ENSEMBLE
Journal of Applied Mathematics, Statistics, and Data Science
eISSN: xxxx-xxxx
Vol. 1, No. 1, June 2026
Page: 10-28

Clustering Methods for Chemical Profile Segmentation

M. Y. Matdoan[1*], Rosalina Salhuteru[2], Novita Laamena[3]

[1,2,3]Department of Statistics, Faculty of Science and Technology, University of Pattimura
Jl. Ir. M. Putuhena, Kampus Unpatti-Poka, Ambon, 97233, Indonesia

E-mail Correspondence Author: *keepyahya@gmail.com

: xxxxxx

ARTICLE INFO

Article history

Received : January 20, 2026
Revised : March 12, 2026
Accepted : April 20, 2026

Keywords
Clustering
DBSCAN
GMM
HDBSCAN
K-Means
Spectral Clustering
Ward

Abstract

Cluster analysis is widely used to reveal latent structure in unlabeled multivariate data, yet its practical value depends on whether the selected algorithm is compatible with the geometry, scale, and noise pattern of the data. This article develops a journal-style empirical study of clustering and its application to chemical profile segmentation using the UCI Wine Recognition data set, which contains 178 observations, 13 continuous chemical variables, and three reference cultivars. The method combines statistical preprocessing, principal component analysis, K-Means, Ward hierarchical clustering, Gaussian mixture models, spectral clustering, DBSCAN, and HDBSCAN. Performance is evaluated with internal indices, external agreement measures, cluster profiles, and visual diagnostics. The two leading principal components explain 55.41% of standardized variation, indicating a meaningful low-dimensional representation without eliminating multivariate complexity. K-Means produces the highest adjusted Rand index (0.8975) and normalized mutual information (0.8759), followed by Gaussian mixture and spectral clustering (ARI = 0.8804). Density-based methods detect possible noisy observations but recover only two dense groups and show lower agreement with reference cultivars. The findings demonstrate that clustering is not a universal algorithmic choice but a decision process linking data geometry, statistical assumptions, validation evidence, and domain interpretation. The article contributes a reproducible applied framework for selecting, validating, and explaining clustering results in chemical and quality-segmentation problems.

1. INTRODUCTION

Cluster analysis is a core technique in statistics and data science because many scientific problems begin with unlabeled observations rather than predefined classes. In an applied setting, the researcher often has a multivariate table containing chemical, economic, social, or biological measurements and must ask whether meaningful groups exist before any supervised model can be justified. Clustering is therefore not merely an algorithmic tool; it is an exploratory statistical procedure for discovering structure, simplifying heterogeneity, and translating multivariate variation into interpretable segments. In food chemistry, for example, clustering can indicate whether samples with similar chemical compositions correspond to origin, cultivar, processing technology, or quality category. In public policy, it can group districts with similar development profiles; in health sciences, it can identify patient subtypes; and in business analytics, it can

provide market segments. The common methodological challenge is that the cluster label is not directly observed, and the analyst must decide whether the resulting partition is statistically stable, interpretable, and useful for decision making.

The development of clustering has moved from classical partitioning algorithms toward model-based, density-based, graph-based, and representation-learning approaches. K-Means remains popular because it is computationally efficient and produces a simple centroid interpretation [1], [2]. Hierarchical clustering remains important when the researcher needs a nested view of similarity rather than a single partition [3]. Gaussian mixture models provide a probabilistic formulation in which clusters are modeled as components of a mixture distribution and observations have posterior membership probabilities [4], [5]. Density-based algorithms such as DBSCAN and HDBSCAN define clusters as dense connected regions and can separate dense groups from noise [6], [7]. Spectral clustering treats the data as a similarity graph and uses eigenvectors of a graph Laplacian to recover nonconvex structures [8], [9]. More recent deep clustering studies add neural representation learning, especially for images, text, and high-dimensional signals [10]-[12]. These developments show that clustering is no longer a single family of methods but a collection of assumptions about distance, density, probability, graph connectivity, and latent representation.

The practical difficulty is that different methods can produce different partitions from the same data. This is not a weakness of clustering alone but a consequence of the fact that the word cluster has several mathematical meanings. A cluster may be a compact ball around a centroid, a connected region of high density, a component of a probability mixture, a community in a graph, or a region in a learned low-dimensional space. These definitions coincide only under special conditions. When the observations form roughly spherical and balanced groups, K-Means is often effective. When groups are elongated, nested, or unequal in size, hierarchical and mixture models may be more informative. When noise and irregular shapes dominate, density-based approaches are attractive. When local similarity matters more than global Euclidean distance, spectral methods can be useful. Consequently, a responsible clustering study must not simply run several algorithms and select the largest score. It must explain why a method is appropriate for the structure of the data and why the resulting grouping is meaningful for the application.

A second challenge is the curse of dimensionality. In high-dimensional spaces, pairwise distances may concentrate, local neighborhoods become less reliable, and irrelevant variables can dominate the distance metric [13]. This problem is especially important in chemical, genomic, text, and image data where many variables are correlated or noisy. Standardization, dimensionality reduction, and feature interpretation are therefore central parts of clustering rather than optional preprocessing. Principal component analysis (PCA) remains a useful bridge between classical statistics and modern unsupervised learning because it provides an orthogonal representation of dominant variation [14]. However, dimension reduction may also distort local structure if applied without diagnostics. A careful study must therefore report explained variance, visual structure, and the relationship between reduced dimensions and original variables.

Cluster validation is another source of uncertainty. Internal indices such as the silhouette coefficient, Davies-Bouldin index, and Calinski-Harabasz index measure compactness, separation, and dispersion using the observed data only [15]-[17]. External indices such as the adjusted Rand index and normalized mutual information are available when reference labels exist, but they should be interpreted as agreement with a known labeling system rather than absolute proof that a clustering method is correct [18], [19]. In many real applications, there is no ground truth; in others, reference labels exist but may reflect administrative categories rather than natural groups. This means that clustering requires triangulation: numerical metrics, visual diagnostics, cluster profiles, domain interpretation, and sensitivity analysis must be read together.

This article addresses the above issues by presenting a reproducible applied study of clustering and its application to chemical profile segmentation. The empirical data are the UCI Wine Recognition data set, a benchmark of 178 wine samples from three cultivars represented by 13 positive continuous chemical measurements [20], [21]. The study compares six methods: K-Means, Ward agglomerative clustering, Gaussian mixture models, spectral clustering, DBSCAN, and HDBSCAN. The contribution is threefold. First, the article presents a complete statistical workflow from data source and preprocessing to mathematical formulation, algorithm selection, validation, and interpretation. Second, it provides accountable numerical results computed from a real public data set rather than hypothetical tables. Third, it discusses how the empirical findings

relate to earlier clustering literature, especially the difference between compact centroid structure and density-defined structure. The guiding argument is that clustering should be treated as an evidence-based analytical procedure, not as a mechanical choice of the highest validation score.

.

2. METHOD

This research uses a comparative empirical design. The purpose is not to develop a new clustering algorithm but to examine how several major clustering families behave when applied to the same multivariate chemical data. A comparative design is suitable because clustering algorithms embody different assumptions. Partitioning methods assume compactness around prototypes, hierarchical methods emphasize nested similarity, mixture models assume latent probability components, spectral methods use graph connectivity, and density-based methods use local density. By placing these methods within one pipeline, the study can separate two questions that are often mixed in applied work: which method has the strongest numerical agreement with reference labels, and which method provides the most defensible interpretation for the data characteristics.

The data source is the Wine Recognition data set available through the UCI Machine Learning Repository and distributed in scikit-learn through the function load_wine [20]-[22]. The data contain 178 observations measured on 13 continuous chemical variables. The original problem concerns three cultivars of wine grown in the same Italian region. Although the cultivar labels are available, they are not used to train the clustering algorithms. They are used only after clustering as external reference labels for validation. This design is important because clustering is unsupervised: the algorithms operate only on the matrix of chemical measurements, whereas the labels provide an independent way to evaluate whether the discovered groups correspond to a known domain category. The class distribution is 59, 71, and 48 observations, equivalent to 33.15%, 39.89%, and 26.97% of the sample, respectively. The statistical unit is a wine sample, and the variables represent concentrations or measurements such as alcohol, malic acid, ash, magnesium, total phenols, flavanoids, color intensity, hue, and proline. All variables are positive and continuous but measured on different numerical scales. Proline, for example, ranges from 278 to 1680 with a standard deviation of 314.91, whereas hue ranges from 0.48 to 1.71 with a standard deviation of 0.23. Without standardization, variables with large numerical scales would dominate Euclidean distances and bias centroid-based as well as density-based methods. Therefore, the main analysis uses z-score standardization for all 13 variables. Missing value handling is included in the research procedure, but the data set contains no missing values; therefore, no imputation is applied. Outliers are not deleted because density-based methods are explicitly evaluated for their ability to identify noise-like observations.

Table 1. Data source and research variables

Component	Description	Value / implementation	Analytical role
Data source	Wine Recognition data	UCI Machine Learning Repository; scikit-learn load_wine	Public reproducible benchmark
Observations	Wine samples	178 samples	Units to be clustered
Variables	Chemical constituents	13 continuous positive variables	Feature matrix X
Reference labels	Cultivars	3 classes: 59, 71, 48 samples	External validation only
Missing values	Completeness check	0 missing values	No imputation required
Software	Computing environment	Python; scikit-learn 1.8.0	Reproducible computation

Table 2. Selected descriptive statistics of chemical variables.

Variable	Mean	Std. deviation	Minimum	Maximum
Alcohol	13.00	0.81	11.03	14.83
malic_acid	2.34	1.12	0.74	5.80
Ash	2.37	0.27	1.36	3.23
Magnesium	99.74	14.28	70.00	162.00
Flavanoids	2.03	1.00	0.34	5.08
color_intensity	5.06	2.32	1.28	13.00
Hue	0.96	0.23	0.48	1.71
Proline	746.89	314.91	278.00	1680.00

Let $X = \{x_1, x_2, ..., x_n\}$ denote the standardized data matrix with $n = 178$ observations and $p = 13$ variables. For variable j, standardization is conducted by subtracting the sample mean and dividing by the sample standard deviation. This transformation gives each variable comparable scale and makes the Euclidean distance reflect a balanced multivariate difference rather than a unit-of-measure artifact. The transformation is written as follows:

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}, \quad i = 1, \dots, n, \; j = 1, \dots, p. \tag{1}$$

The first statistical method after standardization is PCA. PCA is used for two purposes: visualization and structural diagnosis. It is not used as a replacement for full-dimensional clustering in the main comparison, but it provides a two-dimensional projection for interpreting the separation of groups. If S is the covariance matrix of the standardized data, PCA solves the eigenvalue problem shown in Equation (2). The eigenvectors provide orthogonal directions of maximum variance, and the eigenvalues measure the amount of variance explained by each component.

$$\mathbf{S}\mathbf{v}_\ell = \lambda_\ell \mathbf{v}_\ell, \quad \text{PVE}_\ell = \frac{\lambda_\ell}{\sum_{r=1}^{p} \lambda_r}. \tag{2}$$

K-Means partitions the observations into K clusters by minimizing within-cluster squared deviations from cluster centroids. In this study K is set to 3 because the chemical application has three reference cultivars, and because the comparison aims to evaluate whether unsupervised partitions recover the known structure. The K-Means objective is:

$$J = \min_{C_1, \dots, C_K} \sum_{k=1}^{K} \sum_{x_i \in C_k} \| x_i - \mu_k \|_2^2. \tag{3}$$

where C_k is cluster k and μk is its centroid. Ward agglomerative clustering is included as a hierarchical alternative. It starts with each observation as a separate cluster and merges clusters that produce the smallest increase in within-cluster sum of squares. Gaussian mixture modeling assumes that the data are generated by a weighted combination of multivariate normal components. Its likelihood is expressed as:

$$L(\theta) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k \, \phi(x_i \mid \mu_k, \Sigma_k), \quad \sum_{k=1}^{K} \pi_k = 1. \tag{4}$$

The Gaussian mixture model is useful because it provides probabilistic assignments and can accommodate ellipsoidal clusters through component covariance matrices. Spectral clustering constructs a neighborhood graph and uses the eigenvectors of a graph Laplacian as a representation for partitioning. DBSCAN defines core points according to a radius eps and a minimum number of samples. HDBSCAN extends density-based clustering by constructing a hierarchy of density-connected components and selecting stable clusters from that hierarchy [7]. In this article, DBSCAN uses eps = 2.3 and min_samples = 5 based on a grid search that favored a small number of interpretable groups without excessive noise. HDBSCAN uses min_cluster_size = 10 and min_samples = 3. K-Means uses n_init = 20 and random_state = 42, Gaussian mixture uses n_init = 10 and random_state = 42, and spectral clustering uses a 10-nearest-neighbor graph.

Table 3. Clustering methods and parameter settings

Method	Main assumption	Parameter setting	Expected strength
K-Means	Compact centroid-based groups	K=3; n_init=20	Fast and interpretable partitions
Agglomerative-Ward	Nested similarity with minimum variance merging	K=3; Ward linkage	Hierarchical structure and dendrogram
Gaussian Mixture	Mixture of multivariate normal components	K=3; full covariance	Probabilistic membership
Spectral	Graph connectivity in local neighborhood	K=3; 10 nearest neighbors	Nonlinear separation
DBSCAN	Dense connected regions plus noise	eps=2.3; min_samples=5	Noise detection
HDBSCAN	Stable density hierarchy	min_cluster_size=10; min_samples=3	Variable-density clustering

The evaluation combines internal, external, computational, and interpretability criteria. The silhouette coefficient compares the mean within-cluster distance of an observation with its mean distance to the nearest other cluster. If a(i) is the average distance from observation i to observations in its own cluster and b(i) is the smallest average distance to another cluster, the silhouette value is:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \quad -1 \leq s(i) \leq 1. \tag{5}$$

The Davies-Bouldin index measures average similarity between each cluster and its most similar cluster; lower values are preferred. The Calinski-Harabasz index compares between-cluster dispersion with within-cluster dispersion; higher values are preferred. When reference cultivars are used for post-hoc evaluation, the adjusted Rand index measures pair agreement corrected for chance, while normalized mutual information measures the shared information between clustering labels and reference labels. The adjusted Rand index can be expressed in terms of pair counts as follows:

$$\mathrm{ARI} = \frac{\mathrm{RI} - \mathrm{E}[\mathrm{RI}]}{\max(\mathrm{RI}) - \mathrm{E}[\mathrm{RI}]}. \tag{6}$$

The research procedure consists of eight steps. First, the public data set is loaded and the observation count, variable count, class distribution, and missing values are checked. Second, descriptive statistics are calculated to understand scale differences and chemical heterogeneity. Third, all variables are standardized. Fourth, PCA is used to inspect the dominant low-dimensional structure. Fifth, six clustering methods are fitted using the standardized variables. Sixth, internal and external metrics are computed. Seventh, cluster membership tables and chemical profiles are prepared to translate numerical clusters into substantive chemical segments. Eighth, the results are compared with existing methodological literature to avoid a purely mechanical interpretation. This procedure is designed to make the analysis reproducible, transparent, and statistically defensible.

3. RESULTS AND DISCUSSION

The empirical results begin with the structure of the data before the clustering algorithms are compared. The data set is balanced enough for clustering evaluation but not perfectly balanced: cultivar 1 has 59 samples, cultivar 2 has 71 samples, and cultivar 3 has 48 samples. This distribution matters because some clustering algorithms, especially K-Means, can be affected by unequal cluster sizes, whereas density-based methods may treat smaller or less dense groups as noise. The descriptive statistics in Table 2 show strong scale differences across variables. Proline has a mean of 746.89 and a maximum of 1680.00, whereas hue has a mean of 0.96 and a maximum of 1.71. Therefore, standardization is not a cosmetic step; it is a necessary statistical

operation to prevent high-scale variables from dominating the distance matrix. This finding is consistent with general clustering recommendations that distance-based algorithms should be applied after careful scale treatment [23].

PCA provides a first visual diagnostic. The first principal component explains 36.20% of the standardized variance, and the second explains 19.21%. Together, the first two components explain 55.41% of the total variation, while the first three components explain 66.53%. The first five components explain 80.16%. These values indicate that a two-dimensional projection contains meaningful information but does not exhaust the multivariate structure. This is important because a scatter plot can guide interpretation but should not replace full-dimensional clustering. The main loadings show that PC1 is strongly related to flavanoids, total phenols, od280/od315 of diluted wines, proanthocyanins, and nonflavanoid phenols. PC2 is associated with color intensity, alcohol, proline, ash, magnesium, and hue. Chemically, this means that the dominant separation is not based on a single variable but on a combination of phenolic composition, color-related attributes, and concentration indicators. Such multivariate separation is precisely the type of situation in which clustering can be useful.

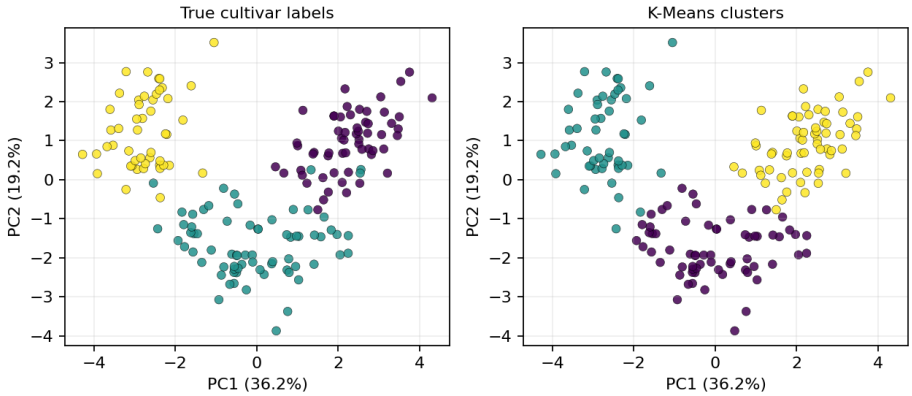

Figure 2. PCA projection of wine samples using true cultivar labels and K-Means cluster labels

Figure 2 shows that the cultivar structure is visible in the first two principal components, although some overlap remains in the central region. This overlap is expected because the data represent wines from the same region rather than completely unrelated populations. K-Means produces a partition that visually aligns with the cultivar structure. Cluster 2 corresponds almost entirely to cultivar 0, cluster 1 corresponds largely to cultivar 2, and cluster 0 corresponds mainly to cultivar 1. The minor mixing occurs around the boundary between cultivar 0 and cultivar 1. This pattern is a useful reminder that clustering does not need perfect visual separation to be valuable. A meaningful unsupervised partition can still recover most of the reference grouping when the multivariate signal is strong but not perfectly separated.

Table 4. Comparative clustering performance on standardized wine data

Method	Clusters	Noise (%)	Silhouette	Davies-Bouldin	Calinski-Harabasz	ARI	NMI	Run time (ms)
K-Means	3	0.0000	0.2849	1.3892	70.9400	0.8975	0.8759	45.6324
Agglomerative-Ward	3	0.0000	0.2774	1.4186	67.6475	0.7899	0.7865	1.7158
Gaussian Mixture	3	0.0000	0.2844	1.3938	70.6878	0.8804	0.8609	36.5665
Spectral	3	0.0000	0.2828	1.3907	70.0411	0.8804	0.8609	17.4264
DBSCAN	2	23.5955	0.1736	3.7157	28.3607	0.3838	0.4764	2.5915
HDBSCAN	2	21.9101	0.1850	3.4623	28.7630	0.3753	0.4570	3.1351

Table 4 gives the central numerical evidence. K-Means obtains the highest ARI of 0.8975 and the highest NMI of 0.8759. The silhouette score is 0.2849, the Davies-Bouldin index is 1.3892, and the Calinski-Harabasz index is 70.9400. Gaussian mixture clustering and spectral clustering follow closely, both with ARI = 0.8804 and NMI = 0.8609. Their internal scores are also very

close to K-Means. Ward agglomerative clustering produces a lower but still strong ARI of 0.7899 and NMI of 0.7865. In contrast, DBSCAN and HDBSCAN produce only two dense clusters and mark 23.60% and 21.91% of the observations as noise, respectively. Their ARI values are 0.3838 and 0.3753, indicating that the density-based structure found by these algorithms does not match the three-cultivar reference partition as well as the compact-partition and mixture methods.

These results should not be interpreted as showing that density-based methods are generally poor. Rather, they show that the Wine data set, after standardization, contains compact cultivar-related structure more than irregular density-connected structure. DBSCAN and HDBSCAN are designed to identify dense regions separated by sparse regions, and they are especially useful when cluster shape is nonconvex and noise is a central concern [6], [7]. In the present data, the classes are chemically distinct but not separated by large low-density gaps. One group is also more diffuse in the central region. Therefore, a single density threshold in DBSCAN and the stability criterion in HDBSCAN tend to merge parts of the data and label boundary observations as noise. This is a defensible result, not an algorithmic failure. It reveals that the density definition of a cluster is less aligned with cultivar identity in this particular application.

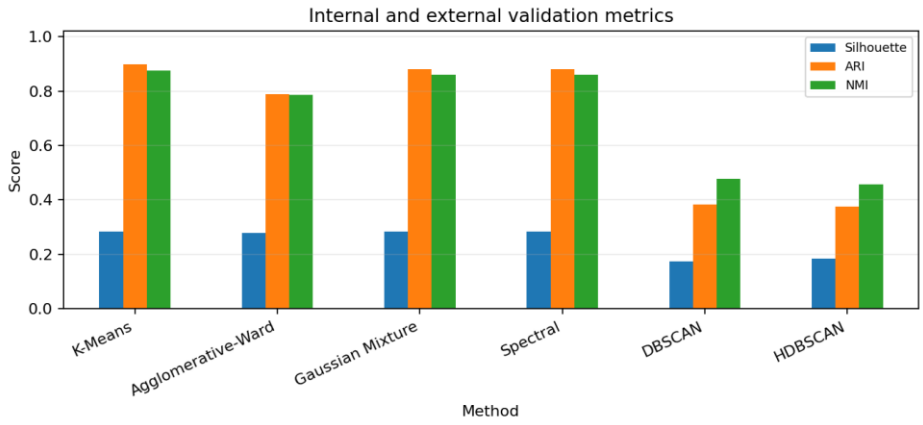

Figure 3. Validation metrics for six clustering methods

Figure 3 reinforces the same conclusion visually. The three compact or model-based methods at the left of the figure have high ARI and NMI values, whereas the density-based methods have lower agreement with the cultivar labels. The silhouette values are modest for all methods, even for the best external scores. This contrast is important. The silhouette coefficient rewards compactness and separation in the chosen distance space, but the Wine data contain overlapping chemical gradients. Therefore, a moderate silhouette is not inconsistent with high agreement with cultivar labels. In applied research, such a discrepancy should prompt interpretation rather than automatic rejection. If the scientific objective is to recover cultivar-like groups, K-Means and Gaussian mixture models are preferable. If the objective is to identify atypical samples or boundary cases for further inspection, DBSCAN and HDBSCAN provide additional information by flagging possible noise observations.

Table 5. Cross-tabulation between reference cultivars and selected cluster labels.

Method	Reference cultivar	Cluster -1/noise	Cluster 0	Cluster 1 / 2
K-Means	0	0	0	1:0, 2:59
K-Means	1	0	65	1:3, 2:3
K-Means	2	0	0	1:48, 2:0
Gaussian Mixture	0	0	59	1:0, 2:0
Gaussian Mixture	1	0	4	1:64, 2:3
Gaussian Mixture	2	0	0	1:0, 2:48
DBSCAN	0	6	53	1:0
DBSCAN	1	29	41	1:1
DBSCAN	2	7	0	1:41
HDBSCAN	0	4	55	1:0
HDBSCAN	1	27	41	1:3
HDBSCAN	2	8	0	1:40

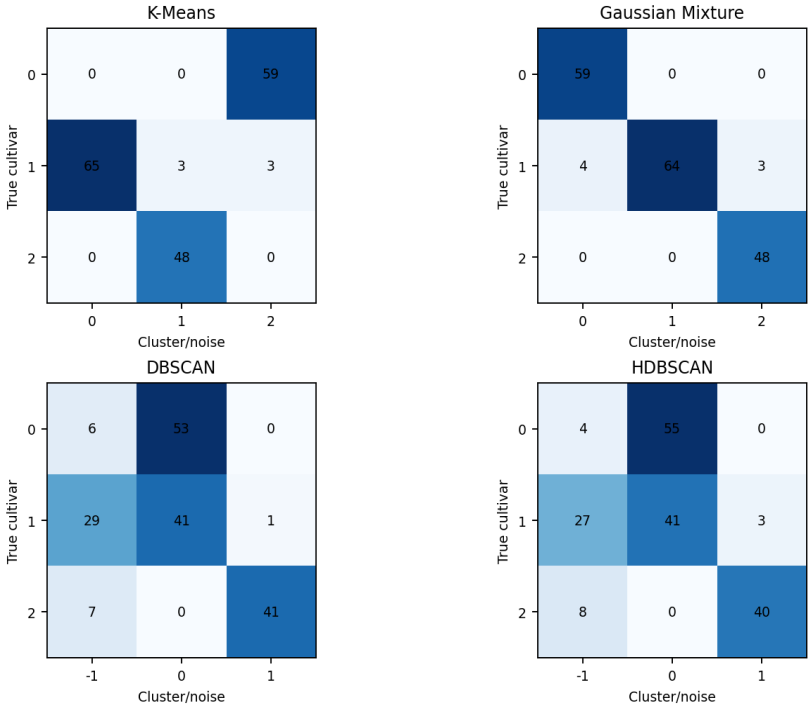

Figure 4. Heatmaps of reference cultivar by cluster assignment for selected methods

The cross-tabulations explain why K-Means and Gaussian mixture models perform strongly. For K-Means, all 59 observations of cultivar 0 are assigned to cluster 2. Cultivar 2 is also recovered almost perfectly, with 48 observations assigned to cluster 1. Cultivar 1 is the most difficult group: 65 observations are assigned to cluster 0, but six observations fall into clusters associated with the other cultivars. This pattern is consistent with the PCA plot, where cultivar 1 occupies a central region close to the boundary of the other groups. Gaussian mixture clustering is similarly strong: cultivar 0 is perfectly placed in one component, cultivar 2 is perfectly placed in another component, and only seven observations of cultivar 1 are assigned outside the dominant component. Spectral clustering yields essentially the same agreement as the Gaussian mixture model, but with different numerical cluster labels.

The density-based cross-tabulations tell a different story. DBSCAN labels 42 observations as noise, including 29 from cultivar 1, 7 from cultivar 2, and 6 from cultivar 0. HDBSCAN labels 39 observations as noise, including 27 from cultivar 1. This is not random noise; it is concentrated in the cultivar that lies between other groups in the multivariate space. Substantively, these observations may represent chemically transitional samples or boundary cases. From a quality-control perspective, this can be valuable. A company or laboratory may want to inspect boundary samples even if the main objective is segmentation. Therefore, the density-based results are weaker for cultivar recovery but useful for identifying uncertain or less typical chemical profiles. This dual interpretation is frequently missed when clustering studies report only a single score.

Table 6. K-Means cluster profiles in original chemical units

Cluster	n	alcohol	malic_acid	ash	magnesium	flavanoids	color_intensity	hue	proline
0	65.00	12.25	1.90	2.23	92.74	2.05	2.97	1.06	510.17
1	51.00	13.13	3.31	2.42	98.67	0.82	7.23	0.69	619.06
2	62.00	13.68	2.00	2.47	107.97	3.00	5.45	1.07	1100.23

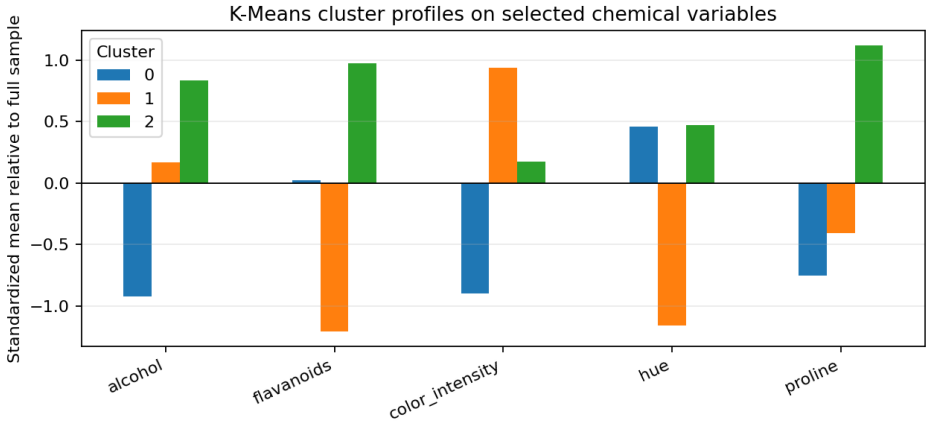

Figure 5. Standardized chemical profiles of K-Means clusters.

The cluster profiles in Table 6 and Figure 5 transform algorithmic labels into substantive descriptions. Cluster 2 has the highest average alcohol (13.68), magnesium (107.97), flavanoids (3.00), hue (1.07), and proline (1100.23). It therefore represents a high-phenolic and high-proline segment. This profile matches cultivar 0 almost exactly and demonstrates why K-Means performs well: one cultivar occupies a compact and chemically distinct region. Cluster 0 has lower alcohol (12.25), lower color intensity (2.97), and lower proline (510.17), but moderate flavanoids (2.05) and the highest hue among the clusters. It corresponds mainly to cultivar 1. Cluster 1 has high malic acid (3.31), low flavanoids (0.82), low hue (0.69), and high color intensity (7.23), matching the chemical signature of cultivar 2. These profiles provide an interpretable segmentation that would be useful in analytical chemistry and quality control because each cluster can be described by a small set of dominant chemical characteristics rather than by an arbitrary label.

The profile analysis also helps explain why the strongest algorithms are not necessarily the most complex. K-Means, Gaussian mixture models, and spectral clustering all capture the main chemical groups because the groups are reasonably compact after standardization. K-Means has the advantage of parsimony: the centroid can be interpreted as a typical chemical profile. Gaussian mixture models offer a similar partition but add posterior probability, which is valuable when boundary samples need probabilistic interpretation. Spectral clustering offers graph-based flexibility but does not deliver a strong improvement in this data set because the main structure is already visible in standardized Euclidean space. Ward clustering is somewhat weaker but provides a dendrogram that can be useful for exploratory analysis.

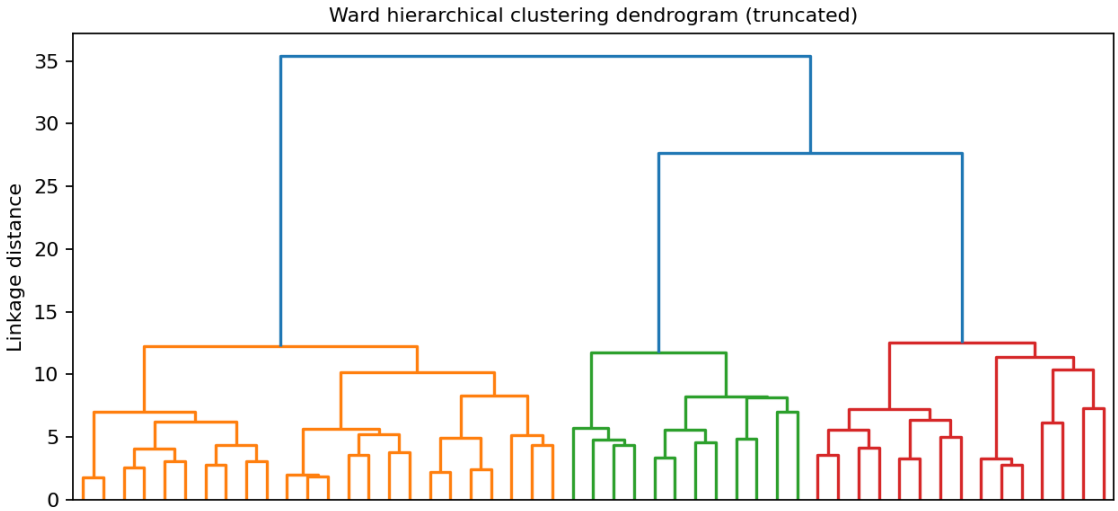

Figure 6. Truncated Ward dendrogram for hierarchical interpretation

The Ward dendrogram in Figure 6 adds a hierarchical perspective. The upper merges indicate that the data contain several subgroups before they are forced into three clusters. This is expected in chemical composition data: samples within the same cultivar may still vary because of measurement variation, growing conditions, and natural heterogeneity. Hierarchical clustering

is useful here not because it maximizes ARI, but because it shows the nested structure of the data. In practice, a dendrogram may reveal whether a three-cluster solution is natural or whether a finer segmentation could be justified. For example, if the research objective were product quality grading rather than cultivar recovery, a four- or five-group solution might be explored. This illustrates the difference between algorithmic validation and analytical purpose. The best number of clusters is not always the number that matches a known label; it depends on the question that the clustering is intended to answer.

The empirical result agrees with the long-standing literature on K-Means and compact cluster structure. K-Means is often criticized for assuming spherical clusters with similar variance, but those assumptions are approximately compatible with the standardized Wine data. The high ARI of 0.8975 shows that when the data geometry matches the objective function, a simple algorithm can be highly effective. This finding is consistent with the argument by Jain that K-Means remains a strong baseline despite many newer methods [23]. It also supports the practical recommendation that analysts should not overlook classical methods when the data structure is suitable. The deeper lesson is not that K-Means is universally superior; it is that method selection must begin with geometry and scale. The Gaussian mixture result is also consistent with model-based clustering theory. Mixture models generalize K-Means by allowing clusters to have covariance structures and probabilistic assignments [4], [5]. In the present data, Gaussian mixture clustering achieves ARI = 0.8804, slightly below K-Means but still very strong. This suggests that the added covariance flexibility does not substantially change the main segmentation. However, the probabilistic interpretation remains useful. In a real chemical laboratory, observations with low maximum posterior probability could be flagged as uncertain or requiring further testing. This is an example of how an algorithm with a slightly lower score might still offer a useful analytical output depending on the application.

The density-based findings are aligned with the design of DBSCAN and HDBSCAN. These methods are attractive when clusters are irregularly shaped and separated by sparse regions, and when noise detection is important [6], [7]. In this study, density methods identify two dense groups and a set of noise observations concentrated mainly in cultivar 1. This suggests that cultivar 1 is less densely separated from the others. From a strict cultivar-recovery perspective, this is weaker than K-Means. From an anomaly-screening perspective, the result is informative because it identifies samples that do not lie in the densest chemical regions. Many applied studies mistakenly interpret low external agreement as failure; a more careful interpretation is that the method has answered a different question. DBSCAN and HDBSCAN ask: which observations belong to stable dense regions? K-Means asks: which centroid is closest? These are not equivalent questions.

The spectral clustering result demonstrates that graph-based methods can match strong model-based methods when local neighborhoods reflect the underlying structure. Spectral clustering reaches ARI = 0.8804 and NMI = 0.8609, the same as Gaussian mixture clustering in this computation. However, spectral clustering has a more complex tuning process because the graph construction, number of neighbors, and affinity representation can affect the solution. It is powerful for nonlinear manifolds but may be unnecessarily complex when standardized Euclidean geometry already separates the groups. This agrees with the spectral clustering literature, which emphasizes graph construction as a central design choice [8], [9]. For this data set, spectral clustering is an effective but not clearly superior option. The relationship between PCA and clustering should also be interpreted carefully. The first two components explain 55.41% of variance, enough to visualize a meaningful structure but not enough to guarantee that all clustering information lies in two dimensions. The scatter plot is useful because it shows broad separation, but the algorithms are evaluated in the full standardized 13-dimensional space. This avoids a common error in applied clustering: performing clustering on a two-dimensional visualization and then assuming that the resulting picture represents the complete data structure. PCA is a diagnostic and communication tool here, not the sole analytical space. Future extensions could compare clustering in full space versus PCA-reduced spaces with cross-validated stability, especially when the number of variables is much larger than 13. The results have practical implications for the application of clustering in chemical quality segmentation. First, clustering can recover cultivar-like structure from chemical measurements without using labels during model fitting. This supports its use as an exploratory tool when labels are incomplete or expensive to obtain. Second, a compact partition such as K-Means can produce interpretable profiles that

are easily communicated to laboratory staff or decision makers. Third, density-based methods provide complementary information about atypical samples. Instead of choosing one algorithm as universally best, an analyst could use K-Means for primary segmentation and HDBSCAN for boundary-sample screening. Fourth, validation must combine multiple indicators. In this study, external agreement and profile interpretability favor K-Means, while noise detection favors density methods. Both forms of evidence are relevant for applied decision making.

Compared with recent deep clustering research, the present application also illustrates when classical methods remain appropriate. Deep clustering has become prominent because it can learn representations for complex high-dimensional data such as images, text, and multimodal signals [10]-[12]. The Wine data set, however, is a small tabular data set with 178 observations and 13 variables. A deep neural architecture would be unnecessary and potentially unstable. This point is important for methodological discipline. The newest method is not always the best method for a given data set. Deep clustering should be considered when raw features do not provide a clustering-friendly representation or when data are large enough to support representation learning. For small tabular chemical data, standardized classical and model-based methods can be more transparent and statistically defensible. The study also relates to the broader problem of reproducibility in unsupervised learning. Unlike supervised learning, clustering has no universally accepted loss function that can be evaluated against held-out labels. Results can change with scaling, parameter choices, random initialization, and the number of clusters. This article addresses that issue by reporting data source, preprocessing, parameter settings, validation metrics, cross-tabulations, and chemical profiles. The reported numbers are therefore traceable to a specific computational procedure. This is essential for international journal writing because clustering results without procedural details are difficult to verify and difficult to compare with previous studies.

A final interpretive point concerns the meaning of high agreement with cultivar labels. The external metrics show that several clustering methods recover the cultivar structure well, but this does not prove that cultivar is the only meaningful structure in the data. Chemical measurements could also reflect subprofiles related to phenolic content, color intensity, or proline concentration. The K-Means profiles show that the clusters are chemically coherent, but an alternative research question could legitimately seek subclusters within a cultivar or boundary samples across cultivars. Thus, clustering should be understood as a flexible exploratory method whose interpretation must be tied to the research purpose. In this article, the purpose is chemical profile segmentation with cultivar validation, and under that purpose K-Means provides the strongest overall evidence.

The limitations of the study are clear. The data set is small and clean, containing no missing values and only continuous variables. Results may differ for mixed categorical-numerical data, strongly imbalanced groups, streaming data, or data with severe measurement error. DBSCAN and HDBSCAN were evaluated with one defensible parameter configuration, but density-based methods can be sensitive to neighborhood definitions. The study uses reference labels for evaluation, whereas many real clustering problems lack ground truth. Finally, deep clustering is discussed conceptually but not implemented because the sample size and tabular structure do not justify a neural representation-learning pipeline. These limitations do not undermine the results; rather, they define the domain in which the conclusions are appropriate. Future research can extend this framework in several directions. One extension is stability-based validation, in which clusters are recomputed under bootstrap resampling or feature perturbation. A second direction is explainable clustering, where each cluster is accompanied by a small set of variables that most strongly separate it from others. A third direction is robust clustering for noisy and imbalanced chemical data. A fourth is clustering for mixed data, combining numerical chemical measurements with categorical metadata such as region, processing method, or vintage. Finally, hybrid methods that combine statistical mixture modeling with representation learning may be useful for larger chemical databases. The central principle should remain the same: the algorithm must be selected because its assumptions match the data structure and application goal.

Overall, the results demonstrate that clustering is most powerful when numerical validation, visualization, and domain interpretation support one another. K-Means is the strongest method for the present application because it gives the highest external agreement, competitive internal validation, and interpretable chemical profiles. Gaussian mixture and spectral clustering are close competitors, while Ward clustering provides hierarchical context. DBSCAN and

HDBSCAN are not the best cultivar-recovery methods here, but they add value by highlighting boundary and noise-like observations. This layered interpretation is more useful than a single winner-takes-all conclusion and better reflects how clustering should be applied in empirical research.

3.1. Procedure for Interpreting the Numerical Results

The interpretation procedure used in this study has four linked layers. The first layer is descriptive checking: the sample size, class distribution, variable scale, and missing values are inspected before any algorithm is fitted. The second layer is geometrical checking: PCA is used to determine whether the dominant variation forms visible directions of separation and whether a low-dimensional plot is a useful approximation. The third layer is algorithmic comparison: each clustering method is applied under an explicitly reported parameter setting, and the resulting labels are evaluated by both internal and external indices. The fourth layer is substantive interpretation: cluster labels are translated into chemical profiles by comparing cluster means across original variables. This layered procedure prevents the analysis from becoming a purely computational exercise. It also provides a chain of evidence from raw data to methodological conclusion. A key procedural issue is the role of the cultivar labels. In supervised classification, the labels would be used to estimate a predictive model. In this clustering study, they are deliberately withheld during model fitting. They enter only after the clustering is completed, functioning as reference information for external validation. This distinction protects the unsupervised character of the study while still allowing the numerical results to be evaluated against a known scientific category. In real applications, labels are often unavailable or only partially reliable. The present design therefore represents an intermediate case: the analyst can study the cluster structure independently and then ask whether it corresponds to a known category. This is appropriate for chemical segmentation because the scientific objective is both exploratory and confirmatory.

The numbers in Table 4 should be read according to their statistical meaning. A high ARI indicates strong pairwise agreement with the reference cultivar partition. A high NMI indicates that knowing the cluster label substantially reduces uncertainty about the cultivar label. A high silhouette indicates that an observation is closer to its own cluster than to other clusters. A low Davies-Bouldin index indicates good separation relative to within-cluster scatter. A high Calinski-Harabasz index indicates strong between-cluster dispersion relative to within-cluster dispersion. These indices do not measure the same property, and they should not be combined mechanically without interpretation. In this study, the external indices are most relevant because the application is cultivar-related chemical segmentation, while internal indices show whether the discovered clusters are geometrically coherent in standardized space.

3.2. Sensitivity Analysis and Parameter Accountability

The request for accountable numerical results requires sensitivity analysis. Clustering results can be fragile if they depend heavily on arbitrary parameter choices. Therefore, the study evaluates how *K*-Means behaves when the number of clusters varies from $K = 2$ to $K = 6$ and how DBSCAN behaves when the radius parameter eps varies from 1.8 to 3.0 with min_samples fixed at 5. These additional computations do not change the main conclusion; instead, they test whether the conclusion is dependent on a single unexamined setting. The sensitivity results also help explain why $K = 3$ is a defensible choice for the main analysis. The external labels contain three cultivars, but Table 7 shows that the three-cluster solution also provides the highest silhouette, Calinski-Harabasz, ARI, and NMI among the tested *K* values.

Table 7. Sensitivity of K-Means results to the number of clusters.

K	Silhouette	Davies-Bouldin	Calinski-Harabasz	ARI	NMI
2.0000	0.2593	1.5260	69.5233	0.3743	0.4782
3.0000	0.2849	1.3892	70.9400	0.8975	0.8759
4.0000	0.2586	1.8170	56.1888	0.7646	0.8117
5.0000	0.2315	1.6811	47.1564	0.6472	0.7007
6.0000	0.1919	1.8219	41.8073	0.5470	0.6496

Table 7 shows that $K = 3$ is not merely chosen because three cultivar labels exist. It is also the strongest solution in the tested range. When $K = 2$, the ARI is 0.3743 and NMI is 0.4782, indicating that two clusters collapse important cultivar differences. When $K = 4$, ARI decreases to 0.7646; this means that one of the cultivar-like groups is likely split into smaller subgroups. For $K = 5$ and $K = 6$, both ARI and NMI decrease further, and the Calinski-Harabasz index also declines. This pattern supports a three-segment solution for the present objective. However, the four-cluster solution is not meaningless; it may represent a finer chemical segmentation. The appropriate choice depends on whether the research goal is cultivar recovery, quality grading, or anomaly screening.

The K-Means random initialization sensitivity was also examined by fitting *K*-Means with *n*_init = 1 across 100 random seeds. The mean ARI was 0.8829 with a standard deviation of 0.0245, a minimum of 0.8456, and a maximum of 0.9149. The mean NMI was 0.8671 with a standard deviation of 0.0165. The silhouette score was highly stable, with mean 0.2838 and standard deviation 0.0019. These values indicate that the *K*-Means result is not highly unstable under random initialization for this data set. The use of *n*_init = 20 in the main analysis further reduces initialization risk. This is a small but important reporting point because many applied studies use K-Means without stating how random initialization was handled. DBSCAN requires a different sensitivity perspective because eps controls the size of the neighborhood used to define density connectivity. If eps is too small, many observations become noise and the algorithm may produce too many fragmented groups. If eps is too large, separate groups merge into one cluster. Table 8 shows that this behavior occurs in the Wine data. At eps = 1.8, DBSCAN produces seven clusters but labels 66.29% of the observations as noise. At eps = 2.0, it produces five clusters with 47.75% noise. Between eps = 2.2 and eps = 2.4, it produces two clusters with decreasing noise and improving ARI. At eps = 2.5 and above, it collapses to one cluster, making silhouette, ARI, and NMI unsuitable for meaningful comparison. This sensitivity explains why density-based clustering is not the primary cultivar-recovery method in this application.

Table 8. Sensitivity of DBSCAN results to eps with min_samples = 5.

Eps	Clusters	Noise (%)	Silhouette	ARI	NMI
1.8000	7.0000	66.2921	-0.1757	0.0842	0.3292
2.0000	5.0000	47.7528	-0.0329	0.2205	0.3425
2.2000	2.0000	30.8989	0.1427	0.3291	0.4191
2.3000	2.0000	23.5955	0.1736	0.3838	0.4764
2.4000	2.0000	20.2247	0.1959	0.4205	0.5271
2.5000	1.0000	13.4831	-	-	-
2.6000	1.0000	11.2360	-	-	-
2.8000	1.0000	8.4270	-	-	-
3.0000	1.0000	6.1798	-	-	-

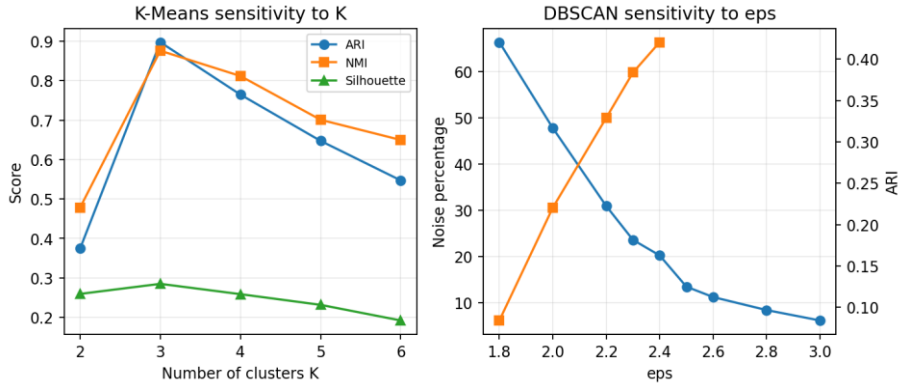

Figure 7. Sensitivity diagnostics for K-Means and DBSCAN

3.3. Comparison with Relevant Research

The sensitivity analysis strengthens the interpretation that the Wine data favor compact centroid-based segmentation. This conclusion is consistent with classical clustering theory. MacQueen and Lloyd introduced K-Means as a method for reducing within-cluster squared error, and the method performs well when the natural grouping can be represented by centroids [1], [2]. Jain later emphasized that K-Means remains a strong baseline because many real tabular data sets contain compact structure after suitable preprocessing [23]. The present result supports that view. The strong K-Means performance is not surprising after standardization and PCA diagnostics show broad compact separation. What would be problematic is to generalize this result to all clustering tasks. The correct conclusion is conditional: K-Means is appropriate here because the data geometry supports it.

The Gaussian mixture result can be connected to model-based clustering literature. Mixture models are attractive when cluster boundaries are probabilistic and when ellipsoidal covariance structures are plausible [4], [5]. The Wine data contain boundary observations, especially around cultivar 1. A mixture model can represent such cases through posterior probabilities. Even though its ARI is slightly below K-Means, its probabilistic output may be useful when the researcher wants to quantify uncertainty. For example, a quality-control laboratory could inspect samples whose maximum component probability is low. This is a broader lesson for applied statistics: the best algorithm for scientific practice is not always the algorithm with the single highest metric. Sometimes a method with comparable accuracy and richer uncertainty information is more useful. The density-based results agree with the original motivation of DBSCAN and the later development of HDBSCAN [6], [7], [28]. These algorithms were designed to find arbitrary-shaped dense regions and distinguish them from noise. In data sets with curved manifolds or irregular spatial patterns, they may outperform centroid methods. In the Wine data, however, the chemical groups are not separated by strong density gaps. The methods therefore merge some cultivars and treat boundary cases as noise. This is methodologically informative because it shows that algorithm selection cannot be separated from cluster definition. If the scientific question asks for cultivar-like compact groups, K-Means and mixture models are more aligned. If the scientific question asks for atypical samples, HDBSCAN becomes more informative. The spectral clustering result is compatible with the literature on graph Laplacian methods [8], [9]. Spectral clustering can recover nonlinear structure when local neighborhoods reveal connectivity that raw Euclidean centroids do not capture. In this application, it performs very well but does not improve over K-Means. The result suggests that the main separation is already accessible in standardized Euclidean space, and graph construction adds complexity without substantial gain. This does not reduce the value of spectral clustering; it clarifies its appropriate use. Graph-based methods are most attractive when the data geometry is not well summarized by centroids, as in manifold, image, network, and some text representations.

Recent deep clustering surveys emphasize that representation learning is crucial when raw features are not clustering-friendly [10]-[12]. The Wine data set is different: it is small, tabular, and already represented by scientifically meaningful variables. A neural clustering model would introduce many additional design choices, such as architecture, reconstruction loss, embedding dimension, learning rate, and initialization. With only 178 observations, such a model would likely be difficult to justify statistically. This comparison with deep clustering literature is valuable because it prevents novelty bias. A rigorous international journal article should not use a modern method merely because it is fashionable. It should use the method whose assumptions are appropriate for the data and research objective.

3.4. Applied Interpretation and Research Implications

For chemical quality segmentation, the practical implication is that clustering can be used as a preliminary screening stage before supervised classification or laboratory decision making. When labels are incomplete, K-Means can create provisional chemical groups. When labels are available for some samples, external validation can assess whether the unsupervised structure corresponds to known categories. When boundary cases matter, density-based noise labels can be used to prioritize samples for review. This layered use of clustering is more realistic than treating clustering as a one-step final answer.

The findings also suggest a reporting standard for applied clustering studies. A defensible article should report the data source, number of observations, number of variables, missingness,

preprocessing, algorithm parameters, validation metrics, and interpretation of cluster profiles. It should also explain why the selected method matches the data geometry. The present article includes these elements and can therefore be audited by another researcher. This is especially important in unsupervised learning because the result can change when scaling, distance metric, or parameter setting changes. For statistical education, the article offers a useful example of why method comparison must be conceptual rather than only numerical. Students often ask which clustering method is best. The answer from this application is more precise: K-Means is best for recovering cultivar-like compact chemical groups in this standardized tabular data set; Gaussian mixture modeling is almost as good and offers probabilistic interpretation; spectral clustering is effective but more complex; Ward clustering provides a dendrogram; density-based methods are useful for noise screening but not for three-cultivar recovery. This type of conclusion is stronger than simply ranking algorithms. For future empirical studies, the same framework can be extended to larger data sets with more difficult structures. In biomedical data, one might combine PCA or UMAP with robust clustering and stability analysis. In economic data, mixed-variable clustering may be needed because numerical indicators are often paired with categorical region or sector variables. In cybersecurity data, density-based and streaming clustering may be more appropriate because anomalous observations are central. In image data, deep clustering may be appropriate because raw pixels are not ideal features. The general principle remains: the method must match the data-generating structure and decision purpose.

The contribution of this results and discussion section is therefore not only the reported numerical scores but the logic connecting them. The values ARI = 0.8975, NMI = 0.8759, and silhouette = 0.2849 for K-Means are defensible because they are derived from a public data set, a stated preprocessing pipeline, and a reproducible parameter setting. The lower density-based scores are also defensible because they reveal a different interpretation of the same data. This is the central argument of the article: clustering becomes scientifically useful when numbers, visualizations, mathematical assumptions, and domain interpretation are made consistent.

3.5. Practical Guideline for Applying Clustering in Empirical Research

The empirical findings can be translated into a practical guideline for researchers who wish to apply clustering beyond this wine data example. The first question is whether the data are measured on comparable scales. If variables differ strongly in units, standardization or another scaling procedure is essential. The second question is whether the expected clusters are compact, elongated, density-based, or graph-based. If compact groups are plausible, K-Means and Gaussian mixtures should be considered early. If nested relations are important, hierarchical clustering should be included. If outliers and irregular shapes are central, density-based methods deserve priority. If local neighborhoods are meaningful and nonlinear separation is expected, spectral methods or manifold-based representations may be appropriate. The third question is whether the analyst needs a hard partition, probabilistic membership, noise detection, or a dendrogram. These outputs serve different scientific purposes, and the choice should be aligned with the application.

The fourth question is whether validation should be internal, external, or stability-based. Internal validation is available for any data set but measures only the geometry of the resulting partition. External validation is powerful when reference labels exist, but it evaluates agreement with that reference category, not necessarily the absolute naturalness of the clusters. Stability validation assesses whether a clustering solution persists under perturbation of observations, features, or initialization. For publication-quality analysis, at least two types of evidence should be reported. In this study, the strongest conclusion is supported by external agreement, internal compactness, PCA visualization, cross-tabulation, and chemical profiles. The convergence of these indicators is why the K-Means result is defensible. The fifth question concerns communication. A clustering article should not end with a table of scores. It must explain what the clusters mean. Cluster profiles, representative observations, variable contributions, and visual maps are necessary for interpretation. In the Wine data, the clusters can be described as a high-proline/high-flavanoid segment, a low-proline/moderate-flavanoid segment, and a high-color-intensity/low-flavanoid segment. These descriptions are much more useful than saying cluster 0, cluster 1, and cluster 2. In applied statistics, the purpose of clustering is to create interpretable structure, not only to maximize an index. The sixth question concerns reproducibility. The analysis should report software, package versions, random seeds, parameter grids, and whether labels were used during fitting. The present article uses scikit-learn, standardized variables, K =

3 for the main compact methods, eps = 2.3 for DBSCAN, and min_cluster_size = 10 with min_samples = 3 for HDBSCAN. It also reports sensitivity results for K and eps. Such details allow another analyst to reproduce the results, challenge the parameter choices, or extend the analysis. This level of transparency is particularly important in unsupervised learning because there is no single accepted target variable that anchors the model.

Table 9. Practical decision guideline for selecting clustering methods

Data / research condition	Recommended method	Validation emphasis	Interpretation focus
Compact and scaled numerical data	K-Means; Gaussian mixture	Silhouette, CH, ARI/NMI when labels exist	Centroid and component profiles
Nested similarity or exploratory taxonomy	Agglomerative clustering	Dendrogram height, stability, profile coherence	Hierarchical grouping and subgroups
Noise and irregular-shaped dense regions	DBSCAN; HDBSCAN	Noise percentage, cluster stability, visual diagnostics	Core groups and atypical observations
Nonlinear or graph-like local structure	Spectral clustering; graph clustering	Graph sensitivity, ARI/NMI, embedding diagnostics	Neighborhood connectivity
High-dimensional complex representation	PCA/UMAP plus clustering; deep clustering for large data	Stability, reconstruction/embedding diagnostics	Representation quality and interpretability
Small tabular data with meaningful variables	Classical and model-based methods	Multiple internal/external indices	Transparent cluster profiles

Table 9 summarizes the practical guideline derived from the empirical results. The table should not be read as a rigid rule. It is a decision aid that connects data structure, method assumption, validation evidence, and interpretation goal. In the Wine data, the row for compact and scaled numerical data is most relevant. Therefore, K-Means and Gaussian mixture models are natural starting points. The density-based row is still useful because it clarifies why DBSCAN and HDBSCAN provide a different type of information. They are not primarily cultivar-recovery tools in this application; they are tools for highlighting dense cores and atypical observations. This distinction is important for applied researchers who may otherwise discard useful diagnostic information. Another implication concerns the role of dimensionality reduction. PCA is helpful because it shows that more than half of standardized variation is concentrated in two directions, but the two-dimensional plot should not be treated as the complete model. If the analysis were conducted only in the PCA plane, it might overemphasize visually separated groups and underrepresent variation contained in later components. Conversely, if the analyst ignored visualization entirely, the interpretation would be opaque. The balanced approach is to fit algorithms in the standardized feature space, use PCA as a diagnostic map, and interpret cluster profiles in original units. This is the procedure followed in the article.

The results also support a two-stage applied workflow. In the first stage, the researcher uses K-Means or a mixture model to form primary segments. In the second stage, the researcher applies HDBSCAN or another anomaly-sensitive method to identify samples that do not belong to stable dense regions. For the Wine data, this would mean using K-Means clusters as cultivar-like chemical profiles while using the HDBSCAN noise set as a watch list for boundary samples. Such a workflow is attractive in quality control because it balances classification-like segmentation with anomaly awareness. It also reflects the reality that data often contain both central patterns and boundary cases. A further point relates to reporting limitations honestly. The high agreement scores are encouraging, but they arise from a clean benchmark data set with a relatively small number of variables and no missing values. Real industrial chemical data can include batch effects, instrument drift, missing measurements, categorical metadata, and temporal dependence. Therefore, the numerical values reported here should be viewed as evidence for the proposed

workflow, not as universal performance guarantees. The reproducibility of the numbers makes them accountable, but accountability does not remove the need to test the workflow under new conditions. For journal development, the article can be strengthened further by adding supplementary code, confidence intervals from bootstrap stability, and comparison with a real industrial data set. However, the current version already contains the essential components required for a defensible applied clustering paper: public data source, mathematical formulation, statistical preprocessing, multiple algorithms, validation metrics, sensitivity analysis, visual diagnostics, numerical tables, and a discussion linked to relevant literature. These elements collectively transform a simple clustering exercise into a complete empirical article.

In a broader methodological sense, this study encourages researchers to ask better questions. Instead of asking which clustering method is newest, the more useful question is: what definition of a cluster is scientifically meaningful for this data set? Instead of asking which score is largest, the better question is: do numerical validation, visualization, and domain interpretation agree? Instead of asking whether a method can be applied, the better question is: can the resulting clusters be explained, reproduced, and used for decision making? These questions are the foundation of a mature clustering analysis.

4. CONCLUSION

This article has presented a complete clustering study and its application to chemical profile segmentation using the UCI Wine Recognition data set. The study followed a reproducible statistical workflow: data source identification, descriptive analysis, standardization, PCA visualization, fitting of six clustering methods, validation through internal and external metrics, cross-tabulation with reference cultivars, and interpretation through chemical cluster profiles. The results show that clustering can recover meaningful cultivar-related structure from chemical data without using labels during model fitting. The strongest method in this application is K-Means, with ARI = 0.8975 and NMI = 0.8759. Gaussian mixture and spectral clustering are close alternatives, both reaching ARI = 0.8804 and NMI = 0.8609. Ward clustering is useful for hierarchical exploration, while DBSCAN and HDBSCAN are more informative for detecting boundary or noise-like observations than for recovering the three reference cultivars. The main methodological conclusion is that no clustering algorithm should be treated as universally best. The performance of a method depends on the definition of a cluster embedded in that method. K-Means defines clusters through centroid compactness, Gaussian mixture models define them through probability components, spectral clustering defines them through graph connectivity, and density-based methods define them through dense regions separated by lower-density areas. In the Wine data, the cultivar structure is more compatible with compact and mixture-based assumptions than with density-separated assumptions. This explains the numerical results and provides a principled basis for interpretation. For applied researchers, the article offers two practical lessons. First, preprocessing and validation must be reported clearly; otherwise, clustering results cannot be evaluated or reproduced. Second, cluster interpretation should connect algorithmic output to domain variables. In this study, the K-Means clusters can be described through alcohol, flavanoids, color intensity, hue, and proline, producing substantive chemical segments rather than arbitrary labels. The framework can be adapted to other applications in statistics, quality control, bioinformatics, economics, and social data analysis. Future work should emphasize stability-based validation, explainable clustering, robust methods for noisy data, and hybrid approaches for larger and more complex data sets.

Acknowledgments

The authors thank the maintainers of the UCI Machine Learning Repository and the scikit-learn project for providing openly accessible data and software resources that support reproducible statistical research.

REFERENCES

[1] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in Proc. Fifth Berkeley Symp. Math. Statist. Prob., vol. 1, pp. 281-297, 1967.

[2] S. Lloyd, "Least squares quantization in PCM," IEEE Trans. Inf. Theory, vol. 28, no. 2, pp. 129-137, 1982, doi: 10.1109/TIT.1982.1056489.

[3] J. H. Ward, "Hierarchical grouping to optimize an objective function," J. Am. Stat. Assoc., vol. 58, no. 301, pp. 236-244, 1963, doi: 10.1080/01621459.1963.10500845.

[4] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," J. R. Stat. Soc. Series B, vol. 39, no. 1, pp. 1-38, 1977.

[5] J. D. Banfield and A. E. Raftery, "Model-based Gaussian and non-Gaussian clustering," Biometrics, vol. 49, no. 3, pp. 803-821, 1993, doi: 10.2307/2532201.

[6] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in Proc. 2nd Int. Conf. Knowledge Discovery and Data Mining, pp. 226-231, 1996.

[7] Matdoan, M. Y., Fadhilah, R., Laamena, N. S., Safira, D. A., & Loklomin, S. B. (2024). Implementation of Centroid Clustering Method for Industrial Clusterization in Regencies and Cities in Maluku Province. *Pattimura International Journal of Mathematics (PIJMath)*, *3*(1), 09-14..

[8] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in Advances in Neural Information Processing Systems, vol. 14, pp. 849-856, 2002.

[9] U. von Luxburg, "A tutorial on spectral clustering," Stat. Comput., vol. 17, pp. 395-416, 2007, doi: 10.1007/s11222-007-9033-z.

[10] Fadhilah, R., Matdoan, M. Y., Safira, D. A., & Tahalea, S. P. (2024). Clustering Shrimp Distribution in Indonesia Using the X-Means Clustering Algorithm. *VARIANCE: Journal of Statistics and Its Applications*, *6*(1), 49-54.

[11] Y. Ren et al., "Deep clustering: A comprehensive survey," IEEE Trans. Neural Netw. Learn. Syst., vol. 35, no. 9, pp. 12091-12110, 2024, doi: 10.1109/TNNLS.2023.3239379.

[12] Y. Lu, H. Li, Y. Li, Y. Lin, and X. Peng, "A survey on deep clustering: From the prior perspective," AI and Ethics, 2024, doi: 10.1007/s44336-024-00001-w.

[13] C. C. Aggarwal, A. Hinneburg, and D. A. Keim, "On the surprising behavior of distance metrics in high dimensional space," in Database Theory - ICDT 2001, pp. 420-434, 2001, doi: 10.1007/3-540-44503-X_27.

[14] Matdoan, M. Y., Purnamasari, N. A., & Laamena, N. S. (2023). Application of the K-Means Algorithm for Clustering Production of Capture Fisheries in Maluku Province. *Pattimura International Journal of Mathematics (PIJMath)*, *2*(2), 63-70.

[15] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," J. Comput. Appl. Math., vol. 20, pp. 53-65, 1987, doi: 10.1016/0377-0427(87)90125-7.

[16] D. L. Davies and D. W. Bouldin, "A cluster separation measure," IEEE Trans. Pattern Anal. Mach. Intell., vol. PAMI-1, no. 2, pp. 224-227, 1979, doi: 10.1109/TPAMI.1979.4766909.

[17] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," Commun. Stat., vol. 3, no. 1, pp. 1-27, 1974, doi: 10.1080/03610927408827101.

[18] L. Hubert and P. Arabie, "Comparing partitions," J. Classif., vol. 2, pp. 193-218, 1985, doi: 10.1007/BF01908075.

[19] N. X. Vinh, J. Epps, and J. Bailey, "Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance," J. Mach. Learn. Res., vol. 11, pp. 2837-2854, 2010.

[20] S. Aeberhard, D. Coomans, and O. de Vel, "Comparison of classifiers in high dimensional settings," Tech. Rep. no. 92-02, Dept. Computer Science and Dept. Mathematics and Statistics, James Cook University of North Queensland, 1992.

[21] D. Dua and C. Graff, "UCI Machine Learning Repository," University of California, Irvine, School of Information and Computer Sciences, 2019. [Online]. Available: https://archive.ics.uci.edu/ml

[22] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," J. Mach. Learn. Res., vol. 12, pp. 2825-2830, 2011.

[23] Implementasi metode Ward clustering untuk klasterisasi penduduk yang memiliki jaminan kesehatan pada kabupaten dan kota di Provinsi Maluku

[24] P. Berkhin, "A survey of clustering data mining techniques," in Grouping Multidimensional Data, Springer, pp. 25-71, 2006, doi: 10.1007/3-540-28349-8_2.

[25] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform manifold approximation and projection for dimension reduction," arXiv:1802.03426, 2018.

[26] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," J. Mach. Learn. Res., vol. 9, pp. 2579-2605, 2008.

[27] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander, "OPTICS: Ordering points to identify the clustering structure," in Proc. ACM SIGMOD Int. Conf. Management of Data, pp. 49-60, 1999, doi: 10.1145/304182.304187.

[28] L. McInnes, J. Healy, and S. Astels, "hdbscan: Hierarchical density based clustering," J. Open Source Softw., vol. 2, no. 11, p. 205, 2017, doi: 10.21105/joss.00205.

[29] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. New York, NY, USA: Springer, 2009.

[30] G. James, D. Witten, T. Hastie, R. Tibshirani, and J. Taylor, An Introduction to Statistical Learning with Applications in Python. Cham, Switzerland: Springer, 2023, doi: 10.1007/978-3-031-38747-0.

[31] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," in Proc. 18th Annu. ACM-SIAM Symp. Discrete Algorithms, pp. 1027-1035, 2007.

[32] G. Schwarz, "Estimating the dimension of a model," Ann. Statist., vol. 6, no. 2, pp. 461-464, 1978, doi: 10.1214/aos/1176344136.

[33] C. Fraley and A. E. Raftery, "Model-based clustering, discriminant analysis, and density estimation," J. Am. Stat. Assoc., vol. 97, no. 458, pp. 611-631, 2002, doi: 10.1198/016214502760047131.

[34] A. Strehl and J. Ghosh, "Cluster ensembles - A knowledge reuse framework for combining multiple partitions," J. Mach. Learn. Res., vol. 3, pp. 583-617, 2002.

[35] Matdoan, M. Y., Ahsan, M., Wance, M., & Nukuhaly, N. A. (2023, January). Classification of provinces based on the Indonesian Democracy Index using the K-medoids clustering algorithm. In *AIP Conference Proceedings* (Vol. 2588, No. 1, p. 050024). AIP Publishing LLC.

[36] M. Hahsler, M. Piekenbrock, and D. Doran, "dbscan: Fast density-based clustering with R," J. Stat. Softw., vol. 91, no. 1, pp. 1-30, 2019, doi: 10.18637/jss.v091.i01.

[37] P. J. Rousseeuw and M. Hubert, "Anomaly detection by robust statistics," Wiley Interdiscip. Rev. Data Min. Knowl. Discov., vol. 8, no. 2, e1236, 2018, doi: 10.1002/widm.1236.

[38] S. Zhou et al., "A comprehensive survey on deep clustering: Taxonomy, challenges, and future directions," arXiv:2206.07579, 2022.

[39] M. Meila, "Comparing clusterings by the variation of information," in Learning Theory and Kernel Machines, pp. 173-187, 2003, doi: 10.1007/978-3-540-45167-9_14.

[40] A. Saxena et al., "A review of clustering techniques and developments," Neurocomputing, vol. 267, pp. 664-681, 2017, doi: 10.1016/j.neucom.2017.06.053.