Eurographics Conference on Visualization (EuroVis) 2016
K-L. Ma, G. Santucci, and J. J. van Wijk
(Guest Editors)

Volume 35 (2016), Number 3

Exploring and Comparing Clusterings of Multivariate Data Sets Using Persistent Homology

B. Rieck[1,2] and H. Leitte[1]

[1]TU Kaiserslautern, Germany
[2]Heidelberg University, Germany

Abstract

Clustering algorithms support exploratory data analysis by grouping inputs that share similar features. Especially the clustering of unlabelled data is said to be a fiendishly difficult problem, because users not only have to choose a suitable clustering algorithm but also a suitable number of clusters. The known issues of existing clustering validity measures comprise instabilities in the presence of noise and restrictive assumptions about cluster shapes. In addition, they cannot evaluate individual clusters locally. We present a new measure for assessing and comparing different clusterings both on a global and on a local level. Our measure is based on the topological method of persistent homology, which is stable and unbiased towards cluster shapes. Based on our measure, we also describe a new visualization that displays similarities between different clusterings (using a global graph view) and supports their comparison on the individual cluster level (using a local glyph view). We demonstrate how our visualization helps detect different—but equally valid—clusterings of data sets from multiple application domains.

Categories and Subject Descriptors (according to ACM CCS): I.3.6 [Computer Graphics]: Methodology and Techniques—Interaction techniques

1. Introduction

Clustering algorithms play an important role in exploratory data analysis. Their partitions help users detect interesting patterns in their data. This leads to a better understanding of salient properties and supports creating mental models of even complex multivariate data sets. Production-quality clustering libraries [PVG*11] make it easy to obtain different clusterings of a data set—the challenging part lies in assessing them. Clustering assessment consists of two tasks: First, a suitable clustering algorithm needs to be chosen. Second, the parameters of the algorithm need to be configured. A well-known adage in the clustering community states that "There is no best clustering algorithm. [...] When there is a good match between the model and the data, good partitions are obtained." [Jai10]. Users hence often employ multiple clustering algorithms to their data and need to compare them with each other. Most clustering algorithms, such as the k-means algorithm, use a single parameter k that determines the number of clusters. Finding suitable values for k is still an active topic of research within the clustering community. Commonly, different clustering algorithms are run with varying parameters. For each run, a *clustering validity index* such as the Dunn index [HBV01] is evaluated and k is selected such that the index shows the best value. While clustering validity indices are useful for comparing multiple clusterings from the same algorithm, their utility for exploratory data analysis is limited—complex cluster geometries often lead to unstable results so that a suitable k fails to be found even for small data sets like the "Iris" data [ZM14]. Furthermore, existing clustering validity indices are unable to assess individual clusters of a clustering without referring to labels, which are often unavailable in real-world data sets.

Especially when dealing with multivariate unlabelled data sets, users need to be supported in choosing a suitable clustering algorithm and a suitable amount of clusters. Since clustering is a method for detecting interesting patterns in data, visualizations for comparison and evaluation need to play an integral part in this process. In this paper, we present visualizations of clusterings from two different perspectives. Globally, our *clustering similarity graph* arranges clusterings by similarity to provide an overview. Locally, our *cluster map* shows the individual clusters making up a clustering, arranged as glyphs among a common reference embedding of the data. The glyphs enable users to see how a given clustering partitions their data. This information permits the comparison of clusters among each other and among different clusterings of the data. Our visualizations are driven by an underlying clustering assessment measure, based on *persistent homology*, a method from computational topology. By focusing on the topology of a data set, our measure is robust, unbiased concerning cluster shapes—i.e. it shows no preference for e.g. convex or concave clusters—and hence less prone to the shortcomings of existing clustering validity measures. Furthermore, our measure provides a well-defined way

of comparing clusterings, both on a global and on a local level. We demonstrate the utility of our framework by analysing clusterings of data sets with varying complexities.

2. Related work

Clustering algorithms & clustering evaluation. A recent survey [XT15] shows that there is a nigh-uncountable amount of clustering algorithms available nowadays. Clustering remains one of the most important techniques for understanding multivariate data sets. Now, several decades after the first clustering algorithms, many challenging issues remain. For instance, there still is no answer to the question of what constitutes a good cluster [Jai10]. While there are many indices available for assessing aspects of a given clustering [HBV01], they often suffer from unstable results in the presence of noise, cluster boundary overlaps or complicated geometries [AGM*13].

Clustering & visualization. The visualization community has already taken a great interest in helping users better understand the results of clustering algorithms. The *hierarchical clustering explorer* [SS02] enables biologists to interact with different hierarchical clusterings of microarray experiment data. Various model-based indices and auxiliary visualizations aid in understanding not only the data space but also the decisions made by the clustering algorithm. While some methods aim for evaluating clusters within auxiliary visualizations, e.g. parallel coordinate plots [HVW10] or scatterplots [EDF08], others let users modify the results of clustering algorithms by e.g. splitting data into subgroups, which are then clustered separately [LSP*10]. This helps detect relations that would otherwise be obscured by classical methods. A similar idea of cluster modification is also prevalent in visual analytics tools. Nam et al. [NHZI07], for example, let users "sculpt" clusters by changing the relevance of different attributes in the data, while using several auxiliary visualizations to make sense of the current clustering. Schreck et al. [SBvLK09] embed clustering analysis in a general visual analytics workflow. Their system permits modifying clustering results as well as verifying their validity, but forgoes traditional clustering algorithms in favour of *self-organizing maps*. Hence, the system does not permit the comparison of multiple clustering results. Tatu et al. [TMF*12] support the clustering process by pre-selecting interesting subspaces in the data, which are then clustered using hierarchical clustering. Users may then interact with dissimilar subspaces to understand their structures. This approach is somewhat orthogonal to ours. It helps explore patterns in the data prior to applying any clustering algorithms. Pilhöfer et al. [PGU12] developed a method for re-ordering categorical variables to improve visualizations of multiple clusterings. This permits tracking similarities of partitions over different clusterings. Their approach is complementary to our method because we focus more on exploring the shapes of individual clusters.

Computational topology. Topological methods have become mature in recent years and are now offering another view on multivariate data sets. Singh et al. [SMC07] developed MAPPER, a visualization combining clustering methods with representations of the connectivity of data. Lum et al. [LSL*13] demonstrated how topology helps in analysing the shape of data. They presented improvements of the MAPPER technique that lead to insights into data sets from various domains. Carlsson [Car14] outlined a general workflow for topological data analysis, formalizing the clustering shape descriptor we are using in this paper. Persistent homology has also been employed successfully in very diverse contexts, including shape description and comparison [BdFF*08], visualization [RML12, RL14], as well as machine learning [RHBK15].

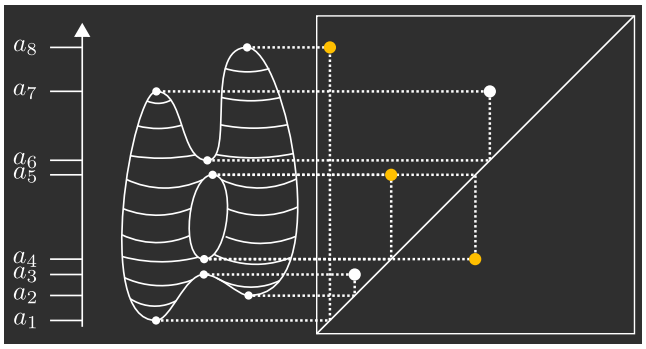


Figure 1: *Calculating persistent homology of a manifold* $\mathbb{M}$ *(left) with a height function* f*. The extended persistence diagram (right) serves as a fingerprint.*

3. Methods

In the following, we give a brief overview of the most important details of *persistent homology*. We strive for clarity of exposition and refer to Edelsbrunner and Harer [EH10] for an in-depth introduction.

3.1. Persistent homology

The basic idea of persistent homology is to describe the geometrical and topological properties of a function defined on a manifold, i.e. a space that locally has the structure of some $\mathbb{R}^n$. In the following, we assume that we have a manifold $\mathbb{M} \subseteq \mathbb{R}^n$ and a function $f : \mathbb{M} \to \mathbb{R}$. We need f to have a finite number of critical points, i.e. points at which its gradient vanishes. Furthermore, the function values $a_1, \ldots, a_m$ at the critical points need to be pairwise distinct. Figure 1, left, shows an example for $\mathbb{M}$ and f. Since $\mathbb{M}$ is two-dimensional, its critical points are local extrema or saddles.

When sweeping the values of f from low to high values, the topology of its *sublevel sets* changes only at critical values $a_1, \ldots, a_m$. The critical values give rise to topological features of different dimensions: *Connected components* (dimension 0), *tunnels* or *holes* (dimension 1), and *voids* (dimension 2). Higher-dimensional features can be similarly described but only occur in higher-dimensional manifolds. During the sweep, a saddle either merges two connected components, thereby destroying the one with the larger height, or creates a new hole. A minimum always creates a new connected component in the corresponding sublevel set. A maximum destroys a hole by closing it or, in the case of the global maximum, creates a new void.

Having accounted for all types of critical values, we may now pair the values of "creators" and "destroyers". The *persistence diagram* $\mathcal{D}_f$ of a function f contains points from $\mathbb{R}^2$ that correspond to the pairings of critical values; see Figure 1, right, for an example. Given a point $(c,d) \in \mathcal{D}_f$, its *persistence* is defined as $\operatorname{pers}(c,d) := |c-d|$. Points with a high persistence represent large-scale features of f, while points with a low persistence represent small-scale features. The component created at a_2 and merged at a_3 in Figure 1, for example, is a typical small-scale feature.

If we follow the sweep as defined above, we observe that we cannot pair the global minimum at a_1, the two saddles at a_4 and a_5, and the global maximum at a_6. In the classical persistence algorithm [ELZ02], these points remain unpaired, which makes calculating $\operatorname{pers}(\cdot)$ difficult. To resolve this issue, we first pair a_1 with a_8. This denotes the range of the function f on $\mathbb{M}$. Next, we pair the two saddles with each other—in both directions, so that we have $(a_4,a_5),(a_5,a_4) \in \mathcal{D}_f$. These pairs are obtained by calculating the *extended persistence diagram* [CSEH09]. In Figure 1, they are depicted as yellow points. The extended persistence diagram ensures that no critical point remains unpaired. Its calculation requires a second sweep through the *superlevel sets* of f.

We can see that small perturbations in the values of f only slightly change the pairing and the persistence diagram $\mathcal{D}_f$. This intuitive notion of stability has been formally [CSEH07] and experimentally [BdFF*08] proven, making persistent homology an appealing technique.

Computational aspects & implementation. For calculating persistent homology on real-world multivariate data sets, we use a *Rips graph* $\mathcal{R}_\varepsilon$ and a distance function $\mathrm{d}(\cdot,\cdot)$ to obtain connectivity information. $\mathcal{R}_\varepsilon$ contains a vertex v for each data point and an edge (u,v) between two points if $\mathrm{d}(u,v) \leq \varepsilon$. The parameter ε controls the coarseness of $\mathcal{R}_\varepsilon$. The following heuristic [RML12] has proven to be useful and robust: For each data point, we estimate the average of the distances to its nearest neighbours and use the median of these averages as ε. Because $\mathcal{R}_{\varepsilon'} \subseteq \mathcal{R}_\varepsilon$ for $\varepsilon' \leq \varepsilon$, we get information about all scales up to the selected threshold when calculating $\mathcal{R}_\varepsilon$, making this a stable construction [CdSO14].

Given a function f on the data (see Section 3.2), we assign every vertex v the value $f(v)$ and every edge (u,v) the value $\max\{f(u),f(v)\}$. By traversing vertices and edges in order of ascending function values, we can calculate persistent homology in dimensions zero and one. Extended persistence requires a second pass in descending order. This computation is highly efficient; it has a complexity of only $\mathrm{O}(n \log n + n \cdot \alpha(n))$, where n denotes the number of data points and $\alpha(\cdot)$ is the extremely slow-growing inverse of the Ackermann function.

Total persistence. Since persistence diagrams summarize the geometrical-topological behaviour of functions, we need a method for comparing them. A very useful summary statistic in this context is given by the *total persistence* [CSEHM10] of a function f. Given the persistence diagram $\mathcal{D}_f$ of the function f, total persistence is defined as the sum of all squared persistence values, i.e.

$$\operatorname{Pers}(\mathcal{D}_f) := \sum_{(c,d)\in\mathcal{D}_f} \operatorname{pers}(c,d)^2. \tag{1}$$

Conceptually, the total persistence of a function is similar to the concept of *total variation* in mathematics. Like the total variation, $\operatorname{Pers}(\cdot)$ measures the amount of changes that are characteristic of a function. It thus serves as a coarse characterization of function behaviour.

3.2. A shape descriptor function for multivariate data

Persistent homology requires a scalar function f of the data set. Lum et al. [LSL*13] present a function that is particularly useful for clustering analysis because it permits the detection of interesting features in real-world data sets [Car14]. We have

$$f(x) = \sum_{y\in\mathbb{X}} \mathrm{d}(x,y)^2, \tag{2}$$

where $\mathbb{X}$ denotes the input data and $\mathrm{d}(\cdot,\cdot)$ denotes a distance measure such as the Euclidean distance. The squaring ensures that the effects of small distance values are attenuated. Outlying points are indicated by high values in f. In practice, we want higher function values to indicate central points, so we modify the previous equation to

$$\hat{f}(x) = \left(\max f - f(x)\right) / \left(\max f - \min f\right), \tag{3}$$

where $\max f$ and $\min f$ are the extremal values of f.

The expressive power of a shape descriptor largely hinges on its discriminative properties. Biasotti et al. [BdFF*08] analyse to what extent these properties are present in certain functions. Useful functions include density estimators [RL15], Laplacian eigenfunctions [SMC07], and kernel regression estimators such as the Nadaraya–Watson estimator [Nad64]. While our workflow can handle any function, we focus on a single one in this paper for reasons of clarity.

3.3. Assessing clusterings

In order to assess clusterings of a data set, we first calculate a Rips graph $\mathcal{R}_\varepsilon$ as described above on the unclustered data and assign it the values of the shape descriptor f from the previous section. A clustering $\mathrm{C} = \{\mathrm{C}_1,\ldots,\mathrm{C}_k\}$ of the data induces a partition of the vertex indices of $\mathcal{R}_\varepsilon$. A partition is a finite number of sets that are pairwise disjoint—no point may occur in more than one set—such that their union contains all indices in the data. C induces a partition on the Rips graph $\mathcal{R}_\varepsilon$ by connecting vertices u and v if $(u,v) \in \mathcal{R}_\varepsilon$ and $(u,v) \in \mathrm{C}_i$ for some i. Hence, a weighted edge is only kept if both of its vertices are in the same cluster. The result is a set of Rips graphs, each corresponding to a cluster $\mathrm{C}_i \in \mathrm{C}$. Calculating persistent homology on each of the Rips graphs results in a set of persistence diagrams $\mathcal{D}_\mathrm{C} = \{\mathcal{D}_1,\ldots,\mathcal{D}_k\}$. Each diagram $\mathcal{D}_i$ measures the topological features of the shape descriptor f present in the partition induced by C.

Our measure is based on the following assumption: If C is suitable, it should cluster data such that the most interesting features measured by f are retained. Figure 2 demonstrates this for some examples. The right-most clusterings are retaining all features of f. Optimally, we would like the persistence diagram $\mathcal{D}_f$ of the unpartitioned data to be the disjoint union of the individual diagrams $\mathcal{D}_i$, i.e. $\mathcal{D}_f = \mathcal{D}_1 \cup \cdots \cup \mathcal{D}_k$, with $\mathcal{D}_i \cap \mathcal{D}_j = \varnothing$ for $i \neq j$. In this case, all

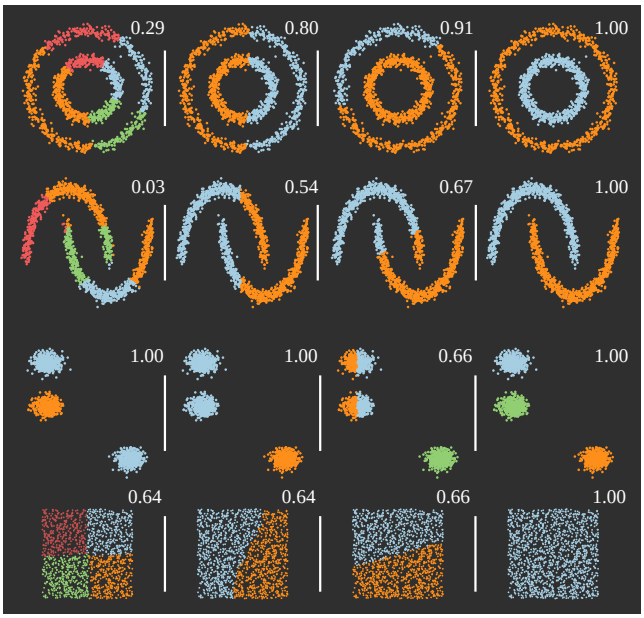


Figure 2: *Different clusterings for several test data sets and the values of our global measure σ_{Global}. Higher values indicate that more geometrical-topological variation has been retained.*

features of f remain represented in one unique cluster of C. We can use *total persistence* to measure the amount of features that are lost. Ideally, the total persistence of the data remains unchanged, so that we have $\operatorname{Pers}\mathcal{D}_f = \sum_{i=1}^{k} \operatorname{Pers}\mathcal{D}_i$.

Global assessment of a clustering. Following the previous equations, we can assess the global quality by calculating

$$\sigma_{\text{Global}} = \frac{\sum_{i=1}^{k} \operatorname{Pers}(\mathcal{D}_i)}{\operatorname{Pers}(\mathcal{D}_f)}, \tag{4}$$

i.e. the ratio of total persistence that is being retained by the clustering. σ_{Global} has a range of $[0,1]$, with 1 meaning that the amount of topological variation has been fully retained by the partition. Note that we need to treat zero-dimensional persistent homology differently when calculating $\operatorname{Pers}(\cdot)$: For each connected component, we ignore the pair containing the extremal function values. For example, in Figure 1, we ignore (a_1, a_8). This is similar to the concept of *reduced homology* [EH10, p. 83] and ensures that we do not overestimate the total persistence when splitting a connected component. Else, a split would result in at least one additional pair of extremal function values. The remaining vertices belonging to a connected component are still taken into account, though. We use the values of σ_{Global} to colour-code clusterings in our clustering similarity visualization (Section 3.4.1). Figure 2 shows some example clusterings along with the values for σ_{Global}. The data sets depicted in this figure are known to exhibit interesting behaviour of clustering algorithms that already occurs in two dimensions [PVG*11]. In the supplementary materials, we also show that existing clustering validity indices are either incapable or unstable with respect to their assessment of these data sets.

Local assessment of a clustering. Assessing a clustering locally, i.e. on the level of individual clusters, is more complex and to the best of our knowledge, our measure is the only one capable of such an assessment without requiring class labels. We first calculate $\mathcal{D}_{f,i} := \mathcal{D}_f \cap \mathcal{D}_i$. The persistence diagram $\mathcal{D}_{f,i}$ contains topological features of the original data that are retained in the clustering. This does not account for topological features that are slightly changed by the partition because the connectivity of the partitioned Rips graph changes as well. For each unaccounted point p in $\mathcal{D}_i \setminus \mathcal{D}_{f,i}$ we find its nearest neighbour q, measured using the L_∞ distance, in $\mathcal{D}_f$. If $2\|p-q\|_\infty \leq \operatorname{pers}(p)$, we match the two points and add q to $\mathcal{D}_{f,i}$. The point q hence becomes the new representative for the topological feature described by p. We now calculate

$$\sigma_{\text{Local}} = \frac{\operatorname{Pers}(\mathcal{D}_{f,i})}{\operatorname{Pers}(\mathcal{D}_i)}, \tag{5}$$

i.e. the ratio of total persistence both present in the cluster and the complete data set (numerator) to the amount of the total persistence in the cluster (denominator). σ_{Local} also has a range of $[0,1]$, with 1 meaning that all features found in the cluster are also present in the original data. The idea behind this measure is that a cluster should contain only features of the function f that are present in the original data set. We use the values of σ_{Local} for colour-coding our local cluster visualization (Section 3.4.2).

Properties of total persistence. Both σ_{Global} and σ_{Local} may be thought of as topological equivalents of the *explained variance* or *explained variation* measures from statistical modelling. We deem a clustering C to be suitable when it explains much of the geometrical-topological variation present in the data. *Total persistence* has several beneficial properties in comparison to methods that only assess the geometry of the data: (i) As a topological measure, it even works for complex cluster shapes. (ii) It is unbiased with respect to the size of clusters—a small cluster may still contain large-scale features. (iii) It is stable because it considers the scale of features—$\operatorname{Pers}(\cdot)$ barely changes if only few low-persistence points in a persistence diagram are changed. (iv) It permits the assessment of clusterings on a local level without requiring label information.

Limitations. Our measure is incapable of distinguishing clusterings when individual clusters are well-separated on a large scale and similarly-shaped. For example, some clusterings in the third row of Figure 2 cannot be told apart. Three of the four splits are considered equally valid by our measure; one split destroys much topological information by splitting two of the "blobs", however, so it gets penalized with $\sigma_{\text{Global}} = 0.66$. Similarly, different clusterings shown in the fourth row will also not be rated highly by our measure because the clusters are too close and do not contain any prominent topological features. Our measure penalizes these splits and we feel that for these kinds of data, any split should be penalized. Locally, each of the individual clusters is considered to be a good fit, though. In the supplementary materials, we compare the behaviour of our measure on these data sets with existing clustering validity indices. For the last data set, existing measures are incapable of reaching a consensus, tending towards more clusters rather than fewer.

3.4. Visualization

Our visualization provides two views on the data. First, given multiple clusterings of a data set, we group them using the *clustering similarity graph*. The graph visualizes the similarity of different clusterings. This visualization indicates the complexity of the underlying data set. If many clusterings agree, for example, the data may exhibit a simple structure. Second, we enable the comparison and assessment of individual clusters of a data set through the *cluster map*. The map helps understand the patterns that occur in individual clusters, making it possible to assess whether they are interesting and informative.

In the following, we will colour-code clusterings using the values of σ_{Global} and individual clusters using the values of σ_{Local}. Since both measures have the same range, we can use the same colours to indicate the amount of explained topological variation in the data: Green indicates values in $[0.80, 1.00]$, yellow indicates values in $[0.60, 0.80)$, and red indicates values less than 0.60. Similar ranges are being used in statistical modelling, where a model is considered bad when it cannot account for more than 60% of the variation.

3.4.1. Clustering similarity graph

To handle the global comparison of multiple clusterings, we require a similarity measure. As the comparison of different clusterings is an integral part of data analysis, numerous similarity measures already exist [Mei07]. Since our goal is to compare multiple partitions among each other, we prefer similarity measures that are also metrics in the mathematical sense. In particular, we require the *triangle inequality* to be satisfied, i.e. $\mathrm{d}(x,y) \leq \mathrm{d}(x,z) + \mathrm{d}(z,y)$, for clusterings x, y, z. The triangle inequality ensures that two clusterings x, y that are similar to the same cluster z must be similar to each other as well. Most similarity measures do not satisfy this inequality, thereby yielding inconsistent similarity values. We use the *Mirkin metric* [Mei07], which is a metric on the space of clusterings. It is defined by

$$\mathrm{d}(x,y) := 2\,(n_{01} + n_{10})\,, \tag{6}$$

where n_{01} is the number of pairs of points that are in different clusters under x but in the same cluster under y, and n_{10} is defined vice-versa. The Mirkin metric is known to work less well when comparing clusterings with different amounts of clusters among each other. This poses no problem for our visualization because we only use the metric to compare clusterings with the same number of clusters.

We build the *clustering similarity graph* visualization using the Mirkin metric by showing each individual clustering as a node. For each node, our application stores the corresponding clustering algorithm and its parameters, which we will subsequently only refer to if it helps explain a given clustering better. The Mirkin metric yields a matrix of pairwise distances between the points, and we use force-directed graph visualization techniques [BETT99, Chapter 10] to embed the points in 2D. Edges between nodes indicate the two nearest neighbours of a given clustering in order to support the orientation of the user. The edge opacity reflects the amount of overlap between two clusters, measured using the *Rand index*, i.e.

$$s_{\text{Rand}}(x,y) := \frac{n_{11} + n_{00}}{\binom{n}{2}}, \tag{7}$$

where n is the number of data points, n_{11} is the number of pairs of points that are in the same clusters under both x and y and n_{00} is the number of pairs of points that are in different clusters under both x and y. While the Rand index is not a proper metric, it has the advantage of being simple to understand. Node colours correspond to the values of σ_{Global} (Section 3.3) and indicate how well a clustering overall retains features in the data.

3.4.2. Cluster map

To permit the exploration of individual clusters in a data set, we provide a combination of a glyph-based view and a simplified density projection of the data. We first use a dimensionality reduction algorithm such as *principal component analysis* to obtain a two-dimensional embedding of the original data set. This embedding will serve as an invariant map of the data. The map provides a shared reference coordinate system on which different clusterings can be compared. We visualize the embedding as a hexagonally-binned plot (using Sturges' formula to find an approximate number of bins), because it is known to have better data aggregation properties than a rectangularly-binned plot [CLNL87]. We colour each cell according to the number of points it contains. Lighter colours indicate few points, while more saturated colours indicate many points. This serves as a simple visualization of the density distribution within the embedding of the data, which supports reasoning about clusterings [TAE*09].

In addition to the map, a second component of our visualization is a set of glyphs for representing individual clusters. Each glyph contains a *star plot* [CCKT83] that depicts a simplified representation of the data points within the cluster. Data points are represented using a band that shows the minimum, median, mean, and maximum values of each attribute (i.e. each dimension) for all data points in the cluster. Mean and median have been selected to indicate whether the distribution of values of a specific dimension within a given cluster is skewed. In conjunction with visualizing the minimum and maximum values, this improves understanding the "profile" of a cluster. The background of each glyph is colour-coded according to the respective value of σ_{Local} (Section 3.3). It indicates how well a cluster matches the geometrical-topological features present in the shape descriptor f. The glyphs are placed automatically along the map to minimize clutter. Each glyph is then connected to the centroid of the cluster it represents in order to highlight cluster placements. Our implementation optionally employs *semantic zooming* to offer more details on demand. In addition, the user may trigger the visualization of the extents of a cluster within the map. This helps when comparing the boundaries of multiple clusters within the data space.

4. Results

In the subsequent sections, we analyse different clusterings of multivariate data sets of varying complexities. We selected data sets that are known to be challenging to cluster in order to highlight the benefits of our analysis pipeline. We use different clustering algorithms from the SCIKIT-LEARN toolkit [PVG*11]. Since the goal of this paper is not the evaluation of different clustering algorithms but rather their clusterings, we do not describe every clustering re-

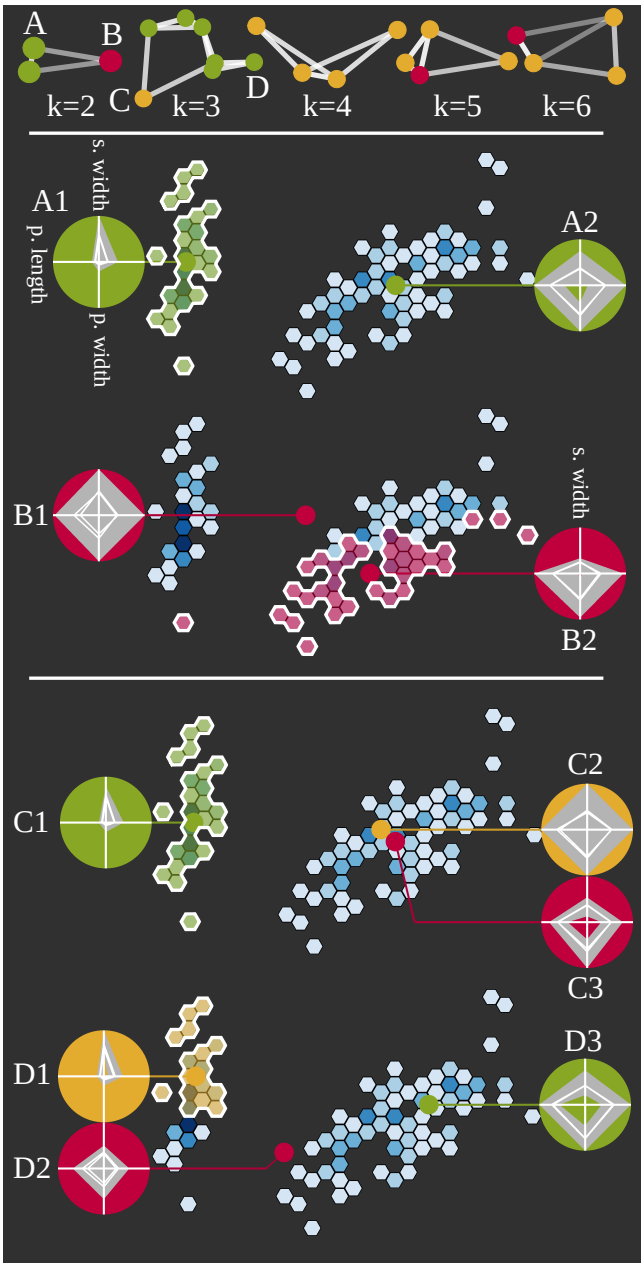


Figure 3: *Clustering similarity graphs (top) and cluster maps for several "Iris" data clusterings.*

sult in detail. We instead explain interesting properties of selected clusterings by means of our visualizations and our measures.

4.1. Iris

The "Iris" data set [Lic13] contains 150 measurements of 4 attributes of 3 different "Iris" flower species, *I. setosa*, *I. virginica*, and *I. versicolor*. It is challenging because the flowers cannot be clustered correctly without knowing the species information. There are two pronounced clusters in the data, one for *I. setosa*, the other

	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$
BetaCV	**0**	0.187	0.215	0.253	0.235
C-index	**0.056**	0.069	0.059	0.062	0.057
WCS	89.90	90.60	81.76	72.67	**67.74**
Dunn index	**0.339**	0.098	0.105	0.117	0.133
NC	1.652	2.763	3.792	4.812	**5.820**
Silhouette	**0.630**	0.480	0.434	0.353	0.383
σ_{Global}	**1.0**	0.967	0.627	0.561	0.561

Table 1: *Common clustering validity indices and our global measure σ_{Global} for the "Iris" data set. For every k, we have used the best possible clustering—measured using the Rand index—with respect to the species labels.*

one for the remaining two species. Splitting the *I. virginica* and *I. versicolor* cluster is hard because its boundaries are unclear. Figure 3, top, shows the clustering similarity graphs for selected clusterings with different amounts of clusters. We can see that starting with $k = 4$, most clusterings are incapable of retaining large amount of important features in the data. Without using any class labels, our topological measure thus suggests that $k = 2$ and $k = 3$ are more suitable for the number of clusters than $k \geq 4$.

Table 1 shows how different clustering validity indices assess clusterings of these data for an increasing number of clusters. The partitions with $k = 2$ and $k = 3$ are optimal with respect to the species labels. The best value for each measure is shown in bold. The first three rows contain measures for which smaller values indicate a better fit—for the remaining rows, larger values are better. None of the measures suggests $k = 3$ as an optimal number of clusters. Nevertheless, our measure indicates that $k = 3$ is still a viable option with less than 4% of geometrical-topological information being lost. It is the only measure capable of indicating that clusterings with $k \geq 4$ are significantly less suitable than clusterings for $k = 2, 3$. Please refer to the supplementary materials for a more detailed comparison.

In the following, we will refer to the individual clusters as depicted in Figure 3.

Two clusters. We first analyse two clusterings with $k = 2$ in order to get an intuition for the different visualizations. Clustering A contains the correct species assignments. We have $\sigma_{\text{Local}} = 1.0$ for both A1 and A2, meaning these clusters retain all features of the data. By showing the *cluster extents* of A1 using a thicker boundary for its hex cells, we can confirm the simple shape of this cluster, which contains all *I. setosa* flowers. The glyph shows that flowers in A1 have small petal lengths/widths and extremely large sepal widths. By contrast, A2 contains flowers with a significantly larger petal lengths/widths. Clustering B is shown to be very different. It is distant from the other clusterings in the clustering similarity graph and has a very low σ_{Global} value. The cluster map shows that B1 and B2 also have low σ_{Local} values, making these clusters dubious. The disconnected cluster extents (shown for B2) and the centroid placement of B1 further confirm this. The extremal bands of B1 indicate that it also contains flowers with smaller sepal widths, just like B2—this clustering is thus far from optimal, just as indicated by our measures.

Three clusters. We next analyse clusterings with $k = 3$. The spatial proximity of many clusterings and high edge opacities in the clustering similarity graph indicate a strong overlap between most of the partitions. All clusterings are assigned values of $\sigma_{\text{Global}} \geq 0.90$, except C. We now compare C with D because the clustering similarity graph shows them to be most dissimilar. From the cluster glyphs, we observe that C1 and D1 are very similar. However, the cluster extents of D1 show that it misses lots of *I. setosa* flowers—breaking up this pronounced cluster results in a lower σ_{Local} value in comparison to C1. This demonstrates how our measure σ_{Local} helps assess the individual clusters.

More clusters. The clustering similarity graph also helps evaluate the behaviour of clusterings with a larger amount of clusters. Figure 3, top, shows clusterings for increasing values of k. Already for $k = 4$, all clusterings have $\sigma_{\text{Local}} < 0.80$, meaning that less than 80% of the geometrical-topological features are retained globally by the clustering. For $k = 5$ and $k = 6$, a few clusterings remain stable (with respect to σ_{Global}) because they prefer splitting up *I. versicolor* and *I. virginica* prior to splitting up the more compact and concise *I. setosa* cluster. This retains some features of the data. We also observe that clusterings become progressively dissimilar, as indicated by edges with higher transparency, because there are more possibilities for partitioning the data points. This demonstrates how the clustering similarity graph, in combination with σ_{Global}, can be used to quickly explore the overall suitability of different partitions without having to explore them on a local level.

4.2. Olive oils

We use the "Olive oils" data from the UCI Machine Learning Repository [Lic13]. It contains the ratios of 8 fatty acids (palmitic, palmitoleic, stearic, oleic, linoleic, arachidic, linolenic and eicosenoic) for 572 olive oils, produced in 9 different regions of Italy. All variables have been scaled to represent a ratio between zero and one. We explore the data set using different clusterings. Figure 4, top, shows several clustering similarity graphs for the "Olive oils" data. We can see that after $k = 9$ clusters, σ_{Global} values start to decay, i.e. clusterings only retain 60%–80% of the geometrical-topological features in the data. Existing clustering validity indices cannot detect that either $k = 3$ or $k = 9$ are suitable values for k. Please refer to the supplementary materials for more details.

Three clusters. We first compare clusterings A and B with each other. The cluster glyphs show that A1 and A2 contain oils without any eicosenoic acid. The bands also show how A1 differs from A2. Oils in A1 have e.g. lower amounts of oleic acid and higher amounts of linoleic acid than oils in A2. A3 is characterized by non-zero amounts of eicosenoic acid. The cluster extents—coloured according to the cluster label—show that A1 and A2 are smaller than A3. Few overlaps occur and the boundaries are placed in sparse areas, whereas centroids are placed near dense areas—-see e.g. A2.

Clustering B has a lower σ_{Global} value. Its cluster map indicates that lots of overlaps exist between the clusters. One partition, B3, has a high σ_{Local} value that makes it a potentially interesting partition. In total, however, this clustering captures less geometrical-

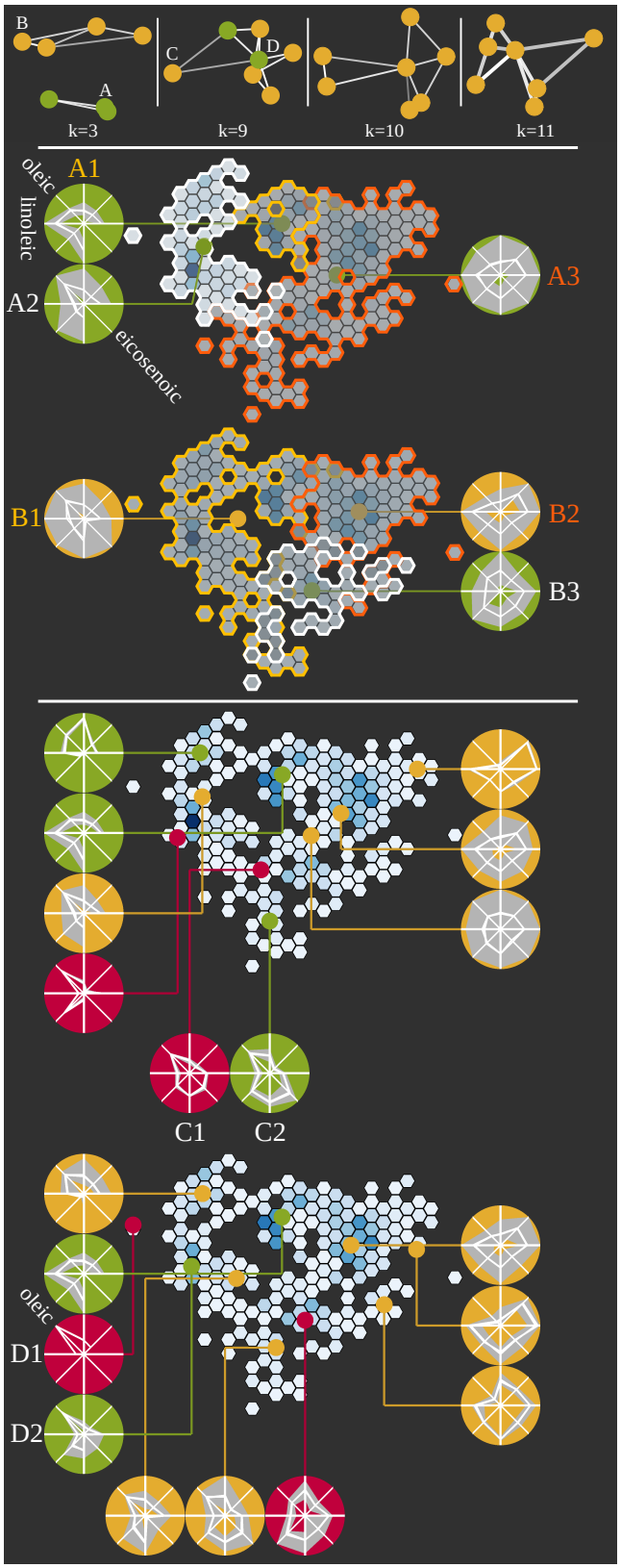


Figure 4: *Clustering similarity graphs (top) and selected cluster maps for the "Olive oils" data.*

topological variation in the data and demonstrates the benefits of using both σ_{Global} and σ_{Local} to assess clusterings.

Nine clusters. The clustering similarity graph for $k = 9$ (Figure 4, top) indicates that all clusterings satisfy $\sigma_{\text{Global}} \geq 0.60$. The distances in the graph show that the clusterings are rather similar, except for a single outlying one (C). Cluster C1 contains only few oils and shares similar characteristics to oils in cluster C2, whose centroid is located nearby in the map—the split between C1 and C2 thus seems arbitrary and is penalized by σ_{Local} because these oils are connected on all scales in the Rips graph due to their similar composition. This clustering was created by the DBSCAN algorithm; a slight perturbation of its parameters results in merging these clusters, which shows that the split was not justified in the first place. Clustering D has the best σ_{Global} value of 0.951. Here, we observe two new subgroups (D1 and D2) that do not appear anywhere else. Their oils have a substantial amount of oleic acid. However, the oils in D1 are too similar to some oils in D2, leading our measure to consider this as a problematic cluster. Interestingly, neither D1 nor D2 are consistent with respect to the original classes of the data—their oils come from three different regions in Italy. In summary, this analysis shows how our geometrical-topological assessment helps (i) detect informative clusterings that go beyond class label information (ii) and may even be able to support the detection and correction of instabilities in clustering algorithms.

4.3. El Niño

El Niño refers to a powerful pattern in world climate that is characterized by a distinct anomaly in sea surface temperatures in the Pacific Ocean. The formation of El Niño is still not fully understood, but it is known that the phenomenon causes catastrophic weather in many parts of the earth. It occurs at irregular intervals (3–7 years) and may last up to 2 years. In the following, we will analyse the "El Niño" data set from the UCI Machine Learning Repository [Lic13]. The data set contains 178080 buoy measurements of five continuous attributes in the Pacific Ocean, comprising a period of 18 years. The complex parameter space and its size make this data set very challenging to cluster. All clustering algorithms that we employed exhibited problems when handling the data. These are caused by the large number of measurements, which often differ only by small amounts. We observed decreases in both σ_{Local} and σ_{Global} values already for $k \geq 3$ (see Figure 5, top).

First, we exemplarily discuss a clustering with two clusters (A). With $\sigma_{\text{Global}} \approx 0.978$, it retains more than 97% of the geometrical-topological features of the data. Figure 5 shows the corresponding cluster maps. The data set is displayed as having a high-density core with density decreasing towards the boundary. The visualized cluster extents indicate that the partitions barely overlap. The cluster glyphs show that A2 contains measurements with, on average, much warmer air temperatures (AT) and sea surface temperatures (SST) than A1. Their σ_{Local} values are high, so we consider both clusters to be trustworthy. Since El Niño is commonly associated with abnormally warm sea surface temperatures, clustering A is highly informative. Referring back to the data set, we found that measurements from A2 indeed predominantly arose in El Niño years.

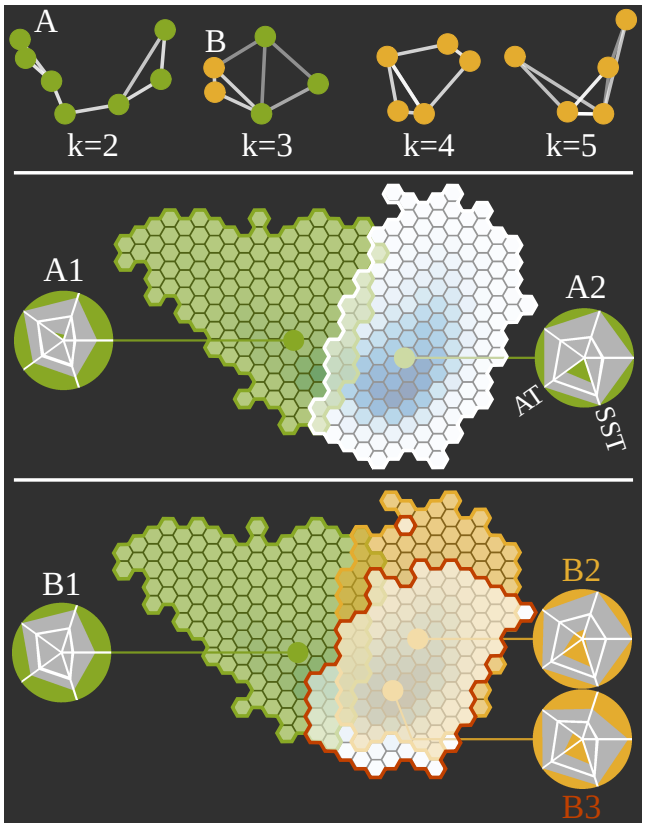


Figure 5: *Selected clustering similarity graphs and cluster maps for the "El Niño" data set.*

Next, we discuss a clustering with three clusters (B). It has $\sigma_{\text{Global}} \approx 0.77$, hence almost 80% of the features of the data are retained. The cluster extents in clustering B overlap more than the ones for clustering A. In particular, B2 and B3 have a large overlap. The glyph colours show that B2 and B3 have $\sigma_{\text{Local}} < 0.80$. Each cluster loses about 25% of the geometrical-topological variation. This loss is apparent in our glyph visualization. The glyph bands depict mean values and spread, but this does not explain why data points have been assigned to either B2 or B3 instead of remaining in one cluster. The clear distinction between "extraordinary measurements" and "regular measurements" as present in clustering A is not apparent here. This clustering also shows the advantages of assessing clusters individually: The glyph for B1 indicates that it retains at least 80% of the features in the data. B1 has $\sigma_{\text{Local}} \approx 0.999$, meaning that it fits the local structure of the data extremely well. In summary, even though clustering algorithms found this data set challenging, their results still reveal useful information about patterns in the data. Our visualization, combined with the values for σ_{Global} and σ_{Local} guides our attention and ensures that we do not have to treat clustering results as "black boxes".

5. Conclusion

We presented two visualizations for supporting users in exploring and comparing different clusterings of multivariate data sets.

Globally, our *clustering similarity graphs* permit the rapid exploration of different clusterings by arranging them, using a clustering similarity measure. Locally, our *cluster maps* create a shared reference coordinate system coupled with glyphs for representing individual clusters that supports the comparison of clusters among each other and among different clusterings. Our visualizations are driven by two measures based on persistent homology that assess the geometrical-topological properties of a clustering and individual clusters. We demonstrated the utility of our visualizations by analysing three data sets of varying complexities.

Both our visualization and our measures can be realized efficiently but the visualizations do not scale to substantially more than about 10 dimensions. The cluster glyphs become unwieldy for more dimensions. A similar issue pertains to the cluster map when more than 10 clusters are present. These cases require a *details-on-demand* approach. Nonetheless, these limitations still leave room for interesting data sets. For higher-dimensional data sets, the selection of a suitable dimensionality reduction algorithm for generating the cluster map is also important. An evaluation algorithm based on persistent homology was recently proposed [RL15].

For future work, the integration of higher-dimensional topological features could be investigated. This requires a generalization of the Rips graph (the *Vietoris–Rips complex*) whose computation is more complicated. The methods described in this paper still remain fully applicable. Furthermore, it would be interesting to augment the calculation of our measures with metrics between persistence diagrams, such as the *bottleneck distance* [EH10, pp. 180–185]. This is challenging because the metrics are not designed to quantify partial matches between persistence diagrams.

References

[AGM*13] ARBELAITZ O., GURRUTXAGA I., MUGUERZA J., PÉREZ J. M., PERONA I.: An extensive comparative study of cluster validity indices. *Pattern Recognition 46*, 1 (2013), 243–256. 2

[BdFF*08] BIASOTTI S., DE FLORIANI L., FALCIDIENO B., FROSINI P., GIORGI D., LANDI C., PAPALEO L., SPAGNUOLO M.: Describing shapes by geometrical-topological properties of real functions. *ACM Computing Surveys 40*, 4 (2008), 1–87. 2, 3

[BETT99] BATTISTA G. D., EADES P., TAMASSIA R., TOLLIS I. G.: *Graph drawing: Algorithms for the visualization of graphs*. Prentice Hall, 1999. 5

[Car14] CARLSSON G.: Topological pattern recognition for point cloud data. *Acta Numerica 23* (2014), 289–368. 2, 3

[CCKT83] CHAMBERS J. M., CLEVELAND W. S., KLEINER B., TUKEY P. A.: *Graphical methods for data analysis*. Wadsworth & Brooks/Cole Publishing Company, 1983. 5

[CdSO14] CHAZAL F., DE SILVA V., OUDOT S. Y.: Persistence stability for geometric complexes. *Geometriæ Dedicata 173*, 1 (2014), 193–214. 3

[CLNL87] CARR D. B., LITTLEFIELD R. J., NICHOLSON W. L., LITTLEFIELD J. S.: Scatterplot matrix techniques for large N. *JASA 82*, 398 (1987), 424–436. 5

[CSEH07] COHEN-STEINER D., EDELSBRUNNER H., HARER J.: Stability of persistence diagrams. *DCG 37*, 1 (2007), 103–120. 3

[CSEH09] COHEN-STEINER D., EDELSBRUNNER H., HARER J.: Extending persistence using Poincaré and Lefschetz duality. *FoCM 9*, 1 (2009), 79–103. 3

[CSEHM10] COHEN-STEINER D., EDELSBRUNNER H., HARER J., MILEYKO Y.: Lipschitz functions have L_p-stable persistence. *FoCM 10*, 2 (2010), 127–139. 3

[EDF08] ELMQVIST N., DRAGICEVIC P., FEKETE J.-D.: Rolling the dice: Multidimensional visual exploration using scatterplot matrix navigation. *IEEE TVCG 14*, 6 (2008), 1539–1148. 2

[EH10] EDELSBRUNNER H., HARER J.: *Computational topology: An introduction*. AMS, 2010. 2, 4, 9

[ELZ02] EDELSBRUNNER H., LETSCHER D., ZOMORODIAN A.: Topological persistence and simplification. *DCG 28*, 4 (2002), 511–533. 3

[HBV01] HALKIDI M., BATISTAKIS Y., VAZIRGIANNIS M.: On clustering validation techniques. *Journal of Intelligent Information Systems 17*, 2–3 (2001), 107–145. 1, 2

[HVW10] HOLTEN D., VAN WIJK J. J.: Evaluation of cluster identification performance for different PCP variants. *Computer Graphics Forum 29*, 3 (2010), 793–802. 2

[Jai10] JAIN A. K.: Data clustering: 50 years beyond *k*-means. *Pattern Recognition Letters 31*, 8 (2010), 651–666. 1, 2

[Lic13] LICHMAN M.: UCI machine learning repository, 2013. URL: http://archive.ics.uci.edu/ml. 6, 7, 8

[LSL*13] LUM P. Y., SINGH G., LEHMAN A., ISHKANOV T., VEJDEMO-JOHANSSON M., ALAGAPPAN M., CARLSSON J., CARLSSON G.: Extracting insights from the shape of complex data using topology. *Scientific Reports 3* (2013), 1–8. 2, 3

[LSP*10] LEX A., STREIT M., PARTL C., KASHOFER K., SCHMALSTIEG D.: Comparative analysis of multidimensional, quantitative data. *IEEE TVCG 16*, 6 (2010), 1027–1035. 2

[Mei07] MEILĂ M.: Comparing clusterings—an information based distance. *Journal of Multivariate Analysis 98*, 5 (2007), 873–895. 5

[Nad64] NADARAYA E. A.: On estimating regression. *Theory of Probability & Its Applications 9*, 1 (1964), 141–142. 3

[NHZI07] NAM E. J., HAN Y., ZELENYUK K. M. A., IMRE D.: ClusterSculptor: A visual analytics tool for high-dimensional data. In *IEEE VAST* (2007), pp. 75–82. 2

[PGU12] PILHÖFER A., GRIBOV A., UNWIN A.: Comparing clusterings using Bertin's idea. *IEEE TVCG 18*, 12 (2012), 2506–2515. 2

[PVG*11] PEDREGOSA F., VAROQUAUX G., GRAMFORT A., MICHEL V., THIRION B., GRISEL O., BLONDEL M., PRETTENHOFER P., WEISS R., DUBOURG V., VANDERPLAS J., PASSOS A., COURNAPEAU D., BRUCHER M., PERROT M., DUCHESNAY É.: Scikit-learn: Machine learning in Python. *JMLR 12* (2011), 2825–2830. 1, 4, 5

[RHBK15] REININGHAUS J., HUBER S., BAUER U., KWITT R.: A stable multi-scale kernel for topological machine learning. In *IEEE CVPR* (2015). 2

[RL14] RIECK B., LEITTE H.: Structural analysis of multivariate point ccloud using simplicial chains. *Computer Graphics Forum 33*, 8 (2014), 28–37. 2

[RL15] RIECK B., LEITTE H.: Persistent homology for the evaluation of dimensionality reduction schemes. *Computer Graphics Forum 34*, 3 (2015), 431–440. 3, 9

[RML12] RIECK B., MARA H., LEITTE H.: Multivariate data analysis using persistence-based filtering and topological signatures. *IEEE TVCG 18*, 12 (2012), 2382–2391. 2, 3

[SBvLK09] SCHRECK T., BERNARD J., VON LANDESBERGER T., KOHLHAMMER J.: Visual cluster analysis of trajectory data with interactive Kohonen maps. *Information Visualization 8*, 1 (2009), 14–29. 2

[SMC07] SINGH G., MÉMOLI F., CARLSSON G.: Topological methods for the analysis of high dimensional data sets and 3D object recognition. In *Proceedings of the Eurographics Symposium on Point-Based Graphics* (2007). 2, 3

[SS02] SEO J., SHNEIDERMAN B.: Interactively exploring hierarchical clustering results. *Computer 35*, 7 (2002), 80–86. 2

[TAE*09] TATU A., ALBUQUERQUE G., EISEMANN M., SCHNEIDEWIND J., THEISEL H., MAGNOR M., KEIM D.: Combining automated analysis and visualization techniques for effective exploration of high-dimensional data. In *IEEE VAST* (2009), pp. 59–66. 5

[TMF*12] TATU A., MAASS F., FÄRBER I., BERTINI E., SCHRECK T., SEIDL T., KEIM D.: Subspace search and visualization to make sense of alternative clusterings in high-dimensional data. In *IEEE VAST* (2012), pp. 63–72. 2

[XT15] XU D., TIAN Y.: A comprehensive survey of clustering algorithms. *Annals of Data Science 2*, 2 (2015), 165–193. 2

[ZM14] ZAKI M. J., MEIRA JR. W.: *Data mining and analysis: Fundamental concepts and algorithms*. Cambridge University Press, 2014. 1

Eurographics Conference on Visualization (EuroVis) 2016
K-L. Ma, G. Santucci, and J. J. van Wijk
(Guest Editors)

Volume 35 (2016), Number 3

Exploring and Comparing Clusterings of Multivariate Data Sets Using Persistent Homology: Supplementary materials

B. Rieck[1,2] and H. Leitte[1]

[1]TU Kaiserslautern, Germany
[2]Heidelberg University, Germany

1. Notation

We assume that we are given a data set X of n points in some $\mathbb{R}^m$. Furthermore, we assume that we may calculate the distance $\mathrm{d}(x,y)$ between two points x and y. This permits us to calculate the $n \times n$ distance matrix

$$D = \left\{D_{ij}\right\}_{i,j=1}^{n} \tag{1}$$

with

$$D_{ij} = d(x_i, x_j) \tag{2}$$

being the distance between the ith and the jth data point.

Let C be a *clustering* with k clusters, i.e. $\mathrm{C} = \{\mathrm{C}_1, \ldots, \mathrm{C}_k\}$, where cluster C_i contains $n_i = |\mathrm{C}_i|$ points. Given two subsets U and V of our input data, we define $D(U,V)$ as the sum of distances with one index in U and the other in V, i.e.

$$D(U,V) = \sum_{x \in U} \sum_{y \in V} d_{ij}, \tag{3}$$

which is always well-defined. We denote the *complement* of a set U by $\overline{U}$.

There are two specific sets of distances we are interested in. First, the *intracluster distances* are given as

$$D_{\text{intra}} = \frac{1}{2} \sum_{i=1}^{k} D(\mathrm{C}_i, \mathrm{C}_i), \tag{4}$$

where we need the division by two because we count every pair of distances twice. Second, the *intercluster distances* are similarly given as

$$D_{\text{inter}} = \frac{1}{2} \sum_{i=1}^{k} D(\mathrm{C}_i, \overline{\mathrm{C}_i}), \tag{5}$$

with the same division as above.

Since the distance matrix D is symmetric and has a diagonal of zero, we may also consider D to be the *weighted adjacency matrix* of the complete graph over our data points. This makes it possible to count the number of intracluster edges N_{intra} and the number of intercluster edges N_{inter} as

$$N_{\text{intra}} = \frac{1}{2} \sum_{i=1}^{k} n_i \left(n_i - 1\right) \tag{6}$$

and

$$N_{\text{inter}} = \sum_{i=1}^{k} \sum_{j=i+1}^{k} n_i n_j, \tag{7}$$

respectively.

2. Clustering validity indices

If no ground truth information in the form of labels for the data points is available, there are numerous *clustering validity indices* that measure certain properties of a clustering C by means of the distance matrix D. Subsequently, we briefly introduce several common clustering validity indices. We will later compare them with the *global clustering assessment measure* σ_{Global} that we describe in the paper. A comparison with our local measure σ_{Local} is impossible because no clustering validity index is capable of assessing a single cluster on its own.

BetaCV. The BetaCV measure calculates the ratio between the mean intracluster distance to the mean intercluster distance, i.e.

$$\text{BetaCV} = \frac{D_{\text{intra}}/N_{\text{intra}}}{D_{\text{inter}}/N_{\text{inter}}} = \frac{N_{\text{inter}}\, D_{\text{intra}}}{N_{\text{intra}}\, D_{\text{inter}}}, \tag{8}$$

where small values are considered to be better because they indicate that intracluster distances are, on average, smaller than intercluster distances. In this case the clusters are well-separated.

C-index. The C-indexrelates the intracluster distances to the sum of the largest distances in the distance matrix. We

have

$$\text{C-index} = \frac{D_{\text{intra}} - D_{\min}(N_{\text{intra}})}{D_{\max}(N_{\text{intra}}) - D_{\min}(N_{\text{intra}})}, \tag{9}$$

where D_{intra} is again the sum of all intracluster distances, $D_{\min}(N_{\text{intra}})$ is the sum of the N_{intra} smallest distances in the distance matrix D (not including the diagonal), and $D_{\max}(N_{\text{intra}})$ is the sum of the N_{intra} largest distances. The C-index has values in $[0,1]$. Smaller values are considered to be better because they indicate compact clusters.

Within-cluster-scatter. The *within-cluster-scatter WCS* is another name for the intracluster distances D_{intra} that we already encountered above. Small values are considered good. The *k*-means algorithm attempts to minimize this measure.

Dunn index. The Dunn index *Dunn* measures the ratio between the minimum distance between points from different clusters and the maximum distance between points from the same cluster. We have

$$\text{Dunn} = \frac{D_{\text{inter}}^{\min}}{D_{\text{intra}}^{\max}}, \tag{10}$$

where

$$D_{\text{inter}}^{\min} = \min_{i \neq j} \{ \mathrm{d}(x,y) \mid x \in \mathrm{C}_i, y \in \mathrm{C}_j \} \tag{11}$$

is the *minimum intercluster distance* and

$$D_{\text{intra}}^{\max} = \max_{i} \{ \mathrm{d}(x,y) \mid x,y \in \mathrm{C}_i \} \tag{12}$$

is the *maximum intracluster distance*. A large Dunn index corresponds to a good clustering because it indicates that even the closest distance between points in different clusters is larger than the maximum distance within a cluster. Hence, the Dunn index is maximized when we have very compact clusters that are extremely far from each other.

NC. The *normalized cut* measure is motivated by graph-theoretic cuts. If we take a single cluster C_i from the clustering C, the distances of all edges with at least one vertex in the cluster is an indicator of the volume of the C_i. We denote this sum of distances by $D(\mathrm{C}_i, X)$. If we consider C_i to induce a cut in the graph, the weight of the cut is given by all edges that go outside the cluster C_i. Hence, C_i induces a cut whose weight is $D(\mathrm{C}_i, \overline{\mathrm{C}_i})$. The normalized cut measure *NC* now measures the total sum of the ratio between the cut weight and the volume of the cluster, i.e.

$$\text{NC} = \sum_{i=1}^{k} \frac{D(\mathrm{C}_i, \overline{\mathrm{C}_i})}{D(\mathrm{C}_i, X)}, \tag{13}$$

where higher values indicate better clusterings because they imply that the intercluster edges have larger distances than the intracluster edges. Again, small intracluster distances in comparison to intercluster distances are indicative of a good clustering.

Silhouette coefficient. The *silhouette coefficient s* measures both the separation of clusters as well as their internal connectivity. We first calculate a *silhouette coefficient* s_x as

$$s_x = \frac{b_x - a_x}{\max\{a_x, b_x\}}, \tag{14}$$

where a_x is the average distance of point x to all other points within its cluster, and b_x is the average of all distances of points x to points in the closest other cluster. We have $s_x \in [-1,+1]$, where $+1$ shows that x is much closer to points in its own cluster and removed from other clusters, 0 indicates that x is on a cluster boundary, and -1 indicates that x is closer to another cluster than its own—which may indicate a mis-clustered point. The silhouette coefficient of a clustering C is defined as the mean value of s_x across all points.

3. Expressive power

We observed that the expressive power of our persistence-based measure, in conjunction with a suitable shape descriptor for the data set, often outperforms existing clustering validity indices. To this end, we added some tables showing the values for indices introduced above.

3.1. Synthetic data: "Nested circles"

Table 1 shows the performance of clustering validity indices on the "nested circles" data set. For the first three indices, lower values are generally better. For the subsequent indices, higher values indicate better clusterings. We have marked the best value in each column. For the first data set, only the *Dunn index* is capable of detecting a good clustering. In Section 5 we will see that this index is prone to severe instabilities, whereas our persistence-based measure σ_{Global} remains stable.

Note that our measure is able to detect the correct clustering over extremely large scales in the data. Even if the two circles are being connected with some edges, these edges will only introduce short-lived topological features. The value of σ_{Global} thus remains unchanged.

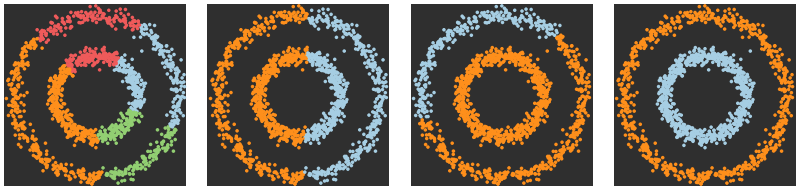
Figure 1 shows the edges of the Rips graph $\mathcal{R}_\varepsilon$ for different thresholds. Our measure is stable over the whole range of these thresholds.

3.2. Synthetic data: "Nested arcs"

For the "nested arcs" data, shown in Table 2, the *Dunn index* and our measure are the only measures capable of detecting a suitable clustering. Again, the Dunn index is not stable. Even if only three points—0.2% of the data—are misclassified, the index drops by a factor of 40 (see Section 5).

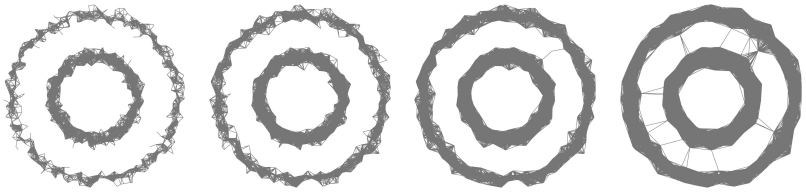
Again, our measure is stable over large scales and even permits the two arcs to be connected by some edges. It is interesting to note that so many clustering validity indices

ID	BetaCV	C-index	WCS	Dunn index	NC	Silhouette	σ_{Global}
A	0.305	0.080	619.96	0.0099	4.76	0.35	0.141
B	0.0	0.221	1052.08	0.0064	1.52	0.35	0.373
C	0.0	0.279	1170.42	0.0182	1.35	0.27	0.328
D	0.897	0.436	1283.52	0.0637	2.39	−0.09	1.0



A B C D

Table 1: Different clusterings of the "nested circles" data. From the common clustering validity indices, only the Dunn index is capable of finding the correct partition. Its value is unstable, though (Section 5).



(a) $\varepsilon = 0.15$ (b) $\varepsilon = 0.20$ (c) $\varepsilon = 0.30$ (d) $\varepsilon = 0.50$

Figure 1: Edge sets of the Rips graph $\mathcal{R}_\varepsilon$ of the "nested circles" data set, for varying values of ε. Our heuristic (see the paper) suggests using $\varepsilon = 0.20$, but we can see that there is still a lot of leeway in both directions for good values for ε.

ID	BetaCV	C-index	W	Dunn index	NC	Silhouette	σ_{Global}
A	0.255	0.0559	499.96	0.008	4.80	0.456	0.034
B	0.0	0.1824	966.32	0.005	1.56	0.388	0.104
C	0.47	0.0858	856.01	0.011	1.62	0.499	0.069
D	0.559	0.1741	957.39	0.157	1.56	0.389	1.0

A B C D

Table 2: Different clusterings of the "nested arcs" data. For the perfect clustering, the Dunn index performs very well. In Section 5, we will see that its value drops by a factor of 40 when we change the assignment of only three points.

consider the first clustering, i.e. the ones with largest amount of clusters, to be best. We feel that an index should rather be biased towards *fewer* clusters because clusters should explain global as well as local aspects of a data set.

Very small clusters may fit the data locally very well, but they often do not yield global information—as is the case for clustering A of the "nested arcs" data.

3.3. Synthetic data: "Gaussian blobs"

Table 3 shows the results of several clustering validity indices for the "Gaussian blobs" data. We argue that only clusterings B and D are "meaningful" in the sense that they properly express spatial proximity. Most validity indices tend to favour clustering D.

Our measure is incapable of detecting differences between clustering A, clustering B, and clustering D because the blobs are well-separated. When approximating the connectivity of our data using the Rips graph, as detailed in the paper, our heuristic will never create edges that go from one "blob" to another "blob". Hence, our persistence-based measure cannot detect any differences between different splits of these three components.

For this data, our visualizations will be useful in showing differences between clusterings A and B, for example: Both the *clustering similarity graph* and the *cluster map* will show that the clusterings differ significantly. This example also stresses the importance of using suitable shape descriptor functions.

3.4. Synthetic data: "Uniform distribution"

We also include a somewhat controversial data set consisting of uniformly-distributed points in $\mathbb{R}^2$. Table 4 shows the numerical results for several clustering validity indices. We argue that only clustering D—which assigns all points to a single cluster—is true to the structure of the data. If parts of a data set are truly random, a clustering validity index should not rate statistically arbitrary partitions to be suitable. As the numerical results show, none of the existing clustering validity indices is capable of assessing data set D properly.

The definition of some of the indices does not permit us to calculate them on a single partition. Even if we slightly change the assignment of some points to a dummy cluster and leave the majority of the points in a larger cluster, the results do not change. In particular, all indices except for BetaCV considers clustering A to be the most suitable.

3.5. Real-world data: "Iris"

As we state in the paper, we use the "Iris" data set as an example because its clusters are already sufficiently challenging. Since we have labels available, we can use them to calculate the *Rand index* of the clustering. This number indicates the percentage of correct cluster assignments made by an algorithm. We calculated numerous clusterings of the data set, including the correct label assignment.

Table 5 shows the results for $k = 3$, sorted by ascending clustering quality. We can see that our measure is the only one that is able to detect the correct clustering. We also note that we are unable to retain all topological information. Precisely because the cluster boundaries are not well-defined, we will invariably lose some information. Furthermore, the table also shows that a second good candidate is given by the clustering in the third row. It does not follow the original cluster boundaries, though.

A similar behaviour is observable for the other clustering indices as well. This again demonstrates the challenges with the "Iris" data in particular and with clustering analysis in general: If the original definition of the labels clashes with the original definition of the features one is looking for, clustering validity indices will not perform well. Our measure is less prone to these issues because it looks for large-scale features in the shapes of the different clusters.

3.6. Real-world data: "Olive oils"

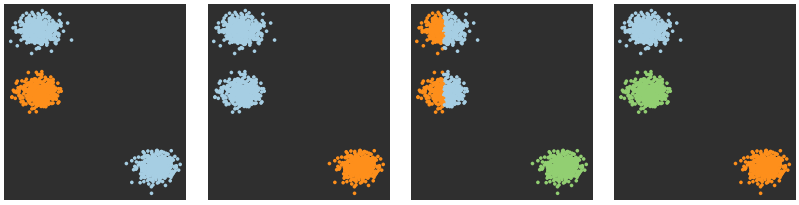
We can perform the same analysis for the "Olive oils" data, for $k = 3$ and $k = 9$. For both numbers of clusters, we calculated different partitions and sorted them according to their *Rand index*, which indicates the percentage of correct cluster assignments made by the algorithm.

Table 7 shows the results for $k = 3$ clusters. We note that only the *Dunn index* and our measure are capable of detecting the "best" clustering. Note that the Dunn index is again very unstable—the clustering with a Rand index of 0.986, which gets assigned $\sigma_{\text{Global}} = 0.96$ is rated even worse as the clustering with a Rand index of 0.759.

For $k = 9$, the different clusterings become more similar to each other. The boundaries of the "real" clusters do not always follow the geometry of the data. Hence, our measure is incapable of detecting the "correct" label assignment—along with all other clustering validity indices. Table 8 shows the results for all measures. We can see, however, that our measure is consistent in its evaluation of the clusterings. Starting from a Rand index of approximately 0.89, we consider all clusterings to describe the data equally well. This is where our visualizations, coupled with our σ_{Local} measure can be used to find out in what ways the clusterings differ.

These clusterings also illustrate a general problem with clustering algorithms: With an increasing number of clusters, it gets easier to find some reasonable structure in the data. Hence, the Rand indices of different algorithms is more or less similar. Getting the clustering algorithm to cluster the last 10% of the data correctly cannot always be done—often, this requires *supervised* clustering algorithms.

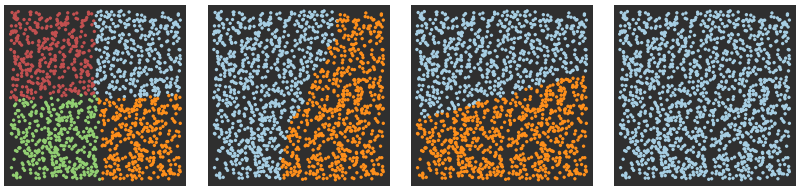
ID	BetaCV	C-index	W	Dunn index	NC	Silhouette	σ_{Global}
A	0.0	0.4360	295.94	0.08	1.476	0.2662	1.0
B	0.0	0.0019	127.79	0.92	1.787	0.8010	1.0
C	0.1530	0.1235	507.58	0	2.772	0.2795	0.66
D	0.1140	0.0013	61.84	0.29	2.914	0.8041	1.0



A B C D

Table 3: Different clusterings of the "Gaussian blobs" data. This data set has a very simple geometry and the individual clusters are well-separated. Hence, almost all clustering indices are able to detect useful clusterings. Since the "blobs" are separated on a large scale and do not contain any prominent topological features, our measure cannot discern between three of the clusterings.

ID	BetaCV	C-index	W	Dunn index	NC	Silhouette	σ_{Global}
A	0.4234	0.079	669.55	0.0129	3.737	0.4132	0.64
B	0	0.222	1052.54	0.0020	1.514	0.3456	0.64
C	0	0.213	1045.64	0.0045	1.522	0.3527	0.66
D	NaN	1.0	1354.86	NaN	0	NaN	1.0



A B C D

Table 4: Different clusterings of the "Uniform distribution" data. We argue that only clustering D is true to the structure in the data. While the split in clustering A is uniform with respect to the cluster sizes, it is somewhat arbitrary. Note that σ_{Local} of the individual clusters will still be very high because they are good subsets of the data.

Rand index	BetaCV	C-index	W	Dunn index	NC	Silhouette	σ_{Global}
0.741	0.267	0.154	117.64	0.048	2.648	0.349	0.356
0.743	0.263	0.151	104.97	0.052	2.677	0.367	0.809
0.821	0.251	0.098	94.40	0.090	2.712	0.429	0.949
0.824	0.215	0.110	98.94	0.041	2.705	0.415	0.790
0.825	0.233	0.091	93.07	0.098	2.721	0.446	0.842
0.828	0.385	0.089	90.60	0.026	2.733	0.458	0.857
0.857	0.394	0.092	90.95	0.058	2.731	0.450	0.861
1.0	0.419	0.118	97.23	0.074	2.710	0.380	0.967

Table 5: Clustering validity indices for several partitions of the "Iris" data set for $k = 3$ clusters. It is interesting to note that most clustering validity indices do not exhibit better values as the Rand index increases.

Rand index	BETACV	C-INDEX	W	DUNN INDEX	NC	SILHOUETTE	σ_{Global}
0.803	0.363	0.106	83.76	0.059	3.740	0.370	0.895
0.814	0.244	0.102	92.14	0.059	3.720	0.257	0.695
0.823	0.266	0.083	81.76	0.034	3.765	0.412	0.603
0.828	0.263	0.084	82.35	0.105	3.763	0.399	0.697

Table 6: Clustering validity indices for several partitions of the "Iris" data set for $k = 4$ clusters. Since we only have $k = 3$ labels, we cannot achieve a Rand index of 1.0 here. Our measure rates a refinement of a hierarchical clustering best. Note that there is still a significant difference between the ratings for $k = 3$ and $k = 4$ for our measure. This does not hold for the other measures. Most of the measures get *better* for $k = 4$. For higher values of k, the effects get even worse.

Rand index	BETACV	C-INDEX	W	DUNN INDEX	NC	SILHOUETTE	σ_{Global}
0.695	0.429	0.195	771.91	0.0892	2.57	0.280	0.77
0.698	0.407	0.141	777.52	0.0476	2.57	0.307	0.82
0.720	0.360	0.154	769.35	0.0564	2.58	0.301	0.86
0.759	0.348	0.166	754.69	0.0873	2.59	0.314	0.88
0.825	0.361	0.133	761.47	0.0187	2.58	0.312	0.92
0.986	0.495	0.209	783.43	0.0809	2.55	0.251	0.96
1.0	0.476	0.197	778.26	0.1506	2.56	0.256	1.0

Table 7: Clustering validity indices for several partitions of the "Olive oils" data set with $k = 3$ clusters. Only two measures, the Dunn index and ours, are capable of detecting the best clustering. For $k = 3$, the cluster boundaries follow the geometry very well.

4. Limitations

We already alluded in the paper—and in the "Gaussian blobs" example data—that our measure cannot distinguish between clusterings where parts of a cluster are disconnected on large scales. This implies that we cannot use our measure to assess the *similarity* of clusterings. This limitation does not imply, however, that we are biased with respect to the number of clusters. To show this, we conducted a series of experiments on data sets such as the one shown in Table 9. We varied the number of circles between 2–100, and perturbed their coordinates.

We can see that our measure considers clusterings where "nearby" circles are in a different cluster, such as clustering B, in a similar manner than clusterings where "nearby" circles are in wholly different clusters, such as clustering D.

In practice, this means that when calculating clusterings for different values of k, our measure does not necessarily decrease with an increasing number of clusters. In our experiments, the σ_{Global} changes only in the decimal place of magnitude around 10^{-4}, meaning that the difference between a single cluster with $\sigma_{\text{Global}} = 1.0$ and k clusters for k linked circles is of the order of 10^{-4} and thus negligible.

Rand index	BETACV	C-INDEX	W	DUNN INDEX	NC	SILHOUETTE	σ_{Global}
0.820	0.413	0.162	704.55	0.0719	8.64	0.112	0.86
0.890	0.410	0.082	507.51	0.0670	8.75	0.288	0.99
0.908	0.405	0.062	545.15	0.0871	9.72	0.303	0.97
0.915	0.444	0.096	513.32	0.0392	8.74	0.287	0.97
0.917	0.414	0.055	508.05	0.1184	8.75	0.331	0.99
0.921	0.366	0.051	507.00	0.1016	8.75	0.332	0.98
0.929	0.363	0.071	553.97	0.1090	8.72	0.203	0.97
1.0	0.406	0.075	10153.30	0.0827	8.75	0.320	0.97

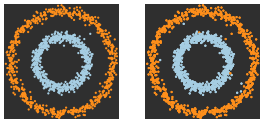
Table 8: Clustering validity indices for several partitions of the "Olive oils" data set with $k = 9$ clusters. No measure is capable of detecting the correct label assignment. Our measure assesses almost all partitions with a high Rand index similarly. This is caused by very small clusters that do not contribute any geometrical-topological information.

ID	BETACV	C-INDEX	W	DUNN INDEX	NC	SILHOUETTE	σ_{Global}
A	NaN	1.0	190.75	NaN	0.0	NaN	1.0
B	0.0	0.56	178.64	0.0059	1.34	0.036	0.9999
C	0.85	0.43	175.50	0.0043	1.40	0.139	0.8091
D	0.40	0.09	95.62	0.0591	2.72	0.379	0.9997

A B C D

Table 9: An excerpt of a series of experiments with a data set of "linked circles". This sort of data poses no significant challenge for most clustering algorithms. We can see that the values of our measure barely differ for clustering A, clustering B, and clustering D. We thus consider all of these splits to be equally valid.

Measure	A	B
BETACV	0.897	0.901
C-INDEX	0.436	0.437
WCS	1283.52	1285.31
DUNN INDEX	0.0637	0.0123
NC	2.39	1.39
SILHOUETTE	-0.09	0.11
σ_{Global}	1.00	0.997



A B

Table 10: Stability behaviour for the validity measures on the "nested circles" data.

Measure	A	B
BETACV	0.559	0.557
C-INDEX	0.1741	0.1713
WCS	957.39	954.22
DUNN INDEX	0.157	0.0042
NC	1.563	1.1565
SILHOUETTE	0.389	0.3923
σ_{Global}	1.0	0.993

A B

Table 11: Stability behaviour for the validity measures on the "nested arcs" data.

5. Stability

As the previous tables indicate, the clustering validity indices are not stable with respect to their assessment of a clustering. The re-assignment of a small number of points may result in large changes in the measure.

Our measure is not prone to these instabilities. We show this for two of the example data sets only—however, we observed this in all of the data sets that we were working with.

As an experiment, we slightly modified the perfect clusterings of the synthetic data sets to show the effects of noise in the data: We randomly changed the assignment of a fraction of the points in the data set. We were somewhat surprised by the results: Even if less than 0.5% of the points are being assigned incorrectly, most clustering validity indices changed drastically.

5.1. Synthetic data: "Nested circles"

Table 10 shows one example result for the "nested circles" data. When we add more noise to the data set, our persistence-based measure remains stable at approximately 99% of explained topological variation. The *Dunn index*—previously capable of determining that the given clustering was suitable—drops to around 20% of its previous value. Similarly, the *silhouette coefficient* changes by 0.2, which is a shift of 10% of its value range. The remaining indices remain somewhat stable but are still incapable of determining this to be a suitable clustering.

5.2. Synthetic data: "Nested arcs"

Table 11 shows one example result for the "nested arcs" data. Again, our measure remains stable and changes only by 0.7% of its value range. Again, the *Dunn index* changes drastically by dropping to only 3% of its previous value. The remaining indices also exhibit some instabilities.