Accuracy Metrics for Explainable Unsupervised Methods in Remote Sensing: A Scorecard for Validity and Explanation Quality

Arnab Bhowmik
German Aerospace Center (DLR)
Oberpfaffenhofen, Germany
arnab.bhowmik@dlr.de

Chandrabali Karmakar
German Aerospace Center (DLR)
Oberpfaffenhofen, Germany
chandrabali.karmakar@dlr.de

Corneliu Octavian Dumitru
German Aerospace Center (DLR)
Oberpfaffenhofen, Germany
corneliu.dumitru@dlr.de

Mihai Datcu
CEOSpaceTech
University POLITEHNICA of Bucharest (UPB)
mihai.datcu@upb.ro

***Abstract*—Unsupervised learning is widely used in remote sensing when labels are sparse, delayed, or heterogeneous across regions and seasons. Its operational use is limited when cluster outputs are not accompanied by transparent quantitative reporting. We propose a compact scorecard for explainable unsupervised pipelines that separates cluster validity, explanation quality, and seed-based reproducibility in no-label settings. We focus on explainable k-means (XKMeans) and explainable Gaussian mixture models (XGMM) and link each scorecard quantity to concrete explanation artifacts: model-tied cluster feature sets and surrogate-based summaries. Validity is quantified through silhouette and, when benchmark labels exist, purity and normalized mutual information. Explanation quality is evaluated through fidelity, coverage, sparsity, and stability. We demonstrate the scorecard on a Sentinel-2 wildfire scene without pixel-level ground truth. The experiment is intentionally a compact proof-of-concept in a low-dimensional feature space, while broader validation on higher-dimensional and multimodal remote-sensing settings is a natural next step.**

***Index Terms*—Explainable AI, unsupervised learning, clustering, evaluation metrics, remote sensing, interpretability.**

I. Introduction

Remote sensing workflows frequently rely on unsupervised learning for retrieval, mapping, monitoring, and exploratory analysis. This is common in situations where labels are limited, arrive late in time, or vary across regions and seasons due to changes in land cover, illumination, sensor geometry, and acquisition conditions. A clustering model can reveal structure, yet for analyst-facing and decision-facing settings *a partition alone is not sufficient*: users require quantitative reporting that clarifies (i) whether the structure is coherent, (ii) what features explain the discovered groups, and (iii) whether these explanations remain consistent under small changes such as random initialization. Explainable machine learning is therefore increasingly treated as a practical requirement in geo- and biosciences because it connects model outputs to domain reasoning and supports auditability beyond a single performance number [1]. In unsupervised settings, the reporting challenge is sharper: classical cluster validity indices measure separation and compactness, but they do not measure whether an explanation is faithful to the assignments or stable across perturbations. Conversely, an explanation can be compact and visually attractive while not reflecting the model assignments in a reliable way.

A. Problem statement

A complete evaluation for explainable unsupervised remote sensing should make three aspects explicit:

- **Validity:** Is the discovered structure coherent in the chosen feature space and, when benchmark labels exist, aligned with reference classes?
- **Explanation quality:** Do the explanations match the assignments (fidelity), exist for the evaluated samples (coverage), remain concise (sparsity), and remain consistent across re-initializations (stability)?
- **No-label reproducibility:** When ground truth is absent, does the partition remain similar across seeds, indicating sensitivity or robustness to initialization?

B. Contributions

We provide a compact reporting scorecard that separates validity from explanation quality and includes seed-based reproducibility for no-label scenes. We define operational metrics that match the mechanisms of explainable clustering—model-tied cluster feature sets and surrogate-based summaries—and we demonstrate reporting on a Sentinel-2 wildfire scene where pixel-level ground truth is not available.

II. Related Work

Explainable machine learning for remote sensing has been positioned as a bridge between model outputs and domain

reasoning, with a strong emphasis on evaluation and transparency beyond standard accuracy-style reporting [1]. Explainable clustering methods connect partitions to concise descriptions such as rules, prototypes, or feature subsets [4], [5]. For unsupervised models more broadly, label-free explanation objectives study which features drive discovered structure without requiring labels [6]. Robustness is central because interpretability artifacts can vary under small changes, motivating stability as a first-class axis of evaluation [11]. Within remote sensing, interactive and analyst-centered explainability has been explored for SAR and multi-modal image understanding, highlighting the need for methods whose explanation artifacts remain usable for inspection and comparison [2], [3].

III. Methodology

A. Unsupervised models and explanation artifacts

We consider two clustering families that are widely used in remote sensing feature spaces:

XKMeans. k-means produces hard assignments $\hat{c}_i \in \{1, \ldots, K\}$ and cluster centroids $\{\boldsymbol{\mu}_k\}_{k=1}^{K}$. Explainability can be supported by centroids because each centroid coordinate corresponds to an interpretable input feature (band, index, SAR statistic, texture descriptor). A cluster can therefore be described by the features in which its centroid differs most from the global mean.

a) Hierarchical / tree-style interpretability for XKMeans.: In addition to feature-based centroid descriptions, XKMeans can be presented in a tree-style form by ordering clusters using centroid distances and reporting successive "split" directions (dominant features separating parent/child groups). In this paper, we keep the reporting compact and operational by (i) describing each cluster via its top-m centroid-salient features (Eq. (2)) and (ii) using the surrogate decision tree to provide an explicit, human-readable partition summary. In this work, the tree representation is provided by the surrogate decision tree used for explaining assignments; we do not require a hierarchical clustering algorithm.

XGMM. A Gaussian mixture model (GMM) provides responsibilities r_{ik} and component means $\{\boldsymbol{\mu}_k\}$; hard labels can be obtained by $\hat{c}_i = \arg\max_k r_{ik}$. The EM algorithm is the standard estimation procedure [7]. As with k-means, component means admit feature-based descriptions in an interpretable feature space.

Model-tied explanation artifact (cluster feature sets). For either model, define the global mean vector $\boldsymbol{\mu}_{\cdot}$ in the standardized feature space (computed over valid samples). For each cluster k and feature j, define a salience score

$$\Delta_{k,j} = \left|\mu_{k,j} - \mu_{\cdot,j}\right|. \tag{1}$$

The top-m explanatory feature set for cluster k is

$$A_k = \text{Top-}m(\Delta_{k,1:d}), \tag{2}$$

where d is feature dimension and $m^\star = \min(m, d)$ is used in practice. This yields a transparent explanation artifact (feature subsets) that is directly tied to the clustering model.

Algorithm 1 Scorecard computation for explainable unsupervised RS.

Require: Features $\{\mathbf{x}_i\}$, method $\mathcal{M} \in \{\text{XKMeans,XGMM}\}$, clusters K, seeds $\mathcal{S}$, top-m, subset size n_{sil}
Ensure: Scorecard metrics $\mathcal{R}$

```
Standardize features and select fixed subset I of size n_sil
for all s ∈ S do
    Fit clustering (k-means or GMM) with seed s
    Compute silhouette on I
    if benchmark labels exist then
        Compute purity and NMI
    else
        Store labels for seed-consistency
    end if
    Extract top-m cluster features from centroids/means
    Train depth-3 surrogate tree
    Compute fidelity, coverage, and sparsity
end for
Align clusters across seeds and compute stability
Aggregate metrics as mean ± std
```

Surrogate explanation artifact. We also fit a compact surrogate classifier $h(\cdot)$ (depth-3 decision tree) to predict $\hat{c}_i$ from standardized features $\tilde{\mathbf{x}}_i$. The surrogate serves two roles: (i) it provides a fidelity score for how well a compact explanation matches the assignments, and (ii) it provides a sparsity measure via the number of unique features used in the tree (or mean rule length).

B. Scorecard computation across seeds

We evaluate stability and reproducibility by running each method across several random seeds. Algorithm 1 summarizes the full protocol, including: internal validity, optional external agreement, surrogate-based fidelity and sparsity, model-tied feature-set extraction, cluster alignment, and stability computation.

IV. Metric Suite (Validity vs. Explanation Quality)

The scorecard reports validity and explanation quality side-by-side. External agreement is used only when benchmark labels exist.

A. Validity

Given assignments C and reference classes G, purity is

$$\text{Purity}(C, G) = \frac{1}{N} \sum_k \max_g |C_k \cap G_g|. \tag{3}$$

Normalized mutual information (NMI) is

$$\text{NMI}(C, G) = \frac{2I(C; G)}{H(C) + H(G)}, \tag{4}$$

following the information-theoretic normalization used for clustering comparison [9]. We always report internal structure via the silhouette coefficient [8]

$$\text{Silhouette} = \frac{1}{N} \sum_i \frac{b(i) - a(i)}{\max(a(i), b(i))}, \tag{5}$$

TABLE I
SCORECARD METRICS (VALIDITY VS. EXPLANATION QUALITY).

Metric	Needs G?	Type	Captures
Purity	Yes	External	Class alignment
NMI [9]	Yes	External	Information agreement
Silhouette [8]	No	Internal	Separation/compactness
NMI consistency	No	Reproduc.	Seed agreement
ARI consistency [12]	No	Reproduc.	Chance-corrected
Fidelity [6], [10]	No	Explain.	Surrogate agreement
Coverage	No	Explain.	Explanation availability
Sparsity	No	Explain.	Explanation conciseness
Stability [11]	No	Explain.	Feature-set consistency

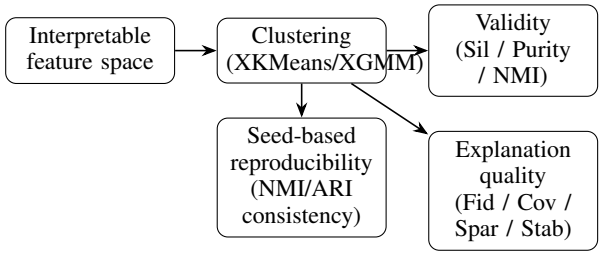

Fig. 1. Scorecard structure: validity, explanation quality, and seed-based reproducibility are reported together.

where $a(i)$ is the mean intra-cluster distance and $b(i)$ is the smallest mean distance to other clusters.

B. No-label reproducibility (seed-consistency)

When ground truth is absent, we report reproducibility across random seeds via pairwise agreement between runs on the same evaluation subset. We use (i) NMI-consistency and (ii) ARI-consistency, where ARI follows the corrected-for-chance definition [12]. These quantities indicate whether the partition is sensitive to initialization in the chosen feature space.

C. Explanation quality

Fidelity. A compact surrogate $h(\cdot)$ is trained to predict unsupervised labels, and we report

$$\text{Fidelity} = \frac{1}{N}\sum_{i} \mathbb{1}[h(\tilde{\mathbf{x}}_i) = \hat{c}_i]\,, \tag{6}$$

as a proxy for agreement between the surrogate explanation and the clustering assignments [6], [10].

Coverage. Fraction of evaluated samples for which explanations are defined (valid features after masking):

$$\text{Coverage} = \frac{1}{|I|}\sum_{i\in I} \mathbb{1}[\tilde{\mathbf{x}}_i \text{ valid}]\,. \tag{7}$$

Sparsity. Conciseness of the surrogate explanation, reported as the number of unique features used by the depth-3 tree (or mean rule length).

Stability. Jaccard overlap of cluster-level top-m feature sets across seeds after cluster alignment (Algorithm 1), consistent with robustness considerations for interpretability methods [11].

V. Demonstration on a Sentinel-2 Wildfire Scene (No Ground Truth)

We demonstrate scorecard reporting on a Sentinel-2 wildfire scene over a Greece test area (L2A reflectance). Pixel-level ground truth is not available; therefore the goal is not to report accuracy, but to report reliability indicators: internal separability, reproducibility across seeds, and explanation quality.

A. Feature space and resolution handling

We use a compact feature space derived from Sentinel-2 reflectance: (i) NIR reflectance $B08$, (ii) SWIR reflectance $B12$, and (iii) Normalized Burn Ratio (NBR),

$$\text{NBR} = \frac{B08 - B12}{B08 + B12}. \tag{8}$$

To evaluate features on a common grid, $B12$ is mapped to the $B08$ grid via nearest-neighbor resampling. All features are standardized via z-score over valid pixels (after masking nodata and saturated pixels).

B. Protocol and reproducibility setup

We run XGMM with $K = 10$ components for 5 seeds. Silhouette and seed-consistency are computed on a fixed evaluation subset I of valid pixels with size

$$n_{\text{sil}} = 100{,}000, \tag{9}$$

kept fixed across seeds to make per-seed comparison meaningful. Seed-consistency is reported as pairwise NMI/ARI agreement across the 5 runs on the same subset.

C. Explanation reporting

Explanation quality is reported in two complementary forms: (i) model-tied cluster feature sets A_k obtained from component means (used for stability), and (ii) a depth-3 decision-tree surrogate (used for fidelity and sparsity). Coverage is 1.0 when all evaluated pixels have valid features. Because the feature space is compact (three features), high stability is expected when the dominant separating directions remain unchanged; in higher-dimensional feature spaces (e.g., multi-band plus textures), stability typically becomes more discriminative and highlights which explanatory features persist across runs.

D. Scorecard summary

Table II reports the scorecard for the wildfire scene. The results indicate strong cluster separability (high silhouette), high reproducibility across seeds (high NMI/ARI consistency), and compact, consistent explanations (high surrogate fidelity with low sparsity, and high top-m stability). A high silhouette is consistent with the chosen feature space: NIR/SWIR reflectance and NBR strongly separate burned vegetation, unburned vegetation, and background materials in many wildfire scenes, yielding stable mixture modes in the standardized space.

TABLE II
SCORECARD ON THE SENTINEL-2 WILDFIRE SCENE (MEAN ± STD ACROSS 5 SEEDS; $K = 10$, TOP-$m = 5$; SEED 3 REFERENCE).

Metric	Value
Silhouette ↑	0.826 ± 0.000
NMI consistency ↑	0.950 ± 0.025
ARI consistency ↑	0.983 ± 0.009
Fidelity ↑	0.970 ± 0.003
Coverage ↑	1.000 ± 0.000
Sparsity ↓ (unique features)	3.00 ± 0.00
Stability ↑ (Jaccard top-5)	1.000 ± 0.000

TABLE III
PER-SEED RESULTS ON THE WILDFIRE SCENE ($K = 10$, TOP-$m = 5$, $n_{\text{sil}} = 100{,}000$).

Seed	Sil	Fid	Spar	Stab
0	0.827	0.966	3	1.000
1	0.826	0.973	3	1.000
2	0.826	0.972	3	1.000
3	0.827	0.967	3	1.000
4	0.826	0.969	3	1.000

E. Scorecard plot (from Table II)

The scorecard plot provides a compact visual summary that is easy to compare across methods or scenes. We plot only unit-scaled metrics (0–1) to avoid mixing incomparable units; metrics where "lower is better" (e.g., sparsity) are reported in the table.

F. Per-seed breakdown

Table III reports per-seed internal validity and surrogate metrics. Stability is computed relative to the reference seed (seed 3) after mixture alignment via Hungarian matching [13]. Per-seed reporting is included to make sensitivity to initialization transparent in no-label settings. Consistent values across seeds indicate that both the partition and the explanation artifacts are reproducible under re-initialization, which is essential for operational use where analysts expect stable semantic summaries.

G. Qualitative outputs

a) How the probability map is computed and used.: For XGMM, the EM algorithm returns per-pixel responsibilities $r_k(p) = p(z = k \mid \tilde{\mathbf{x}}(p))$ for mixture component k given standardized features $\tilde{\mathbf{x}}(p)$. We visualize either (i) the maximum responsibility $\max_k r_k(p)$ as a confidence score for the hard label, or (ii) the responsibility of a semantically chosen component $r_{k^\star}(p)$ to inspect spatial extent and mixing. This view supports analyst verification by showing gradual transitions and ambiguous regions that are not visible in the hard label map. We use the term *certainty* in the sense of high posterior support, i.e., higher $r_{k^\star}(p)$ indicates stronger assignment to component $k^\star$. For visualization, we use a 3-level legend derived from the displayed score $s(p) \in [0, 1]$ (here: the selected-component responsibility). We map values to three bins: *high* (green) if $s(p) > 0.7$, *medium* (yellow) if $0.3 \le s(p) \le 0.7$, and *low* (red) if $s(p) < 0.3$. This discretization provides an immediate, analyst-friendly summary while the underlying map remains continuous in $[0, 1]$.

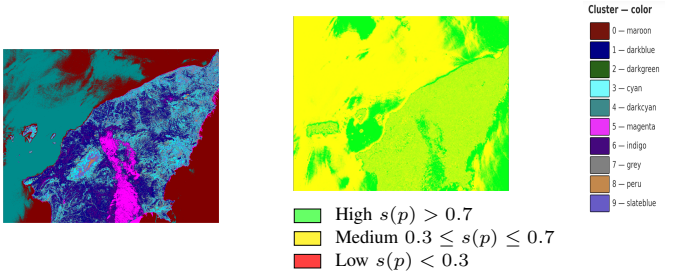

Fig. 2. Example outputs over a Greece test area (Sentinel-2). Left: XGMM hard cluster map $\hat{c}(p) = \arg\max_k r_k(p)$ (colors denote mixture components). Middle: responsibility/probability map for a selected component $k^\star$, showing $s(p) = r_{k^\star}(p) \in [0, 1]$. For quick interpretation we use a 3-level legend: green if $s(p) > 0.7$, yellow if $0.3 \le s(p) \le 0.7$, and red if $s(p) < 0.3$. Right: legend/color key for the displayed bins. All panels are rendered on the same $B08$ grid and identical spatial extent (no shift); any resampled feature (e.g., $B12$) is mapped to the $B08$ grid by nearest-neighbor to preserve alignment.

Fig. 2 includes an example cluster map and a component responsibility map; both should be produced on the same grid as the input features for consistent inspection.

VI. DISCUSSION

The scorecard makes analyst-relevant trade-offs explicit. A method may yield a coherent partition while producing unstable explanations, or stable explanations with weaker internal structure. Reporting validity and explanation quality together therefore supports method selection based on mapping consistency, explanation reliability, and seed robustness. The current validation should be interpreted with appropriate scope. It is intentionally simple: a single no-label wildfire scene and a compact three-feature representation. This is useful for showing that the scorecard is operational and reproducible, but it does not yet exhaust harder remote-sensing cases such as high-dimensional, noisier, weakly structured, or multimodal settings. In such cases, stability, sparsity, and surrogate fidelity may become even more discriminative. The scorecard is not itself another clustering algorithm. Rather, it is an evaluation layer that can accompany centroid-based, mixture-based, and other interpretable unsupervised pipelines, including remote-sensing-specific approaches such as latent-topic models and interactive interpretation frameworks.

VII. CONCLUSION

The scorecard links metrics to concrete explanation artifacts and demonstrates how reliability can be reported when accuracy metrics are not justified by missing ground truth. The Sentinel-2 wildfire experiment should be viewed as an operational proof-of-concept rather than a complete benchmark. Its role is to make the scorecard explicit, reproducible, and practical. A natural next step is broader validation on higher-dimensional, noisier, and multimodal remote-sensing settings.

REFERENCES

[1] R. Roscher, B. Bohn, M. F. Duarte, and J. Garcke, "Explain it to me – facing remote sensing challenges in the bio- and geosciences with explainable machine learning," *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, vol. V-3-2020, pp. 817–824, 2020, doi: 10.5194/isprs-annals-V-3-2020-817-2020.

[2] C. Karmakar, C. O. Dumitru, G. Schwarz, and M. Datcu, "Feature-free explainable data mining in SAR images using latent Dirichlet allocation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 676–689, 2021, doi: 10.1109/JSTARS.2020.3039012.

[3] C. Karmakar and M. Datcu, "A framework for interactive visual interpretation of remote sensing data," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022, doi: 10.1109/LGRS.2022.3161959.

[4] M. Moshkovitz, C. Rashtchian, and S. Dasgupta, "Explainable k-means and k-medians clustering," in *Proc. 37th Int. Conf. Machine Learning (ICML)*, PMLR 119, 2020, pp. 7055–7065.

[5] N. Frost, M. Moshkovitz, and C. Rashtchian, "ExKMC: Expanding explainable k-means clustering," *arXiv preprint arXiv:2006.02399*, 2020.

[6] J. Crabbé and M. van der Schaar, "Label-free explainability for unsupervised models," in *Proc. 39th Int. Conf. Machine Learning (ICML)*, PMLR 162, 2022, pp. 4391–4420.

[7] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Statist. Soc. Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[8] 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.

[9] 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.

[10] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?: explaining the predictions of any classifier," in *Proc. ACM SIGKDD*, 2016, pp. 1135–1144.

[11] D. Alvarez-Melis and T. S. Jaakkola, "On the robustness of interpretability methods," *arXiv preprint arXiv:1806.08049*, 2018.

[12] L. Hubert and P. Arabie, "Comparing partitions," *J. Classification*, vol. 2, pp. 193–218, 1985.

[13] H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, no. 1–2, pp. 83–97, 1955.

[14] P. Helber, B. Bischke, A. Dengel, and B. Borth, "EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 7, pp. 2217–2226, 2019, doi: 10.1109/JSTARS.2019.2918242.