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.

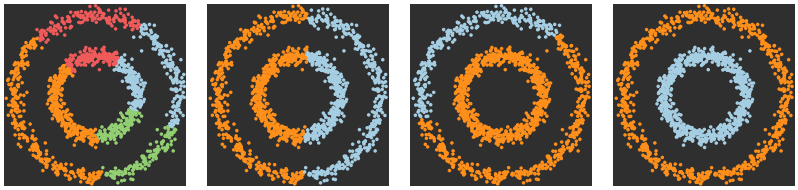
Figure 1 shows the edges of the Rips graph $\mathcal{R}_\varepsilon$ for different thresholds. Our measure is stable over the whole range of these thresholds.

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).

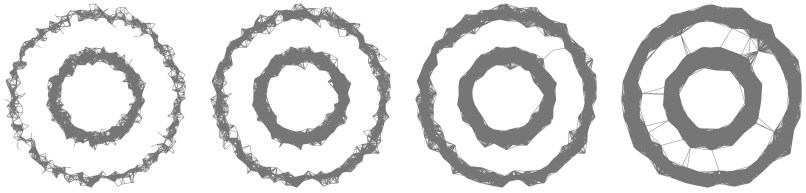
Again, our measure is stable over large scales and even permits the two arcs to be connected by some edges. It is interesting to note that so many clustering validity indices

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"

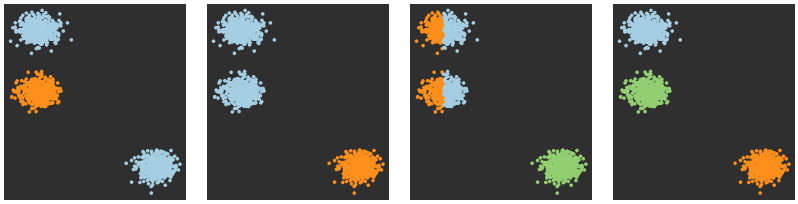
We can perform the same analysis for the "Olive oils" data, for $k = 3$ and $k = 9$. For both numbers of clusters, we calculated different partitions and sorted them according to their *Rand index*, which indicates the percentage of correct cluster assignments made by the algorithm.

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.

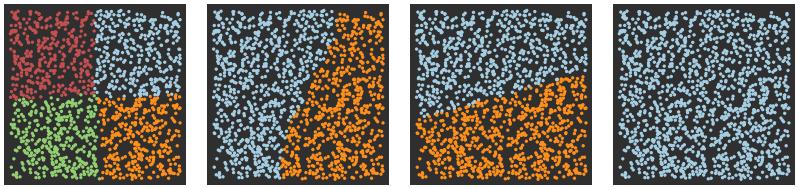
ID	BetaCV	C-index	W	Dunn index	NC	Silhouette	σ_{Global}
A	0.0	0.4360	295.94	0.08	1.476	0.2662	1.0
B	0.0	0.0019	127.79	0.92	1.787	0.8010	1.0
C	0.1530	0.1235	507.58	0	2.772	0.2795	0.66
D	0.1140	0.0013	61.84	0.29	2.914	0.8041	1.0



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	10 153.30	0.0827	8.75	0.320	0.97

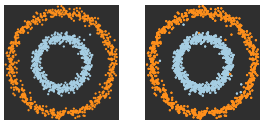
Table 8: Clustering validity indices for several partitions of the "Olive oils" data set with $k = 9$ clusters. No measure is capable of detecting the correct label assignment. Our measure assesses almost all partitions with a high Rand index similarly. This is caused by very small clusters that do not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