.

[57] J. Trusty, B. Thompson, and J. V. Petrocelli. "Practical Guide for Reporting Effect Size in Quantitative Research in the Journal of Counseling & Development". In: *Journal of Counseling & Development* 82.1 (2004), pp. 107–110. DOI: 10.1002/j.1556-6678.2004.tb00291.x.

[58] N. X. Vinh, J. Epps, and J. Bailey. "Information theoretic measures for clusterings comparison: Is a Correction for Chance Necessary?" In: *Proceedings of the 26th Annual International Conference on Machine Learning*. ICML '09. ACM Press, 2009, pp. 1073–1080. DOI: 10.1145/1553374.1553511.

[59] N. X. Vinh, J. Epps, and J. Bailey. "Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance". In: *Journal of Machine Learning Research* 11.95 (2010), pp. 2837–2854.

[60] S. Wagner and D. Wagner. *Comparing clusterings: an overview*. Tech. rep. Universität Karlsruhe, 2007.

[61] M. J. Warrens and H. van der Hoef. *Understanding partition comparison indices based on counting object pairs*. Tech. rep. Groningen Institute for Educational Research, 2019. arXiv: 1901.01777 [stat.ML].

[62] J. Wu, H. Xiong, and J. Chen. "Adapting the right measures for K-means clustering". In: *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '09. ACM Press, 2009, pp. 877–886. DOI: 10.1145/1557019.1557115.

[63] Q. Xiang, Q. Mao, K. M. A. Chai, H. L. Chieu, I. W.-H. Tsang, and Z. Zhao. "A Split-merge Framework for Comparing Clusterings". In: *Proceedings of the29th International Conference on Machine Learning*. ICML'12. Omnipress, 2012, pp. 1259–1266. URL: http://dl.acm.org/citation.cfm?id=3042573.3042735.

[64] P. Zhang. "Evaluating accuracy of community detection using the relative normalized mutual information". In: *Journal of Statistical Mechanics: Theory and Experiment* 2015.11 (2015), P11006. URL: https://iopscience.iop.org/article/10.1088/1742-5468/2015/11/P11006.

[65] S. Zhang, Z. Yang, X. Xing, Y. Gao, D. Xie, and H.-S. Wong. "Generalized Pair-Counting Similarity Measures for Clustering and Cluster Ensembles". In: *IEEE Access* 5 (2017), pp. 16904–16918. DOI: 10.1109/access.2017.2741221.

A Additional Results

Figure 5: Significance of the results regarding the comparison of the segment heights performed in Section 5.2 over all pairs of transformations, considered for each measure and parameter set. For instance, the top four matrices correspond to Figure 4.a. Green (resp. red) cells represent significant (resp. non-significant) differences between the considered transformations, with a significance level of $\alpha = 0.05$. Figure available at 10.6084/m9.figshare.13109813 under CC-BY license.

B Evaluation Measures

In this section, we give the formal definition of the evaluation measures used in this work. A common point of those measures is that they can be computed using the so-called *confusion matrix* (also called *association matrix* or *contingency table*) based on the two partitions.

We note n the numbers of elements of a dataset D. Also, let $P = \{C_1, ..., C_k\}$ $(1 \leq k \leq n)$ be a k-partition of D, i.e. a division of D into k non-overlapping and non-empty clusters C_i $(1 \leq i \leq k)$. Let have another partition P' formed by k' clusters, where k' may be different from k. Then, the *confusion matrix* is a $k \times k'$ integer matrix, whose ii'th cell is the number of elements in the intersection of clusters C_i and $C_{i'}$, as shown in Table 5.

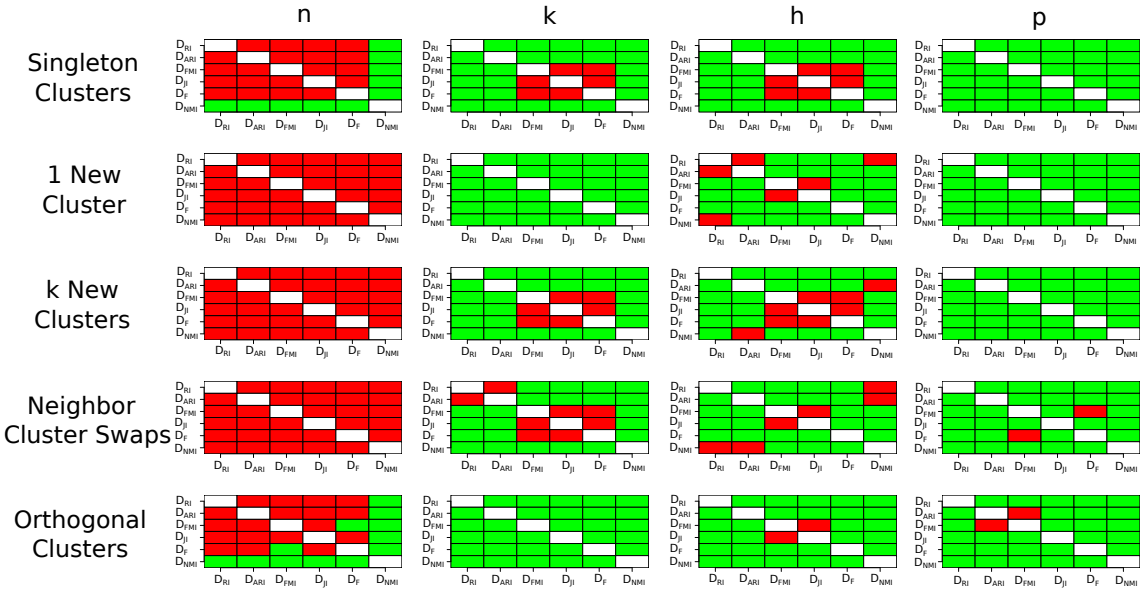

Figure 6: Significance of the results regarding the comparison of the segment heights performed in Section 5.2 over all pairs of measures, considered for each transformation and parameter set. For instance, the top four matrices correspond to the last stacked bar in each barplot of Figure 4 (*Singleton Clusters*). Green (resp. red) cells represent significant (resp. non-significant) differences between the considered measures, with a significance level of $\alpha = 0.05$. Figure available at 10.6084/m9.figshare.13109813 under CC-BY license.

Table 5: The confusion matrix for two partitions $P = \{C_1, ..., C_k\}$ and $P' = \{C'_1, ..., C'_{k'}\}$ of n elements, where $n_{ij} = |C_i \cap C'_j|$ are the number of elements in both clusters $C_i \in P$ and $C'_i \in P'$.

		Partition P'			
	Cluster	C'_1	...	$C'_{k'}$	Marginal sum
Partition P	C_1	n_{11}	...	$n_{1k'}$	$n_{1.}$
	.	.		.	.
	.	.		.	.
	C_k	n_{k1}	...	$n_{kk'}$	$n_{k.}$
	Marginal sum	$n_{.1}$	...	$n_{.k'}$	$n_{..} = n$

B.1 Rand Index, *RI*

The formulation of all pair-counting measures can be expressed in terms of four types of element pairs. The *positive agreement* N_{11} corresponds to the number of element pairs which are in the *same* cluster in *both* partitions P and P'. The *negative agreement* N_{00} is the number of element pairs which are in *different* clusters in *both* P and P'. The partitions *disagree* on the remaining element pairs, as N_{10} (resp. N_{01}) corresponds to the number of element pairs which are in the same cluster in P (resp. P'), but not in P' (resp. P). The formula of each term is shown in Table 6.

The *Rand Index* (RI) [47] is the proportion of total agreement, i.e. when counting both positive and negative agreement:

$$RI(P, P') = \frac{N_{11} + N_{00}}{N_{..}}. \tag{2}$$

Its values lie between 0 and 1, where 0 occurs for the absence of any positive and negative agreements, whereas 1 corresponds to the case where the partitions are perfectly identical.

B.2 Adjusted Rand Index, *ARI*

The *Adjusted Rand Index* (ARI) [24] is a well-known extension of the Rand Index, with additional correction for chance. It aims at dealing with the statistical independence of two partitions (see Section 2.1.2). Its formula is

$$ARI(P, P') = \frac{RI(P, P') - \mathbb{E}[RI(P, P')]}{1 - \mathbb{E}[RI(P, P')]}. \tag{3}$$

Table 6: Formulae for the number of (unordered) element pairs of the four types

Type	Formula
N_{11}	$\sum_{i=1}^{k}\sum_{j=1}^{k'}\binom{n_{ij}}{2} = \frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{k'} n_{ij}(n_{ij}-1)$
N_{00}	$\binom{n}{2} - (N_{11}+N_{10}+N_{01})$
N_{10}	$\frac{1}{2}\Big(\sum_{i=1}^{k} n_{i\cdot}^2 - \sum_{i=1}^{k}\sum_{j=1}^{k'} n_{ij}^2\Big)$
N_{01}	$\frac{1}{2}\Big(\sum_{j=1}^{k'} n_{\cdot j}^2 - \sum_{i=1}^{k}\sum_{j=1}^{k'} n_{ij}^2\Big)$
$N_{\cdot\cdot} = N_{11} + N_{10} + N_{01} + N_{00}$	$\binom{n}{2} = n(n-1)/2$

where $\mathbb{E}[RI(P,P')]$ corresponds to the estimated score of $RI(P,P')$ for independent partitions under hypergeometric assumption (so-called permutation model). This term is defined as

$$\mathbb{E}(\sum_{ij}\binom{n_{ij}}{2}) = \sum_{i}\binom{n_{i\cdot}}{2}\sum_{j}\binom{n_{\cdot j}}{2}/\binom{n}{2}. \tag{4}$$

The ARI takes a value of 1 for identical partitions, whereas 0 indicates a case of statistical independence. Moreover, ARI can take a negative value for very dissimilar partitions [39], when the observed RI is smaller than expected.

B.3 Jaccard Index, *JI*

The Jaccard Index (JI) was originally defined to compare sets [25], but it is also used as an external measure [7]. As reported in [39], the negative agreement N_{00} can be often almost as large as the maximum number of element pairs $\binom{n}{2}$. The Jaccard Index is an improved version of RI on this aspect, as it does not take N_{00} into account. It is defined as

$$JI(P,P') = \frac{N_{11}}{N_{11}+N_{01}+N_{10}}. \tag{5}$$

The Jaccard Index ranges from 0 (absence of any positive agreement) to 1 (identical partitions). Note that one minus the Jaccard Index is a metric on the finite sets [36].

B.4 Fowlkes-Mallows Index, *FMI*

The Fowlkes-Mallows Index [14] is the final pair-counting measure that we consider in this work. It was originally introduced to ease the comparison of hierarchical dendrograms. Like the Jaccard Index, it ignores negative agreements. It can be described as the geometric mean of two asymmetric forms of positive agreement: the proportion of positive agreements relative to the number of pairs belonging to the same cluster in P vs. those in P'. Its formal description is

$$FM(P,P') = \frac{N_{11}}{\sqrt{(N_{11}+N_{10})(N_{11}+N_{01})}}. \tag{6}$$

B.5 F-measure, *F*

In the category of set-matching measures, we select the F-measure (F). Note that this name is sometimes used in the literature as a synonym of *harmonic mean*, and therefore covers several distinct measures (e.g. [17, 46]). We use the definition of Artiles *et al.* [6], according to which the F-measure is the harmonic mean of two quantities called *Purity* and *Inverse Purity*.

The formal definition of *Purity* is as follows:

$$Purity(P,P') = \sum_{i}\frac{n_{i\cdot}}{n}\max_{j}\frac{n_{ij}}{n_{i\cdot}}. \tag{7}$$

The Inverse Purity is simply the Purity of the second partition relative the first, i.e. $Purity(P', P)$. Finally, the F-measure is the harmonic mean of the Purity and Inverse Purity

$$F(P, P') = 2\frac{Purity(P, P') \times Purity(P', P)}{Purity(P, P') + Purity(P', P)}. \tag{8}$$

B.6 Normalized Mutual Information, *NMI*

The last measure that we consider is the Normalized Mutual Information (NMI), which belongs to the category of information-theoretical measures. It is based on the notions of *entropy* and *Mutual Information* [10]. The principle behind these notions is to consider each partition as a categorical random variable, whose possible values are the clusters.

In the context of clustering, entropy in the sense of Shannon is defined as

$$H(P) = -\sum_{i=1}^{k} \frac{n_{i\cdot}}{n} \log \frac{n_{i\cdot}}{n}. \tag{9}$$

Each element in dataset D has an equal probability of being picked, so its probability of being in cluster C_i is n_i/n. Thus, we have a discrete random variable taking k values, which is associated to the partition P. If the partition P has only 1 cluster containing all the points, then $H(P)$ will be zero, since there is no uncertainty in the clustering structure. If the partition P consists of as many clusters as n, it will reach its maximum value. Note that $H(P)$ does not depend on n, but on the relative proportions of the clusters.

The Mutual Information can be described as the mutual dependence between these variables, and it can then be interpreted as the similarity between the partitions. It is formally described as

$$MI(P, P') = \sum_{i=1}^{k}\sum_{j=1}^{k'} \frac{n_{ij}}{n} \log \frac{\frac{n_{ij}}{n}}{\frac{n_{i\cdot}}{n}\frac{n_{\cdot j}}{n}}. \tag{10}$$

There are a number of variants of the notion of mutual information, in particular several normalizations have been proposed (see for instance [59]). In this work, we focus on the sum normalization as defined in [29, 56], which is very widespread. The resulting NMI is

$$NMI(P, P') = \frac{2MI}{H(P) + H(P')}. \tag{11}$$

C Experimental details about the heterogeneity of cluster sizes

There are many ways to make clusters imbalanced. In this work, we opt for a sequence based on an arithmetic progression. Consider the sizes of the clusters in a partition as a sequence of values S_k whose sum is equal to the number of nodes n, as in (12). In this equation, α corresponds to the first value and β corresponds to the constant increment value

$$\begin{aligned} n &= \alpha + (\alpha + \beta) + (\alpha + 2\beta) + ... + (\alpha + (k-1)\beta) \\ &= \alpha k + \frac{\beta k(k-1)}{2}. \end{aligned} \tag{12}$$

Note that the sequence contains as many terms as the number of clusters.

In such a sequence, each term is a constant increment value β larger than the previous term (e.g. $\beta = 2$ for the sequence 3, 5, 7, ..). This β is computed based on the parameter h (heterogeneity of cluster sizes). When $h = 1$, it reaches its maximal value that we note β_{max}. In the case of $h < 1$, β is proportional to β_{max} to the extent of h (i.e. $\beta = h \times \beta_{max}$). The value of β_{max} can be determined in different ways. In order not to introduce an additional parameter for this, our approach is to assign the first term α and the constant increment β_{max} to the same value, i.e. $\alpha = \beta_{max}$. Then, we compute β as follows

$$\begin{aligned} n &= \frac{\beta_{max} k(k+1)}{2} \\ \beta &= \lfloor h\beta_{max} \rfloor. \end{aligned} \tag{13}$$

Note that $\lfloor . \rfloor$ denotes the floor function (returning the greatest integer less than or equal to the input value). Finally, we obtain the value of α as follows

$$\alpha = \frac{n - \frac{\beta k(k-1)}{2}}{k}. \tag{14}$$