# Clustering Validity Assessment: Finding the optimal partitioning of a data set

Maria Halkidi Michalis Vazirgiannis
*Dept of Informatics,*
*Athens University of Economics & Business*
*Email: {mhalk, mvazirg }@aueb.gr*

## Abstract

*Clustering is a mostly unsupervised procedure and the majority of the clustering algorithms depend on certain assumptions in order to define the subgroups present in a data set. As a consequence, in most applications the resulting clustering scheme requires some sort of evaluation as regards its validity.*
*In this paper we present a clustering validity procedure, which evaluates the results of clustering algorithms on data sets. We define a validity index, S_Dbw, based on well-defined clustering criteria enabling the selection of the optimal input parameters' values for a clustering algorithm that result in the best partitioning of a data set. We evaluate the reliability of our index both theoretically and experimentally, considering three representative clustering algorithms ran on synthetic and real data sets. Also, we carried out an evaluation study to compare S_Dbw performance with other known validity indices. Our approach performed favorably in all cases, even in those that other indices failed to indicate the correct partitions in a data set.*

## 1. Introduction and Motivation

In the literature a wide variety of algorithms have been proposed for different applications and sizes of data sets [14]. The application of an algorithm to a data set aims at, assuming that the data set offers such a tendency (clustering tendency), discovering its real partitions. This implies that i. all the points that naturally belong to the same cluster will eventually be attached to it by the algorithm, ii. no additional data set points (i.e., outliers or points of another cluster) will be attached to the cluster.

In most algorithms' experimental evaluations [1, 6, 10, 11, 12, 17], 2D-data sets are used in order the reader is able to visually verify the validity of the results (i.e., how well the clustering algorithm discovered the clusters of the data set). It is clear that visualization of the data set is a crucial verification of the clustering results. In the case of large multidimensional data sets (e.g. more than three dimensions) effective visualization of data can be difficult. Moreover the perception of clusters using available visualization tools is a difficult task for the humans that are not accustomed to higher dimensional spaces.
The various clustering algorithms behave in a different way depending on i)the features of the data set (geometry and density distribution of clusters), ii) the input parameters values.

Assuming that the data set includes distinct partitions (i.e., inherently supports clustering tendency), the second issue becomes very important. In Figure 1 we can see the way an algorithm (e.g. DBSCAN[6]) partition a data set having different input parameter values. It is clear that only some specific values for the algorithms' input parameters lead to optimal partitioning of the data set. As it is evident, if there is no visual perception of the clusters it is impossible to assess the validity of the partitioning. What is then needed is a visual-aids-free assessment of some objective criterion, indicating the validity of the results of a clustering algorithm application on a potentially high dimensional data set. In this paper we define and evaluate a cluster validity index (S_Dbw). Assuming a data set S, the index enables the selection of optimal input parameter values for a clustering algorithm that best partition S.

The rest of the paper is organized as follows. Section 2 surveys the related work. We motivate and define the validity index in Section 3, while in Section 4 we provide a theoretical and experimental evaluation of S_Dbw using different algorithms and data sets. Furthermore we compare our approach to other validity indices. In Section 5 we conclude by briefly presenting our contributions and indicate directions for further research.

## 2. Related Work

The fundamental clustering problem is to partition a given data set into groups (clusters), such that the data points in a cluster are more similar to each other than points in different clusters [10]. In the clustering process, there are no predefined classes and no examples that would show what kind of desirable relations should be valid among the data [2]. This is what distinguishes clustering from classification [7, 8].

There is a multitude of clustering methods available in the literature, which can be broadly classified into the following types [11, 14]: i) *Partitional clustering ii)Hierarchical clustering, iii) Density Based clustering,* iv) *Grid-based clustering*.

For each of these types there exists a wealth of subtypes and different algorithms [1, 11, 12, 13, 14, 17, 19, 22, 24] for finding the clusters. In general terms, the clustering algorithms are based on a criterion for judging the validity of a given partitioning. Moreover, they define a partitioning of a data set based on certain assumptions and *not* the optimal one that fits the data set.

Since clustering algorithms discover clusters, which are not known a priori, the final partition of a data set requires some sort of evaluation in most applications [18]. A


Eps=2, Nps=4 Eps=6, Nps=4

Figure 1: The different partitions resulting from running DBSCAN with different input parameter values.

particularly difficult problem, which is often ignored in clustering algorithms is "how many clusters are there in the data set?".

Previously described requirements for the evaluation of clustering results is well known in the research community and a number of efforts have been made especially in the area of pattern recognition [22]. However, the issue of cluster validity is rather under-addressed in the area of databases and data mining applications, even though recognized as important. In general terms, there are three approaches to investigate cluster validity [22]. The first is based on *external criteria*. This implies that we evaluate the results of a clustering algorithm based on a pre-specified structure, which is imposed on a data set and reflects our intuition about the clustering structure of the data set. The second approach is based on *internal criteria*. We may evaluate the results of a clustering algorithm in terms of quantities that involve the vectors of the data set themselves (e.g., proximity matrix). The third approach of clustering validity is based on *relative criteria*. Here the basic idea is the evaluation of a clustering structure by comparing it with other clustering schemes, resulting by the same algorithm but with different parameter values. A number of validity indices have been defined and proposed in the literature for each of above approaches [22]. A cluster validity index for crisp clustering proposed in [4], attempts to identify "compact and well-separated clusters". Other validity indices for crisp clustering have been proposed in [3] and [16]. The implementation of most of these indices is very computationally expensive, especially when the number of clusters and number of objects in the data set grows very large [25]. In [15], an evaluation study of thirty validity indices proposed in the literature is presented. The results of this study place Caliski and Harabasz(1974), Je(2)/Je(1) (1984), C-index (1976), Gamma and Beale among the six best indices. However, it is noted that although the results concerning these methods are encouraging they are likely to be data dependent. For fuzzy clustering [22], Bezdek proposed the partition coefficient (1974) and the classification entropy (1984). The limitations of these indices are [3]: i) their monotonous dependency on the number of clusters, and ii) the lack of direct connection to the geometry of the data. Other fuzzy validity indices are proposed in [9, 25, 18]. We should mention that the evaluation of proposed indices and the analysis of their reliability are limited.

Another approach for finding the best number of cluster of a data set proposed in [21]. It introduces a practical clustering algorithm based on Monte Carlo cross-validation. This approach differs significantly from the one we propose. While we evaluate clustering schemes based on widely recognized validity criteria of clustering, the evaluation approach proposed in [21] is based on density functions considered for the data set. Thus, it uses concepts related to probabilistic models in order to estimate the number of clusters, better fitting a data set, while we use concepts directly related to the data.

## 3. Validity index definition

In this research effort we focused on *relative* criteria where the algorithm is running repetitively using different input values and the resulting clusters are compared as for their validity.

The criteria widely accepted for partitioning a data set into a number of clusters are: i. the *separation* of the clusters, and ii. their *compactness*. However, the data set is falsely partitioned in most of the cases, whereas only specific values for the algorithms' input parameters lead to optimal partitioning of the data set. Here the term "optimal" implies parameters that lead to partitions that are as close as possible (in terms of similarity) to the real partitions of the data set.

Therefore our o*bjective* is the definition of a relative [22] algorithm-independent validity index, for assessing the quality of partitioning for each set of the input values. Such a validity index should be able to select for each algorithm under consideration the optimal set of input parameters with regard to a specific data set.

The criteria (i.e., compactness and separation) on which the proposed index is partially based are the fundamental criteria of clustering. However, the algorithms aim at satisfying these criteria based on initial assumptions (e.g. initial locations of the cluster centers) or input parameter values (e.g. the number of clusters, minimum diameter or number of points in a cluster). For instance the algorithm DBSCAN[6] defines clusters based on density variations, considering values for the cardinality and radius of an object's neighborhood. It finds the best partitions for the given input values but we don't know if the resulting partitions are the optimal or even the ones presented in the underlying data set.

The above motivated us to take in account density variations among clusters. We formalize our clustering validity index, S_Dbw, based on: i. clusters' compactness (in terms of intra-cluster variance), and ii.density between clusters (in terms of inter-cluster density).

Let $D=\{v_i| i=1,\ldots,c\}$ a partitioning of a data set S into *c* convex clusters where $v_i$ is the center of each cluster as it results from applying a clustering algorithm to S.

Let *stdev* be the average standard deviation of clusters defined as: $stdev = \frac{1}{c}\sqrt{\sum_{i=1}^{c}\|\sigma(v_i)\|}$.

Further the term ||x|| is defined as : $\|x\| = (x^T x)^{1/2}$, where x is a vector.

Then the overall inter-cluster density is defined as:

**Definition 1.** *Inter-cluster Density* (ID) - It evaluates the average density in the region among clusters in relation with the density of the clusters. The goal is the density among clusters to be significant low in comparison with the density in the considered clusters. Then, we can define inter-cluster density as follows:

$$Dens\_bw(c) = \frac{1}{c\cdot(c-1)}\sum_{i=1}^{c}\left(\sum_{\substack{j=1\\ i\neq j}}^{c}\frac{density(u_{ij})}{\max\{density(v_i), density(v_j)\}}\right), \quad (1)$$

where $v_i$, $v_j$ centers of clusters $c_i$, $c_j$, respectively and $u_{ij}$ the middle point of the line segment defined by the clusters' centers $v_i$, $v_j$ . The term density(u) defined in equation(2):

$$density(u) = \sum_{l=1}^{n_{ij}} f(x_l, u), \text{ where } n_{ij} = \text{number of tuples} \quad (2)$$

that belong to the clusters $c_i$ and $c_j$, *i.e.*, $x_l \in c_i \cup c_j \subseteq S$

represents the number of points in the neighborhood of u. In our work, the neighborhood of a data point, u, is defined to be a hyper-sphere with center u and radius the average standard deviation of the clusters, *stdev*. More specifically, the function *f(x,u)* is defined as:

$$f(x,u) = \begin{cases} 0, & \text{if } d(x,u) > stdev \\ 1, & \text{otherwise} \end{cases} \quad (3)$$

It is obvious that a point belongs to the neighborhood of u if its distance from u is smaller than the average standard deviation of clusters. Here we assume that the data have been scaled to consider all dimensions (bringing them into comparable ranges) as equally important during the process of finding the neighbors of a multidimensional point [2].

**Definition 2.** *Intra-cluster variance* - Average scattering for clusters. The average scattering for clusters is defined as:

$$Scat(c) = \frac{1}{c}\sum_{i=1}^{c}\|\sigma(v_i)\| \Big/ \|\sigma(S)\| \quad (4)$$

The term σ(S) is the variance of a data set; and its $p_{th}$ dimension is defined as follows:

$$\sigma_x^p = \frac{1}{n}\sum_{k=1}^{n}\left(x_k^p - \bar{x}^p\right)^2 \quad (4a)$$

where $\bar{x}^p$ is the $p_{th}$ dimension of $\bar{X} = \frac{1}{n}\sum_{k=1}^{n} x_k, \forall x_k \in S$ (4b)

The term $\boldsymbol{\sigma}(v_i)$ is the variance of cluster $c_i$ and its $p_{th}$ dimension is given by

$$\sigma_{v_i}^p = \sum_{k=1}^{n_i}\left(x_k^p - v_i^p\right)^2 \Big/ n_i \quad (4c)$$

Then the validity index S_Dbw is defined as:

$$S\_Dbw(c) = Scat(c) + Dens\_bw(c) \quad (5)$$

The definition of S_Dbw indicates that both criteria of "good" clustering (i.e., compactness and separation) are properly combined, enabling reliable evaluation of clustering results. The first term of S_Dbw, *Scat(c)*, indicates the average scattering within c clusters. A small value of this term is an indication of compact clusters. *Dens_bw(c)* indicates the average number of points between the *c* clusters (i.e., an indication of inter-cluster density) in relation with density within clusters. A small *Dens_bw(c)* value indicates well-separated clusters. The number of clusters, *c*, that minimizes the above index can be considered as an optimal value for the number of clusters present in the data set.

## 4. Validity Index Evaluation

In this section we evaluate the proposed validity index S_Dbw both theoretically (at some points we offer intuitive proof sketches) and experimentally.

### 4.1 Theoretical Evaluation

Let a data set S containing convex clusters (as in Figure 2a) and various ways to partition it using different clustering algorithms (Figure 2b-e). Assume the optimal partitioning of data set S (as it is appeared in Figure 2a) in three clusters. The number of clusters as it emanates from the case of optimal partitioning is further called "correct number of clusters". We assume that the data set is evenly distributed, i.e., on average similar number of data points are found for each surface unit in the clusters.

**Lemma 1:** *Assume a data set S with convex clusters and a clustering algorithm A applied repetitively to S, each time with different input parameter values $P_i$, resulting in different partitions $D_i$ of S. The value of S_Dbw is minimized when the correct number of clusters is found.*

**Proof:** Let *n* the correct number of clusters of the data set S corresponding to the partitioning $D_1$ (optimal partitioning of S): $D_1(n, S) = \{c_{D1i}\}$, i=1,..,n and *m* the number of clusters of another partitioning $D_2$ of the same data set: $D_2(m, S) = \{c_{D2j}\}$, j=1,..,m.

Let $S\_Dbw_{D1}$ and $S\_Dbw_{D2}$ be the values of the validity index for the respective partitioning schemes. Then, we consider the following cases:

i) Assume $D_2$ to be a partitioning where more than the actual clusters are formed (i.e., m>n). Moreover, parts of the actual clusters (corresponding to $D_1$) are grouped into clusters of $D_2$ (as in Figure 2d). Let $fC_{D1} = \{fc_{D1p} \mid p=1, \ldots, nfr1\ fc_{D1p} \subseteq c_{D1i}, i=1,\ldots,n\}$ a set of fractions of clusters in $D_1$. Similarly, we define $fC_{D2} = \{fc_{D2k} \mid k=1, \ldots, nfr2, fc_{D2k} \subseteq c_{D2j}, j=1,\ldots, m\}$. Then:

a) $\exists\ c_{D2j}$: $c_{D2i} = \cup fc_{D1p}$ , where $p=p_1,\ldots., p_n$, $p_1 >= 1$ and $p_n <= nfr1$, nfr1 is the number of considered fractions of clusters in $D_1$, and b) $\exists\ c_{D1i}$: $c_{D1i} = \cup fc_{D2k}$ , where $k=k_1,\ldots., k_n$, $k_1 >= 1$ and $k_n <= nfr2$, where nfr2 is the number of considered fractions of clusters in $D_2$,

In this case, some of the clusters in $D_2$ include regions of low density (for instance cluster 3 in Figure 2d). Thus, the value of the first term of the index related to intra-cluster variance of $D_2$ increases as compared to the intra-cluster variance of $D_1$ (i.e., Scat(m) >Scat(n)). On the other hand,

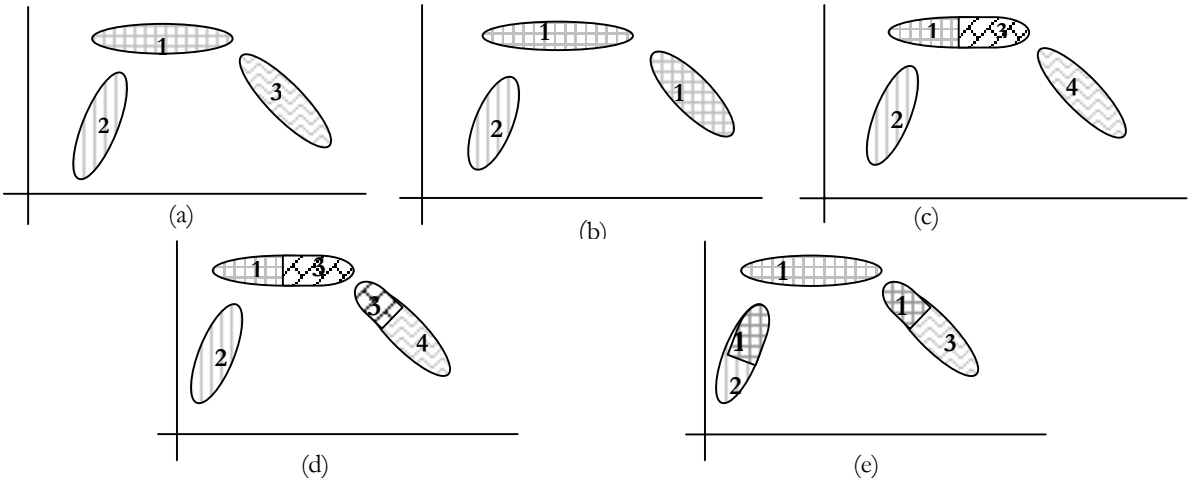

Figure 2: A data set S partitioned in (a) the actual three clusters, (b) two, (c,d) in four clusters and (e) falsely in three clusters

the second term (inter-cluster density) is also increasing as compared to the corresponding term of index for $D_1$ (i.e., Dens_bw(m)>Dens_bw(n)). This is because some of the clusters in $D_1$ are split and therefore there are border areas between clusters that are of high density (e.g., clusters 1 and 3 in Figure 3d). Then, since both S_Dbw terms regarding $D_2$ partitioning increase we conclude that $S\_Dbw_{D1} < S\_Dbw_{D2}$.

ii) Let $D_2$ be a partitioning where more clusters than in $D_1$ are formed (i.e., m>n). Also, we assume that at least one of the clusters in $D_1$ is split to more than one in $D_2$ while no parts of $D_1$ clusters are grouped into $D_2$ clusters (as in Figure 2c), i.e., $\exists\ c_{D1i} : c_{D1i} = \cup\ c_{D2j}$, $j=k_1,\ldots,$ k and $k_1>=1$, k<=m. In this case, the value of the first term of the index related to intra-cluster variance slightly decreases compared to the corresponding term of $D_1$ since the clusters in $D_2$ are more compact. As a consequence Scat(m)<=Scat(n). On the other hand, the second term (inter-cluster density) is increasing as some of the clusters in $D_1$ are split and therefore there are borders between clusters that are of high density (for instance clusters 1 and 3 in Figure 2c). Then Dens(m)>>Dens(n). Based on the above discussion and taking in account that the increase of inter-cluster density is significantly higher than the decrease of intra-cluster variance we may conclude that $S\_Dbw_{D1} < S\_Dbw_{D2}$.

iii) Let $D_2$ be a partitioning with less clusters than in $D_1$ (m<n) and two or more of the clusters in $D_1$ are grouped to a cluster in $D_2$ (as in Figure 2b.). Then, $\exists\ c_{D2j}$: $c_{D2j}=\cup c_{D1i,}$ where $i=p_1, \ldots,$ p and $p_1>=1$, p<=n. In this case, the value of the first term of the index related to intra-cluster variance increases as compared to the value of corresponding term of $D_1$ since the clusters in $D_2$ contain regions of low density. As a consequence, Scat(m)>>Scat(n). On the other hand, the second term of the index (inter-cluster density) is slightly decreasing or remains vaguely the same as compared to the corresponding term of $D_1$ (i.e., Dens_bw(n)≅Dens_bw(m)). This is because similarly to the case of the $D_1$ partitioning (Figure 3a) there are no borders between clusters in $D_2$ that are of high density. Then, based on the above discussion and considering that the increase of intra-cluster variance is significantly higher than the decrease of inter-cluster density, we may conclude that $S\_Dbw_{D1} < S\_Dbw_{D2}$.

**Lemma 2:** *Assume a data set S containing convex clusters and a clustering algorithm A applied repetitively to S, each time with different parameter values $P_i$, resulting in different partitions $D_i$ of S. For each $D_i$ it is true that the correct number of clusters is found. The value S_Dbw is minimized when the optimal partitions are found for the correct number of clusters*.

**Proof:**. We consider $D_2$ to be a partitioning with the same number of clusters as the optimal one $D_1$ (Figure 3a), (i.e., m=n). Furthermore, we assume that one or more of the actual clusters corresponding to $D_1$ are split and their parts are grouped into different clusters in $D_2$ (as in Figure 2e). That is, if $fC_{D1} = \{fc_{D1p} \mid p=1, \ldots, nfr1\ fc_{D1p} \subseteq c_{D1i}, i=1,\ldots,n\}$ a set of clusters fractions in $D_1$ then $\exists\ c_{D2j}$: $c_{D2j} = \cup fc_{D1i}$ , $i=p_1, \ldots,$ p and $p_1>=1$, p<=n. In this case, the clusters in $D_2$ contain regions of low density and as a consequence the value of the first term of the index, intra-cluster variance, increases as compared to the corresponding term of $D_1$, i.e., Scat(m)>Scat(n). On the other hand, some of the clusters in $D_2$ are split and therefore there are border areas between clusters that are of high density (for instance clusters 1, 3 and 1, 2 in Figure 2e). Therefore, the second term (inter-cluster density) of $D_2$ is also increasing as compared to the one of $D_1$, i.e., Dens_bw(m)>Dens_bw(n). Based on the above discussion it is obvious that $S\_Dbw_{D1} < S\_Dbw_{D2}$.

## 4.2 Time Complexity

The complexity of the validity index S_Dbw, is based on the complexity of its two terms as defined in (1) and ( 4). Assuming *d* the number of attributes (data set dimension), c is the number of clusters; *n* is the number of database tuples. Then the intra-cluster variance complexity is O(*ndc*) while the complexity of inter-cluster density is $O(ndc^2)$. Then S_Dbw complexity is $O(ndc^2)$. Usually, *c*, *d* << *n*, therefore the complexity of our index for a specific clustering scheme is O(*n*). The graphs in Figure 3 show the results of an experimental study referring to the execution time of our approach. The considered datasets for these experiments are synthetically generated

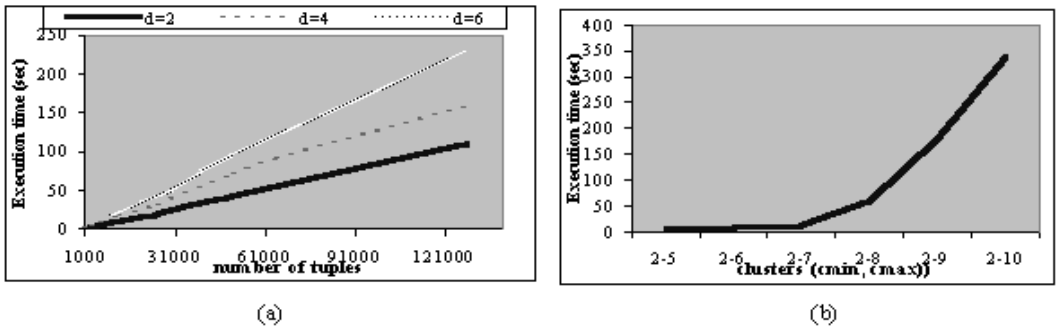


Figure 3: Execution time in seconds as function of (a) the number of points (b) the number of clusters

according to the normal distribution. Figure 3a demonstrates that the execution time is almost linear to the number of points as expected from the preceding complexity study. Furthermore, Figure 3b shows the execution time as function of the number of clusters. The execution time, as expected, is nearly quadratic with respect to the number of clusters but as *c* is usually a small integer, it creates no problem.

## 4.3 Experimental evaluation

In this section S_Dbw is experimentally tested using representative clustering algorithms of different categories: partitional, hierarchical and density-based.
We experiment with real and synthetic multidimensional data sets containing different number of clusters. In all cases our approach performs favorably selecting the best partitioning among these proposed by an algorithm. Additionally we compare S_Dbw to other validity indices found in the literature. In the sequel, due to lack of space, we present only some representative examples of our experimental study.

**4.3.1. Selection of the optimal partitioning defined by a clustering algorithm.** The goal of this experiment is to evaluate our index with regards to the selection of the optimal clustering scheme by a specific clustering algorithm. More specifically, we consider a 2-dimensional data set consisting of four clusters (see Figure 4a). We define a number of different clustering schemes of our data set using the K-means algorithm, with its input parameters (number of clusters) ranging between 2 and 8. The behavior of S_Dbw is depicted in Figure 5. It is clear that the correct number of clusters is proposed (i.e., four), as at this value the index reaches its minimum.
Similarly, we assume the clustering schemes of DataSet3 (see Figure 4c) as defined by CURE when the number of clusters ranges between 2 and 8. Then, we evaluated the defined clustering schemes based on the S_Dbw index so as to find which of them best fits the underlying data. As Figure 6 shows the clustering scheme of seven clusters is proposed as the best partitioning of DataSet3.

*A multidimensional data set.* In the sequel, we demonstrate that our index works properly in multidimensional data sets. The validity of clustering results (i.e., that the set has been optimally partitioned) can be visually verified only in 2D or 3D cases. In higher dimensions it is difficult to verify the resulting clusters. The proposed index, S_Dbw, offers a solution to this problem giving an indication of the best clustering scheme without visualization of the data set. We consider a synthetic six-dimensional data set, containing two distinct clusters. This is also verified by S_Dbw. As Figure 7

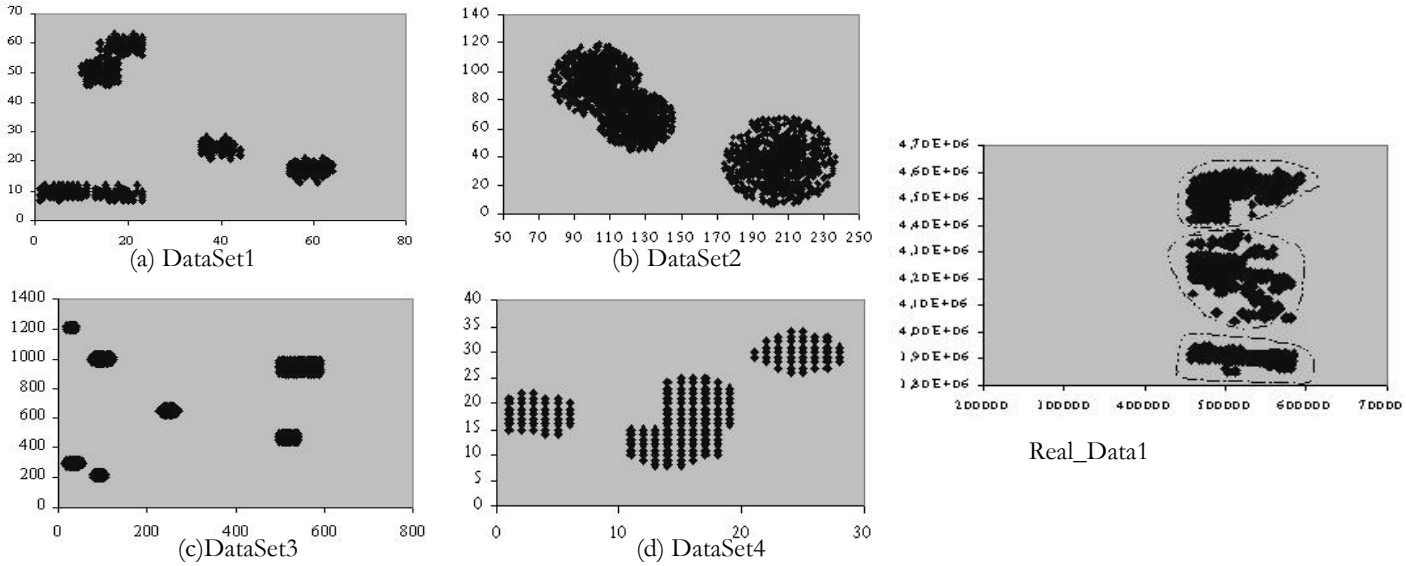


Figure 4: Sample Synthetic (a,b,c,d) datasets and a real dataset (e).

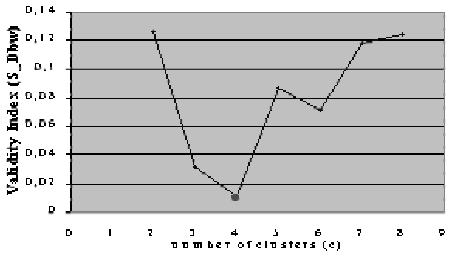


Figure 5:S_Dbw as a function of number of clusters for DataSet1.

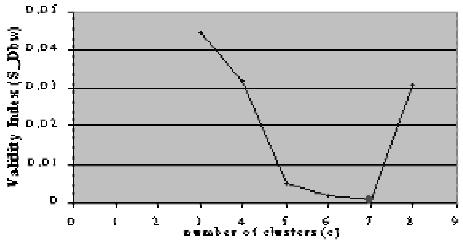


Figure 6: S_Dbw as a function of number of clusters for DataSet3.

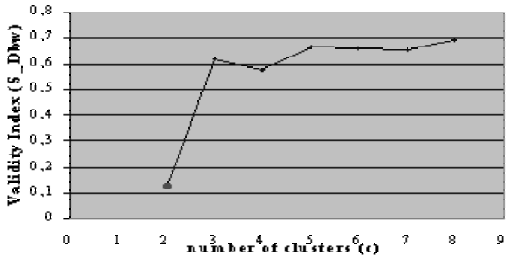


Figure 7:S_Dbw as a function of the number of clusters for a six-dimensional data set consisting of two clusters.

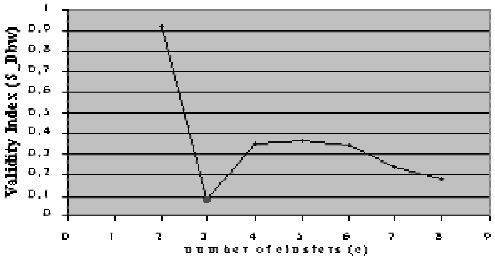


Figure 8: S_Dbw as a function of number of clusters for real data representing a part of Greek road network.

depicts, S_Dbw finds the correct number of clusters as it takes its minimum value when c=2.

*Real data sets.* Clustering spatial databases is an important problem and many applications require the management of spatial data. Therefore evaluate our approach using representative real spatial data sets [26]. One of the data sets, we studied, contains three parts of Greek roads network (Figure 4e). The roads are represented by their MBR approximations' vertices. The behavior of S_Dbw regarding the different clustering schemes (i.e., number of clusters) defined by K-means is depicted in Figure 8. It is clear that S_Dbw indicates the correct number of clusters (three) as the best partitioning for the data set.

**4.3.2. The index is independent of clustering algorithm.** As mentioned in previous sections, different input values for clustering algorithms applied to a dataset result in different partitioning schemes. In the following we show that S_Dbw selects the optimal partitioning among those defined by a clustering algorithm independently of the algorithm used. We use three well-known algorithms, one from each of the popular algorithm categories: K-means (partitional), DBSCAN (density based) and CURE (hierarchical).

Table 1a and Table 1b present S_Dbw values for the resulting clustering schemes for DataSet1, and Real_Data1 (see Figure4a and Figure4e respectively) found by K-means, DBSCAN and CURE respectively. More specifically, we consider the clustering schemes revealed by the algorithms mentioned above while their input parameter values are depicted in Table 1. In the case of DataSet1, all three algorithms propose four clusters as the optimal clustering schemes (see Table 1a). Similarly, considering any of the three algorithms, the proposed validity index selects three clusters as the best partitioning for Real_Data1 (see Table1b).

In some cases, however, an algorithm may partition a data set into the correct number of clusters but in a wrong way. For instance, Figure 9(a) and (b) present the partitioning of Dataset5 into three clusters by K-means and CURE respectively. Though the correct number of clusters is three (3), both algorithms partition it falsely. On the other hand, DBSCAN finds the correct three partitions of the data set (Figure 9(c)). Table 1c presents the behavior of S_Dbw in each of the above cases. If we consider the K-means clustering results, the index proposes four clusters as the best partitioning for our data set. Given the inability of K-means to handle skewed geometries, this result somehow makes sense. The results from running CURE can be interpreted in a similar way. In the case of DBSCAN the index finds the correct number of clusters that is three. The result is that S_Dbw selects the best partitioning among those proposed by different algorithms.

**4.3.3.Comparison to other validity indices.** We consider the known validity indices proposed in the literature, such as RS-RMSSTD[20], DB[22] and the recent one SD[23]. RMSSTD and RS have to be taken into account simultaneously in order to find the correct number of clusters. The optimal values of the number of clusters are those for which a significant local change in values of RS and RMSSTD occurs. As regards DB, an indication of the optimal clustering scheme is the point at which it takes its minimum value. We carried an evaluation study comparing S_Dbw to the indices mentioned above. We used four synthetic two-dimensional data sets further referred to as DataSet1, DataSet2, DataSet3 and DataSet4 (see Figure 4a-d) and the real data set Real_Data1 (Figure 4e), which contains three clusters.

Table 2 summarizes the results of the validity indices (RS, RMSSDT, DB, SD and S_Dbw), for different clustering schemes of the above-mentioned data sets as resulting from a clustering algorithm. For our study, we use the results of the algorithms K-Means and CURE with their input value (number of clusters), ranging between 2 and 8. Indices RS, RMSSTD propose the partitioning of

Table 1: Optimal partitioning found by S_Dbw for each algorithm

| | K-means | | DBSCAN | | CURE r=10, a=0.3) | |
|---|---|---|---|---|---|---|
| **No clusters** | **Input** | **S_Dbw Value** | **Input** | **S_Dbw Value** | **Input** | **S_Dbw Value** |
| 6 | C=6 | 0.0712 | Eps=2, MinC=8 | 0.087 | C=6 | 0.082 |
| 5 | C=5 | 0.0866 | Eps=2, MinC=4 | 0.0865 | C=5 | 0.091 |
| **4** | **C=4** | **0.0104** | **Eps=10, MinC=15** | **0.0104** | **C=4** | **0.0104** |
| 3 | C=3 | 0.0312 | Eps=15, MinC=15 | 0.0312 | C=3 | 0.031 |
| 2 | C=2 | 0.1262 | Eps=20, MinC=15 | 0.1262 | C=2 | 0.126 |

(a) DataSet1

| | K-means | | DBSCAN | | CURE (r=10, a=0.3) | |
|---|---|---|---|---|---|---|
| **No clusters** | **Input** | **S_Dbw Value** | **Input** | **S_Dbw Value** | **Input** | **S_Dbw Value** |
| 6 | C=6 | 0.3434 | - | - | C=6 | 0.3434 |
| 5 | C=5 | 0.367 | - | - | C=5 | 0.3670 |
| 4 | C=4 | 0.35 | Eps=20000, MinC =10 | 0.1925 | C=4 | 0.3501 |
| **3** | **C=3** | **0.083** | **Eps=30000, MinC=10** | **0.084** | **C=3** | **0.0831** |
| 2 | C=2 | 0.9189 | Eps=50000, MinC=10 | 0.891 | C=2 | 0.9188 |

(b) Real_Data1

| | K-Means | | DBSCAN | | CURE (r=10, a=0.3) | |
|---|---|---|---|---|---|---|
| **No clusters** | **Input** | **S_Dbw Value** | **Input** | **S_Dbw Value** | **Input** | **S_Dbw Value** |
| 6 | C=6 | 0.1585 | - | - | C=6 | 0.1767 |
| 5 | C=5 | 0.1354 | _ | _ | C=5 | 0.1648 |
| 4 | **C=4** | **0.0329** | _ | _ | **C=4** | **0.0446** |
| 3 | C=3 | 0.1094 | **Eps=2,MinC=4** | **0.0404** | C=3 | 0.0597 |
| 2 | C=2 | 0.4374 | Eps=5,MinC=4 | 1.7228 | C=2 | 0.4096 |

(c) DataSet5

DataSet1 into three clusters while DB selects six clusters as the best partitioning. On the other hand, SD and S_Dbw select four clusters as the best partitioning for DataSet1, which is also the correct number of clusters fitting the underlying data. Moreover, the indices S_Dbw and DB select the correct number of clusters (i.e., seven) as the optimal partitioning for DataSet3 while RS, RMSSTD and SD select the clustering scheme of five and six clusters respectively. Also, all indices propose three clusters as the best partitioning for Real_Data1. In the case of DataSet2, DB and SD select three clusters as the optimal scheme, while RS-RMSSDT and S_Dbw select two clusters (i.e., the correct number of clusters fitting the data set).

Here, we have to mention that S_Dbw *is not a clustering algorithm itself* but a measure to evaluate the results of clustering algorithms and gives an indication of a partitioning that best fits a data set. The essence of clustering is not a totally resolved issue and depending on the application domain we may consider different aspects as more significant. For instance, for a specific application it may be important to have well separated clusters while for another to consider more the compactness of the clusters. In this case, the relative importance of the two terms on which the S_Dbw definition is based can be adjusted. Having an indication of a good partitioning as proposed by the index, the domain experts may analyze further the validation procedure results. Thus, they could select some of the partitioning schemes proposed by S_Dbw, and select the one better fitting their demands for crisp or overlapping clusters. For instance DataSet2 can be considered as having three clusters with two of them slightly overlapping or having two well-separated clusters. In this case we observe that S_Dbw values for two and three clusters are not significantly different (0,12, 0,22 respectively). This is an indication that we may select either of the two partitioning schemes depending on the clustering interpretation. Then, we compare the values of *Scat* and *Dens_bw* terms for the cases of two and three clusters. We observe that the two clusters scheme

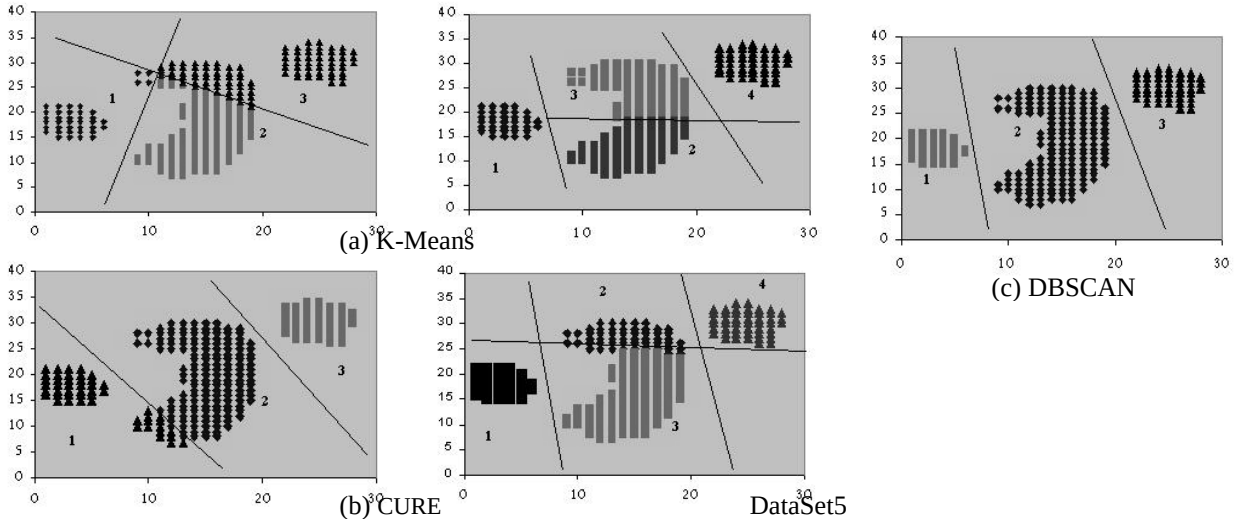


Figure 9: A 2D data set partitioned into (a) three and four clusters using K-Means, (b) three and four clusters using CURE and (c) three clusters using DBSCAN.

Table 2: Optimal number of clusters proposed by validity indices compared with S_Dbw

| | DataSet1 | DataSet2 | DataSet3 | DataSet4 | RealData1 |
|---|---|---|---|---|---|
| | Optimal number of clusters | | | | |
| ***RS, RMSSTD*** | **3** | **2** | **5** | **4** | **3** |
| ***DB*** | **6** | **3** | **7** | **4** | **3** |
| ***SD*** | **4** | **3** | **6** | **3** | **3** |
| ***S_Dbw*** | **4** | **2** | **7** | **3** | **3** |

corresponds to well-separated clusters (Dens_bw(2)= 0.09762<Dens_bw(3)=0.2154) while the three-clusters scheme contains more compact clusters (Scat(2)= 0.0215>Scat(3)= 0.0109).

Moreover, S_Dbw finds the correct number of clusters (three) for DataSet4, on the contrary to RS – RMSSTD and DB indices, which propose four clusters as the best partitioning. In all cases S_Dbw finds the correct number of clusters fitting a data set, while other validity indices fail in some cases.

## 5. Conclusions and Further Work

In this paper we addressed the important issue of assessing the validity of clustering algorithms' results. We have defined a new validity index (S_Dbw) for assessing the results of clustering algorithms. The index is optimized for data sets that include compact and well-separated clusters. The compactness of the data set is measured by the cluster variance whereas the separation by the density between clusters.

We have proved S_Dbw reliability and value i. theoretically, by illustrating the intuition behind it and ii.experimentally, using various data sets of non-standard (but in general non-convex) geometries covering also the multidimensional case. The index results, as indicated by experiments, are not dependent on the clustering algorithm used, and always indicate the optimal input parameters for the algorithm used in each case. It performs better than the most recent validity indices proposed in the literature as it was indicated by experimental evaluation.

**Further work.** As we mentioned earlier the validity assessment index we proposed in this paper does not work properly in the case of clusters of non-convex (i.e., rings) or extraordinarily curved geometry. We are going to work on this issue as the density and its continuity is not any more sufficient criteria. We plan to use sets of representative points, or even multidimensional curves rather than a single center point.

## Acknowledgements

This work was supported by the General Secretariat for Research and Technology through the PENED ("99EΔ 85") project. We would like to thank Y. Batistakis for his help in the experimental study. We are also thankful to C. Rodopoulos and C. Amanatidis for the implementation of CURE algorithm as well as to Drs Joerg Sander and Eui-Hong (Sam) Han for providing information and the source code for DBSCAN and CURE algorithms respectively.

## References

[1] Rakesh Agrawal, Johannes Gehrke, Dimitrios Gunopulos, Prabhakar Raghavan,"Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications", *Proceedings of SIGMOD*, 1998.

[2] Michael J. A. Berry, Gordon Linoff. Data Mining Techniques For marketing, Sales and Customer Support. John Willey & Sons, Inc, 1996.

[3] Rajesh N. Dave. "Validating fuzzy partitions obtained through c-shells clustering", *Pattern Recognition Letters*, Vol .17, pp613-623, 1996

[4] J. C. Dunn. "Well separated clusters and optimal fuzzy partitions", *J. Cybern.* Vol.4, pp. 95-104, 1974

[5] Martin Ester, Hans-Peter Kriegel, Jorg Sander, Michael Wimmer, Xiaowei Xu. "Incremental Clustering for Mining in a Data Warehousing Environment", *Proceedings of 24th VLDB Conference*, New York, USA, 1998.

[6] Martin Ester, Hans-Peter Kriegel, Jorg Sander, Xiaowei Xu. "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise", *Proceedings of 2nd Int. Conf. On Knowledge Discovery and Data Mining*, Portland, OR, pp. 226-231, 1996.

[7] Usama M. Fayyad, Gregory Piatesky-Shapiro, Padhraic Smuth and Ramasamy Uthurusamy. *Advances in Knowledge Discovery and Data Mining*. AAAI Press 1996

[8] Usama Fayyad, Ramasamy Uthurusamy. "Data Mining and Knowledge Discovery in Databases", *Communications of the ACM*. Vol.39, No11, November 1996.

[9] I. Gath, B. Geva. "Unsupervised Optimal Fuzzy Clustering". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol 11, No7, July 1989.

[10] Sudipto Guha, Rajeev Rastogi, Kyueseok Shim. "CURE: An Efficient Clustering Algorithm for Large Databases", *Published in the Proceedings of the ACM SIGMOD Conference*, 1998.

[11] Sudipto Guha, Rajeev Rastogi, Kyueseok Shim. "ROCK: A Robust Clustering Algorithm for Categorical Attributes", *Published in the Proceedings of the IEEE Conference on Data Engineering*, 1999.

[12] Alexander Hinneburg, Daniel Keim. "An Efficient Approach to Clustering in Large Multimedia Databases with Noise". *Proceeding of KDD '98*, 1998.

[13] Zhexue Huang. "A Fast Clustering Algorithm to Cluster very Large Categorical Data Sets in Data Mining", *DMKD*, 1997

[14] A.K Jain, M.N. Murty, P.J. Flyn. "Data Clustering: A Review", *ACM Computing Surveys*, Vol. 31, No3, September 1999.

[15] Milligan, G.W. and Cooper, M.C. (1985), "An Examination of Procedures for Determining the Number of Clusters in a Data Set", *Psychometrika*, 50, 159-179.

[16] Milligan G. W., Soon S.C., Sokol L. M. "The effect of cluster size, dimensionality and the number of clusters on recovery of true cluster structure". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 5, pp. 40-47, 1983

[17] Raymond Ng, Jiawei Han. "Efficient and Effective Clustering Methods for Spatial Data Mining". *Proceeding of the 20th VLDB Conference*, Santiago, Chile, 1994.

[18] Ramze Rezaee, B.P.F. Lelieveldt, J.H.C Reiber. "A new cluster validity index for the fuzzy c-mean", *Pattern Recognition Letters*, 19, pp237-246, 1998.

[19] C. Sheikholeslami, S. Chatterjee, A. Zhang. "WaveCluster: A-MultiResolution Clustering Approach for Very Large Spatial Database". *Proceedings of 24th VLDB Conference, New York*, USA, 1998.

[20] Sharma S.C. *Applied Multivariate Techniques.* John Willwy & Sons, 1996.

[21] Padhraic Smyth. "Clustering using Monte Carlo Cross-Validation". *KDD 1996*, 126-133.

[22] S. Theodoridis, K. Koutroubas. *Pattern recognition*, Academic Press, 1999

[23] M. Halkidi, M. Vazirgiannis, Y. Batistakis. "Quality scheme assessment in the clustering process", *In Proceedings of PKDD*, Lyon, France, 2000.

[24] Tian Zhang, Raghu Ramakrishnman, Miron Linvy. "BIRCH: An Efficient Method for Very Large Databases", *ACM SIGMOD' 96*, Montreal, Canada, 1996.

[25] Xunali Lisa Xie, Genardo Beni. "A Validity measure for Fuzzy Clustering", *IEEE Transactions on Pattern Analysis and machine Intelligence*, Vol13, No4, August 1991.

[26] Spatial Datasets: an "unofficial" collection. http://dias.cti.gr/~ytheod/research/datasets/spatial.html