

STATISTICAL ANALYSIS OF AGREEMENT BETWEEN DBSCAN AND HDBSCAN CLUSTERING RESULTS

IGNATEV NIKOLAY*

NATIONAL UNIVERSITY OF UZBEKISTAN NAMED AFTER M.ULUGBEK, TASHKENT, UZBEKISTAN
n.ignatev@rambler.ru

TOSHPULATOV ABDUHALIL

NATIONAL UNIVERSITY OF UZBEKISTAN NAMED AFTER M.ULUGBEK, TASHKENT, UZBEKISTAN
toshpulatov00701@gmail.com

ABSTRACT. We propose a statistical approach to comparing the DBSCAN and HDBSCAN clustering algorithms. The approach examines transitions in object status (core, border, or noise) and the association between clustering results. The neighborhood radius ε for DBSCAN is selected subject to a constraint defined by the proportion of objects labeled as noise by HDBSCAN. The resulting assignments are analyzed using contingency tables, Pearson's χ^2 statistic, Cramér's V , and Tschuprow's T . A computational experiment is conducted on data with clusters of differing densities. The two algorithms identify nearly identical numbers of noise objects. Cramér's V increases with the number of nearest neighbors k , whereas Tschuprow's T shows an overall increase but is not monotonic. The results support the use of HDBSCAN-derived information for calibrating DBSCAN and for a statistical comparison of density-based clustering results.

MSC (2020): 62H30; 62H17; 62H20.

Key words: comparison methodology, hybrid algorithm, object status, optimal radius selection.

Introduction

A major difficulty in comparing clustering algorithms is the absence of a universal comparison criterion. Theoretical limitations of a unified clustering framework have been discussed in [7]. This difficulty limits the methodological basis for applying density-based clustering to data containing nonspherical clusters. In addition to differences in distance measures and density-estimation parameters, algorithms may differ substantially in both the number and membership of the clusters they identify. External and internal validity criteria for nonspherical clusters formed through object connectivity have been considered in [1].

This study compares two density-based clustering methods, DBSCAN and HDBSCAN, in the context of related clustering approaches [2–5]. Although the two methods belong to the same broad class, they approach clustering differently. Their principal difference concerns the treatment of clusters with varying densities.

A major limitation of DBSCAN [6] is its sensitivity to the neighborhood radius ε . The proposed modification addresses this limitation through the following steps:

- determining an admissible range of ε from the data and a specified number of nearest neighbors k ;
- selecting ε by solving an optimization problem subject to a constraint on the proportion of objects labeled as noise;
- using the objects labeled as noise at the selected ε to subsequently cluster lower-density local regions that do not overlap previously identified clusters.

An important advantage of HDBSCAN is its ability to analyze data with heterogeneous density. The proposed radius-selection approach broadens the applicability of DBSCAN and provides a more systematic basis for comparing it with other density-based methods. Under these conditions, the frequently stated expectation

Received: 02.08.2026, **Accepted:** 22.09.2026

*Corresponding author

that HDBSCAN generally produces better clustering because it automatically analyzes the density structure of the data requires further examination.

The comparison is motivated by two observations:

- HDBSCAN identifies noise objects while accounting for local differences in data density;
- the proportion of noise objects identified by HDBSCAN can serve as a regularization parameter when selecting an optimal radius ε for DBSCAN.

Conventional comparisons typically rely on aggregate clustering-quality indices. Here, we instead examine statistical patterns in object-status transitions between the results of DBSCAN and HDBSCAN. This makes it possible to investigate how differences between the algorithms arise, assess the stability of object statuses as ε changes, determine the role of border objects in cluster formation, and relate the HDBSCAN noise proportion to the selected DBSCAN radius.

Mathematical model

Let $E_0 = \{S_1, \dots, S_m\}$ be a set of objects. Each object is described by n heterogeneous features, $X(n) = (x_1, \dots, x_n)$, of which ξ are measured on interval scales and the remaining $n - \xi$ on nominal scales. The metric $\rho(x, y)$ is used to quantify proximity between objects.

The interpretability of DBSCAN and confidence in its clustering results can be improved by constructing a hybrid model that combines cluster analysis, classification, and fuzzy logic. During clustering, each object receives one of three statuses: core, border, or noise. These statuses are treated as additional nominal features. A constraint on the proportion of objects assigned noise status is used to select an optimal neighborhood radius ε .

The hybrid procedure for selecting an optimal DBSCAN radius ε is implemented as follows [8]. Admissible values of ε are examined sequentially to construct a new data table. Each column represents a nominal feature with either three or two categories corresponding to object statuses.

The problem is thus reduced to optimizing a feature-stability measure for either three categories or two categories, with core and border statuses combined in the latter case. The two-category formulation is equivalent to a two-class recognition problem. The first class, K_1 , represents the DBSCAN clustering results; the second class, K_2 , is an artificially constructed opposing class. The distribution of object categories in the opposing class is initially modeled by a Dirichlet distribution. The proportion of noise objects acts as a regularization parameter.

Within the theory of estimating dependencies from empirical data described in [9], the search for ε can be interpreted as a problem of recovering information from indirect experiments. If the target function $f(t)$ cannot be measured at any point t , another measurable function $F(x)$ may be available that is related to $f(t)$ through the operator equation $Af(t) = F(x)$.

Concepts from fuzzy set theory [10] are used to interpret such indirect experimental results. Let the admissible radius values lie in $[\varepsilon_{\min}, \varepsilon_{\max}]$, where $\varepsilon_{\min}$ corresponds to an assignment in which all objects are labeled as noise, and $\varepsilon_{\max}$ to an assignment in which no object is labeled as noise.

Let category $\mu = 1$ combine core and border statuses, and let $\mu = 2$ denote noise status. Denote by $d(\mu)$ the proportion of objects in class K_1 belonging to category μ , and by $\theta(\mu)$ the corresponding proportion in class K_2 .

The membership function for class K_1 is defined by

$$f(\mu) = \frac{d(\mu)}{d(\mu) + \theta(\mu)}. \quad (1)$$

Using (1), the stability measure is defined as

$$g = \frac{1}{2} \sum_{\mu=1}^2 \begin{cases} (d(\mu) + \theta(\mu)) \cdot f(\mu), & f(\mu) \geq 0.5, \\ (d(\mu) + \theta(\mu)) \cdot (1 - f(\mu)), & f(\mu) < 0.5. \end{cases} \quad (2)$$

For the two-category assignments of the objects in E_0 , the stability values in (2) are evaluated over the admissible interval $[\varepsilon_{\min}, \varepsilon_{\max}]$. The optimal radius ε is obtained by solving the following unimodal minimization problem:

$$g(\varepsilon) - 0.5 \longrightarrow \min_{\varepsilon \in [\varepsilon_{\min}, \varepsilon_{\max}]}.$$

The curve in Fig. 1 has a single pronounced minimum, which is consistent with the assumed unimodal form of the objective.

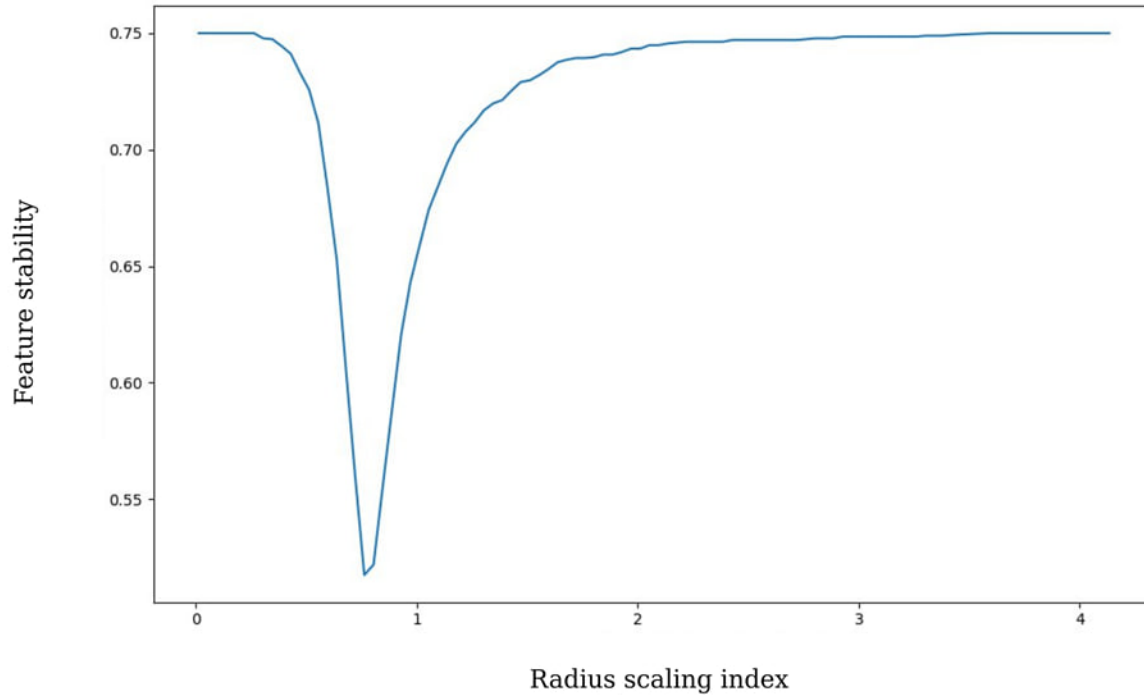


Fig. 1. Stability measure (2) over the admissible radius interval $[\varepsilon_{\min}, \varepsilon_{\max}]$.

Differences in cluster-density structure are reflected in how border objects are assigned to neighboring clusters. This motivates testing the hypothesis that some objects labeled as noise by HDBSCAN become core or border objects under DBSCAN.

The present study reverses the earlier formulation. Previously, the noise proportion was specified as a constraint when selecting ε for DBSCAN. Here, HDBSCAN determines that proportion. The resulting value is then used to select an optimal radius ε and to cluster the data with DBSCAN.

DBSCAN and HDBSCAN share the following properties:

- the ability to identify clusters of arbitrary shape;
- the use of proximity relations in a metric space and nearest-neighbor analysis;
- no requirement to specify the number of clusters in advance;
- the identification of objects that are not assigned to any cluster, referred to here as noise objects.

The aim of the study is to identify structural patterns that reflect differences in the mechanisms of DBSCAN and HDBSCAN. Its contribution is a statistical comparison procedure based on object-status transitions and the use of the HDBSCAN noise proportion as a regularization parameter for selecting the DBSCAN radius. To address this aim, the following hypotheses are proposed.

Hypothesis 1. Object status transitions

Objects labeled as noise by HDBSCAN are hypothesized to form a heterogeneous group. At the selected optimal radius, some of these objects may become border or core objects in DBSCAN. Support for this hypothesis would suggest that an HDBSCAN noise label can reflect a local density deficit rather than the absence of membership in a cluster under an alternative density criterion.

Hypothesis 2. Clustering mechanisms and regularization

Differences between DBSCAN and HDBSCAN are hypothesized to arise primarily from the treatment of cluster boundaries and objects with low local density. The dense interiors of the clusters identified by the two algorithms are expected to overlap substantially, while the main differences involve the reassignment of border and noise objects. The HDBSCAN noise proportion may therefore serve as a regularization parameter when selecting an optimal DBSCAN radius. Support for this hypothesis would indicate a relationship between the local density structure identified by HDBSCAN and the parameters selected for DBSCAN.

Hypothesis 3. Stability of object statuses

The stability of object statuses under changes in ε is hypothesized to differ across categories. Core objects are expected to be highly stable, whereas border and noise objects may exhibit substantially greater variability. This would allow the reliability of each object's category assignment to be assessed quantitatively.

Hypothesis 4. Reassignment of border objects

Changes in ε are hypothesized to affect primarily the classification of border objects. If parameter changes predominantly reassign objects in this category, this would suggest that the main difference between the algorithms lies in how competing neighboring clusters assign objects located in boundary regions.

Hypothesis 5. Association between results and algorithm selection

The association between the algorithms' results is examined by cross-tabulating HDBSCAN assignments with DBSCAN object statuses and cluster memberships. Statistical dependence is assessed using contingency tables and Pearson's χ^2 statistic, together with Cramér's V and Tschuprow's T . Cramér's V is treated as the principal measure of association, while Tschuprow's T provides a supplementary measure. Pearson's χ^2 test of independence can be used to assess the statistical significance of an observed association when its assumptions are satisfied.

The strength of association is proposed as a possible guide to algorithm selection. High values would support considering DBSCAN as a potentially simpler computational alternative, whereas low values would indicate substantial differences and motivate closer examination of the density structure captured by HDBSCAN. Validation of this hypothesis would clarify whether association measures can help characterize the conditions under which each algorithm is applicable.

Computational experiment

The experiment used the Breast Cancer Wisconsin (Diagnostic) dataset [11]. Its features were computed from digitized images of fine-needle aspirates of breast masses. Cluster analysis was performed on 569 objects described by 30 quantitative features. Before clustering, the data were mapped to the interval $[-1, 1]$ using a fractional-linear transformation, after which Euclidean distance was used. For HDBSCAN, the minimum cluster size parameter, `min_cluster_size`, was set to 2.

Table 1 presents DBSCAN results obtained by selecting ε subject to a constraint determined by the HDBSCAN noise proportion. The values $k \in \{3, 7, 10, 15\}$ were chosen to examine changes in the clustering results.

As shown in Table 1, the numbers of objects assigned to clusters by HDBSCAN and DBSCAN differ only slightly. At $k = 15$, HDBSCAN assigns 253 objects to clusters, whereas DBSCAN assigns 254.

This result is consistent with calibrating the DBSCAN radius to the HDBSCAN noise proportion. The two algorithms classify nearly the same number of objects as noise, providing a basis for comparing their cluster assignments. Matching these totals does not imply that the same individual objects receive the same labels.

Table 2 relates the HDBSCAN assignments to the object statuses determined by DBSCAN.

Of the objects labeled as noise by HDBSCAN, 16 are border objects in DBSCAN, and none are core objects (Table 2).

The correspondence between the two algorithms' cluster assignments is shown in Table 3.

Table 1. Clustering results for different numbers of nearest neighbors k .

k	HDBSCAN			DBSCAN	
	Cluster sizes	Clustered objects	Noise (%)	Cluster sizes	Clustered objects
3	2, 3, 2, 2, 304	313	45.0	19, 248, 14, 4, 8, 3, 3, 3, 5, 3, 3	313
7	9, 2, 3, 147	161	71.7	154, 7	161
10	9, 252	261	54.1	14, 247	261
15	7, 246	253	55.5	254	254

Table 2. DBSCAN object statuses at $k = 10$ and an HDBSCAN noise proportion of 54.1%.

HDBSCAN assignment	DBSCAN status		
	Core	Border	Noise
Noise	0	16	292
Cluster 1 (9 objects)	3	5	1
Cluster 2 (252 objects)	164	73	15

Table 3. Cross-tabulation of DBSCAN and HDBSCAN assignments at $k = 10$.

DBSCAN assignment	HDBSCAN assignment		
	Noise	Cluster 1	Cluster 2
Noise	292	1	15
Cluster 1 (14 objects)	6	8	0
Cluster 2 (247 objects)	10	0	237

Tables 2 and 3 show that, among the 16 objects labeled as noise by HDBSCAN but as border objects by DBSCAN, six are assigned to DBSCAN cluster 1 and the remaining ten to DBSCAN cluster 2.

Table 4 reports the contingency-table statistics and measures of association.

Table 4. Contingency-table statistics and measures of association.

Statistic or coefficient	$k = 3$	$k = 7$	$k = 10$	$k = 15$
Pearson's χ^2 statistic	914.1066	634.7085	752.2534	461.4917
Information-based association measure	277.1457	228.1456	313.5447	280.2493
Pearson's contingency coefficient C	0.7851	0.7262	0.7546	0.6690
Tschuprow's T	0.4464	0.6280	0.8130	0.7573
Cram Γ @r's V	0.5669	0.7468	0.8130	0.9006

The statistics and association measures in Table 4 serve different purposes, so their numerical values should not be compared directly. Cram Γ @r's V and Tschuprow's T are particularly useful for describing the strength of association between the clustering results. Figure 2 shows their variation with k , together with Pearson's contingency coefficient C .

The results obtained for the different values of k show the following patterns:

- Cram Γ @r's V increases from 0.567 at $k = 3$ to 0.901 at $k = 15$, indicating stronger association between

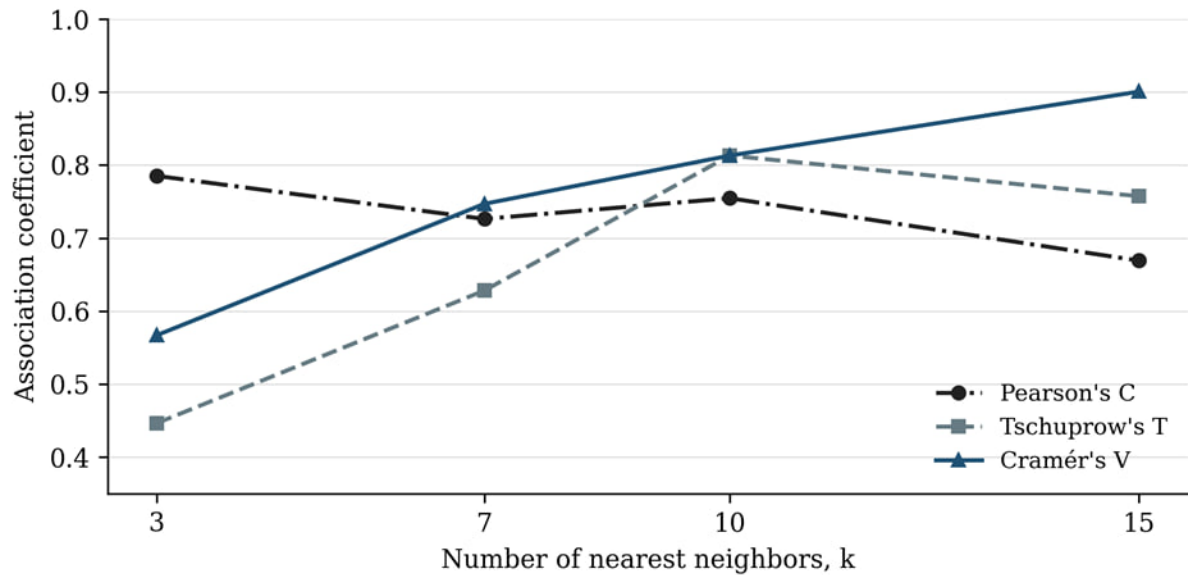


Fig. 2. Association coefficients for different numbers of nearest neighbors k .

the assignments produced by the two algorithms.

- Tschuprow's T also increases overall, from 0.446 at $k = 3$ to 0.757 at $k = 15$, but reaches its maximum of 0.813 at $k = 10$ and then decreases.
- Pearson's contingency coefficient remains relatively high (0.669–0.785), but cannot be compared directly with Cramér's V because its maximum depends on the dimensions of the contingency table.

The χ^2 statistic and the information-based measure depend on the contingency-table dimensions and the number of categories. Their raw values are therefore less suitable for comparing results across different values of k . Cramér's V and Tschuprow's T provide normalized measures of association, although the table structure must still be considered when interpreting them.

For the settings examined, Cramér's V indicates increasing association as k increases, reaching 0.901 at $k = 15$. Tschuprow's T supports an overall increase relative to $k = 3$, but its decrease between $k = 10$ and $k = 15$ shows that the trend is not uniformly increasing across both measures. Thus, the experiment demonstrates a general tendency toward stronger association between DBSCAN and HDBSCAN assignments as k increases, rather than a monotonic increase in every coefficient.

Conclusion

We have proposed a statistical procedure for comparing DBSCAN and HDBSCAN based on object-status transitions and contingency-table analysis of clustering results. In the reported experiment, using the HDBSCAN noise proportion as a constraint on the DBSCAN radius produced nearly identical numbers of noise objects and substantial association between the resulting assignments. The approach may support parameter selection and comparative assessment of density-based clustering algorithms for data with heterogeneous structure. DBSCAN may be considered as a computationally simpler candidate in suitable settings, but any efficiency advantage requires separate validation.

REFERENCES

1. Ignatev N. A., Zguralskaya E. N. Cluster analysis using learning based on connectivity relations and distribution density. Tomsk State University Journal of Control and Computer Science. 2024;68:66–74. <https://doi.org/10.17223/19988605/68/7>. (In Russian.)

2. Zhu Y., Ting K. M., Carman M. J. Density-ratio based clustering for discovering clusters with varying densities. *Pattern Recognition*. 2016;60:983–997.
3. Campello R. J. G. B., Moulavi D., Sander J. Density-based clustering based on hierarchical density estimates. In: *Advances in Knowledge Discovery and Data Mining*. Berlin, Heidelberg: Springer; 2013. pp. 160–172. https://doi.org/10.1007/978-3-642-37456-2_14.
4. Ienco D., Bordogna G. Fuzzy extensions of the DBScan clustering algorithm. *Soft Computing*. 2018;22:1719–1730. <https://doi.org/10.1007/s00500-016-2435-0>.
5. Shi J., Malik J. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2000;22(8):888–905. <https://doi.org/10.1109/34.868688>.
6. Aidagulov R. R., Glavatsky S. T., Mikhalev A. V. Averaging methods in big data clustering problems. *Intelligent Systems: Theory and Applications*. 2021;25(4):12–18. (In Russian.)
7. Kleinberg J. An impossibility theorem for clustering. *Advances in Neural Information Processing Systems*. 2002;15. <https://www.cs.cornell.edu/home/kleinber/nips15.pdf>.
8. Toshpulatov A. O. Problems of parameter selection in the DBSCAN algorithm in multidimensional data analysis. In: *Zamonaviy matematikaning dolzarb muammolari va tatbiqlari [Current Problems and Applications of Modern Mathematics]*. Proceedings of the Republican Scientific Conference. Tashkent; 2026. pp. 509–510.
9. Vapnik V. N. Estimation of Dependencies Based on Empirical Data. Moscow: Nauka; 1979. 447 pp. (In Russian.)
10. Bellman R. E., Zadeh L. A. Decision-making in a fuzzy environment. *Management Science*. 1970;17(4):B-141–B-164.
11. UCI Machine Learning Repository. Breast Cancer Wisconsin (Diagnostic) [dataset]. <https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic>. <https://doi.org/10.24432/C5DW2B>. Accessed 12 July 2026.

[Cite this article](#)

Ignatev N., Toshpulatov A. *Statistical Analysis of Agreement between DBSCAN and HDBSCAN Clustering Results*, **Acta NUUZ**, 2026, 2/3, pp. 102-108.