



Research Article /Review Article:

Evaluation of Internal Cluster Validity Indices for Selecting the Number of K-Means Clusters in Sentinel-2 Imagery

Abdüsselam Kesikoğlu ^{*1}, Ahmet Emin Karkınlı ²

¹ Tomarza MA Vocational School, Kayseri University, Kayseri 38280, Türkiye, e-mail: kesikoglu@kayseri.edu.tr

² Department of Geomatic Engineering, Niğde Ömer Halisdemir University, Niğde, Türkiye

* Correspondence: kesikoglu@kayseri.edu.tr

Article History:

Received: 09.08.2025

Revised 1: 07.10.2025

Revised 2: 14.10.2025

Published: 23.10.2025

Keywords:

K-means;
Cluster number selection;
Internal cluster validity;
Sentinel-2;
LUCAS.



This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution 4.0 International License

(CC BY).

Publisher's Note:

This Journal remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Abstract

Selecting the number of clusters is a central methodological decision in unsupervised analysis of multispectral imagery. This study evaluated the Silhouette coefficient, Calinski-Harabasz index, Davies-Bouldin index, and gap statistic for K-means clustering of harmonised Sentinel-2 Level-2A growing-season composites. Three 50 × 50 km study windows in Spain and Cyprus were analysed using two feature representations: six surface-reflectance bands and the same bands augmented with NDVI, MNDWI, NBR, and BSI. Candidate solutions containing 2-12 clusters were compared with an aggregated LUCAS Copernicus 2018 reference legend using V-measure, assessed for stability across 20 fits to 80% pixel subsamples, and examined with a chance-corrected eight-neighbour agreement coefficient. Silhouette selected two clusters in all six window-by-feature combinations. Calinski-Harabasz selected two clusters in five combinations, whereas Davies-Bouldin produced interior minima in two. The gap statistic did not satisfy its one-standard-error stopping rule within the evaluated range. The V-measure optimum varied among 2, 4, 6, 7, and 11 clusters, and no internal criterion matched it in every combination. The maximum V-measure ranged from 0.213 to 0.590. The three compactness-and-separation criteria had similar median agreement losses (0.115-0.127), but losses varied substantially among windows and reached 0.286. The adjusted Rand index identified the same agreement optimum as V-measure in five of six combinations. Stability was generally high but showed pronounced local reductions in two index-selected solutions. At matched cluster numbers, adding spectral indices increased local neighbour agreement in 27 of 33 comparisons. For the examined scenes, internal validity alone did not consistently identify the partition with the highest thematic agreement. Index curves, thematic agreement, and subsampling stability should therefore be considered together when selecting the number of clusters.

Copyright © 2025 by the author(s).

Contribution to the field statement:

This study advances remote-sensing research by systematically comparing four internal cluster validity indices against thematic agreement, subsampling stability, and spatial continuity in Sentinel-2 imagery. The findings demonstrate that internal indices alone do not consistently identify the most thematically informative K-means partition, highlighting the value of combining complementary validation measures for more reliable cluster-number selection.

Highlights:

- Four clustering measures produced inconsistent recommendations across study areas.
- The best cluster number varied from 2 to 11 across the tested scenes.
- Spectral indices improved local spatial agreement in 27 of 33 comparisons.
- Combining thematic agreement and stability improved cluster selection.

Cite this article?

Kesikoğlu, A., & Karkınlı, A. E. (2025). Evaluation of internal cluster validity indices for selecting the number of K-means clusters in Sentinel-2 imagery. *Advanced Remote Sensing*, 5(1), 21–34.

1. Introduction

The Sentinel-2 constellation delivers multispectral observations at 10-20 m with a revisit interval of five days and is a standard input for land-surface mapping [1]. The availability of analysis-ready Level-2A surface-reflectance archives through cloud platforms such as Google Earth Engine [2] has moved the practical bottleneck



away from data access and towards the methodological choices that turn reflectance into information. Supervised classification is used where labelled samples exist [3], while unsupervised clustering is applied to stratify scenes before sampling, to segment landscapes into spectrally homogeneous units for object-based analysis [4], and to explore data where a thematic legend has not been fixed. In each case the analyst must supply one parameter that the data do not directly provide - the number of clusters.

K-means is one of the most widely used partitional algorithms in remote-sensing applications, on account of its interpretability, its linear scaling in the number of observations and its availability in common software environments [5]. The algorithm originates with MacQueen [6] and Lloyd [7] and remains the subject of an extensive methodological literature [8]. Its known sensitivities - to initialisation, to clusters of unequal size and shape, and to the assumed Euclidean geometry - are partly addressed by seeding strategies such as k-means++ [9], [10]. None of these addresses the choice of k , which remains external to the algorithm.

The common approach is to run K-means over a range of candidate values and select the k that optimises an internal cluster validity index computed from the data and the resulting partition alone. Comparative studies of such indices have a long history [11], [12], and benchmarks continue to appear: Arbelaitz et al. [13] evaluated 30 indices over 720 synthetic and 20 real data sets, and a recent review surveys a large and growing family of criteria [14]. Four commonly used indices were selected for evaluation here. The Silhouette width [15] contrasts, for each object, the mean distance to its own cluster with the mean distance to the nearest competing cluster. The Calinski-Harabasz criterion [16] is a variance ratio between the between-cluster and within-cluster dispersion matrices. The Davies-Bouldin index [17] averages over clusters the worst-case ratio of within-cluster scatter to between-cluster separation. The gap statistic [18] compares the observed within-cluster dispersion with its expectation under a reference distribution containing no clustering structure. The first three score compactness and separation directly; only the fourth is referenced against a null model.

Evidence specific to remote sensing is comparatively limited. Li et al. [19] evaluated 16 validity indices on nine multispectral and hyperspectral data sets and found that no single index performed best across all of them. That comparison asked how often an index recovered a known number of groups. A separate question is how the selected partition behaves when assessed by criteria the index does not encode. Internal indices score a partition using geometric notions closely related to the objective K-means minimises, and Gagolewski et al. [20] show that internal measures can rank highest the partitions their own preference favours rather than those matching a known reference structure.

Two additional properties are not represented by the four internal indices and are seldom included in remote-sensing assessments of cluster-number selection. The first is stability: a partition that changes substantially when K-means is refitted to a subset of the same observations provides a weak basis for subsequent interpretation, and stability-based approaches have therefore been proposed for cluster validation [21]–[24]. The second is spatial continuity. Pixel clustering treats observations as an unordered set, although land-cover patterns are spatially structured; an internally favourable partition may consequently appear locally fragmented when mapped.

The choice of reference data also affects external evaluation. Comparisons with satellite-derived land-cover products partly reuse spectral information related to the clustered imagery. LUCAS Copernicus 2018 provides field-derived land-cover observations over areas resolvable by Sentinel-2 [25], together with harmonised attributes and quality information [26]. In the present analysis, the polygon was treated as the reference unit, consistent with accuracy-assessment practice that bases inference on independently sampled reference units rather than individual pixels [27], [28].

Using a fixed K-means configuration, this study evaluates four commonly used internal cluster validity indices for selecting the number of clusters in Sentinel-2 composites. The cluster numbers indicated by the indices are compared with the partitions that maximise agreement with an aggregated LUCAS legend, and the corresponding solutions are examined for stability under pixel subsampling. The analysis also evaluates whether augmenting reflectance bands with spectral indices changes cluster-number recommendations, thematic agreement, and local spatial continuity.

2. Method

2.1. Study areas and data

Selection criteria were specified before clustering. The 63,287 LUCAS Copernicus 2018 polygons were first restricted to records for which the Copernicus land-cover attribute agreed with the surveyed point observation and the polygon area was at least 0.20 ha, corresponding to five 20 m pixels. Candidate windows were generated on the ETRS89-LAEA grid (EPSG:3035) using three offsets along each axis. Each candidate was summarised by the number of retained polygons, number of broad land-cover classes, dominant-class proportion, and Shannon entropy of the class distribution. Table 1 summarises the selected windows. Figure 1 shows their seasonal true-colour composites and the retained LUCAS reference locations.

The initial target of 80 retained polygons could not be achieved in a 25 km window and was approached only after increasing the window dimension to 50 km. Candidate 50 × 50 km windows were retained when they contained at least 60 polygons from at least four broad classes and no class exceeded 80% of the selection-stage sample. Qualifying windows were ordered by Shannon entropy and spatially filtered to retain regions separated



by at least 200 km. The lowest-, median-, and highest-entropy regions were designated windows A, B, and C. Because entropy reflects the reference sample rather than image structure, the ordering was checked using two image-based summaries: the mean absolute difference between horizontally adjacent pixels in standardized reflectance, averaged across six bands, and the corresponding statistic for NDVI.

Table 1. Study windows and data characteristics.

Window	Country	Centre	Scenes	Median observatio ns	NoData	Final polygons (Art/Crop/Grass/W ood)	Final dominant share	Selection entropy	Reflectance texture	NDVI difference
A	ES	41.50°N, 5.38°W	228	49	0.00%	72 (6/58/2/6)	80.6%	0.701	0.233	0.033
B	ES	42.58°N, 3.18°W	120	25	0.04%	61 (2/41/9/9)	67.2%	0.944	0.238	0.045
C	CY	34.95°N, 33.17°E	110	26	0.17%	86 (2/23/26/35)	40.7%	1.174	0.243	0.048

Note. Art, Crop, Grass, and Wood denote the four broad classes. Entropy refers to the selection-stage polygon set; final counts and dominant share refer to polygons retained for analysis.

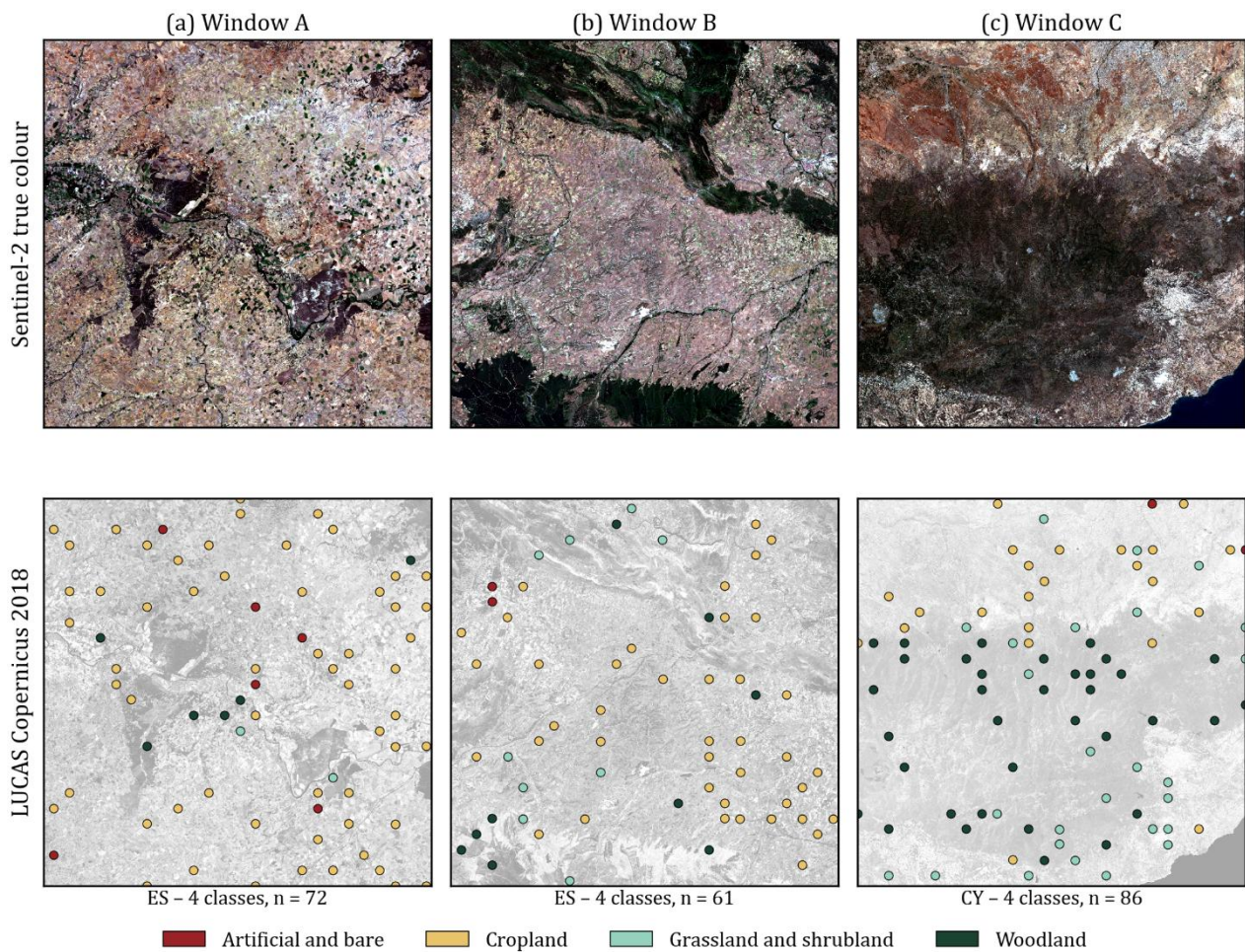


Figure 1. Study windows and LUCAS reference data.

Harmonised Sentinel-2 Level-2A surface-reflectance data were obtained from the Google Earth Engine archive for 1 May–30 September 2018 [2]. Atmospheric correction and the scene classification layer were provided by Sen2Cor [29]. Scenes with catalogued cloud cover of 35% or more were excluded. Within the retained scenes, pixels classified as cloud shadow, medium- or high-probability cloud, cirrus, or snow and ice were masked. A per-pixel median composite was then calculated, together with the number of valid observations contributing to each composite pixel.

The analysis used Sentinel-2 bands B2, B3, B4, and B8 at 10 m and B11 and B12 at 20 m. All variables were aligned to a common 20 m ETRS89-LAEA grid (EPSG:3035). The 10 m reflectance bands were aggregated to 20 m using mean aggregation, while the native 20 m bands were aligned to the same grid without changing their nominal pixel size. Upsampling of the 20 m bands to 10 m was avoided. Each window was exported in tiles and mosaicked locally, and the resulting reflectance distributions and valid-observation counts were checked before analysis.



External comparison used the LUCAS Copernicus 2018 polygon collection. During the Copernicus field module, the homogeneous land-cover extent around each sampled point was recorded in the cardinal directions, and polygons were constructed from those observations [25]. Attributes and quality information were obtained from the harmonised LUCAS database [26]. Level-1 land-cover codes were aggregated before clustering into the five broad classes listed in Table 2, four of which occurred in the selected windows. Polygons were retained when the Copernicus land-cover attribute agreed with the point observation, the area was at least 0.20 ha, and at least three valid pixel centres fell inside the polygon. Rare classes were not merged with other classes.

Table 2. Aggregation of LUCAS land-cover codes.

Broad class	LUCAS level-1 codes	Present in the study windows
Artificial and bare	A (artificial land), F (bare land and lichens/moss)	Yes
Cropland	B (cropland)	Yes
Grassland and shrubland	D (shrubland), E (grassland)	Yes
Woodland	C (woodland)	Yes
Water and wetland	G (water areas), H (wetlands)	No

2.2. Feature construction and K-means clustering

Two feature representations were evaluated. The first contained the six surface-reflectance bands. The second contained the same bands together with NDVI [30], MNDWI [31], NBR [32], and BSI [33], calculated from the seasonal composite (Table 3). The spectral indices are nonlinear transformations of the reflectance bands and therefore alter the weighting and geometry of the feature space rather than introduce an independent data source. Indices expressing the same band contrast were not included together. Because K-means uses Euclidean distance, each variable was standardized separately within each study window and feature representation:

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j} \quad (1)$$

where z_{ij} is the standardized value of variable j for pixel i , x_{ij} is the observed value, and μ_j and σ_j are the mean and standard deviation estimated from the training sample. The same parameters were applied to all valid pixels in the corresponding window.

Table 3. Spectral indices in the band-plus-index feature representation.

Index	Formula	Sentinel-2 bands	Spectral contrast
NDVI	$\frac{B_8 - B_4}{B_8 + B_4}$	B8, B4	Near infrared against red
MNDWI	$\frac{B_3 - B_{11}}{B_3 + B_{11}}$	B3, B11	Green against short-wave infrared
NBR	$\frac{B_8 - B_{12}}{B_8 + B_{12}}$	B8, B12	Near infrared against SWIR 2
BSI	$\frac{(B_{11} + B_4) - (B_8 + B_2)}{(B_{11} + B_4) + (B_8 + B_2)}$	B11, B4, B8, B2	Short-wave infrared and red against near infrared and blue

K-means clustering was evaluated for 2-12 clusters, with the upper limit specified before examination of the results. For each window, 50,000 valid pixels were selected by systematic spatial sampling with a random start. Clustering was performed with scikit-learn 1.8.0 [34] using Lloyd's algorithm, k-means++ initialization, 30 random starts, a maximum of 500 iterations per start, and a convergence tolerance of 10^{-4} . A fixed pseudorandom seed was used throughout. All subsequent measures were calculated from the same fitted partitions, and Table 4 summarises the analysis settings.

Table 4. Data and analysis settings.

Element	Setting
Data source	Harmonised Sentinel-2 Level-2A surface reflectance
Composite period	1 May-30 September 2018
Scene-level cloud threshold	Less than 35% catalogued cloud cover
Pixel-quality mask	Cloud shadow, medium/high-probability cloud, cirrus, snow and ice
Spectral bands	B2, B3, B4, B8 (10 m); B11, B12 (20 m)
Analysis grid	20 m ETRS89-LAEA (EPSG:3035)
Study window	50 × 50 km per window



Element	Setting
Band-only representation	Six surface-reflectance bands
Band-plus-index representation	Six bands plus NDVI, MNDWI, NBR and BSI
Standardization	Zero mean and unit standard deviation within each window and representation
Training sample	50,000 valid pixels per window; systematic spatial sample with random start
Software	scikit-learn 1.8.0
K-means fitting	Lloyd algorithm; k-means++ initialization; 30 random starts; maximum 500 iterations; tolerance 10^{-4}
Candidate cluster numbers	2-12, specified before evaluation
Internal validity indices	Silhouette and Calinski-Harabasz maximized; Davies-Bouldin minimized; Gap one-standard-error rule
Silhouette evaluation	Fixed 10,000-pixel sample per window
Gap reference distribution	20 coordinate-wise uniform reference data sets per cluster number; same sample size and dimensionality as observed data
Gap clustering settings	10 random starts for observed data and 3 for references; maximum 500 and 300 iterations, respectively
Reference data	LUCAS Copernicus 2018 polygons; four broad classes represented
External agreement	V-measure (primary) and adjusted Rand index (robustness)
Stability assessment	20 fits to 80% subsamples without replacement; 5 random starts; fixed 10,000-pixel evaluation sample
Spatial assessment	Chance-corrected eight-neighbour agreement coefficient
Uncertainty	1,000 paired polygon bootstrap samples; 95% percentile intervals

2.3. Cluster validation

The Silhouette coefficient [15] was calculated as $s(i) = \frac{b_i - a_i}{\max(a_i, b_i)}$, where a_i is the mean distance from observation i to other observations in its cluster and b_i is the lowest mean distance to another cluster. Larger values indicate greater within-cluster compactness and between-cluster separation. Silhouette was evaluated on a fixed sample of 10,000 pixels per window, used for every candidate cluster number and both feature representations. The Calinski-Harabasz criterion [16] is a degrees-of-freedom-adjusted ratio of between-cluster to within-cluster dispersion, with larger values preferred. The Davies-Bouldin index [17] averages the largest ratio of within-cluster scatter to centroid separation for each cluster, with smaller values preferred.

The Gap statistic [18] compares the observed within-cluster dispersion W_k with its expectation under a reference distribution:

$$Gap(k) = E_{ref}[\log W_k] - \log W_k \quad (2)$$

The within-cluster dispersion W_k was the sum of squared Euclidean distances between observations and their assigned centroids. The reference expectation was estimated from $B = 20$ independently generated data sets sampled from a coordinate-wise uniform distribution over the hyperrectangle defined by the standardized training data. Each reference set had the same number of observations and variables as the observed sample. Observed partitions used 10 random starts and a maximum of 500 iterations; reference partitions used three random starts and a maximum of 300 iterations. The standard-error term s_k was calculated as the reference standard deviation multiplied by $\sqrt{1 + 1/B}$. Silhouette and Calinski-Harabasz were maximized, Davies-Bouldin was minimized, and the gap statistic selected the smallest k satisfying $Gap(k) \geq Gap(k + 1) - s_{k+1}$. When no candidate satisfied this rule, the result was reported as no stopping point within the evaluated range.

For each retained polygon, the modal cluster label was determined from valid pixels whose centres fell inside the polygon boundary. Invalid pixels were omitted, and polygons with fewer than three valid pixels had already been excluded. Ties were resolved deterministically by assigning the numerically smallest cluster label. Agreement between polygon-level cluster assignments and the broad LUCAS classes was evaluated primarily with V-measure [35], the harmonic mean of homogeneity and completeness. The adjusted Rand index [36] was calculated as a robustness measure. Overall accuracy was not used because it assumes a one-to-one mapping between clusters and reference classes. For each window and feature representation, the candidate with the largest V-measure was identified within the evaluated range. The agreement loss associated with internal criterion j was defined as:

$$L_j = V_{max} - V(k_j) \quad (3)$$

where V_{max} is the largest V-measure in the evaluated range and k_j is the cluster number selected by criterion j . Agreement loss was calculated only for criteria that returned a cluster number. Because the gap statistic did not produce a stopping point, its endpoint difference was reported separately.

For each candidate cluster number, 20 K-means fits were obtained from random 80% subsamples of the 50,000 training pixels, sampled without replacement. Each fit used k-means++ initialization, five random starts, a maximum of 300 iterations, and a convergence tolerance of 10^{-4} . Labels were predicted for the same fixed 10,000-pixel evaluation sample in every fit. The adjusted Rand index was calculated for all 190 pairs of replicate partitions, and stability was summarized as:



$$Stab(k) = median_{a < b} ARI(C_k^{(a)}, C_k^{(b)}) \quad (4)$$

where $C_k^{(a)}$ denotes the partition predicted by replicate a for k clusters. Larger values indicate greater consistency among fits to perturbed samples. Because the 190 pairwise values are derived from 20 overlapping fits, they are not independent observations; the median was therefore used as a descriptive stability measure without a confidence interval. Random pixel subsampling does not evaluate sensitivity to spatially structured perturbations.

Local spatial continuity was summarized on the full cluster raster with a kappa-form eight-neighbour agreement coefficient:

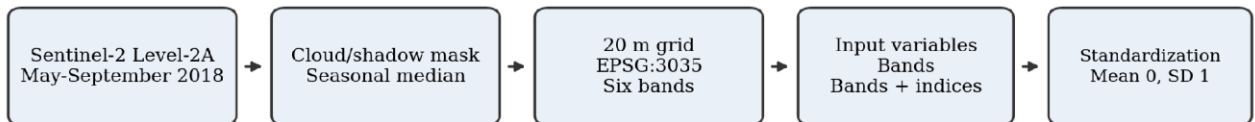
$$\kappa_{NN}(k) = \frac{P_o(k) - P_e(k)}{1 - P_e(k)} \quad (5)$$

where $P_o(k)$ is the observed proportion of valid eight-neighbour pixel pairs assigned to the same cluster and $P_e(k) = \sum_c p_c^2$ is the expected agreement obtained from cluster proportions p_c . Neighbour pairs were enumerated in four directions so that each pair was counted once; pairs containing an invalid pixel or extending beyond the raster boundary were omitted. The coefficient is a descriptive measure of local label continuity. Because the chance correction does not fully remove the structural effect of increasing the number of labels, comparisons were restricted to partitions with the same number of clusters. The correction follows the general form of Cohen's kappa [37].

2.4. Statistical analysis

The polygon was the unit of analysis for thematic agreement. Ninety-five per cent confidence intervals were obtained from 1,000 nonparametric bootstrap samples of polygons drawn with replacement [38]. The same bootstrap samples were used for every candidate cluster number within a window, allowing paired differences from the V-measure maximum to be calculated. A difference was described as not detectable when its 95% percentile interval included zero; this criterion was not interpreted as statistical equivalence. Negative interval limits reflect sampling variation in paired differences. The bootstrap does not account for spatial dependence among polygons within a window. Pixel counts describe data volume and were not treated as independent sample size. Results were reported separately for each window and feature representation, without a pooled estimate. The complete design comprised three windows, two feature representations, and eleven candidate cluster numbers, with additional subsampling fits for the stability analysis. Figure 2 summarises the analytical workflow.

Data preparation



Clustering and independent evaluation



Outcome assessment

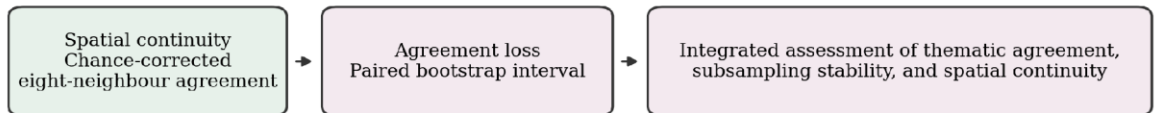


Figure 2. Analytical workflow.

3. Results

3.1. Study-window characteristics

The window-selection audit constrained the design. Across the LUCAS Copernicus 2018 collection (63,287 polygons, of which 41,173 met the homogeneity and 0.20 ha criteria) no 25 x 25 km window in the European Union reached the design target of 80 retained polygons; the best-populated window held 30. Enlarging the window to 40 km raised this to 63 and to 50 km to 86, so the window was fixed at 50 x 50 km.



The three windows (Table 1, Figure 1) are a cereal plain in the Duero basin (window A), a cropland–woodland mosaic on the northern Meseta (window B), and a Mediterranean landscape in central Cyprus (window C). Between 110 and 228 Sentinel-2 scenes with less than 35% scene cloud cover contributed to each composite, giving medians of 49, 25 and 26 valid observations per pixel. The proportions of pixels without a valid observation were 0.00%, 0.04% and 0.17%. After restricting the reference data to homogeneous polygons of at least 0.20 ha containing at least three valid pixel centres, 72, 61 and 86 polygons remained, distributed over four broad classes in each window. No window contained water or wetland polygons, so the effective legend has four classes. Window A is strongly imbalanced: 80.6% of its final polygons are cropland, and grassland and shrubland is represented by two polygons.

The windows were ordered by the Shannon entropy of the selection-stage LUCAS class composition. The mean absolute difference between horizontally adjacent pixels in standardized reflectance, averaged over six bands, was 0.233, 0.238, and 0.243 for windows A, B, and C; the corresponding NDVI values were 0.033, 0.045, and 0.048. Both image summaries reproduced the entropy ordering, although the differences were small (approximately 4% across the reflectance measure). The three windows should therefore be regarded as an ordered set for this experiment rather than as broadly separated levels of landscape complexity.

3.2. Internal validity-index outcomes

Figure 3 shows the four internal-index curves for the band-only representation in the upper row and the band-plus-index representation in the lower row. Values were rescaled only for display; all outcomes were determined from the raw scores, and the Gap uncertainty band represents the standard-error term used in its stopping rule. Silhouette was largest at two clusters in all six window-by-feature combinations and did not exceed that value at any larger cluster number. It declined from 0.467 to 0.237 in window A with the band-only representation, from 0.510 to 0.274 in window B, and from 0.525 to 0.269 in window C. Calinski-Harabasz was also largest at two clusters in five combinations; the exception was window C with bands and indices, where a shallow maximum occurred at five clusters (7,637 compared with 7,034 at two clusters). Davies-Bouldin reached its minimum at two clusters in four combinations, at five clusters in window A with bands and indices, and at four clusters in window C with bands and indices.

The Gap statistic increased throughout the evaluated range in all six combinations. For example, it increased from 2.28 at two clusters to 3.08 at twelve clusters in window A with the band-only representation. The one-standard-error rule was not satisfied in any combination; consequently, the gap statistic produced no stopping point between 2 and 12 clusters. The twelve-cluster result is therefore reported only as the censored endpoint of the evaluated range, not as a selected optimum.

Silhouette and the censored Gap endpoint occupied opposite boundaries of the evaluated range in every combination. Because this boundary pattern occurred in all windows, the magnitude of disagreement among criteria could not be related to the image-texture ordering. Among the three criteria that returned a cluster number, all agreed on two clusters in four combinations, two agreed in window A with bands and indices, and all three differed in window C with bands and indices.

Adding the four spectral indices changed 3 of the 12 window-by-index outcomes, each toward a larger cluster number: Davies-Bouldin changed from two to five clusters in window A, while Calinski-Harabasz and Davies-Bouldin changed from two to five and from two to four clusters, respectively, in window C. The feature representation also changed the V-measure optimum from 11 to 7 clusters in window A and from 4 to 6 clusters in window C. Maximum V-measure increased from 0.213 to 0.261 in window A and from 0.512 to 0.539 in window C, whereas it changed from 0.590 to 0.570 in window B.

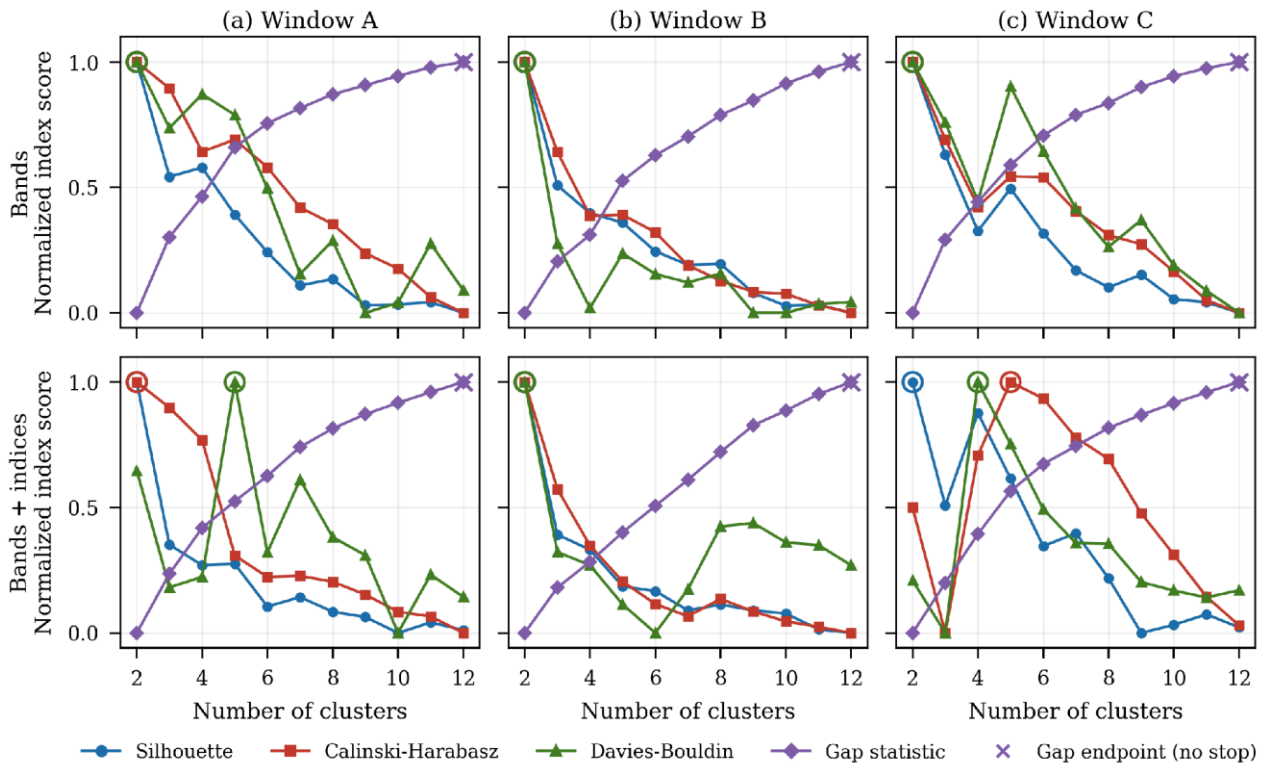


Figure 3. Internal validity-index curves for 2-12 clusters.

3.3. External agreement and robustness

Figure 4 presents V-measure with its 95% bootstrap interval on the left axis and median pairwise adjusted Rand index from the subsampling fits on the right axis. The V-measure optimum occurred at 11 clusters in window A with bands only (0.213), 7 clusters in window A with bands and indices (0.261), 2 clusters in window B with both representations (0.590 and 0.570), 4 clusters in window C with bands only (0.512), and 6 clusters in window C with bands and indices (0.539). The four-cluster partition in window C with bands only gave the largest moderate agreement observed in the study, with homogeneity of 0.509, completeness of 0.515, and an adjusted Rand index of 0.511.

At window B the two-cluster partition gave the highest agreement of the eleven evaluated, and agreement fell to about 0.40 for every larger k . The homogeneity-completeness decomposition accompanies this: at $k = 2$ homogeneity was 0.463 and completeness 0.812, while at $k = 12$ homogeneity had risen to 0.594 and completeness had fallen to 0.269. As additional spectral clusters split observations belonging to the same broad thematic class, completeness tended to decrease faster than homogeneity increased. The pattern is specific to this four-class aggregation of the LUCAS legend.

Agreement was low at window A under both feature sets (V of at most 0.261, adjusted Rand index of at most 0.078), and the V-measure surface there is nearly flat above $k = 3$. Both are consistent with the class imbalance of that window, in which one class accounts for 80.6% of the polygons and two of the four classes are represented by six or fewer. The window was fixed before the analysis and is retained, but the differences between candidate k values there are small relative to their confidence intervals and are interpreted accordingly.

To represent uncertainty in the observed V-measure maximum, the smallest cluster number whose paired bootstrap difference from that maximum had a 95% interval including zero was also recorded. It equalled the observed maximum in windows B and C. In window A, three clusters with bands only and six clusters with bands and indices showed no detectable difference from the nominal maxima at 11 and 7 clusters, respectively. Because no equivalence test was performed, these results indicate insufficient evidence of a difference rather than equivalence.

The adjusted Rand index was used as a robustness measure (Table 5). Its maximum occurred at the same cluster number as the V-measure maximum in five of six combinations. The exception was window A with bands only, where adjusted Rand index was largest at four clusters (0.070) rather than at the eleven-cluster V-measure optimum (0.024), consistent with the flat thematic-agreement profile in that window. Normalized mutual information with arithmetic normalization is mathematically identical to V-measure and was therefore not reported as a separate result.

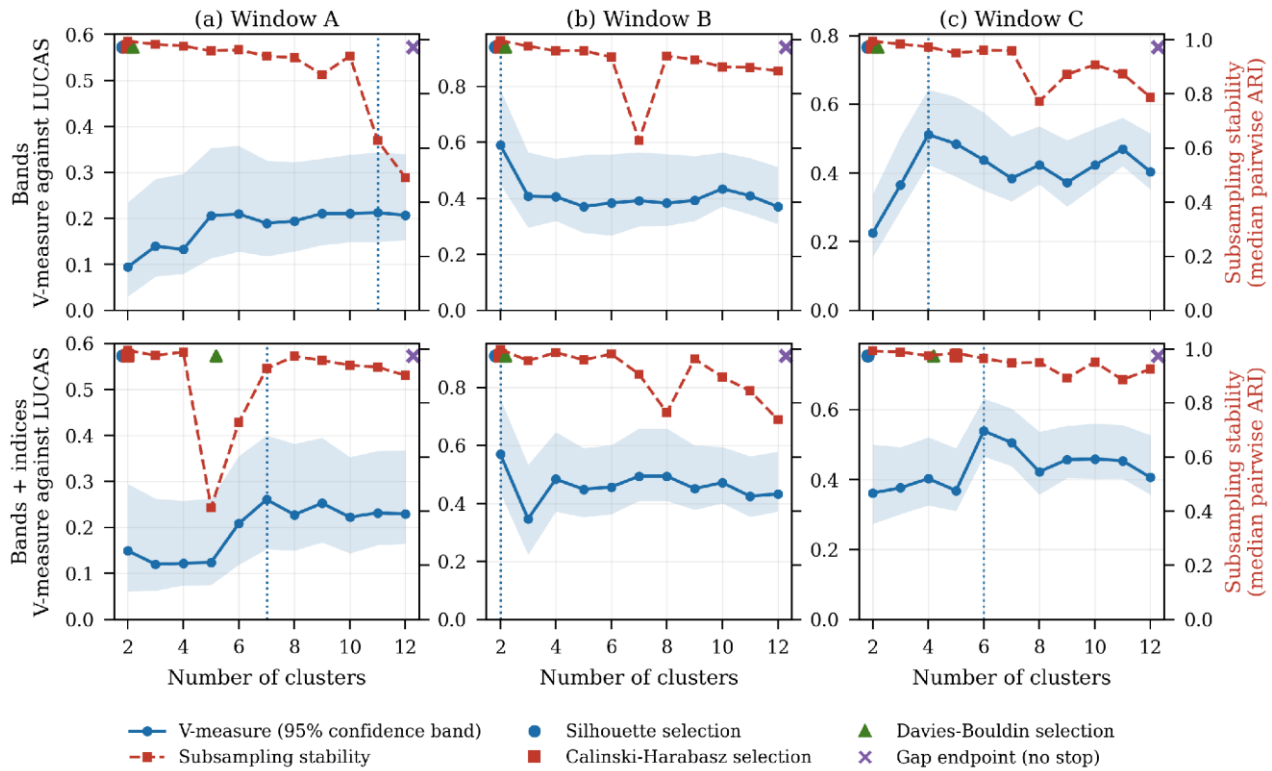


Figure 4. V-measure and subsampling stability across candidate cluster numbers.

Table 5. Cluster numbers maximizing V-measure and adjusted Rand index.

Window	Feature representation	V-measure optimum k	Maximum V-measure	ARI optimum k	Maximum ARI	Same optimum
A	Bands	11	0.213	4	0.070	No
A	Bands + indices	7	0.261	7	0.078	Yes
B	Bands	2	0.590	2	0.654	Yes
B	Bands + indices	2	0.570	2	0.617	Yes
C	Bands	4	0.512	4	0.511	Yes
C	Bands + indices	6	0.539	6	0.440	Yes

Table 6 reports the internal-index outcomes, agreement losses, V-measure optima, and stability values; Figure 6a displays the agreement losses for the three criteria that returned a cluster number. Median loss was 0.115 V-measure units for Silhouette and Calinski-Harabasz and 0.127 for Davies-Bouldin. Variation among windows was larger than variation among these criteria. All three criteria matched the V-measure optimum in window B under both feature representations. In window C with bands only, each selected two clusters and had an agreement loss of 0.286 (95% CI 0.185-0.407), the largest observed loss. In window A, the loss was 0.118 (0.031-0.226) with bands only; with bands and indices it was 0.111 (0.045-0.202) for Silhouette and Calinski-Harabasz and 0.137 (0.035-0.212) for Davies-Bouldin.

Table 6. Internal-index outcomes and agreement loss.

Window	Feature representation	Silhouette k (loss)	Calinski-Harabasz k (loss)	Davies-Bouldin k (loss)	Gap result	V-measure optimum k	Maximum V-measure [95% CI]	Smallest non-different k	Stability at optimum
A	Bands	2 (0.118)	2 (0.118)	2 (0.118)	NS (0.006)	11	0.213 [0.149, 0.344]	3	0.628
A	Bands + indices	2 (0.111)	2 (0.111)	5 (0.137)	NS (0.031)	7	0.261 [0.153, 0.400]	6	0.927
B	Bands	2 (0.000)	2 (0.000)	2 (0.000)	NS (0.220)	2	0.590 [0.451, 0.788]	2	0.997
B	Bands + indices	2 (0.000)	2 (0.000)	2 (0.000)	NS (0.137)	2	0.570 [0.437, 0.762]	2	0.999
C	Bands	2 (0.286)	2 (0.286)	2 (0.286)	NS (0.108)	4	0.512 [0.425, 0.642]	4	0.974
C	Bands + indices	2 (0.178)	5 (0.171)	4 (0.136)	NS (0.133)	6	0.539 [0.466, 0.632]	6	0.964

Note. Loss is the difference from the maximum V-measure. NS denotes no Gap stopping point; parenthesized Gap values are endpoint differences. Stability is the median pairwise adjusted Rand index across 20 subsampling fits.

The gap statistic produced no stopping point, so agreement loss is not defined for it. The endpoint differences were 0.006 (-0.037 to 0.024) and 0.031 (-0.049 to 0.087) in window A, 0.220 (0.053-0.363) and 0.137 (-0.044 to



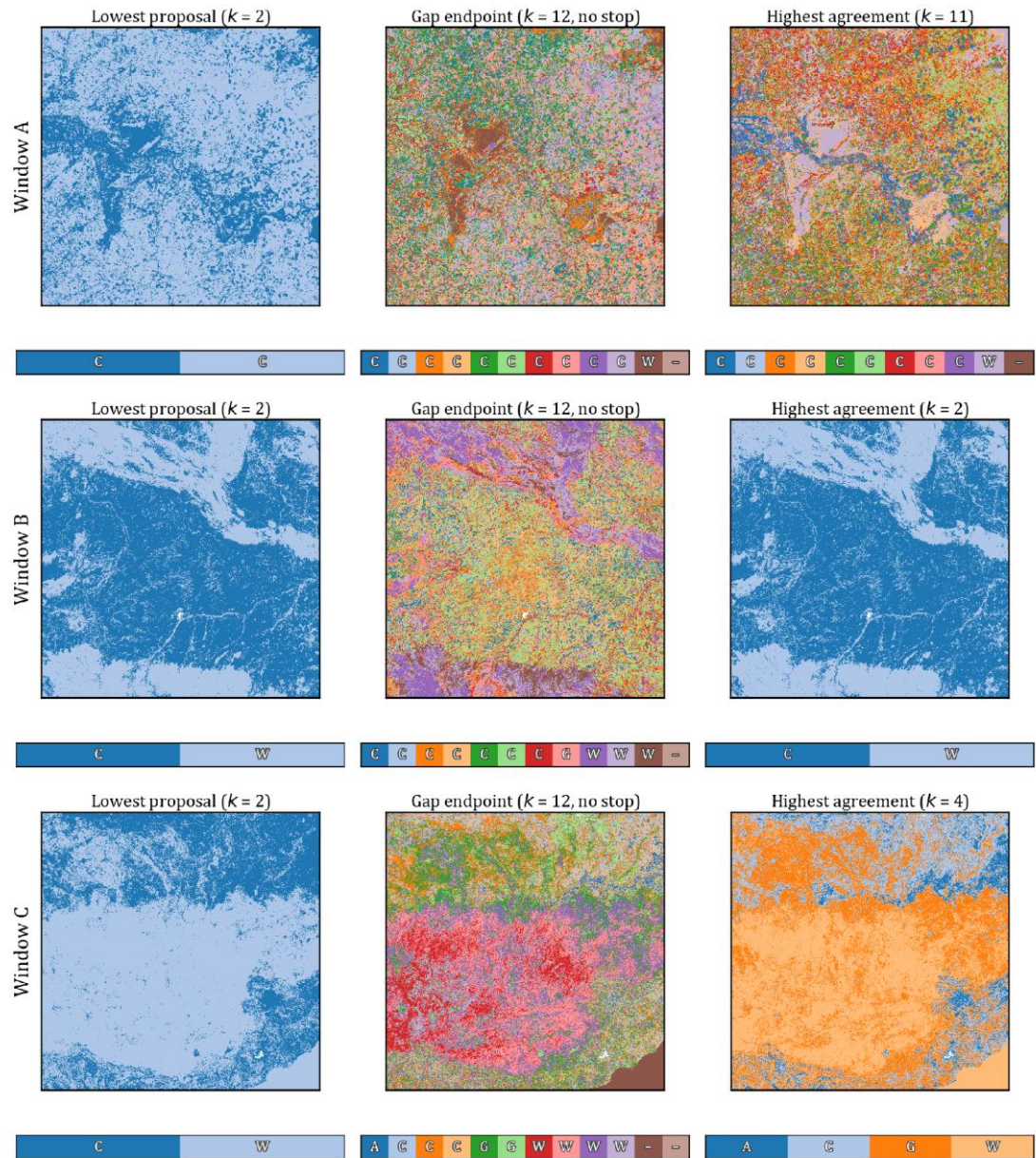
0.296) in window B, and 0.108 (0.014-0.193) and 0.133 (0.032-0.184) in window C for the band-only and band-plus-index representations, respectively. The small endpoint differences in window A reflect its nearly flat agreement profile rather than a cluster-number decision by the gap statistic.

No index coincided with the agreement maximum across all six combinations, and the position of the three compactness-and-separation criteria relative to that maximum changed between windows: a criterion that matched it at window B differed from it by more than a quarter of a V-measure unit at window C, with an interval excluding zero. A median computed across windows conceals this variation.

3.4. Stability and spatial continuity

The median adjusted Rand index across the 190 pairs of subsampling fits was high for many candidates but did not vary monotonically with cluster number. A pronounced local minimum occurred at five clusters in window A with bands and indices (0.413). Other decreases occurred at seven clusters in window B with bands only (0.627) and at 11 and 12 clusters in window A with bands only (0.628 and 0.491). Two low-stability solutions coincided with internal-index outcomes: the Davies-Bouldin minimum at five clusters in window A with bands and indices and the censored Gap endpoint in window A with bands only. Stability therefore supplied information not contained in the internal validity scores.

The chance-corrected eight-neighbour coefficient generally decreased as cluster number increased, but this trend partly reflects the structural effect of dividing a scene into more labels. The coefficient was therefore compared only between feature representations at the same cluster number. Adding spectral indices produced a larger coefficient in 27 of 33 matched comparisons. The median difference was 0.022, with values ranging from -0.021 to 0.054. In Figure 5, the columns show the smallest selected solution, the censored twelve-cluster Gap endpoint, and the V-measure optimum; distinct colours denote spectral clusters and the lower strips indicate their dominant broad LUCAS classes. Figure 6b summarises the matched-cluster differences between feature representations.



Each colour is one spectral cluster; letters give the dominant LUCAS class of that cluster (A artificial and bare, C cropland, G grassland and shrubland, W woodland).

Figure 5. K-means cluster maps for the reflectance-band feature representation.

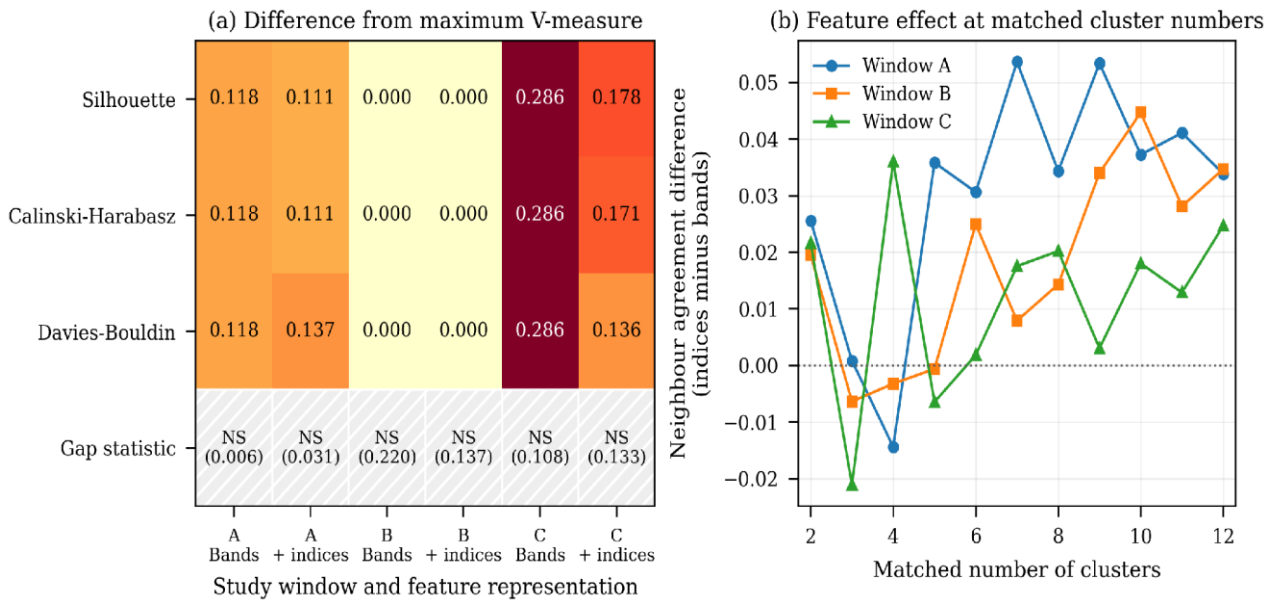


Figure 6. Agreement loss and matched-cluster spatial-continuity differences.

4. Discussion

The four criteria did not converge on a common cluster number in any window-by-feature combination. Silhouette selected the smallest candidate throughout, Calinski-Harabasz did so in five combinations, Davies-Bouldin produced interior minima in two, and the gap statistic reached the upper boundary without satisfying its stopping rule. These curve shapes are consistent with weakly separated or continuous structure in the reflectance feature space. Under such conditions, compactness-and-separation criteria may favour a coarse division, while the coordinate-wise uniform reference distribution used for the gap statistic may be poorly matched to correlated spectral variables. The results are consistent with evidence that internal indices can prefer partitions reflecting their own geometric assumptions rather than an external reference structure [20].

The two-cluster recommendation was not uniformly inappropriate. In window B it coincided with the largest thematic agreement in the study (V-measure 0.590; adjusted Rand index 0.654), whereas in window C with bands only it had an agreement loss of 0.286. The same recommendation therefore had different thematic consequences among windows, and these window-specific differences are more informative than a single median loss.

The V-measure optimum varied from 2 to 11 clusters. In window B, increasing the cluster number reduced thematic agreement because completeness declined more rapidly than homogeneity increased. This pattern depends on the relationship between spectral partitions and the four-class LUCAS aggregation: a different thematic resolution could shift the cluster number that maximises agreement. Spectral clusters and thematic classes should therefore not be treated as equivalent entities.

Previous comparative studies reached compatible conclusions from different experimental settings. Li et al. [19] found that no single internal validity index performed best across nine multispectral and hyperspectral data sets, while Arbelaitz et al. [13] reported similar variability in a broader benchmark. In the present Sentinel-2 composites, only Calinski-Harabasz and Davies-Bouldin produced interior optima, in one and two combinations respectively; Silhouette remained at the lower boundary and the Gap rule did not stop within the evaluated range.

Stability under repeated 80% pixel subsampling required no reference labels and was computationally feasible for the present experiment. Stability was high for many candidates and consequently weakly discriminating over much of the range, but its local minimum at five clusters in window A with bands and indices shows that stability was not determined by cluster number alone. It identified both the Davies-Bouldin outcome at that point and the censored Gap endpoint in window A with bands only as comparatively unstable. The measure describes recurrence under random subsampling of the same pixel pool; because spatial blocks were not resampled, it should not be interpreted as reproducibility under broader spatial perturbations.

The chance-corrected eight-neighbour coefficient describes local agreement of mapped cluster labels. Because part of its decrease with increasing cluster number is structural, its interpretation was restricted to matched-cluster comparisons. The band-plus-index representation produced greater neighbour agreement in 27 of 33 comparisons and a median increase of 0.022, although six negative differences show that the effect was not universal. The coefficient is a descriptive spatial diagnostic rather than a validated measure of map quality.

The conclusions are limited to three study windows, one growing season, two feature representations, and one K-means configuration. Two windows are located in Spain, the image-texture ordering spans a narrow range, and window A contains a strongly imbalanced reference sample. The candidate range ended at 12 clusters; the gap statistic did not stop within this range, and its behaviour beyond the endpoint was not evaluated. A reference



distribution aligned with principal components could behave differently from the coordinate-wise uniform distribution used here. The LUCAS legend was aggregated to four classes, polygon modes omit within-polygon variation, and polygon bootstrap samples do not represent spatial dependence among reference locations. No sensitivity analysis was performed for the 50,000-pixel training sample. Finally, a single seasonal median composite does not contain the phenological information available from multitemporal features.

5. Conclusion

The four internal validity indices produced different cluster-number outcomes across the six window-by-feature combinations. Silhouette selected two clusters throughout; Calinski-Harabasz selected two clusters in five combinations; Davies-Bouldin produced interior minima at four or five clusters in two combinations; and the gap statistic produced no stopping point between 2 and 12 clusters.

The cluster number that maximised agreement with the aggregated LUCAS legend varied among 2, 4, 6, 7, and 11. No internal criterion matched this optimum in every combination. The compactness-and-separation criteria had similar median agreement losses of 0.115-0.127, but their window-specific losses ranged from zero to 0.286. Subsampling stability also revealed local instability in two index outcomes that was not apparent from the internal validity scores.

For the examined Sentinel-2 scenes and K-means configuration, internal validity alone did not consistently identify the partition with the largest thematic agreement. Cluster-number selection should therefore consider the full index curve and, when available, external thematic agreement and stability under data perturbation. Generalisation to other sensors, seasons, thematic legends, or clustering algorithms requires further evaluation.

Acknowledgements

The authors thank the Copernicus programme and the European Commission Joint Research Centre for making the Sentinel-2 and LUCAS Copernicus 2018 data openly available.

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Conflicts of Interest

The authors report no conflicts of interest.

Data Availability Statement

The input data are openly available. Harmonized Sentinel-2 Level-2A surface reflectance (COPERNICUS/S2_SR_HARMONIZED) and the LUCAS Copernicus 2018 polygon collection (JRC/LUCAS_HARMO/COPERNICUS_POLYGONS/V1/2018) were accessed through the Google Earth Engine data catalogue. The derived materials and analysis code are available from the corresponding author upon reasonable request.

Institutional Review Board Statement

Not applicable.

CRediT Author Statement

Abdüsselam Kesikoğlu: Conceptualization; Methodology; Software; Data curation; Formal analysis; Visualization; Writing – original draft; Writing – review & editing. Ahmet Emin Karkinli: Conceptualization; Methodology; Validation; Formal analysis; Visualization; Writing – original draft; Writing – review & editing; Supervision. Both authors have read and approved the final manuscript.

References

- [1] M. Drusch et al., "Sentinel-2: ESA's optical high-resolution mission for GMES operational services," *Remote Sens. Environ.*, vol. 120, pp. 25–36, May 2012, doi: 10.1016/j.rse.2011.11.026.
- [2] N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore, "Google Earth Engine: Planetary-scale geospatial analysis for everyone," *Remote Sens. Environ.*, vol. 202, pp. 18–27, Dec. 2017, doi: 10.1016/j.rse.2017.06.031.
- [3] D. Phiri and J. Morgenroth, "Developments in Landsat land cover classification methods: A review," *Remote Sens.*, vol. 9, no. 9, p. 967, Sep. 2017, doi: 10.3390/rs9090967.
- [4] T. Blaschke, "Object based image analysis for remote sensing," *ISPRS J. Photogramm. Remote Sens.*, vol. 65, no. 1, pp. 2–16, Jan. 2010, doi: 10.1016/j.isprsjprs.2009.06.004.
- [5] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognit. Lett.*, vol. 31, no. 8, pp. 651–666, Jun. 2010, doi: 10.1016/j.patrec.2009.09.011.



- [6] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Math. Statist. Probab.*, vol. 1, Berkeley, CA, USA, 1967, pp. 281–297.
- [7] S. P. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. 28, no. 2, pp. 129–137, Mar. 1982, doi: 10.1109/TIT.1982.1056489.
- [8] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data," *Inf. Sci.*, vol. 622, pp. 178–210, Apr. 2023, doi: 10.1016/j.ins.2022.11.139.
- [9] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," in *Proc. 18th Annu. ACM-SIAM Symp. Discrete Algorithms (SODA)*, New Orleans, LA, USA, 2007, pp. 1027–1035.
- [10] M. E. Celebi, H. A. Kingravi, and P. A. Vela, "A comparative study of efficient initialization methods for the k-means clustering algorithm," *Expert Syst. Appl.*, vol. 40, no. 1, pp. 200–210, Jan. 2013, doi: 10.1016/j.eswa.2012.07.021.
- [11] G. W. Milligan and M. C. Cooper, "An examination of procedures for determining the number of clusters in a data set," *Psychometrika*, vol. 50, no. 2, pp. 159–179, Jun. 1985, doi: 10.1007/BF02294245.
- [12] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "On clustering validation techniques," *J. Intell. Inf. Syst.*, vol. 17, no. 2–3, pp. 107–145, Dec. 2001, doi: 10.1023/A:1012801612483.
- [13] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J. M. Pérez, and I. Perona, "An extensive comparative study of cluster validity indices," *Pattern Recognit.*, vol. 46, no. 1, pp. 243–256, Jan. 2013, doi: 10.1016/j.patcog.2012.07.021.
- [14] A. M. Ikotun, F. Habyarimana, and A. E. Ezugwu, "Cluster validity indices for automatic clustering: A comprehensive review," *Heliyon*, vol. 11, no. 2, p. e41953, Jan. 2025, doi: 10.1016/j.heliyon.2025.e41953.
- [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, Nov. 1987, doi: 10.1016/0377-0427(87)90125-7.
- [16] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Commun. Statist. Theory Methods*, vol. 3, no. 1, pp. 1–27, 1974, doi: 10.1080/03610927408827101.
- [17] 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, Apr. 1979, doi: 10.1109/TPAMI.1979.4766909.
- [18] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 63, no. 2, pp. 411–423, May 2001, doi: 10.1111/1467-9868.00293.
- [19] H. Li, S. Zhang, X. Ding, C. Zhang, and P. Dale, "Performance evaluation of cluster validity indices (PECVI) on multi/hyperspectral remote sensing datasets," *Remote Sens.*, vol. 8, no. 4, p. 295, Apr. 2016, doi: 10.3390/rs8040295.
- [20] M. Gagolewski, M. Bartoszek, and A. Cena, "Are cluster validity measures (in)valid?," *Inf. Sci.*, vol. 581, pp. 620–636, Dec. 2021, doi: 10.1016/j.ins.2021.10.004.
- [21] U. von Luxburg, "Clustering stability: An overview," *Found. Trends Mach. Learn.*, vol. 2, no. 3, pp. 235–274, 2010, doi: 10.1561/22000000008.
- [22] A. Ben-Hur, A. Elisseeff, and I. Guyon, "A stability based method for discovering structure in clustered data," in *Proc. Pacific Symp. Biocomputing*, vol. 7, Lihue, HI, USA, 2002, pp. 6–17.
- [23] T. Lange, V. Roth, M. L. Braun, and J. M. Buhmann, "Stability-based validation of clustering solutions," *Neural Comput.*, vol. 16, no. 6, pp. 1299–1323, Jun. 2004, doi: 10.1162/089976604773717621.
- [24] C. Hennig, "Cluster-wise assessment of cluster stability," *Comput. Statist. Data Anal.*, vol. 52, no. 1, pp. 258–271, Sep. 2007, doi: 10.1016/j.csda.2006.11.025.
- [25] R. d'Andrimont et al., "LUCAS Copernicus 2018: Earth-observation-relevant in situ data on land cover and use throughout the European Union," *Earth Syst. Sci. Data*, vol. 13, no. 3, pp. 1119–1133, Mar. 2021, doi: 10.5194/essd-13-1119-2021.
- [26] R. d'Andrimont et al., "Harmonised LUCAS in-situ land cover and use database for field surveys from 2006 to 2018 in the European Union," *Sci. Data*, vol. 7, p. 352, Oct. 2020, doi: 10.1038/s41597-020-00675-z.
- [27] G. M. Foody, "Status of land cover classification accuracy assessment," *Remote Sens. Environ.*, vol. 80, no. 1, pp. 185–201, Apr. 2002, doi: 10.1016/S0034-4257(01)00295-4.
- [28] P. Olofsson, G. M. Foody, M. Herold, S. V. Stehman, C. E. Woodcock, and M. A. Wulder, "Good practices for estimating area and assessing accuracy of land change," *Remote Sens. Environ.*, vol. 148, pp. 42–57, May 2014, doi: 10.1016/j.rse.2014.02.015.
- [29] M. Main-Knorn, B. Pflug, J. Louis, V. Debaecker, U. Müller-Wilm, and F. Gascon, "Sen2Cor for Sentinel-2," in *Proc. SPIE 10427, Image Signal Process. Remote Sens. XXIII*, Warsaw, Poland, 2017, Art. no. 1042704, doi: 10.1117/12.2278218.
- [30] J. W. Rouse, R. H. Haas, J. A. Schell, and D. W. Deering, "Monitoring vegetation systems in the Great Plains with ERTS," in *Proc. 3rd ERTS Symp.*, NASA SP-351, Washington, DC, USA, 1974, vol. 1, pp. 309–317.
- [31] H. Xu, "Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery," *Int. J. Remote Sens.*, vol. 27, no. 14, pp. 3025–3033, Jul. 2006, doi: 10.1080/01431160600589179.
- [32] C. H. Key and N. C. Benson, "Landscape assessment (LA): Sampling and analysis methods," *USDA Forest Service, Rocky Mountain Res. Station, Fort Collins, CO, USA, Gen. Tech. Rep. RMRS-GTR-164-CD*, 2006, pp. LA-1–LA-55.
- [33] A. Rikimaru, P. S. Roy, and S. Miyatake, "Tropical forest cover density mapping," *Trop. Ecol.*, vol. 43, no. 1, pp. 39–47, 2002.
- [34] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Nov. 2011.
- [35] A. Rosenberg and J. Hirschberg, "V-measure: A conditional entropy-based external cluster evaluation measure," in *Proc. Joint Conf. Empirical Methods Natural Lang. Process. Comput. Natural Lang. Learn. (EMNLP-CoNLL)*, Prague, Czech Republic, 2007, pp. 410–420.
- [36] L. Hubert and P. Arabie, "Comparing partitions," *J. Classification*, vol. 2, no. 1, pp. 193–218, Dec. 1985, doi: 10.1007/BF01908075.
- [37] J. Cohen, "A coefficient of agreement for nominal scales," *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, Apr. 1960, doi: 10.1177/001316446002000104.
- [38] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. New York, NY, USA: Chapman & Hall, 1993.