POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Weighted Conformal Classification for Hyperspectral Mineral Mapping: An Application to the Cuprite AVIRIS Scene

Tesi di Laurea Magistrale in
Mathematical Engineering - Ingegneria Matematica

Marcello Bernardini, 10770570

Advisor:
Prof. Simone Vantini

Co-advisor:
Alfredo Gimenez Zapiola

Academic year:
2025-2026

Abstract: Mineral mapping methods (classification and spectral unmixing) typically lack reliability guarantees, especially when models trained on spectral libraries are applied to large remote scenes affected by mixing and acquisition artifacts. This thesis introduces conformal classification as a practical framework for uncertainty-aware hyperspectral mineral mapping, producing set-valued outputs that control the probability of including, given a random pixel, the true mineral label at a user-chosen level.
Using the Cuprite AVIRIS scene as a case study, we train a probabilistic Random Forest on the USGS spectral library and define the classification task through a data-driven mineral taxonomy.To mitigate distribution shift between library spectra and Cuprite pixels, we adopt weighted conformal classification based on importance reweighting from a low-dimensional density-ratio approximation. Since pixel-level target labels are unavailable at scene scale, we assess conformal outputs through operational summaries (set sizes, empty-set frequency, and spatial patterns).
Operationally, standard conformal classification produces a substantial fraction of empty prediction sets on Cuprite (22.9% of pixels at 95% nominal reliability), while weighted calibration eliminates empty sets at the same nominal level, yielding target-adapted prediction sets that trade decisiveness for robustness in spectrally ambiguous regions. Overall, the framework provides a practical route to uncertainty-aware hyperspectral mineral mapping under covariate shift.

Key-words: hyperspectral mineral mapping, Cuprite (AVIRIS), conformal prediction, weighted conformal classification, covariate shift

1. Introduction

Mineral identification is a core task in a wide range of scientific and industrial applications. In economic geology, it supports the discovery and mapping of strategic resources such as rare earth elements,

which are critical for modern technological development [Bedini, 2017]. Environmental monitoring relies on mineralogical information to manage mining waste and mitigate land degradation [Pereira et al., 2023]. More broadly, in fields ranging from geothermal studies to planetary science [Gavrishin et al., 1992; Thiele et al., 2025], mapping mineral composition over large geographical areas enables the characterization of remote or inaccessible environments where extensive ground-truth sampling (e.g., drilling campaigns) is logistically challenging or economically prohibitive.
In this context, hyperspectral imaging (HSI) has established itself as a key technology for large-scale mineral mapping. Airborne or spaceborne sensors acquire reflectance measurements over hundreds of contiguous wavelength channels, typically spanning the visible to shortwave-infrared range. The resulting hyperspectral data cube provides spatially dense information, where each pixel is described by a high-dimensional spectrum that can be viewed as a smooth spectral function observed on a discrete wavelength grid. These signatures arise from the physical interaction between electromagnetic radiation and surface materials, capturing diagnostic absorption features related to mineral chemistry and structure.
A visual description of remote sensing is shown in Figure 1. However, translating these signals into reliable thematic maps involves significant inferential challenges: compared to laboratory references, scene spectra often differ due to sub-pixel mixing, illumination variability, sensor noise, and residual atmospheric artifacts.

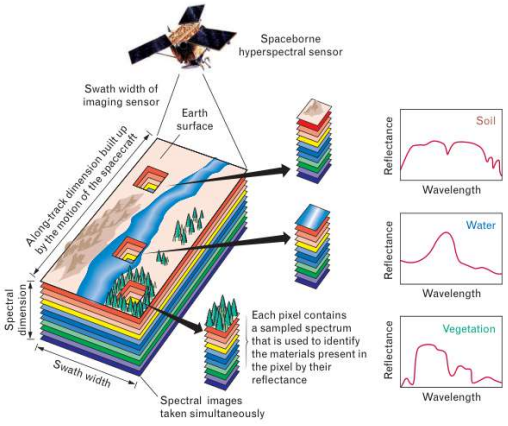


Figure 1: Acquisition scheme of a hyperspectral image: a remote sensor (airborne or satellite) captures spatial data across multiple wavelengths (the colored layers). Each pixel represents a sampled reflectance spectrum used to identify scene materials. The graphs depict the specific spectral variations for three different pixels: soil, water, and vegetation Shaw and Burke [2003].

1.1. Standard approaches : unmixing and classification

Two main approaches have traditionally been used for mineral mapping with HSI: spectral unmixing and supervised classification.

Spectral unmixing. Due to the limited spatial resolution of remote sensors, a single pixel may cover a physical region containing multiple materials. Spectral unmixing aims to decompose the observed spectrum into a set of pure spectral signatures (endmembers) and their fractional abundances [Bioucas-Dias et al., 2012]. A widely used formulation is the Linear Mixing Model (LMM), which assumes that each observed spectrum is a linear combination of endmembers:

$$y_i = Mx_i + \epsilon_i, \tag{1}$$

where M is the matrix of endmember signatures, x_i is the abundance vector, and ϵ_i accounts for noise and modeling error. Abundances are typically constrained to be non-negative and (sometimes) to sum to one, though the latter constraint is often relaxed to accommodate global intensity changes (e.g., shading or illumination effects). Unmixing is attractive because it yields physically interpretable

outputs, but its performance depends critically on the validity of modeling assumptions (e.g., linearity) and on the quality of endmember information. In practice, errors in endmember characterization and model mismatch can propagate to abundance maps and to any downstream mineral interpretation. Some extensions have been developed such as sparse unmixing and augmented unmixing in order to improve robustness by leveraging suitable spectral libraries and by accounting for spectral variability [Boselli, 2022; Hong et al., 2019; Iordache et al., 2011; Zinno, 2024].

Hyperspectral classification. Alongside unmixing, hyperspectral classification assigns a single mineral label to each pixel using supervised learning. This paradigm is natural when reference data provide a unique label per spectrum (as in many spectral libraries) and when the objective is thematic mapping rather than estimating fractional abundances. Classical approaches (e.g., SVM, Random Forest) and more recent learning-based methods often achieve strong empirical performance, but must cope with the Hughes phenomenon (high dimensionality with limited labeled samples) and the strong collinearity induced by contiguous bands. As a consequence, dimensionality reduction and shape-based representations (e.g., derivative spectra or functional features) are commonly employed [Bandos et al., 2009; Bazi and Melgani, 2006; Salem, 2015; Ye et al., 2020].

1.2. Limitations and the need for uncertainty quantification

Despite the sophistication of these pipelines, a critical gap remains: both unmixing and classification typically deliver point estimates—fractional abundances or single labels—without an explicit finite-sample guarantee on predictive reliability.

In spectral unmixing, abundance maps are physically meaningful estimates under a prescribed mixing model, but they are not interpretable as probabilities to find a specific mineral in the selected pixel. In practice, abundances are sometimes interpreted as confidence surrogates, yet their reliability depends on strong modeling assumptions (e.g., linearity, adequate endmembers, stable signatures) that may fail in real scenes.

In supervised classification, the output is a single predicted label, sometimes accompanied by a model-based probability score. However, such scores are not, by themselves, coverage statements: a predicted probability of 0.9 does not imply that predictions will be correct 90% of the time. Miscalibration is exacerbated when training/calibration data (often laboratory spectra) differ from the target scene distribution due to mixing and acquisition effects. As a result, operational use often still requires costly on-site verification to assess whether a "best guess" is trustworthy.

This lack of guarantees is particularly problematic in settings like planetary exploration or interventions in hazardous sites, where field validation (drilling, sampling campaigns) is logistically difficult and economically expensive. In these contexts, uncertainty quantification is not only diagnostic: it supports decision-making about which zones to prioritize for further analysis, reduces false positives, and helps allocate limited validation resources to locations where the algorithm signals genuine ambiguity.

These considerations motivate uncertainty-aware outputs that provide rigorous error control. Rather than returning a single label, the goal is to produce a prediction set of plausible minerals with a prescribed coverage level under the conformal assumptions, offering an explicit bound on the probability that the true mineral identity is included in the reported set.

1.3. Our perspective: conformal prediction under distribution shift

Conformal Classification To address this gap, we adopt conformal classification (CC), a model-agnostic framework that constructs set-valued predictions with finite-sample, distribution-free *marginal* coverage guarantees under exchangeability [Angelopoulos and Bates, 2021; Shafer and Vovk, 2008]. CC is built on top of a probabilistic classifier by converting class scores into nonconformity scores and calibrating a threshold on a held-out calibration set. The resulting prediction set contains the labels that are sufficiently compatible with the observed spectrum.

Formally, let $(X_1, Y_1), \ldots, (X_n, Y_n)$ and (X_{n+1}, Y_{n+1}) be exchangeable pairs. For a target miscoverage level $\alpha \in (0, 1)$, conformal prediction constructs a set-valued predictor $C(\cdot)$ such that

$$P\big(Y_{n+1} \in C(X_{n+1})\big) \;\geq\; 1 - \alpha, \tag{2}$$

without requiring a parametric model for the data-generating process. In mineral mapping, this yields an operational alternative to single-label classification: for each pixel, we output a set of mineral labels at a user-chosen risk level, which can be summarised through set sizes and spatial patterns. Depending on the extent of calibration mismatch, conformal predictors may yield an empty set, interpretable as an abstention, or, conversely, the full set of possible classes, indicating maximal uncertainty.

Weighted Conformal Classification A key challenge in our setting is distribution shift between laboratory spectra used for training/calibration and airborne scene spectra used as test data. Under shift, the exchangeability assumption underlying standard conformal split calibration may be violated, and nominal coverage may fail on the target scene.
To address this issue, we consider **weighted conformal classification** (WCC), which targets coverage under covariate shift by reweighting calibration examples to better represent the test distribution [Tibshirani et al., 2020]. In the covariate shift setting, we assume that the conditional distribution is stable while the covariate marginal changes. Concretely,

$$(X_i, Y_i) \overset{exch}{\sim} P := P_X \times P_{Y|X}, \quad i = 1, \ldots, n, \qquad (X_{n+1}, Y_{n+1}) \sim \widetilde{P} := \widetilde{P}_X \times P_{Y|X}, \tag{3}$$

where $\widetilde{P}_X \neq P_X$ but $P_{Y|X}$ is shared across source and target domains.
Weighted conformal classification uses importance weights $w(x) = \frac{d\widetilde{P}_X}{dP_X}(x)$ to reweight calibration examples so that the weighted empirical distribution of scores better reflects the target covariate distribution, aiming to maintain nominal coverage under covariate shift.

Contributions and Outline In light of these challenges, the main contributions of this thesis are threefold:

- We develop an end-to-end pipeline for library-to-scene mineral mapping that transitions from traditional point estimates to uncertainty-aware, set-valued predictions.
- We provide a comprehensive empirical study of conformal classification under severe library–scene mismatch, highlighting critical failure modes (such as empty prediction sets, or abstentions) and analyzing their spatial structure.
- We apply weighted conformal classification to mitigate covariate shift, introducing practical diagnostics to interpret the conformal outputs when ground-truth target labels are unavailable.

The remainder of this work is organized as follows. Section 2 introduces the USGS spectral library and the Cuprite AVIRIS scene, formalizing the source–target domain setting. Section 3 outlines the preprocessing pipeline required to harmonize laboratory and airborne spectra; this includes channel selection, spectral transformations (e.g., smoothing and derivative analysis), and the definition of the label space. Section 4 details the probabilistic classification models and the conformal calibration procedures, distinguishing between standard marginal coverage under exchangeability and weighted conformal prediction under covariate shift. Section 5 evaluates our methodology on the Cuprite case study. Finally, Section 6 discusses current limitations, practical implications for operational mapping workflows, and promising directions for future research.

2. Data

We use two different datasets for our analysis. For training and calibration we are using the Chapter M from USGS Spectral Library, while as test data we will use a subscene of the Cuprite AVIRIS scene.

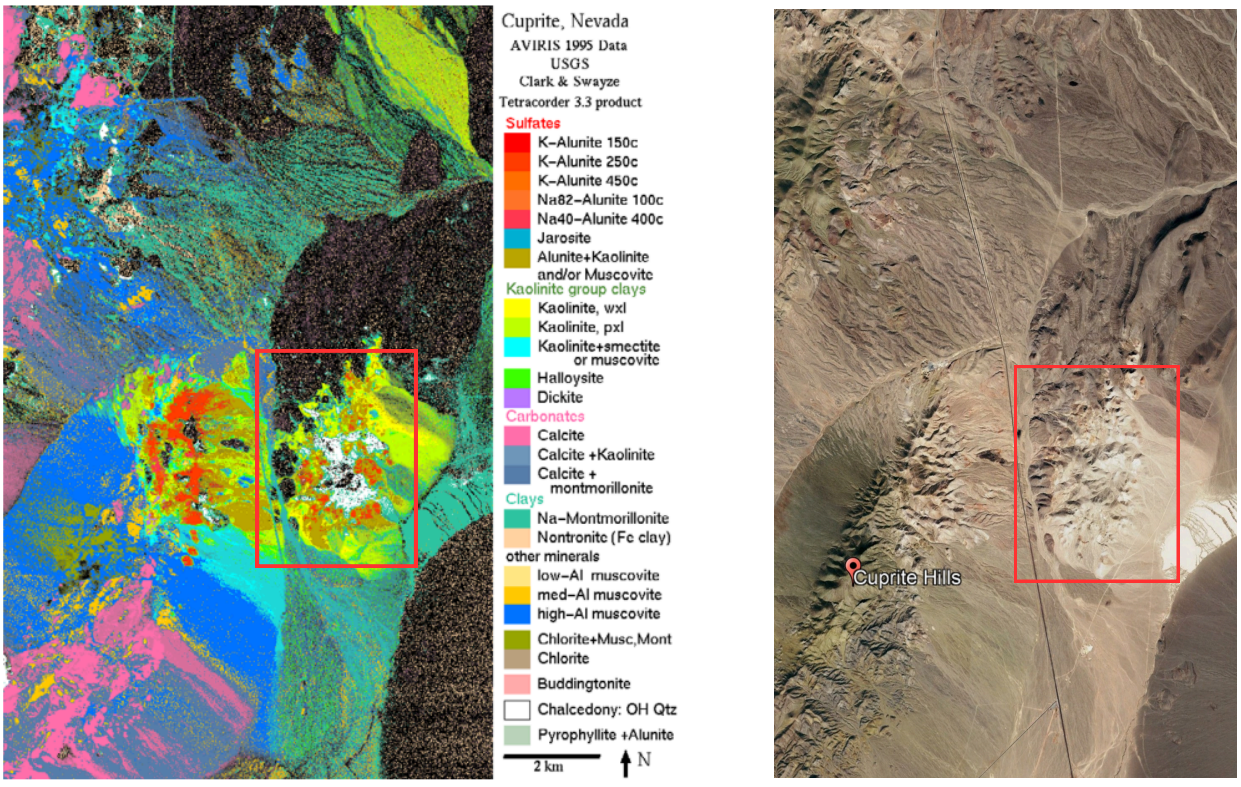
2.1. USGS spectral library

As source data, we use the USGS Spectral Library (Version 7) [Kokaly et al., 2017]. We employ the `s07_97` subset, resampled to the AVIRIS wavelength grid as distributed by NASA (1997). The library is organised into thematic chapters; we focus on Chapter M, which contains laboratory mineral spectra. Chapter M includes 1276 spectra associated with 212 distinct *mineral-name entries* (as provided in filenames), with multiple measurements per entry reflecting different samples and acquisition conditions. Filenames also encode metadata, including a spectral purity flag in {`a`,`b`,`c`,`d`,`u`} summarising

contamination levels, or the reflectance type (*Absolute* or *Relative*). (Pre-processing and any filtering choices are described in Section 3.)

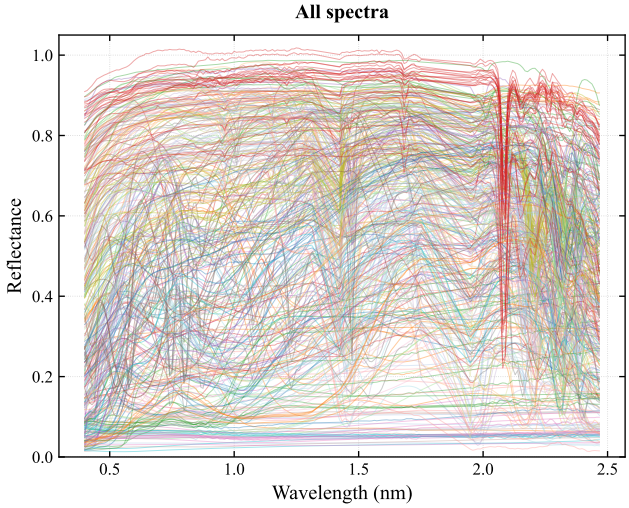
2.2. Cuprite AVIRIS scene

The target domain is a widely used hyperspectral subscene of the Cuprite mining district (Nevada, USA). The scene was acquired in June 1997 by NASA/JPL using the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) and is a standard benchmark for mineral mapping and unmixing due to well-exposed hydrothermal alteration zones and limited vegetation [Chen et al., 2007; Kong et al., 2015; Swayze et al., 1992].
We use a 190×250 subscene, the same considered in Nascimento and Dias [2005]. At the AVIRIS spatial resolution (approximately $20\,\text{m} \times 20\,\text{m}$ per pixel), observed reflectance typically arises from mixtures of surface materials. Relative to laboratory spectra, airborne observations are affected by atmospheric residuals, topography-driven illumination variability and sensor noise, inducing a non-negligible distribution shift between source-library and target-scene spectra. This motivates distribution-shift-aware uncertainty quantification, and in particular the weighted conformal calibration considered here.
Regarding the scene composition, Ren et al. [2020] identify the six most abundant minerals in the Cuprite district as alunite, kaolinite, montmorillonite, muscovite, calcite, and chalcedony.
For qualitative interpretation only, we refer to published mineralogical maps for the Cuprite district [Clark et al., 2003; Swayze et al., 2014]. These products are used solely as contextual references to interpret broad spatial patterns, since exhaustive pixel-level ground truth is generally unavailable at district scale.



(a) Prediction Map (b) Satellite Map

Figure 2: Comparison between the conformal prediction map and the reference satellite view of the Cuprite mining district.



H

Figure 3: All the spectra included in the training library after the preprocessing.

3. Preprocessing

This section summarises the preprocessing applied to the USGS library and the Cuprite scene to obtain a common wavelength support and a stable supervised learning setup. The steps address (i) measurement artefacts and contamination in laboratory spectra and (ii) insufficient sample sizes per mineral-name entry, which can destabilise probabilistic training and conformal calibration.

3.1. Building the training matrix

We construct the training dataset from the USGS spectral library through the following filtering steps.

1. **Reflectance type.** We retain only spectra reported as absolute reflectance (AREF) and discard relative reflectance (RREF) spectra. Compared to AREF, RREF measurements may exhibit scale and baseline distortions induced by reference panels and acquisition conditions, complicating cross-sample comparisons [Kokaly et al., 2017]. A comparison is provided in Figure 14 (Appendix A).
2. **Spectral purity.** We restrict the analysis to purity levels A and B. While level A corresponds to the highest mineralogical purity, its limited size (279 spectra) is insufficient for stable modelling. Including level B increases the dataset to 917 spectra across 173 mineral-name entries, preserving diversity while limiting contamination effects.
3. **Deleted channels.** To ensure consistency with the target-domain wavelength support, we remove spectra containing *deleted channels* within the retained wavelength grid. Deleted channels are encoded by the sentinel value -1.23×10^{34}. Any spectrum containing at least one such value is discarded (see Figure 15 in Appendix A). We discard entire spectra rather than imputing deleted channels to avoid introducing artificial structure in narrow absorption regions. After this step, the dataset consists of 614 spectra associated with 171 mineral-name entries.

The resulting class distribution remains highly imbalanced. We therefore retain only mineral-name entries with at least $n_{min} \geq 5$ spectra, which provides a reasonable compromise between within-class variability and overall sample size. This yields a final source dataset of **51 mineral-name entries and 418 spectra** (Figure 3), ensuring sufficient support for stable probabilistic training and conformal calibration. A detailed analysis of the threshold selection is reported in Appendix A (Figure 16).

To operate in a consistent feature space, all library spectra are aligned to a common wavelength grid. Following standard practice, we remove wavelength regions severely affected by atmospheric water vapor absorption and low signal-to-noise ratios [Iordache et al., 2011]. Specifically, we exclude channels

$$\{1, 2, 3,\ 104\text{–}113,\ 148\text{–}167,\ 221\text{–}224\},$$

resulting in $B = 187$ retained bands. Channel indices are 1-based and follow the AVIRIS ordering. This harmonization is a prerequisite for both training probabilistic classifiers on the source library and estimating importance weights for weighted conformal prediction, which require source and target data to share the same feature support.

Both training and test spectra are observed as high-dimensional vectors indexed by wavelength. However, reflectance spectra are more naturally interpreted as realizations of smooth functions over wavelength. We therefore adopt a functional data analysis (FDA) perspective, modeling each spectrum as a noisy discretization of an underlying smooth reflectance function $f(\lambda)$ [Wang et al., 2016; Ye et al., 2020].
Let $\lambda_1 < \cdots < \lambda_B$ denote the retained wavelength grid. For the i-th spectrum, we observe

$$y_i(\lambda_b) = f_i(\lambda_b) + \eta_{i,b}, \qquad b = 1, \ldots, B, \tag{4}$$

where $\eta_{i,b}$ represents measurement noise and modeling error.
To recover the underlying functional signal, we estimate f_i using a cubic smoothing spline $\hat{f}_i$ obtained via penalized least squares:

$$\hat{f}_i = \arg\min_{g \in \mathcal{H}^2} \left\{ \sum_{b=1}^{B} \big(y_i(\lambda_b) - g(\lambda_b)\big)^2 + \rho \int_{\lambda_1}^{\lambda_B} \big(g''(\lambda)\big)^2 \, d\lambda \right\}, \tag{5}$$

where $\rho > 0$ controls the trade-off between data fidelity and smoothness. The roughness penalty suppresses high-frequency noise while preserving the geometry of diagnostic absorption features. The smoothing parameter ρ is selected via generalized cross-validation (GCV). GCV is used per spectrum to adapt to variable noise levels; this is applied identically in source and target. Smoothing is applied on the common grid of $B = 187$ bands.
Spectral discrimination is often driven more by the *shape* of absorption features than by absolute reflectance levels. Leveraging the functional representation, we compute derivative spectra analytically from the fitted spline:

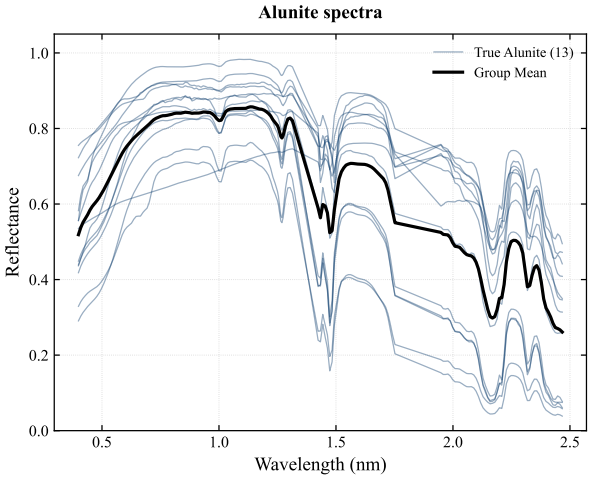
$$\hat{f}_i^{(k)}(\lambda) = \frac{d^k}{d\lambda^k} \hat{f}_i(\lambda). \tag{6}$$

In this work, we focus on the **first derivative** ($k = 1$) as the primary shape-based representation. First derivatives emphasize local slope changes around diagnostic absorption features and reduce sensitivity to global albedo variations and baseline shifts induced by illumination effects. These derivative spectra are used both for data-driven label construction via PAM clustering and as input features for the subsequent classification and conformal uncertainty quantification.
We apply the same channel mask and the same smoothing/derivative extraction to Cuprite AVIRIS scene.

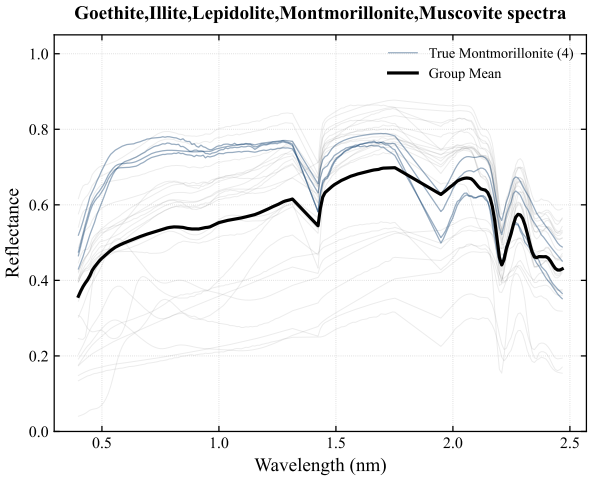
3.2. Building the categories

The USGS library provides *mineral-name entries* via filenames, which we use as baseline identifiers. However, at sensor resolution, grouping spectra solely by nominal mineral name may not reflect empirical spectral similarity: spectra under the same name can be heterogeneous, while different names may exhibit strong spectral overlap. Since diagnostic information is primarily carried by absorption-feature *shape*, we construct a data-driven label space by clustering derivative spectra.
We apply **Partitioning Around Medoids (PAM)** [Kaufman and Rousseeuw, 1990] to first-derivative spectra using the *spectral angle mapper* (SAM) [Yuhas et al., 1992] as the dissimilarity measure. SAM emphasises shape similarity and is invariant to multiplicative scale effects. The number of clusters is set to $K = 40$, selected over a predefined grid using a combination of internal and external diagnostics. Specifically, the silhouette index [Rousseeuw, 1987] summarises within-cluster compactness and separation, while Adjusted Rand Index (ARI) [Hubert and Arabie, 1985] and Normalized Mutual Information (NMI) [Strehl and Ghosh, 2002] quantify agreement with the partition induced by mineral-name entries (used here as a proxy for external coherence). Full details on distance comparisons, cluster selection, and validation analyses are reported in Appendix B (Section B.1).

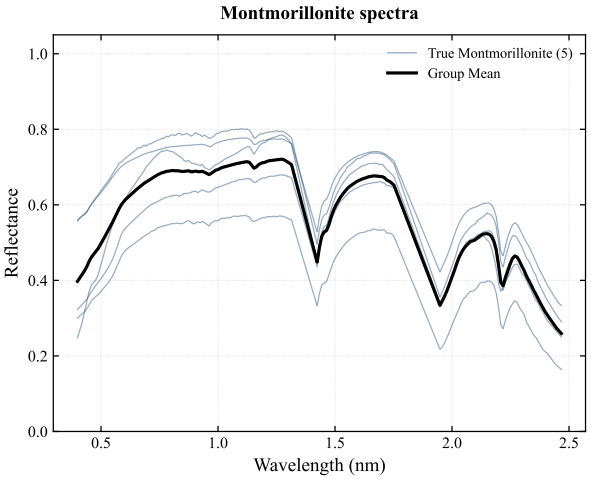
In the remainder of the thesis, the term *class* refers to these PAM clusters (data-driven spectral classes), and the associated cluster-name strings are used only for interpretability.
Figure 4 illustrates two representative outcomes. Alunite forms a single, well-isolated and spectrally coherent cluster, indicating close agreement between the nominal mineral-name entry and empirical spectral similarity. In contrast, montmorillonite spectra split across two clusters: one composite clay cluster grouping minerals with highly similar signatures (goethite, illite, lepidolite, muscovite), and a nearly isolated montmorillonite cluster with limited support. This reflects both spectral overlap among clays and imbalance in the reference library. Such heterogeneity is expected to propagate to downstream classification and uncertainty quantification, where composite or weakly supported clusters typically yield higher predictive uncertainty.



(a) Alunite: single, well-isolated cluster.



(b) Montmorillonite within a composite clay cluster.



(c) Montmorillonite in a nearly isolated cluster with few spectra.

Figure 4: Representative examples of PAM clusters built on first-derivative spectra using the spectral angle mapper (SAM) dissimilarity. Alunite forms a spectrally coherent and well-separated cluster, whereas montmorillonite appears both in a heterogeneous composite cluster and in a nearly isolated cluster with limited support. These examples illustrate how nominal mineral labels may correspond to either coherent or fragmented spectral groups, motivating the use of a data-driven taxonomy in downstream analyses.

4. Methodology

This section describes the end-to-end pipeline used to produce reliable prediction sets under covariate shift. We first specify the probabilistic classifier used to obtain class-probability estimates for the data-driven taxonomy (Section 4.1). We then introduce split conformal classification and its weighted extension, which turn these probability estimates into prediction sets with controlled miscoverage (Sections 4.2 and 4.3).

4.1. Classifiers

As the first step of this pipeline, we train probabilistic classifiers on the preprocessed spectra using the data-driven taxonomy obtained via PAM clustering (Section 3.2). We consider two widely used classifiers for hyperspectral data, which are typically high-dimensional and strongly correlated across bands: **Support Vector Machines with an RBF kernel (SVM-RBF)** [Bazi and Melgani, 2006; Melgani and Bruzzone, 2004] and **Random Forests (RF)** [Xia et al., 2018]. When enabled, PCA is used as a preprocessing step for the classifier inputs to obtain a compact representation and mitigate strong inter-band correlations.

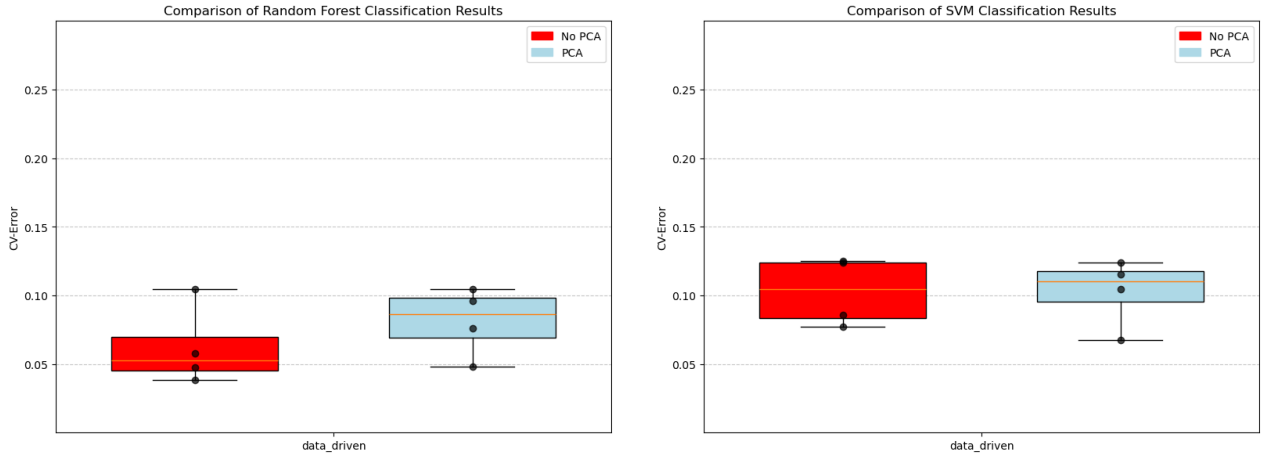
Support Vector Machines with RBF kernel. Support Vector Machines (SVM) are margin-based classifiers that induce nonlinear decision boundaries through kernel functions. We adopt the radial basis function (RBF) kernel,

$$K(x_i, x_j) = \exp\left(-\gamma \|x_i - x_j\|^2\right),$$

a standard choice in hyperspectral classification due to its flexibility [Bazi and Melgani, 2006; Melgani and Bruzzone, 2004]. Multi-class classification is handled via a one-vs-rest strategy. Probabilistic outputs are obtained via post-hoc calibration of the decision scores, and are later converted into nonconformity scores as described in Section 4.2.

Random Forests. Random Forests (RF) aggregate predictions from an ensemble of de-correlated decision trees trained on bootstrap samples. For a spectrum x, class probabilities are estimated as the empirical fraction of trees voting for each class. RF provides a robust non-parametric baseline for hyperspectral classification, with good tolerance to collinearity and noise [Xia et al., 2018].

Cross-validation and hyperparameter search. Model selection is performed using stratified 4-fold cross-validation (shuffled with a fixed random seed), optimizing accuracy. For RF, we tune `max_depth` $\in \{$`None`$, 5, 10, 15, 20, 25\}$, `max_features` $\in \{$`sqrt`$,$ `log2`$, 0.3, 0.5\}$ and `class_weight` $\in \{$`None`$,$ `balanced`$\}$, with `n_estimators` $= 500$. For SVM-RBF, we search over $C \in \{0.1, 1, 10, 100\}$ and $\gamma \in \{$`scale`$,$ `auto`$, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$.
When PCA is enabled, dimensionality reduction is included within a `Pipeline` and fitted within each training fold to prevent information leakage. We retain the smallest number of principal components explaining 90% of the cumulative variance (i.e., `PCA`$(n\_components = 0.9)$), so the retained dimension may vary across folds.
The final hyperparameter values selected via cross-validation for all configurations are reported in Appendix C, Table 5. Based on cross-validation under the data-driven taxonomy (Figure 5), RF achieves lower error than SVM-RBF. We therefore adopt **RF trained on first-derivative spectra, with PCA (90% variance) as the default representation**, as the base predictor in subsequent conformal inference experiments.



(a) Random Forest. (b) SVM-RBF.

Figure 5: Cross-validated classification error (1−Accuracy) under the data-driven taxonomy for (a) Random Forest and (b) SVM-RBF, with and without PCA retaining 90% of the explained variance. Dots indicate fold-wise errors; boxplots summarize their empirical distribution.

4.2. Conformal Prediction

Standard supervised classifiers, such as the Random Forest optimized in Section 4.1, return a point prediction $\hat{y} \in \mathcal{Y}$ for a new spectrum x. Probabilistic outputs $\hat{P}(y \mid x)$ are often used as a proxy for uncertainty, but they may be miscalibrated and do not provide finite-sample guarantees. In mineral mapping, where classification errors can propagate to downstream analyses, uncertainty quantification with statistical validity is therefore required.

We adopt the **Conformal Prediction** (CP) framework [Angelopoulos and Bates, 2021; Shafer and Vovk, 2008], which produces set-valued predictions with finite-sample marginal coverage. Let $Z_1, \ldots, Z_n$ be observation from an unknown distribution and let Z_{n+1} denote a new point from the same distribution. Under exchangeability of $(Z_1, \ldots, Z_{n+1})$ (i.e., invariance of the joint distribution under permutations, guaranteed if the Z_i are i.i.d.), CP constructs a prediction set $C(Z_1, \ldots, Z_n)$ such that, for any $\alpha \in (0, 1)$,

$$P(Z_{n+1} \in C(Z_1, \ldots, Z_n)) \geq 1 - \alpha. \tag{7}$$

The guarantee in (7) is only *marginal* (unconditional): the probability is taken over the joint randomness of $(Z_1, \ldots, Z_{n+1})$, hence also over the new covariate Z_{n+1}.

Operationally, CP relies on a *nonconformity* score that ranks candidate outcomes by how well they agree with the observed sample. A threshold is then calibrated from the empirical distribution of scores on held-out data, yielding a prediction set with the guarantee in (7).

In supervised learning, $Z_i = (X_i, Y_i)$, and the target is to predict Y_{n+1} given $X_{n+1} = x$. When Y is continuous, $C(x)$ is typically an interval (conformal regression); when Y is discrete, $C(x) \subseteq \mathcal{Y}$ is a set of labels **(conformal classification)**. In this work we focus on conformal classification and adopt a split conformal construction to enable scene-scale inference.

4.2.1 Split Conformal Classification

Given the computational constraints of hyperspectral imaging—which requires producing prediction sets for 47,500 pixels in the target scene—we employ **Split Conformal Classification**. In contrast to full conformal methods, which may require refitting or recomputing scores in a leave-one-out fashion for each new observation, split conformal decouples model fitting from calibration and yields substantial computational savings while retaining finite-sample coverage guarantees.

We randomly partition the source dataset (the USGS library) into two disjoint subsets: a *training set* $\mathcal{D}_{train}$ (70% of the source dataset) used to fit the classification model, and a *calibration set* $\mathcal{D}_{cal}$ =

$\{(x_i, y_i)\}_{i=1}^{n}$ (30% of the source dataset) used to form the empirical distribution of nonconformity scores.
We adopt a nonconformity score based on the classifier's probabilistic output:

$$s(x, y) = 1 - \hat{P}(y \mid x), \tag{8}$$

where $\hat{P}(\cdot \mid x)$ denotes the classifier's estimated class probabilities (built in Section 4.1). Larger scores correspond to less conforming labels.
Then, to construct the prediction set for a new target pixel x_{new}, we compute calibration scores $s_i = s(x_i, y_i)$ for all $(x_i, y_i) \in \mathcal{D}_{cal}$ using the classifier trained on $\mathcal{D}_{train}$. Let $\hat{q}$ denote the k-th order statistic of $\{s_i\}_{i=1}^{n}$, with $k = \lceil (n+1)(1-\alpha) \rceil$. The resulting prediction set is

$$C(x_{new}) = \{y \in \mathcal{Y} : s(x_{new}, y) \leq \hat{q}\}, \tag{9}$$

where $s(x_{new}, y)$ is evaluated for each $y \in \mathcal{Y}$. The main idea is to include in the prediction all the classes in $\mathcal{Y}$ whose nonconformity does not exceed the calibrated threshold. By construction, if the calibration sample and the test point are exchangeable, the set in (9) satisfies the marginal coverage guarantee in (7).

4.2.2 The challenge of distribution shift

The marginal coverage guarantee in (7) relies on the **exchangeability** of the calibration sample and the test point. In our remote sensing setting, this assumption is questionable because the classifier is trained on **laboratory spectra** acquired under controlled conditions (source domain), whereas it is deployed on **airborne measurements** affected by atmospheric effects, sensor noise, and subpixel mixing (target domain). As a result, the distribution of the input features may differ between calibration and deployment, a scenario commonly referred to as **covariate shift**. In particular, the marginal distribution of X changes from source to target, while the conditional distribution $Y \mid X$ is typically assumed to be approximately stable.
Under covariate shift, the split conformal construction based on the unweighted empirical quantile of calibration scores is no longer guaranteed to achieve the nominal coverage level. In regions of the feature space that are under-represented in the calibration data, the resulting sets may exhibit under-coverage or become overly conservative [Tibshirani et al., 2020].
To account for this mismatch, we replace the unweighted calibration distribution with an importance-weighted counterpart that reflects the target domain. This leads to Weighted Split Conformal Classification (WCC), described next.

4.3. Weighted Split Conformal Classification

4.3.1 Theoretical formulation

WCC is a particular case of the more general framework of Weighted Split Conformal Prediction (WCP), specialized to multiclass classification. The main difference between the unweighted and weighted versions is the construction of importance weights. WCP adjusts the influence of each calibration sample according to its relevance for the target covariate distribution. Specifically, each calibration point with covariates x is assigned an importance weight proportional to the density ratio

$$w(x) = \frac{d\widetilde{P}_X}{dP_X}(x), \tag{10}$$

where P_X denotes the covariate distribution in the source domain (from which the training and calibration samples are drawn) and $\widetilde{P}_X$ denotes the covariate distribution in the target domain. In practice, $w(x)$ is approximated via a low-dimensional projection, as described below.
Given a test point X_{n+1}, define normalized weights for the calibration set $\mathcal{D}_{\text{cal}} = (X_i, Y_i) \quad \forall i \in 1, ..., n_{cal}$ and for the test point as

$$p_i^w(X_{n+1}) = \frac{w(X_i)}{\sum_{j=1}^{n} w(X_j) + w(X_{n+1})}, \quad p_{n+1}^w(X_{n+1}) = \frac{w(X_{n+1})}{\sum_{j=1}^{n} w(X_j) + w(X_{n+1})}. \tag{11}$$

Let $s_i = s(X_i, Y_i)$ denote the nonconformity scores computed on the calibration set. Instead of the uniform empirical distribution of calibration scores, WCP uses the weighted empirical distribution

$$\sum_{i=1}^{n} p_i^w(X_{n+1}), \delta_{s_i} + p_{n+1}^w(X_{n+1}), \delta_\infty, \tag{12}$$

and constructs prediction sets using its $(1-\alpha)$-quantile. The additional atom at ∞ implements the finite-sample correction ensuring exact marginal validity under known density ratios. In WCC the nonconformity score used is the same in Equation 8.
Tibshirani et al. [2020] show that, when the density ratios are known exactly, this construction achieves finite-sample marginal coverage with respect to the target covariate distribution under covariate shift. Unlike standard split conformal prediction, the normalized weights $p_i^w(X_{n+1})$ depend on the test covariates through the normalization term, which includes $w(X_{n+1})$. Consequently, the calibration quantile becomes test-point–dependent and must, in principle, be recomputed for each X_{n+1}. In scene-scale applications with millions of pixels, this can become computationally demanding, unless the computation is vectorized and/or approximations are used.

4.3.2 Density ratio estimation via PCA

In practice, the covariate marginals P_X and $\widetilde{P}_X$ are unknown and must be approximated from data. Direct density estimation in the full spectral space ($B = 187$) is statistically ill-posed, so we estimate the density ratio in a low-dimensional representation.
Let $z = \Pi_{K_{\text{pca}}}(x) \in R^{K_{\text{pca}}}$ denote the projection of a spectrum x onto the first K_{pca} principal components. In this work we set $K_{\text{pca}} = 1$; results for $K_{\text{pca}} = 2$ are reported in Appendix C.2. The PCA basis is fitted on the source training set $\mathcal{D}_{\text{train}}$ (USGS library) only. Both the calibration set and the Cuprite spectra are then projected using the same principal directions, ensuring a common representation across domains.
We fit separate Gaussian kernel density estimators on the projected samples from the source and target domains, denoted $\widehat{p}_{\text{src}}(z)$ and $\widehat{p}_{\text{tgt}}(z)$, using a bandwidth $h = 0.5$. This choice is not claimed to be optimal; rather, it is selected to provide a stable visual summary without evident overfitting. Although the covariate-shift assumption formally concerns the mismatch between the calibration and target distributions, we estimate $\widehat{p}_{\text{src}}$ using the union of training and calibration spectra (which are drawn from the same USGS library) in order to increase stability of the density estimate. Concretely, pooling these sets increases the source sample size (from 126 calibration spectra to 418 source spectra), reducing variance in the KDE and mitigating spurious ratio inflation.
The importance weight is then defined as the estimated density ratio

$$\widehat{w}(z) = \frac{\widehat{p}_{\text{tgt}}(z)}{\widehat{p}_{\text{src}}(z)}, \tag{13}$$

which serves as a low-dimensional approximation of $w(x) = \frac{d\widetilde{P}_X}{dP_X}(x)$ under the mapping $z = \Pi_{K_{\text{pca}}}(x)$. This procedure yields a marginal approximation of the density ratio along the chosen PCA projection.

Validity considerations. The finite-sample coverage guarantee of weighted conformal prediction holds when the density ratios are known exactly and the calibration and test covariates follow the assumed source/target distributions. In practice, $\widehat{w}$ is estimated from finite samples (and via a low-dimensional PCA projection), so the theoretical guarantee is no longer exact. We therefore treat the weighting step as an approximate covariate-shift correction and assess its behaviour through diagnostics on distribution alignment and weight stability (Section 5.2), rather than as an empirical validation of target-domain coverage.

5. Results

5.1. Classifier results: qualitative assessment of probability maps

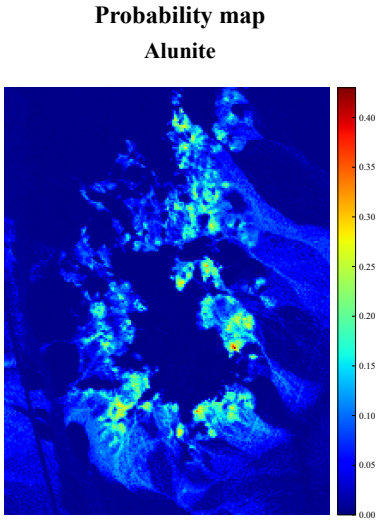
Before presenting conformal and weighted conformal prediction results, we assess whether the probabilistic outputs of the base classifier are spatially and mineralogically consistent. This step is not meant as a standalone quantitative validation of the classifier, but as a sanity check to verify that class-probability maps reflect meaningful spatial structure under the data-driven taxonomy.

Figure 5.1 reports probability maps for representative classes under the data-driven taxonomy, chosen to illustrate a well-isolated mineral (alunite), an intermediate case (kaolinite–halloysite), and a heterogeneous case (montmorillonite). As discussed in Section 3.2 (Figure 4), alunite corresponds to a single well-isolated cluster, whereas montmorillonite spectra split across a composite clay cluster and a nearly isolated cluster with limited support.

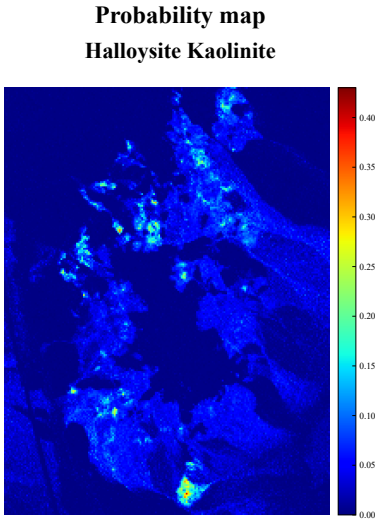
Alunite yields a spatially coherent probability map with pronounced high-probability regions (Figure 6a), broadly consistent with mineral distributions previously reported in Iordache et al. [2014].

Kaolinite forms a nearly isolated cluster jointly with halloysite (Appendix A, Figure 17) and is supported by a moderate number of spectra. The corresponding probability map (Figure 6b) exhibits coherent regions, but with lower and more fragmented probability levels than alunite, consistent with weaker class support and mild overlap among clay signatures.

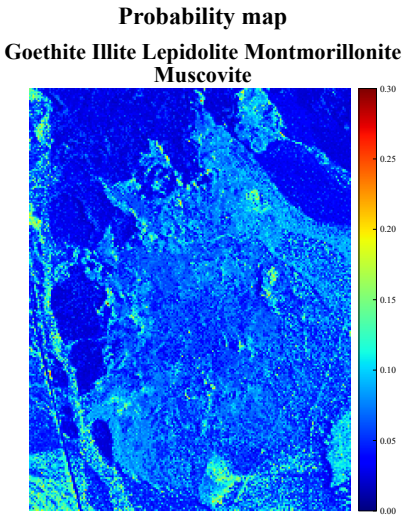
For montmorillonite, the two PAM-derived class definitions lead to qualitatively different spatial probability structures (Figures 6c–6d). The composite clay cluster produces diffuse, lower probability levels, whereas the nearly isolated montmorillonite cluster concentrates probability mass in a more localised region. These mineral-level maps are not directly comparable to those in Iordache et al. [2014], as they reflect the data-driven class definitions used here.



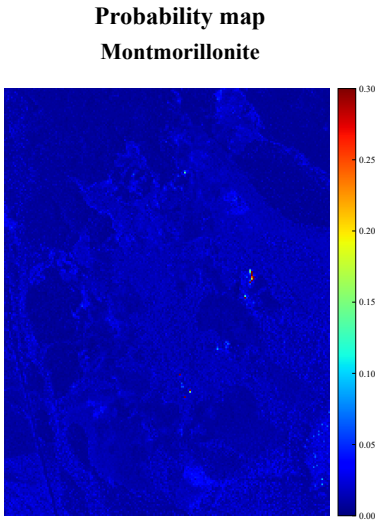
(a) Alunite cluster.



(b) Kaolinite–Halloysite cluster.



(c) Montmorillonite: composite clay cluster.



(d) Montmorillonite: nearly isolated cluster.

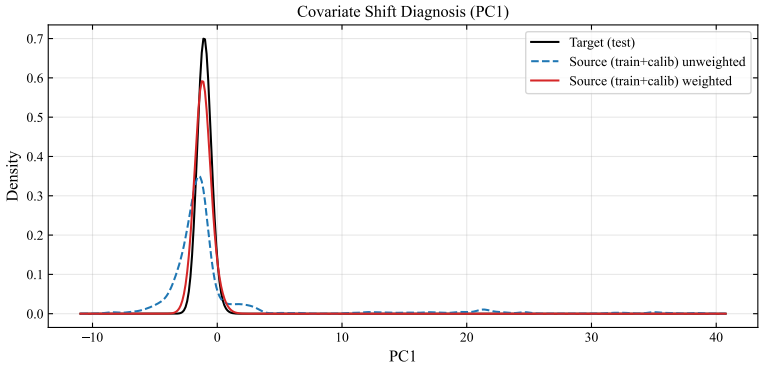
Figure 6: Predicted class-probability maps for representative minerals under the data-driven taxonomy. Alunite corresponds to a well-isolated and well-supported cluster, exhibiting spatially coherent high-probability regions. Kaolinite (clustered jointly with halloysite) represents an intermediate case with moderate spectral support, yielding coherent but less pronounced probability patterns. Montmorillonite illustrates both composite and weakly supported clustering behaviours, leading to more diffuse or spatially localised probability structures.

5.2. Weighted conformal prediction without target labels: operational outputs

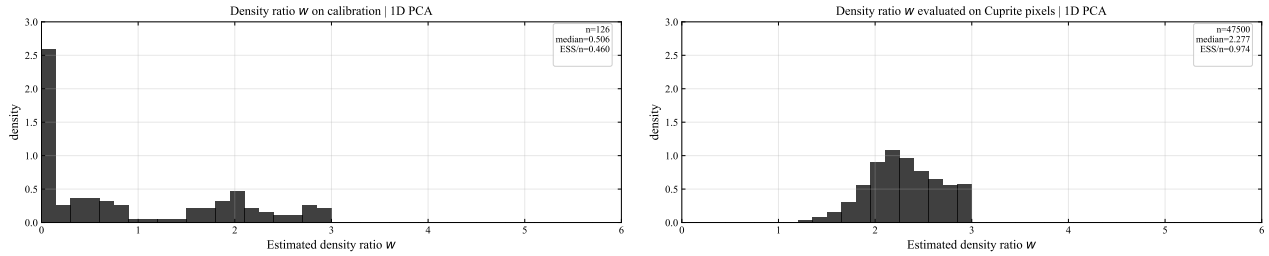
We apply WCC to the Cuprite scene at nominal miscoverage level $\alpha = 0.05$. Since pixel-level labels are unavailable at scene scale, target-domain coverage cannot be evaluated empirically. We therefore report operational summaries of the outputs (set sizes, frequency of empty sets, and spatial patterns), while theoretical validity is understood under the covariate-shift assumptions stated in Section 4.3.

Covariate-shift diagnostic (PC1). We assess distribution shift between the source calibration covariates and Cuprite by comparing their marginals along the first principal component (PC1)

using the PCA–KDE procedure of Section 4.3.2. Figure 7a shows a pronounced mismatch between standard calibration and Cuprite PC1 marginals, while importance reweighting brings the calibration distribution closer to the target. Quantitatively, the overlap coefficient increases from 0.397 to 0.794 (Table 1). The effective sample size is moderate (ESS/$n_{\text{cal}} \approx 0.46$), indicating that the weights are not concentrated on a very small subset of calibration points.



(a) PC1 KDE overlap between calibration and Cuprite covariates, before and after importance reweighting.



(b) Distribution of estimated density ratios $\widehat{w}$ on calibration points (used for WCC) and on Cuprite pixels (diagnostic).

Figure 7: Covariate-shift diagnostics and stability of the estimated importance weights.

Table 1: PC1 covariate-alignment diagnostics between calibration and Cuprite, before and after importance reweighting. Bootstrap intervals are 95%.

Metric	Standard	Weighted
Overlap $\int \min(p, q)$	0.397	0.794
ESS/n_{cal} (weighted)	0.464	

Prediction-set sizes at 95% nominal coverage. Figure 8 summarises per-pixel prediction-set sizes on Cuprite. Under standard split conformal prediction, the mean set size is 0.837, but this is driven by a substantial fraction of empty prediction sets ($|C(x)| = 0$) occurring in 22.9% of pixels (Table 2). Such degenerate outputs are consistent with distribution shift: a source-calibrated threshold may be overly stringent in target covariate regions where the classifier produces low and diffuse probabilities. Under WCC, calibration examples are reweighted towards the Cuprite covariate distribution, yielding target-adapted calibration through importance reweighting. Empty prediction sets are eliminated at $\alpha = 0.05$ (empty fraction 0.0%), at the cost of bigger sets: the mean set size increases to 2.529 and the median to 3 (Table 2). Operationally, weighting trades decisiveness for robustness, replacing "reject" outcomes by bigger ambiguity sets.

To confirm that weighting changes outputs beyond aggregate size statistics, we compute the fraction of pixels whose prediction set differs between standard and weighted conformal prediction. At $\alpha = 0.05$,

89.2% of pixels change their prediction set, and almost all changes correspond to set expansions, consistent with the reduction of empty and singleton outputs.

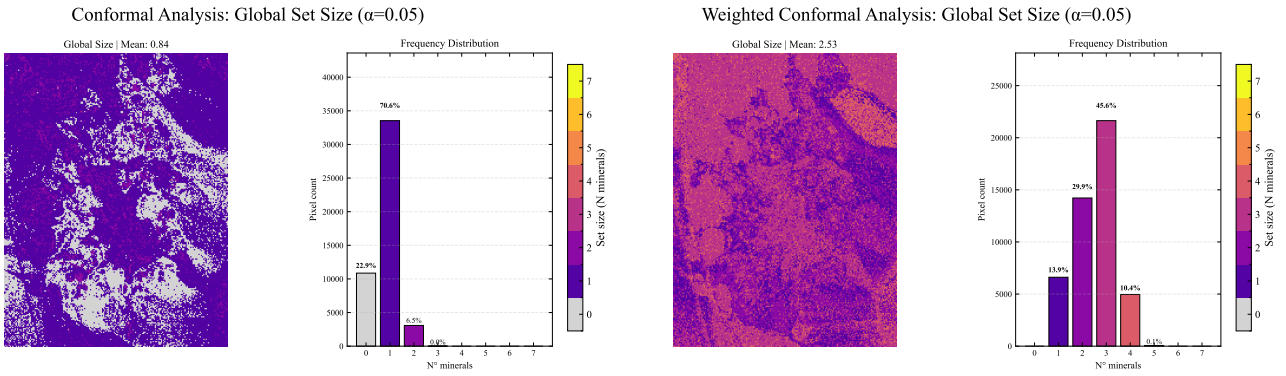


Figure 8: Per-pixel prediction-set size at $\alpha = 0.05$: standard conformal (left) vs weighted conformal (right). Empty sets (size 0) are eliminated under weighting.

Table 2: Prediction-set size summary on Cuprite at nominal coverage $1 - \alpha = 0.95$ ($\alpha = 0.05$).

Metric	Standard	Weighted
Mean $\lvert C \rvert$	0.837	2.529
Max $\lvert C \rvert$	3	5
Empty sets (%)	22.9	0.0
Singletons (%)	70.6	13.9
Mean $\lvert C \rvert$ (non-empty)	1.085	2.529

Class-level inclusion frequencies. To understand how the expansion of prediction sets is distributed across the taxonomy, Table 3 details the frequency with which data-driven clusters are included in the prediction sets. The scene is practically dominated by the first composite class (Allanite–Almandine–...), which appears in 68.20% of pixels under standard calibration and expands to 95.29% under WCC. This ubiquitous presence is expected: because this cluster aggregates numerous mineral types, it is backed by substantial spectral support in the reference library, making it the most frequent baseline prediction for the Random Forest classifier.
Under standard conformal classification, apart from this dominant cluster, secondary classes are heavily penalised by the source-calibrated threshold, leading to the high rate of empty sets. Upon applying WCC, the inclusion rates for these secondary clusters increase dramatically (e.g., the Allanite–Bytownite–Illite–Jadeite cluster jumps from 3.59% to 57.40%). Crucially, we also observe a meaningful increase in the inclusion of minerals known to characterize the Cuprite mining district, such as alunite (from 3.33% to 8.04%) and montmorillonite-bearing clusters. This behaviour aligns with geographical expectations, demonstrating that importance reweighting effectively recovers relevant target-domain signatures that would otherwise be discarded under shift. Ultimately, WCC incorporates multiple plausible overlapping mineral signatures to maintain the desired coverage guarantee while preserving spatial and mineralogical coherence.

Table 3: Top class presence in prediction sets at nominal coverage $1 - \alpha = 0.95$ ($\alpha = 0.05$). Values represent the fraction of Cuprite pixels where each data-driven cluster is included in the conformal prediction set, comparing standard and weighted conformal classification.

Data-driven cluster	Standard (%)	Weighted (%)
Allanite_Almandine_Andradite_Bytownite_Carnallite_ Celestite_Chalcopyrite_Halite_Siderite	68.20	95.29
Allanite_Bytownite_Illite_Jadeite	3.59	57.40
Allanite_Goethite	0.95	40.14
Goethite_Illite_Lepidolite_Montmorillonite_Muscovite	7.16	31.72
Celestite_Lazurite	0.23	10.03
Alunite	3.33	8.04
Illite_Lepidolite_Muscovite	0.15	4.16
Antigorite_Smaragdite	0.00	3.70
Halloysite_Kaolinite	0.08	0.61
Montmorillonite	0.00	0.14

Qualitative examples. Figure 9 shows the spatial structure of *weighted* conformal prediction sets at $\alpha = 0.05$ for two representative minerals, together with the corresponding set-size histograms. Alunite, associated with a nearly pure and well-supported cluster, yields predominantly small prediction sets and spatially coherent patterns, consistent with confident classifier outputs. In contrast, Montmorillonite—often occurring in mixed or weakly supported clusters—produces larger and more heterogeneous sets, reflecting spectral overlap and probability mass spread across competing minerals. To further highlight the effect of importance weighting at the mineral level, Figures 10 and 11 compare *presence/absence* binary maps derived from standard versus weighted conformal prediction. For montmorillonite, which appears in multiple data-driven clusters (Figure 5.1), the binary presence map is constructed by marking a pixel as present whenever *any* cluster containing montmorillonite is included in the conformal prediction set. These examples complement the scene-scale comparison in Figure 8 by showing how weighting changes the spatial extent and localization of mineral detections for specific classes.
Additional standard-versus-weighted mineral-level examples (Kaolinite) are reported in Appendix B (Figure 22).

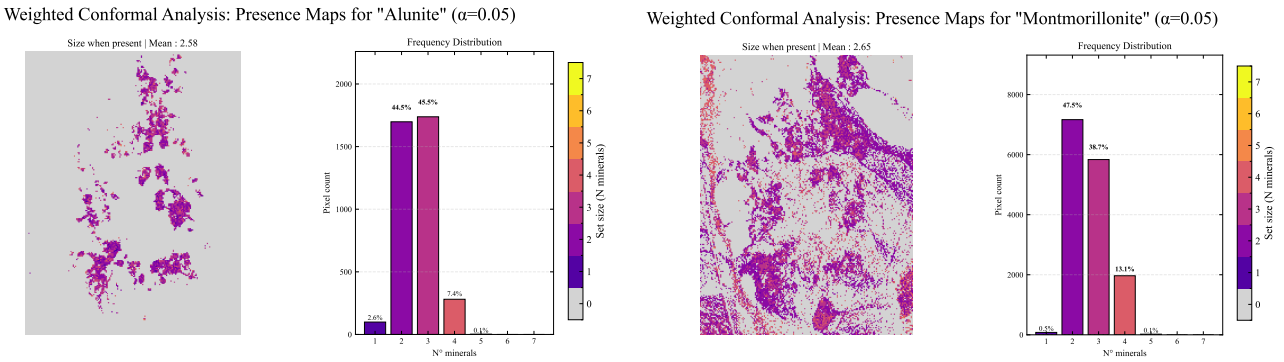


Figure 9: Weighted conformal prediction outputs at $\alpha = 0.05$ for Alunite and Montmorillonite. For each mineral, we show the per-pixel prediction-set size map under weighted conformal calibration and the corresponding distribution of set sizes across pixels. See Figure 8 for the standard-versus-weighted comparison at scene scale.

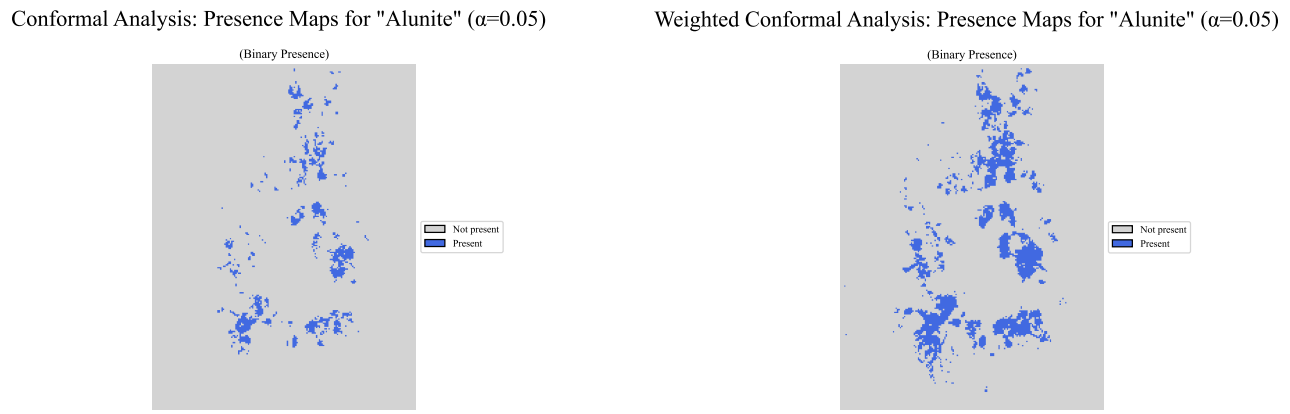


Figure 10: Binary presence/absence maps for Alunite at $\alpha = 0.05$: standard split conformal classification (left) versus weighted conformal classification (right). A pixel is marked as *present* when Alunite is included in the conformal prediction set, and *not present* otherwise.

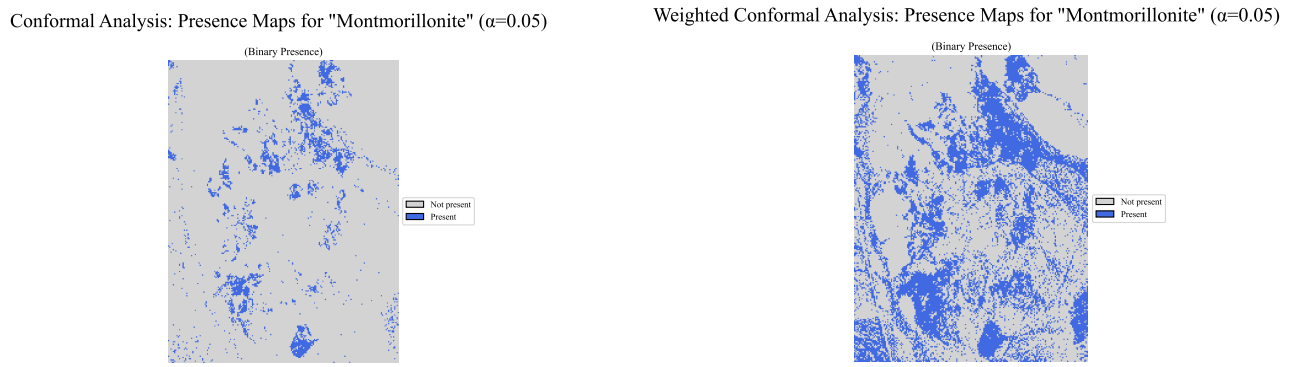


Figure 11: Binary presence/absence maps for Montmorillonite at $\alpha = 0.05$: standard split conformal classification (left) versus weighted conformal classification (right). A pixel is marked as *present* when Montmorillonite is included in the conformal prediction set, and *not present* otherwise.

Set size versus nominal coverage. Figure 12 reports mean prediction-set size as a function of nominal coverage $1 - \alpha$ over $[0.90, 0.96]$ (full range in Appendix C.2, Figure 21). Standard conformal prediction yields small average set sizes over much of the range, partly due to empty outputs under shift. WCC produces larger and more stable sets, consistent with the reduction of degenerate outputs. Since WCC targets validity with respect to the target covariate distribution, the resulting thresholds can be more conservative in covariate regions where the classifier is less confident. In addition, RF probability outputs induce a discrete set of nonconformity values, yielding stepwise changes in calibrated thresholds and occasional sharp transitions in mean set size. The same analysis for a 2D PCA representation is shown in Appendix C (Figure 21).

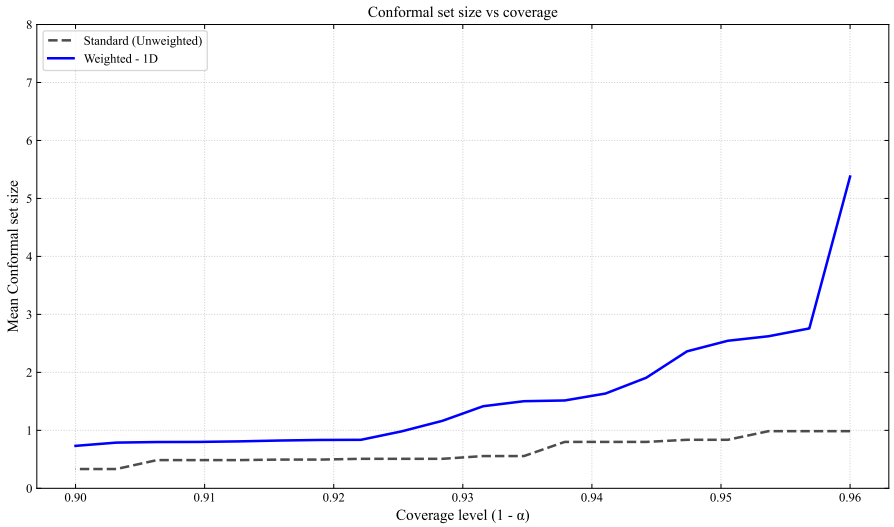


Figure 12: Mean prediction-set size on Cuprite as a function of nominal coverage level $1-\alpha$ (90%–96%), comparing standard and weighted conformal prediction.

Synthesis. Taken together, the results highlight three practical points. First, under pronounced library–scene mismatch, standard split conformal calibration can lead to widespread abstention: at $1-\alpha = 0.95$, empty prediction sets occur in 22.9% of Cuprite pixels (Table 2), and their spatial organization (Figure 8) suggests that this is not an isolated artifact but a systematic effect of applying a source-calibrated threshold in target covariate regions where the classifier outputs low and diffuse probabilities.

Second, importance reweighting materially changes conformal outputs in a predictable way, trading abstention for ambiguity. At $\alpha = 0.05$, WCC eliminates empty sets, but increases the typical set size (mean 2.529 and median 3; Table 2). Crucially, this expansion is not arbitrary: as shown in Table 3, while predictions are heavily anchored to a dominant composite cluster, WCC systematically recovers secondary and geologically pertinent classes—such as alunite and montmorillonite-bearing clusters—that are otherwise suppressed under shift. This reflects a robustness–decisiveness trade-off: fewer "reject" outcomes come at the cost of larger, yet mineralogically meaningful, ambiguity sets.

Third, the resulting prediction-set maps provide an interpretable uncertainty layer. Mineral-level examples (Figures 9–11) show that well-supported and nearly pure clusters (e.g., Alunite) yield smaller sets and spatially coherent patterns, whereas composite or weakly supported minerals (e.g., Montmorillonite) produce larger and more heterogeneous sets. This spatial structure supports the operational use of set sizes to flag regions where additional evidence or validation effort may be most valuable when target labels are unavailable.

6. Conclusions and further developments

In this work, we investigated conformal and weighted conformal prediction for hyperspectral mineral mapping under covariate shift, explicitly addressing the realistic scenario where scene-scale target labels are unavailable.

Our results on the Cuprite scene demonstrate that importance reweighting, achieved via a PCA–KDE density-ratio approximation, yields systematic and interpretable improvements in the conformal outputs. Specifically, at a nominal coverage of $1-\alpha = 0.95$, standard conformal prediction produces empty prediction sets for 22.9% of the pixels. In contrast, weighted conformal calibration successfully adapts to the target covariate distribution, eliminating empty sets entirely (Table 2). Naturally, mitigating overly stringent source-calibrated thresholds resolves degenerate outputs but inevitably introduces a trade-off: the reduction in abstentions comes at the cost of larger prediction sets (e.g., a median size of 3 at $\alpha = 0.05$). This expansion is a desired mathematical behavior, as it correctly reflects the increased ambiguity in regions where the classifier is intrinsically uncertain.

Despite these practical benefits, it is crucial to acknowledge the theoretical and operational limitations

of the proposed framework. First, the validity guarantees provided by (weighted) conformal prediction are strictly *marginal*. Under the required assumptions, the prediction set $C(X)$ satisfies Equation 2 for a randomly drawn target pixel (X, Y). However, this does *not* imply *conditional* coverage given $X = x$; in other words, the framework does not guarantee that

$$P\big(Y \in C(X) \mid X = x\big) \geq 1 - \alpha \qquad \text{for all } x. \tag{14}$$

Furthermore, this marginal guarantee relies on exchangeability (typically framed as an i.i.d. assumption), which is only approximately satisfied in hyperspectral images. Since Cuprite pixels exhibit strong spatial dependence, errors and set sizes may show spatial clustering even when marginal coverage holds under idealized sampling conditions. Accordingly, marginal validity should not be interpreted as a simultaneous, scene-level guarantee over all pixels in a single realization of the image.

Moreover, while weighted conformal prediction explicitly addresses distribution shift, its validity is inherently tied to the accuracy of the density-ratio estimation. Because we approximate this ratio via a low-dimensional PCA–KDE procedure, any resulting guarantee is both approximate and marginal with respect to the chosen projection.

From a practical standpoint, this study suggests two main implications. First, when scene-scale ground truth is unavailable, conformal outputs should be interpreted and reported through diagnostics that remain meaningful without labels. Prediction-set sizes, the frequency of empty sets, and their spatial patterns constitute a compact summary of how uncertainty is expressed across the scene (e.g., Figure 13). In this view, an empty prediction set is best regarded as an abstention signal: it can be informative when it occurs rarely and locally, but a large and spatially structured empty-set rate indicates severe mismatch between calibration and target domains and can limit the operational usability of standard conformal calibration.

Second, importance reweighting can mitigate overly stringent source-calibrated thresholds by adapting calibration to the target covariate distribution, thereby converting widespread abstentions into non-empty ambiguity sets. As visually summarized in Figure 13, this improves the spatial continuity of reliable predictions. However, it inevitably introduces a robustness–decisiveness trade-off: fewer abstentions are obtained at the cost of larger prediction sets, especially in covariate regions where the classifier is intrinsically uncertain. For this reason, weighted conformal outputs should be accompanied by basic stability summaries of the estimated weights (e.g., their distribution and ESS/n_{cal}) and by the sensitivity of set sizes to the chosen risk level α, since the mean set-size versus coverage curve provides a direct handle on the practical cost of operating at higher nominal coverage.

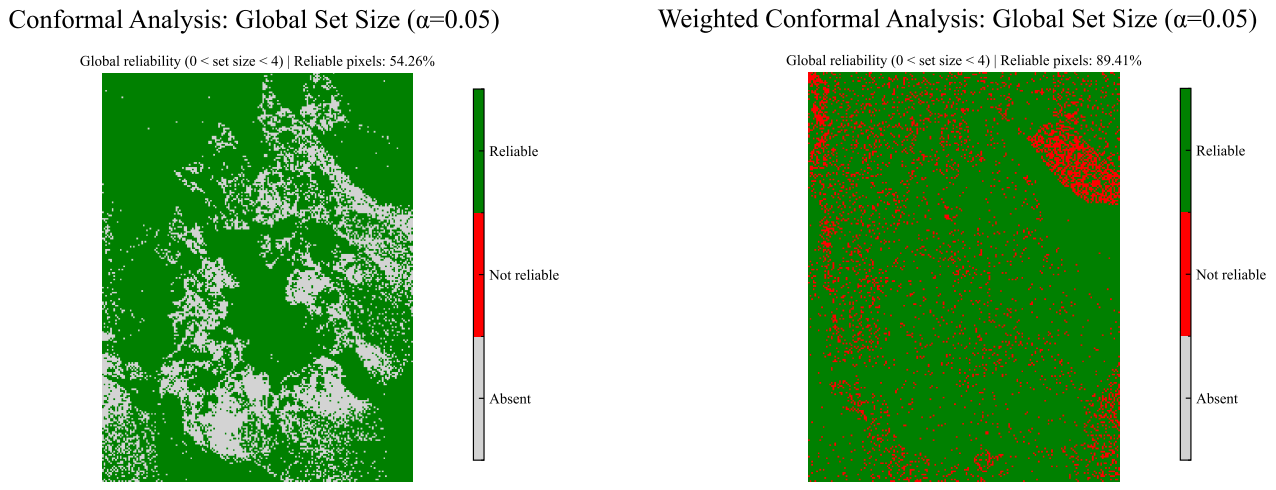


Figure 13: Spatial distribution of prediction reliability for standard (left) and weighted (right) conformal classification. A pixel is deemed *reliable* (green) if its prediction set size strictly bounds between 1 and 4, representing at most 10% of the 40 data-driven classes. Empty sets (size 0, marking abstentions) and highly ambiguous sets (size > 4) are considered *unreliable* (red). Importance reweighting effectively expands the reliable area from 77.13% to 89.41% of the scene, illustrating the practical benefit of adapting to the target covariate distribution.

Third, the proposed workflow is modular and transferable, and can be applied to other scenes and geological settings by adapting the source library, taxonomy definition, and shift-diagnostic step to the new target domain.
Several extensions naturally follow. First, replacing marginal PCA-based density-ratio estimation with richer conditional or locally adaptive schemes could strengthen the connection to target validity. Second, explicitly accounting for spatial dependence, through neighbourhood conditioning or spatially aware post-processing with validity control, may reduce unnecessary set expansion in homogeneous regions and improve interpretability. Third, uncertainty is strongly influenced by the data-driven taxonomy: clusters corresponding to nearly pure minerals yield sharper probability maps and smaller conformal sets, whereas composite clusters induce larger ambiguity. This motivates hybrid strategies combining data-driven clustering with expert-guided refinement or geological priors. In addition, expanding the amount and diversity of reference spectra would improve probabilistic calibration, reduce instability for weakly supported classes, and make density-ratio estimation more reliable. Finally, qualitative assessment by domain experts or comparison with independent geological products would provide valuable insight into the practical utility of weighted conformal outputs beyond operational summaries.

7. Bibliography and citations

References

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

T.V. Bandos, L. Bruzzone, and G. Camps-Valls. Classification of Hyperspectral Images With Regularized Linear Discriminant Analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 47(3): 862–873, March 2009. ISSN 0196-2892, 1558-0644. doi: 10.1109/TGRS.2008.2005729.

Y. Bazi and F. Melgani. Toward an Optimal SVM Classification System for Hyperspectral Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing*, 44(11):3374–3385, November 2006. ISSN 0196-2892. doi: 10.1109/TGRS.2006.880628.

Enton Bedini. The use of hyperspectral remote sensing for mineral exploration: a review. *Journal of Hyperspectral Remote Sensing*, 7(4):189–211, 2017. doi: 10.29150/jhrs.v7.4.p189-211.

José M. Bioucas-Dias, Antonio Plaza, Nicolas Dobigeon, Mario Parente, Qian Du, Paul Gader, and Jocelyn Chanussot. Hyperspectral Unmixing Overview: Geometrical, Statistical, and Sparse Regression-Based Approaches, April 2012.

Boselli. Hyperspectral unmixing algorithms for remote compositional surface mapping: A review of the state of the art with application to AVIRIS Cuprite data and synthetic data. 2022.

Xianfeng Chen, Timothy A. Warner, and David J. Campagna. Integrating visible, near-infrared and short-wave infrared hyperspectral and multispectral thermal imagery for geological mapping at Cuprite, Nevada. *Remote Sensing of Environment*, 110(3):344–356, October 2007. ISSN 00344257. doi: 10.1016/j.rse.2007.03.015.

Roger N. Clark, Gregg A. Swayze, K. Eric Livo, Raymond F. Kokaly, Steve J. Sutley, J. Brad Dalton, Robert R. McDougal, and Carol A. Gent. Imaging spectroscopy: Earth and planetary remote sensing with the USGS Tetracorder and expert systems. *Journal of Geophysical Research: Planets*, 108(E12):2002JE001847, December 2003. ISSN 0148-0227. doi: 10.1029/2002JE001847.

A. I. Gavrishin, A. Coradini, and P. Cerroni. Multivariate classification methods in planetary sciences. *Earth, Moon, and Planets*, 59(2):141–152, November 1992. ISSN 0167-9295, 1573-0794. doi: 10.1007/BF00058728.

Danfeng Hong, Naoto Yokoya, Jocelyn Chanussot, and Xiao Xiang Zhu. An Augmented Linear Mixing Model to Address Spectral Variability for Hyperspectral Unmixing. 28(4):1923–1938, 2019. ISSN 1057-7149, 1941-0042. doi: 10.1109/TIP.2018.2878958. URL https://ieeexplore.ieee.org/document/8528557/.

Kurt Hornik, Ingo Feinerer, Martin Kober, and Christian Buchta. Spherical k-Means Clustering.

Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985. doi: 10.1007/BF01908075.

Marian-Daniel Iordache, José M. Bioucas-Dias, and Antonio Plaza. Sparse Unmixing of Hyperspectral Data. 49(6):2014–2039, 2011. ISSN 0196-2892, 1558-0644. doi: 10.1109/TGRS.2010.2098413. URL http://ieeexplore.ieee.org/document/5692827/.

Marian-Daniel Iordache, Jose M. Bioucas-Dias, and Antonio Plaza. Collaborative Sparse Regression for Hyperspectral Unmixing. *IEEE Transactions on Geoscience and Remote Sensing*, 52(1):341–354, January 2014. ISSN 0196-2892, 1558-0644. doi: 10.1109/TGRS.2013.2240001.

Leonard Kaufman and Peter J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, New York, 1990.

Raymond F. Kokaly, Roger N. Clark, Gregg A. Swayze, K. Eric Livo, Todd M. Hoefen, Neil C. Pearson, Richard A. Wise, William M. Benzel, Heather A. Lowers, Rhonda L. Driscoll, and Anna J. Klein. Usgs spectral library version 7. Data Series 1035, U.S. Geological Survey, 2017. URL https://doi.org/10.3133/ds1035.

Fanqiang Kong, Wenjun Guo, Yunsong Li, Qiu Shen, and Xin Liu. Backtracking-Based Simultaneous Orthogonal Matching Pursuit for Sparse Unmixing of Hyperspectral Data. *Mathematical Problems in Engineering*, 2015:1–17, 2015. ISSN 1024-123X, 1563-5147. doi: 10.1155/2015/842017.

F. Melgani and L. Bruzzone. Classification of hyperspectral remote sensing images with support vector machines. *IEEE Transactions on Geoscience and Remote Sensing*, 42(8):1778–1790, August 2004. ISSN 0196-2892. doi: 10.1109/TGRS.2004.831865.

J.M.P. Nascimento and J.M.B. Dias. Vertex component analysis: A fast algorithm to unmix hyperspectral data. 43(4):898–910, 2005. ISSN 0196-2892. doi: 10.1109/TGRS.2005.844293. URL http://ieeexplore.ieee.org/document/1411995/.

Inés Pereira, S. Alcalde-Aparicio, M. Ferrer-Julià, M. F. Carreño, and E. García-Meléndez. Monitoring sedimentary areas from mine waste products with sentinel-2 satellite images: A case study in the se of spain. *European Journal of Soil Science*, 74(1):e13336, 2023. doi: 10.1111/ejss.13336.

Zhongliang Ren, Lin Sun, and Qiuping Zhai. Improved k-means and spectral matching for hyperspectral mineral mapping. *International Journal of Applied Earth Observation and Geoinformation*, 91: 102154, 2020. doi: 10.1016/j.jag.2020.102154.

Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987. doi: 10.1016/0377-0427(87)90125-7.

et al Salem. Supervised Classification Approaches to Analyze Hyperspectral Dataset. *International Journal of Image, Graphics and Signal Processing*, 7(5):42–48, April 2015. ISSN 20749074, 20749082. doi: 10.5815/ijigsp.2015.05.05.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(12):371–421, 2008.

Gary A. Shaw and Hsiao-hua K. Burke. Spectral imaging for remote sensing. *Lincoln Laboratory Journal*, 14(1):3–28, 2003. URL https://archive.ll.mit.edu/publications/journal/pdf/vol14_no1/14_1remotesensing.pdf.

Alexander Strehl and Joydeep Ghosh. Cluster ensembles – a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3:583–617, 2002.

G. A. Swayze, R. N. Clark, A. F. H. Goetz, K. E. Livo, G. N. Breit, F. A. Kruse, S. J. Sutley, L. W. Snee, H. A. Lowers, J. L. Post, R. E. Stoffregen, and R. P. Ashley. Mapping Advanced Argillic Alteration at Cuprite, Nevada, Using Imaging Spectroscopy. *Economic Geology*, 109(5):1179–1221, August 2014. ISSN 0361-0128, 1554-0774. doi: 10.2113/econgeo.109.5.1179.

Gregg Swayze, Roger N. Clark, Fred Kruse, and Andrea Gallagher. Ground-truthing AVIRIS mineral mapping at Cuprite, Nevada. In *Proceedings of the Third Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) Workshop*, pages 47–49, Pasadena, CA, 1992. U.S. Geological Survey and University of Colorado, JPL Publication 91-28.

Samuel T. Thiele, Gábor Kereszturi, Michael J. Heap, Andréa de Lima Ribeiro, Akshay V. Kamath, Maia Kidd, Matías Tramontini, Marina Rosas-Carbajal, and Richard Gloaguen. Hyperspectral mapping of density, porosity, stiffness, and strength in hydrothermally altered volcanic rocks. *Solid Earth*, 16:1249–1267, 2025. doi: 10.5194/se-16-1249-2025.

Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel J. Candes, and Aaditya Ramdas. Conformal Prediction Under Covariate Shift, July 2020.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

Jane-Ling Wang, Jeng-Min Chiou, and Hans-Georg Müller. Functional Data Analysis. 3(1):257–295, 2016. ISSN 2326-8298, 2326-831X. doi: 10.1146/annurev-statistics-041715-033624. URL `https://www.annualreviews.org/doi/10.1146/annurev-statistics-041715-033624`.

Jun Xia, Pedram Ghamisi, Naoto Yokoya, and Akira Iwasaki. Random forest ensembles and extended multiextinction profiles for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 56(1):202–216, 2018. doi: 10.1109/TGRS.2017.2744662.

Zhijing Ye, Jiaqing Chen, Hong Li, Yantao Wei, Guangrun Xiao, and Jon Atli Benediktsson. Supervised Functional Data Discriminant Analysis for Hyperspectral Image Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 58(2):841–851, February 2020. ISSN 0196-2892, 1558-0644. doi: 10.1109/TGRS.2019.2940991.

Roberta H. Yuhas, Alexander F. H. Goetz, and Joe W. Boardman. Discrimination among semi-arid landscape endmembers using the spectral angle mapper (sam) algorithm. In Robert O. Green, editor, *Summaries of the Third Annual JPL Airborne Geoscience Workshop, Volume 1: AVIRIS Workshop*, pages 147–149, Pasadena, CA, USA, June 1992. Jet Propulsion Laboratory (JPL), California Institute of Technology. JPL Publication 92-14, Vol. 1.

Zinno. Validation and enhancement of hyper-spectral unmixing algorithms. 2024.

A. Appendix A

A.1. Additional preprocessing diagnostics

This appendix provides supplementary diagnostics supporting key preprocessing decisions. Figure 14 illustrates differences between absolute reflectance (AREF) and relative reflectance (RREF) products on a representative mineral (Topaz). Figure 15 shows the impact of deleted channels (sentinel values) on an example mineral (Andradite), motivating their removal.

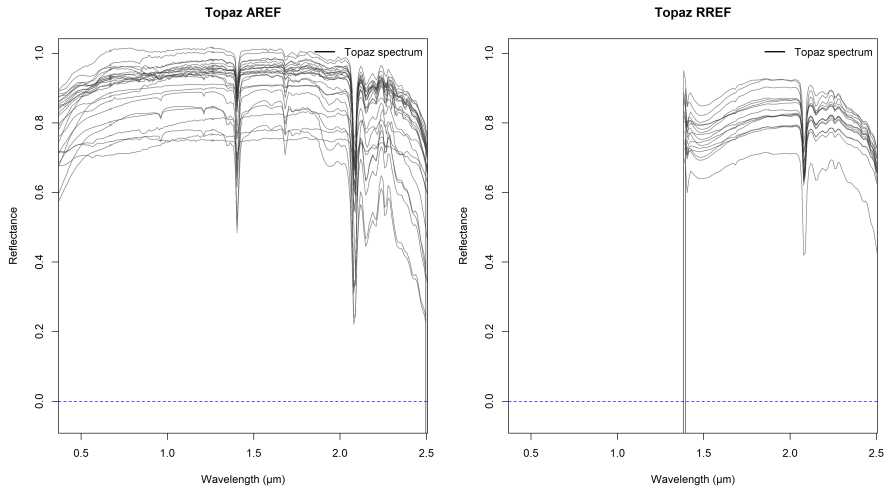


Figure 14: Spectral comparison between Absolute Reflectance (AREF, left) and Relative Reflectance (RREF, right) for Topaz samples. RREF records can exhibit scale and baseline differences, motivating the exclusive use of laboratory-validated AREF spectra in this work.

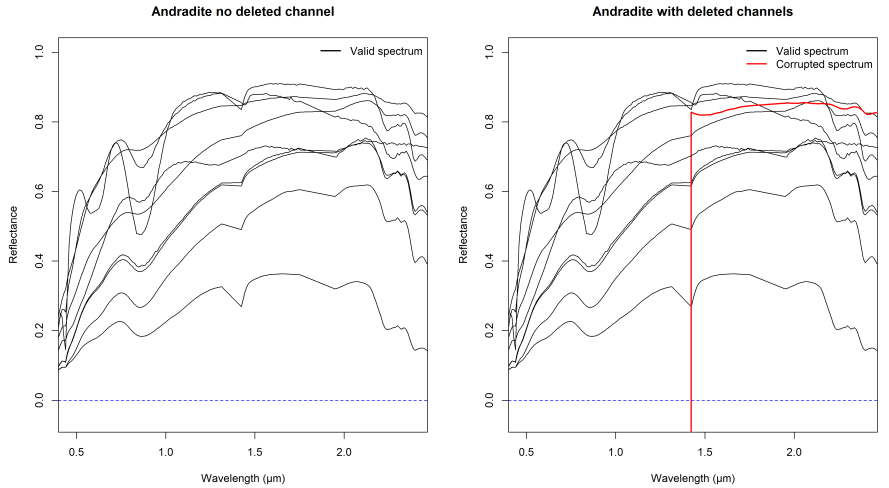
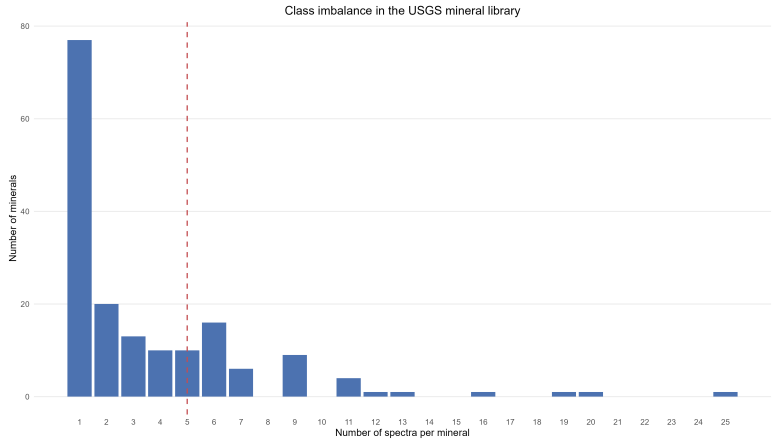


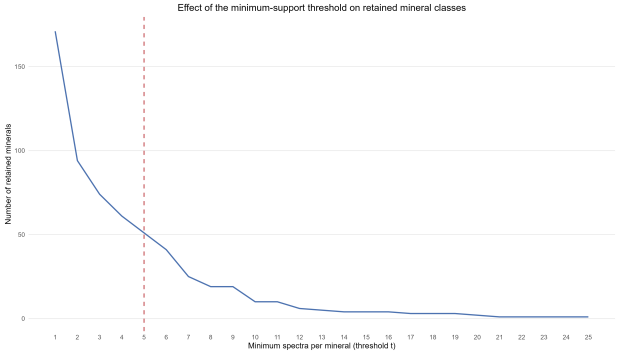
Figure 15: Impact of deleted channels on Andradite spectra. Red curves highlight records containing sentinel values (-1.23×10^{34}) within the retained wavelength grid. Such spectra are removed to avoid discontinuities and artefacts in subsequent smoothing and derivative extraction.

A.2. Threshold choice and stability under sample-size filtering

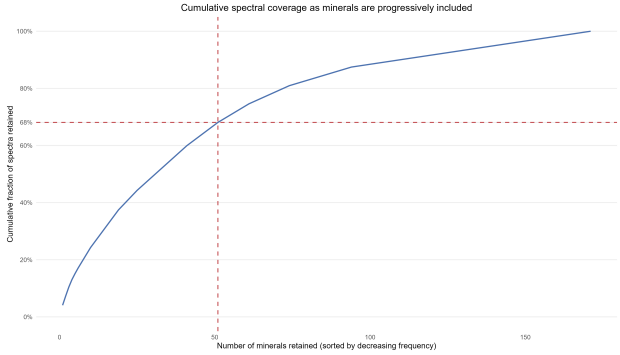
Figure 16 summarises the sample-size distribution across mineral-name entries and the rationale for selecting a minimum support threshold $n_{\min} = 5$.



(a) Sample-size distribution across mineral-name entries. The vertical dashed line marks the selected threshold ($n_{\min} = 5$).



(b) Number of mineral-name entries retained as a function of the minimum sample-size threshold.



(c) Cumulative fraction of spectra retained. The selection ($n_{\min} \geq 5$) retains $\approx 68\%$ of the available data after earlier quality filtering.

Figure 16: Sample-size distribution and threshold selection for mineral-name entries. The chosen threshold $n_{\min} = 5$ retains 51 entries and approximately 68% of spectra after earlier quality filtering, while discarding poorly supported entries.

A.3. Minerals in the library

Table 4 details the 51 mineral classes retained from the USGS Spectral Library (Chapter M) after applying the quality filters described in Section 3.1 (absolute reflectance, purity A–B, no deleted channels). This list represents the source dataset prior to the taxonomic aggregation performed in the methodology. The inclusion threshold was set to $N \geq 5$ to ensure sufficient sample size for cross-validation and probabilistic modeling.

Table 4: Complete distribution of the 51 selected mineral classes ($N \geq 5$). The dataset contains a total of 418 spectra.

Mineral	**Count**	**Mineral**	**Count**	**Mineral**	**Count**
Actinolite	16	Epidote	6	Monazite	9
Allanite	7	Galena	7	Montmorillonite	9
Almandine	11	Gibbsite	7	Muscovite	20
Alunite	13	Goethite	6	Natrolite	6
Andradite	9	Gypsum	7	Nontronite	5
Antigorite	9	Halite	6	Olivine	19
Beryl	7	Halloysite	5	Phlogopite	7
Bytownite	6	Hematite	12	Pyrophyllite	6
Calcite	6	Hornblende	6	Siderite	6
Carnallite	6	Illite	9	Smaragdite	6
Celestite	5	Jadeite	6	Strontianite	5
Chalcopyrite	5	Jarosite	11	Talc	9
Chlorite	11	Kaolinite	11	Topaz	25
Clinochlore	9	Lazurite	5	Ulexite	6
Datolite	5	Lepidolite	9	Uralite	6
Diaspore	6	Magnetite	9	Witherite	5
Dolomite	6	Malachite	5	Zoisite	5

A complete overview of the spectra is highlighted below:

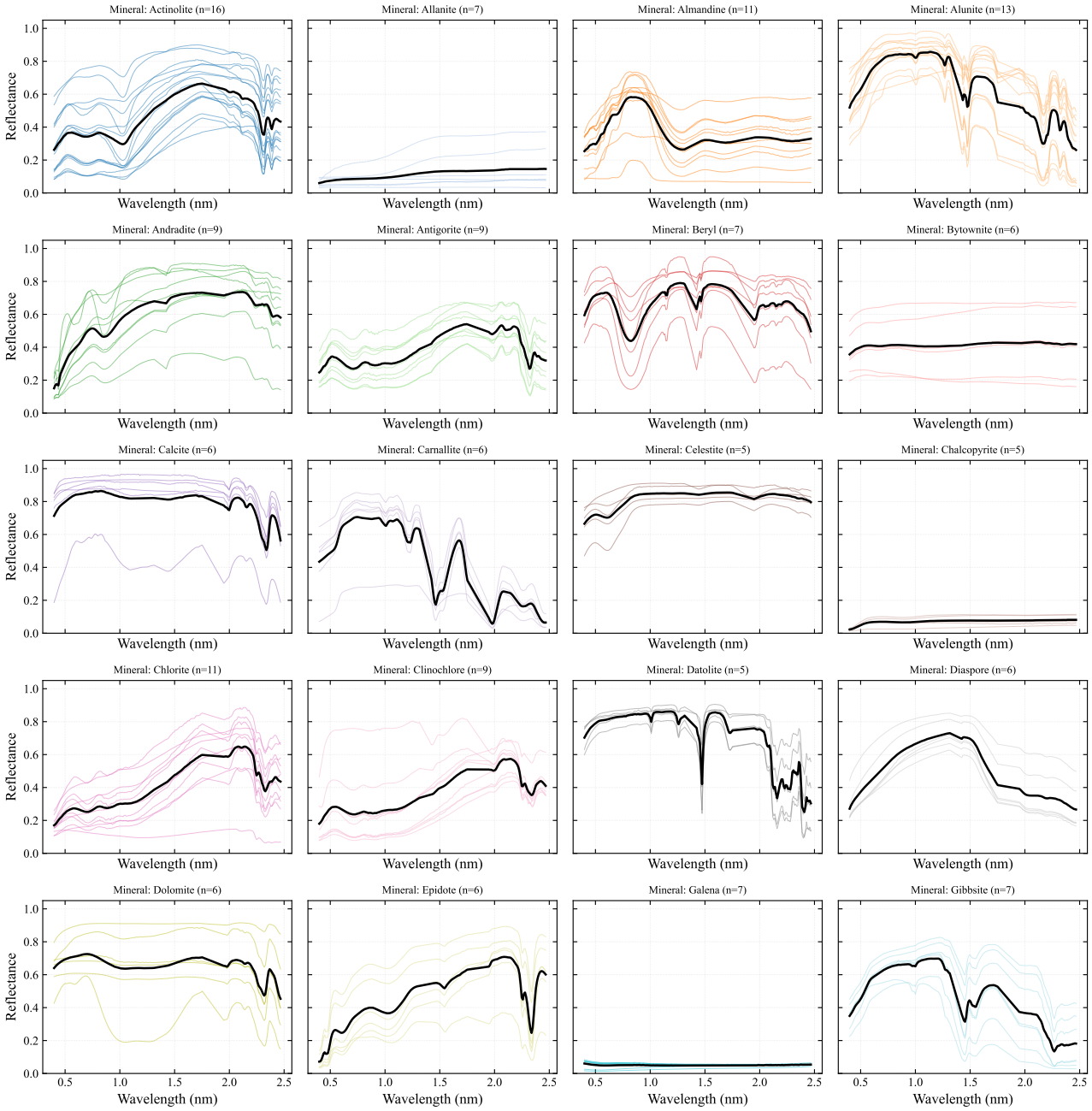
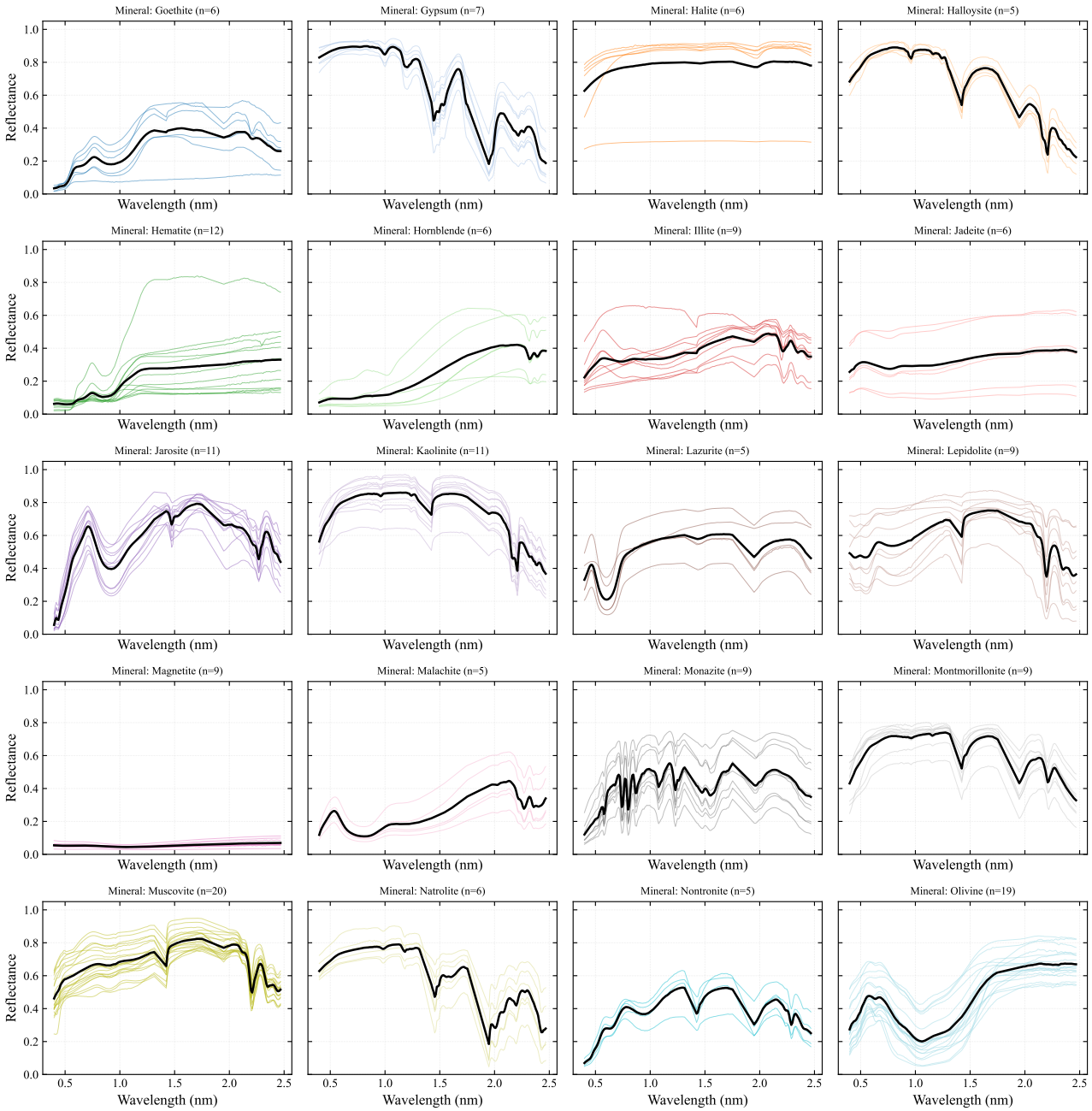
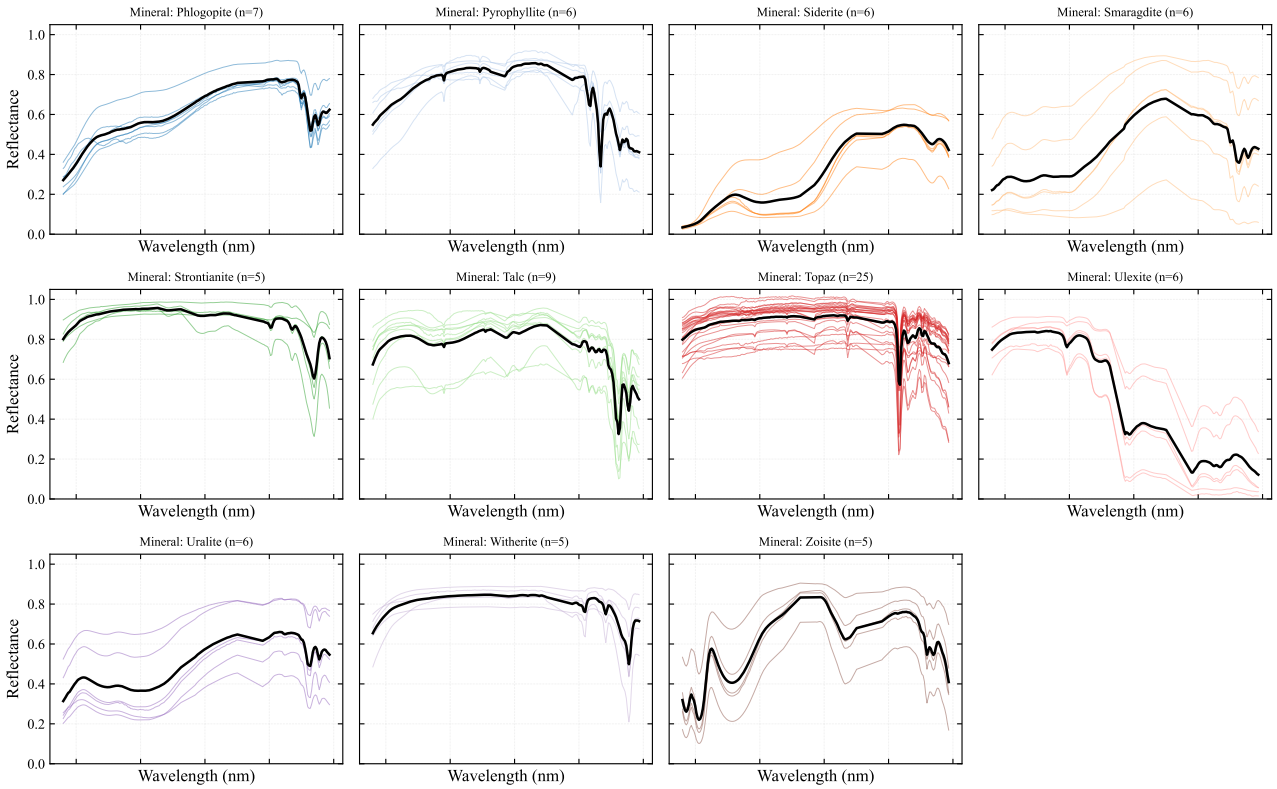


Figure 17: Minerals contained in the source dataset after the preprocessing step.


Mineral: Goethite (n=6)
Mineral: Gypsum (n=7)
Mineral: Halite (n=6)
Mineral: Halloysite (n=5)
Mineral: Hematite (n=12)
Mineral: Hornblende (n=6)
Mineral: Illite (n=9)
Mineral: Jadeite (n=6)
Mineral: Jarosite (n=11)
Mineral: Kaolinite (n=11)
Mineral: Lazurite (n=5)
Mineral: Lepidolite (n=9)
Mineral: Magnetite (n=9)
Mineral: Malachite (n=5)
Mineral: Monazite (n=9)
Mineral: Montmorillonite (n=9)
Mineral: Muscovite (n=20)
Mineral: Natrolite (n=6)
Mineral: Nontronite (n=5)
Mineral: Olivine (n=19)
Reflectance
Wavelength (nm)



B. Appendix B

B.1. Label generation

This appendix provides technical details for the construction and validation of the data-driven taxonomy used in the main pipeline.

PAM algorithm. Partitioning Around Medoids (PAM) [Kaufman and Rousseeuw, 1990] optimizes the objective in Eq. 15 via an iterative greedy search.

$$\min_{\mathcal{M}\subset\mathcal{X},\,|\mathcal{M}|=k} \sum_{x\in\mathcal{X}} \min_{m\in\mathcal{M}} d(x,m), \tag{15}$$

After an initialization step, it enters a *swap* phase where a current medoid $m \in \mathcal{M}$ is tentatively exchanged with a non-medoid object $x \in \mathcal{X} \setminus \mathcal{M}$. A *swap* is accepted if and only if it decreases the global objective. Two practical advantages motivate PAM for spectral data: (i) it supports arbitrary

Figure 16: Sensitivity of clustering performance to the number of clusters K under derivative preprocessing. Rows show (top) ARI, (middle) NMI, and (bottom) silhouette width. External indices (ARI/NMI) display a stability plateau starting around $K \approx 40$; SAM (blue) remains competitive with correlation-based metrics in recovering the original mineral-name labels. Silhouette scores differ across dissimilarities, reflecting geometric effects. We select $K = 40$ as a parsimonious trade-off between spectral discrimination and downstream sample-size constraints.

(possibly non-Euclidean) dissimilarities directly in the optimization, and (ii) cluster representatives are *observed* spectra (medoids), which preserves physical interpretability compared to centroid averages.

B.1.1 Selecting K and evaluation metrics

A key modeling choice is the number of clusters K. Since our downstream objective is to compare candidate dissimilarities under a common clustering granularity, we select a *global* value of K rather than a distance-specific optimum. We use a dual evaluation perspective distinguishing: (i) *external validity* (agreement with the original mineral-name entries) and (ii) *internal validity* (geometric compactness without reference labels).

For external validity we use the adjusted Rand index (ARI) [Hubert and Arabie, 1985] and normalized mutual information (NMI) [Strehl and Ghosh, 2002]. ARI compares a clustering $\mathcal{C}$ to a reference labeling $\mathcal{T}$ via pair counting corrected for chance:

$$\mathrm{ARI} = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right]/\binom{N}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right]/\binom{N}{2}}, \tag{16}$$

where n_{ij} counts the overlap between cluster c_i and reference class t_j, and $a_i = \sum_j n_{ij}$, $b_j = \sum_i n_{ij}$. NMI is defined through entropy H and mutual information I:

$$\mathrm{NMI}(\mathcal{T}, \mathcal{C}) = \frac{2\, I(\mathcal{T}; \mathcal{C})}{H(\mathcal{T}) + H(\mathcal{C})}. \tag{17}$$

For internal validity we use the silhouette width [Rousseeuw, 1987], defined for a sample i as

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \tag{18}$$

where $a(i)$ is the average dissimilarity to points in the same cluster and $b(i)$ is the minimum average dissimilarity to other clusters.
We track these metrics for $K \in [10, 100]$. As illustrated in Figure 16, we select $K = 40$ because it lies on a clear performance plateau for external validity while avoiding over-fragmentation given the limited sample size ($N = 418$). Choosing the numerical maxima at higher K would yield only marginal gains while reducing the average cluster size and destabilizing the subsequent supervised learning stage.

B.1.2 Dissimilarity selection via multi-start benchmarking

Fixing $K = 40$, we benchmark candidate dissimilarities under a multi-start protocol. We consider Euclidean, cosine, correlation, spherical [Hornik et al.], and spectral angle mapper (SAM) [Yuhas et al., 1992]. For each dissimilarity, PAM is run with $R = 10$ random initializations of the medoid set. This repeated-run setup is used because PAM may depend on initialization; comparing distributions of outcomes provides a notion of both performance and stability.

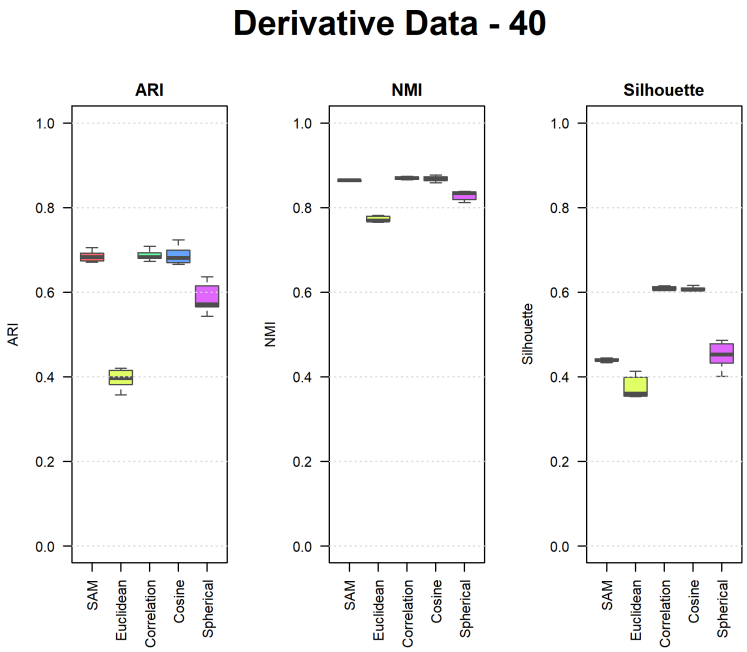


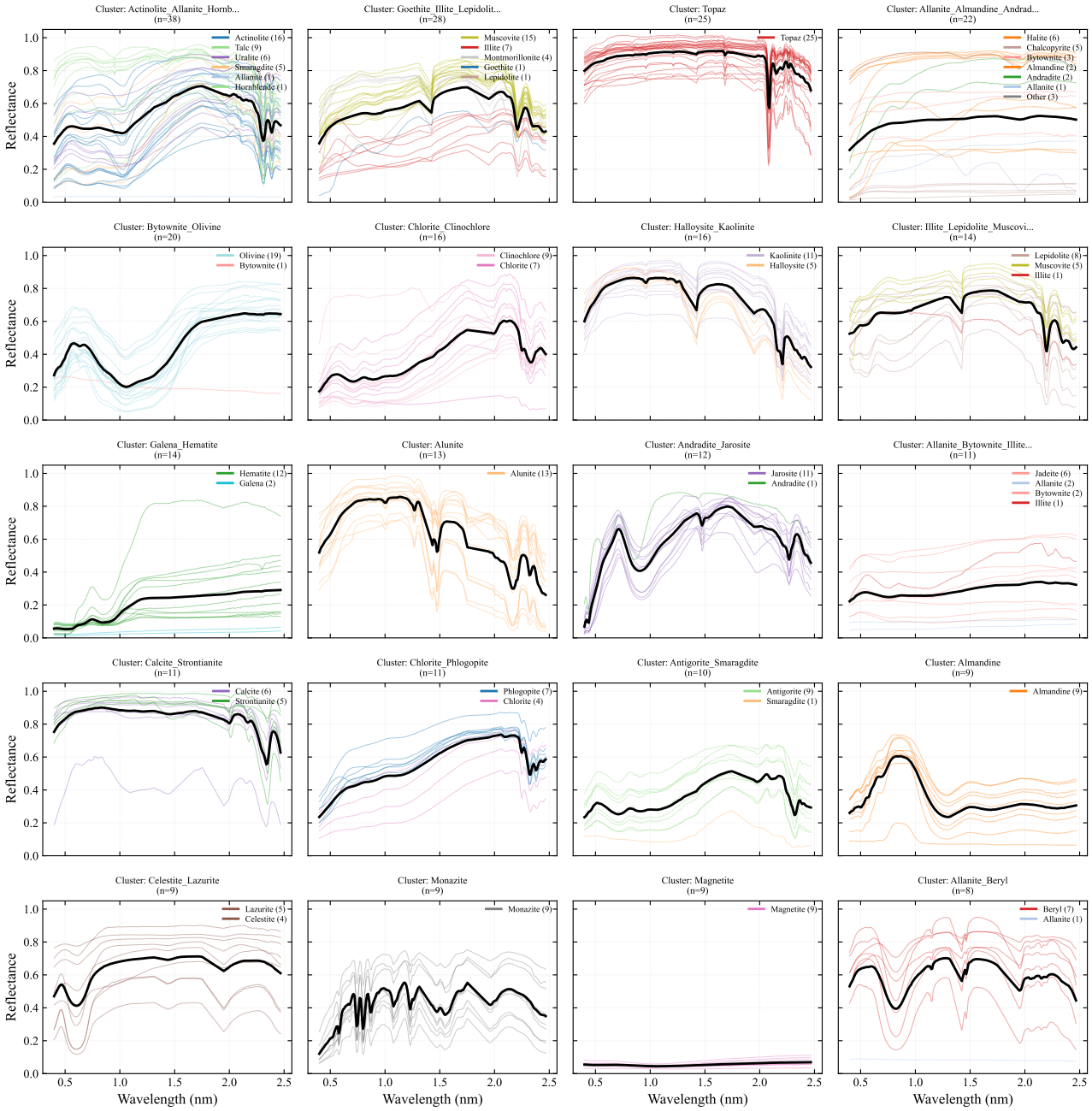
Figure 17: Multi-start benchmarking across dissimilarities under derivative preprocessing with $K = 40$. Boxplots show the distribution of (left) ARI, (middle) NMI, and (right) silhouette width over $R = 10$ PAM restarts. While some distances yield higher internal compactness (silhouette), we prioritize external agreement (ARI/NMI) to maintain interpretability with respect to the original mineral-name entries. SAM achieves high and stable external agreement, motivating its selection.

Figure 17, shows the results of the multi-start benchmarking. In our setting, we prioritize external agreement (ARI/NMI) because the resulting taxonomy should remain meaningfully connected to mineral-name entries, while also exhibiting coherent geometry. Under this criterion, SAM on first-derivative spectra provides consistently high agreement across restarts and is retained for the main pipeline.

B.1.3 Stability protocol and final partition

After selecting (K, d), we must choose a single final partition to define the taxonomy. To avoid circularity, we use external criteria (ARI/NMI) only for selecting K and the dissimilarity family, and we do *not* use them to select a specific run. Concretely, among the $R = 10$ restarts with fixed dissimilarity (SAM), we choose the run that maximizes the average silhouette width computed under the same dissimilarity. This yields a final partition that is internally coherent, while the observed external agreement with mineral-name labels remains an outcome rather than a selection target.

The final visual partition is shown below:



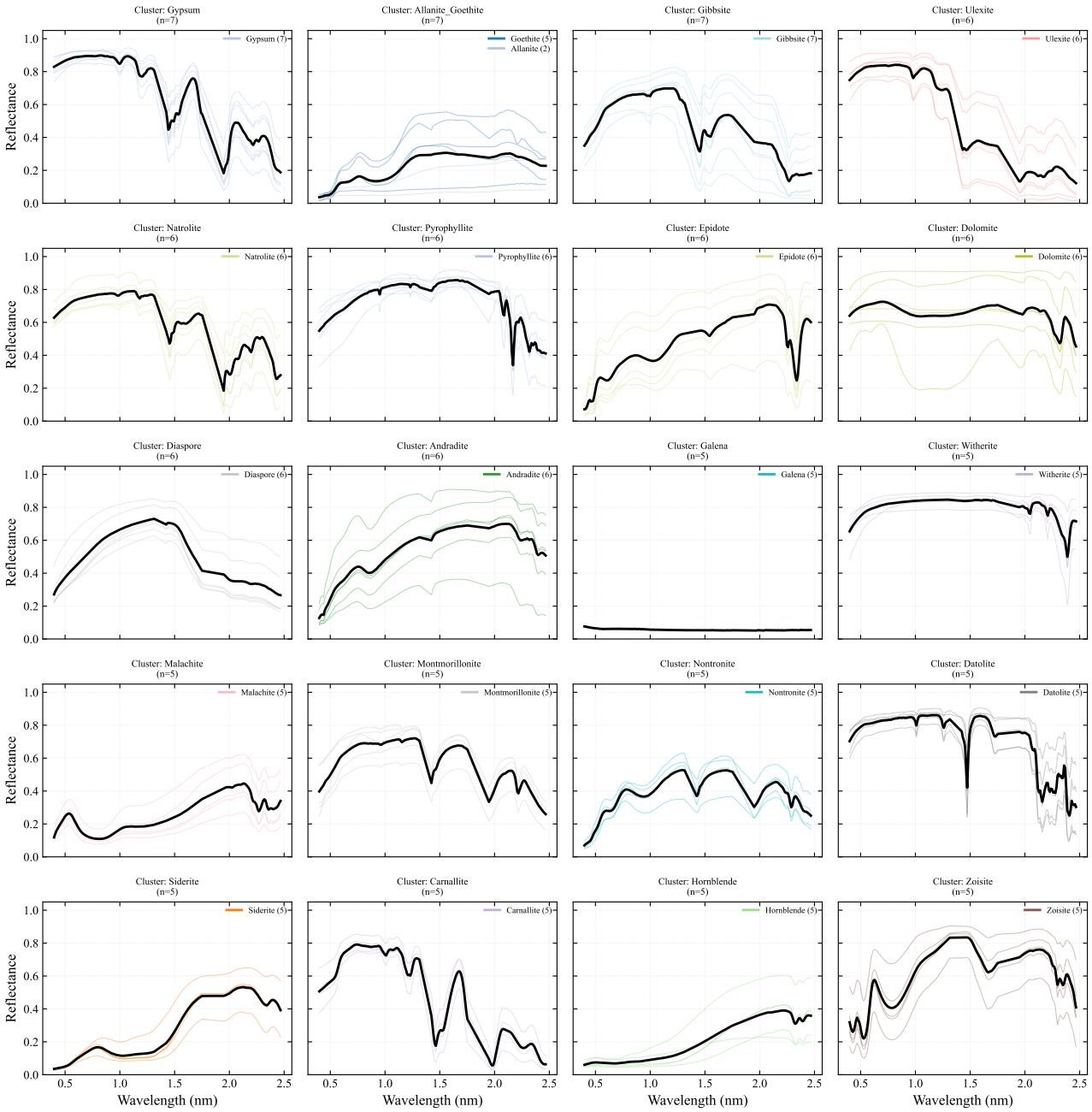
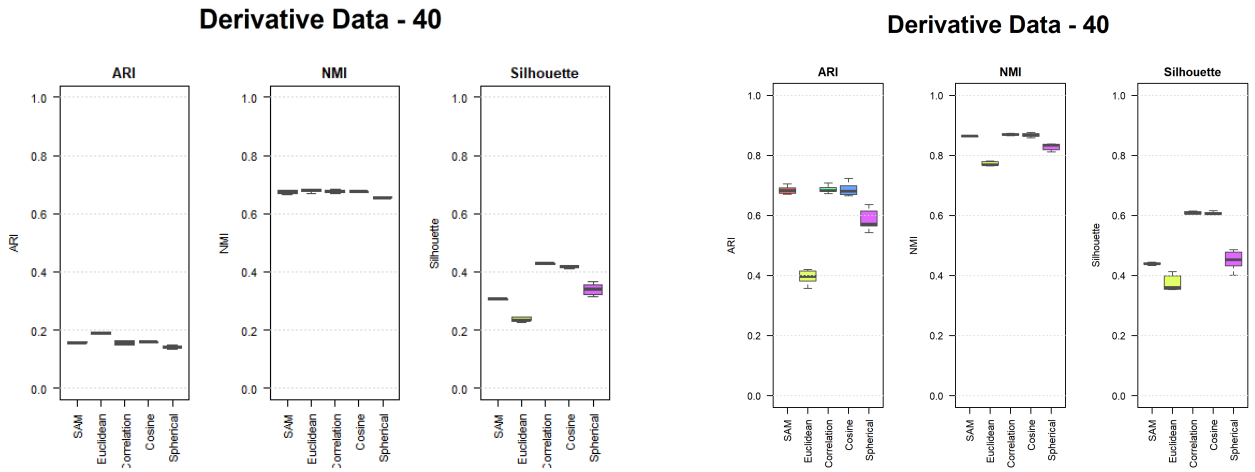


Figure 18: Clusters generated by the data driven process

B.1.4 Effect of sample-size filtering on clustering diagnostics

Having defined the clustering setup and evaluation metrics above, we assess how the minimum sample-size constraint used in Section 3.1 affects the stability of the resulting taxonomy. Figure 19 compares clustering diagnostics on the full derivative library and on the filtered library ($n_{\min} \geq 5$). For direct comparability, both settings are reported at a common number of clusters $K = 40$.



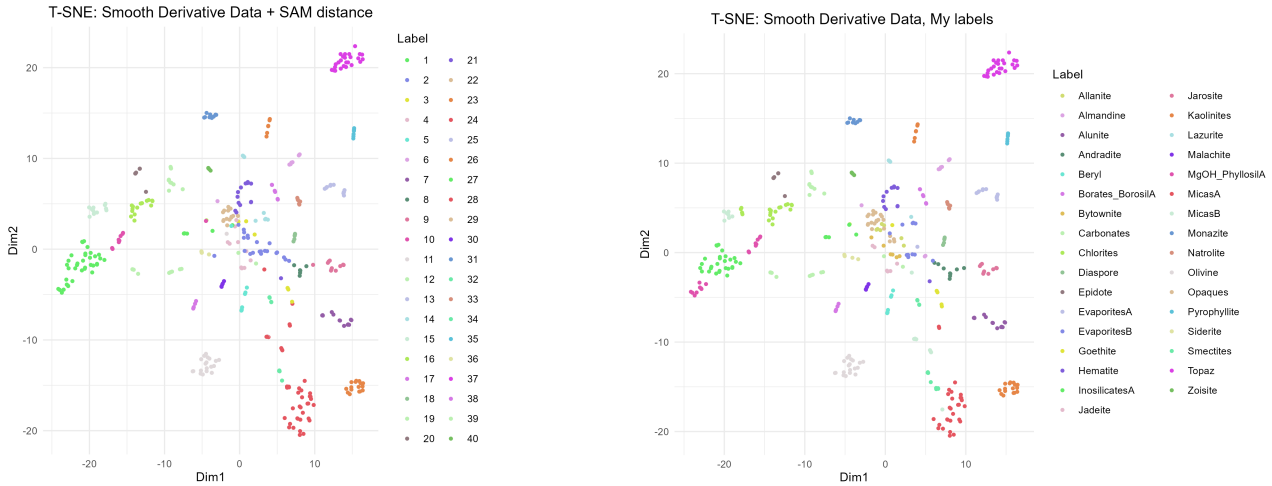
(a) Full library (b) Filtered library ($n_{\min} \geq 5$)

Figure 19: Clustering diagnostics on derivative spectra comparing the full library and the filtered library ($n_{\min} \geq 5$), reported at a common number of clusters $K = 40$. The full library exhibits lower external agreement and internal compactness, consistent with fragmentation induced by poorly supported mineral-name entries.

These diagnostics support the use of a minimum sample-size constraint to stabilise the label space prior to supervised learning and uncertainty quantification, without claiming optimality of the resulting taxonomy.

B.1.5 Visual validation via t-SNE embedding

Finally, we perform a qualitative visual inspection via t-distributed stochastic neighbor embedding (t-SNE) [van der Maaten and Hinton, 2008] on the derivative dataset. t-SNE provides a two-dimensional embedding designed to preserve local neighborhoods and is used here purely as a visualization tool, without any role in model selection or evaluation.
In addition to the PAM data-driven taxonomy, we also display a simulated mineralogical aggregation based on Mindat classes. This auxiliary grouping is not used in the main pipeline and is included solely to provide external context on how broad mineralogical groupings project onto the spectral geometry.



(a) Colored by PAM data-driven clusters ($K = 40$) (b) Colored by simulated Mindat aggregation ($K = 33$)

Figure 20: Two-dimensional t-SNE embedding of the derivative spectra. Panel (a) is colored by the PAM taxonomy ($K = 40$), and panel (b) by the simulated expert aggregation ($K = 33$). The embedding is used for qualitative inspection of neighborhood structure and class coherence.

C. Appendix C

This appendix provides additional implementation details and supplementary diagnostics to support the results in Section 5. We report (i) the final hyperparameters used for the Random Forest and SVM–RBF classifiers, (ii) a sensitivity analysis for the PCA dimension used in density-ratio estimation for weighted conformal calibration, and (iii) additional mineral-level qualitative comparisons between standard and weighted conformal outputs.

C.1. Parameters of Random Forest Classifier and SVM-RBF

The final parameters for the two classification models compared are:

Table 5: Hyperparameters used for model training

Model	Parameter	Value
	`class_weight`	balanced
	`max_depth`	15
	`max_features`	7
	`n_estimators`	500
	`min_samples_split`	2
	`min_samples_leaf`	1
SVC	`C`	100
	`gamma`	0.01
	`probability`	True

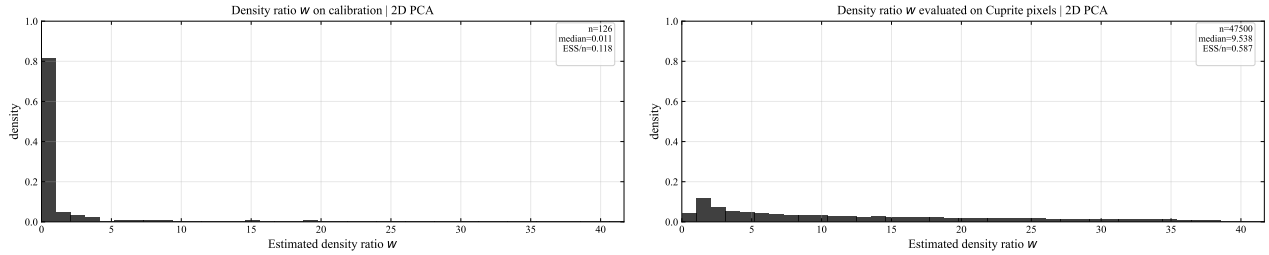
C.2. Sensitivity to the PCA dimension in importance weighting

We investigate the effect of the PCA dimensionality used for importance-weight estimation on the operational outputs of weighted conformal classification. In addition to the one-dimensional representation based on PC1 used in the main analysis, we consider a two-dimensional representation using PC1–PC2. For comparability, we use the same KDE bandwidth ($h = 0.5$) as in the 1D case. The bandwidth choice is guided by marginal alignment diagnostics and is not tuned for the 2D joint distribution; the goal here is qualitative sensitivity assessment rather than parameter optimisation.

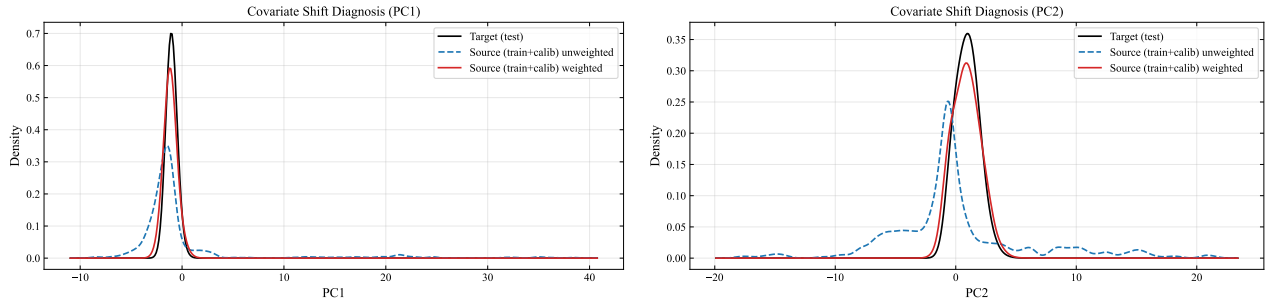
The marginal alignment diagnostics indicate that moving from a 1D to a 2D projection does not change the qualitative picture of covariate shift, and the resulting spatial patterns of prediction-set differences remain broadly consistent. However, the distribution of the estimated importance weights becomes less concentrated when using two principal components, which can affect the stability of the calibrated thresholds and hence the smoothness of operational summaries.

Figure 21 reports the mean prediction-set size as a function of the nominal coverage level $1 - \alpha$, comparing weighted conformal prediction with 1D (PC1) and 2D (PC1–PC2) PCA-based density-ratio estimation, together with standard conformal prediction. Standard conformal yields near-zero mean set sizes over most of the range, consistent with the prevalence of empty prediction sets under covariate shift. Both weighted variants produce substantially larger sets, reflecting the elimination of degenerate outputs. The 2D weighting scheme yields a smoother and more gradually increasing set-size curve across coverage levels, whereas the 1D scheme exhibits more pronounced stepwise changes and sharper increases at high nominal coverage. Overall, increasing the PCA dimension from one to two components primarily affects the numerical stability of the estimated density ratios and hence the smoothness of the calibrated thresholds. While the KDE smoothing is performed jointly in the $(\mathrm{PC1}, \mathrm{PC2})$ space, the bandwidth h is chosen based on marginal considerations and is not optimised for the two-dimensional joint distribution. Under a fixed bandwidth and finite calibration sample size,

density-ratio estimation in higher dimension can be more variable, which in turn may lead to more conservative calibration and larger prediction sets at high nominal coverage levels. Note that, unlike Figure 12, here we report a wider range of coverage levels (corresponding to $\alpha \in [0.01, 0.10]$).



(a) Distribution of estimated density ratios $\widehat{w}$ with 2 PCs on calibration points (used for WCC) and on Cuprite pixels (diagnostic).



(b) Distribution of estimated density ratios $\widehat{w}$ with 2 PCs on calibration points (used for WCC) and on Cuprite pixels (diagnostic).

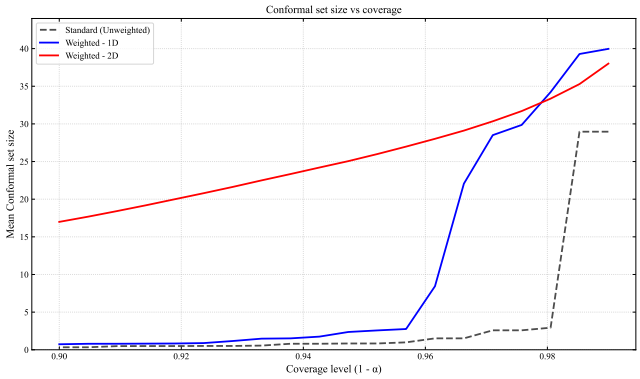
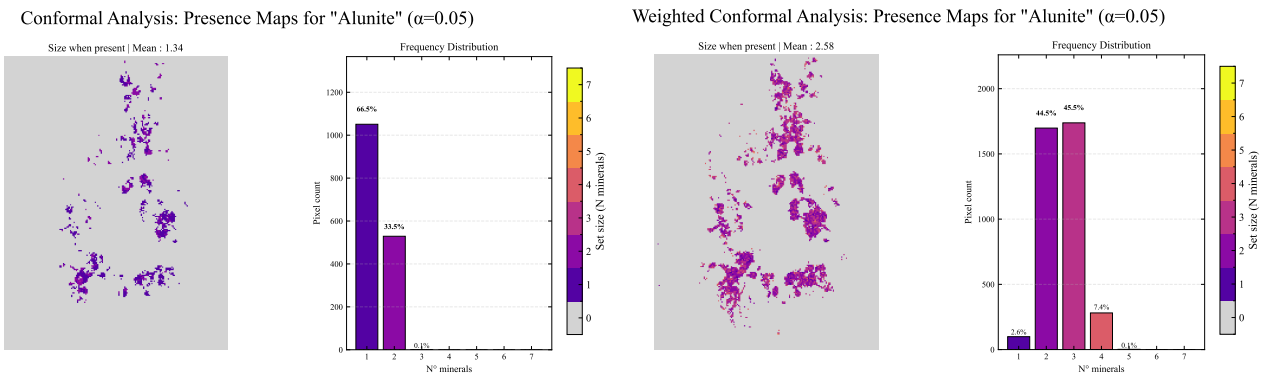


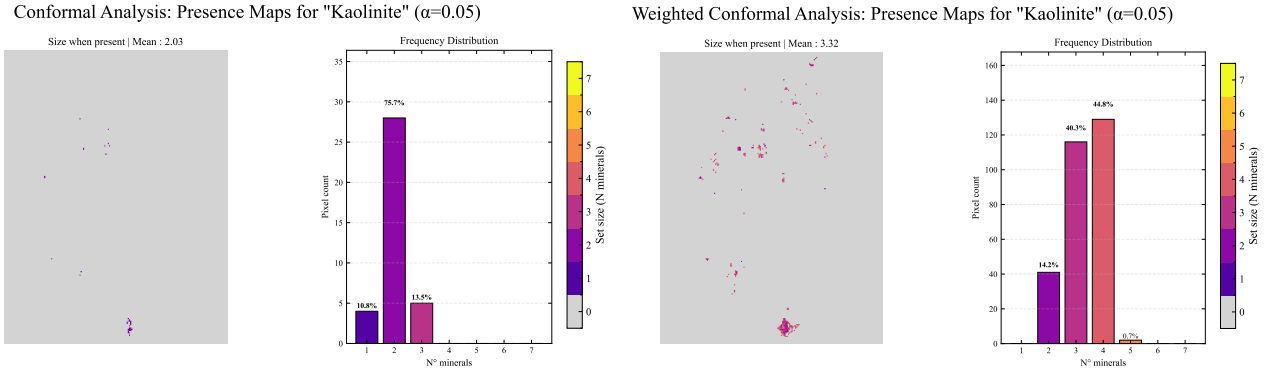
Figure 21: Prediction set sizes vs alpha comparing 2D PCA with 1D PCA and Standard Conformal prediction

C.3. Additional mineral-level examples: standard vs weighted conformal outputs

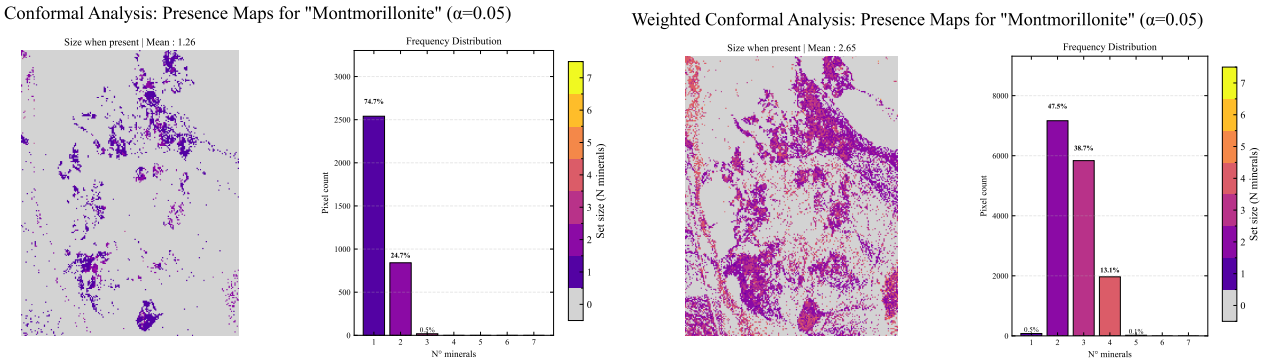
Figure 22 reports mineral-level qualitative comparisons between standard and weighted conformal classification at $\alpha = 0.05$. We include alunite (well-supported, nearly pure cluster), kaolinite–halloysite (intermediate cluster with moderate spectral support), and montmorillonite (composite or weakly supported clusters). Across minerals, weighting systematically reduces degenerate outputs (empty sets) and shifts the distribution towards larger prediction sets, with the magnitude of the change depending on spectral overlap and class support.



(a) Alunite: standard (left) vs weighted (right).



(b) Kaolinite–Halloysite: standard (left) vs weighted (right).



(c) Montmorillonite: standard (left) vs weighted (right).

Figure 22: Mineral-level comparison of conformal classification outputs at $\alpha = 0.05$. For each mineral, we report the per-pixel prediction-set size map and the empirical distribution of set sizes across pixels (bar chart), under standard split conformal calibration (left) and importance-weighted conformal calibration (right).

Abstract in lingua italiana

Il mineral mapping con dati iperspettrali (tramite classificazione o unmixing spettrale) è una tecnica che spesso non fornisce garanzie esplicite sull'affidabilità delle predizioni, soprattutto quando modelli addestrati su librerie spettrali vengono applicati su grandi zone di telerilevamento, dove mix di minerali e artefatti di acquisizione possono alterare le firme spettrali. Questa tesi propone la conformal prediction come approccio pratico per integrare la quantificazione dell'incertezza nella mappatura mineralogica con dati iperspettrali, producendo per ogni pixel un insieme di etichette plausibili e controllando, a un livello scelto dall'utente, la probabilità (scelto a caso un pixel) di includere la vera classe mineralogica.
Come caso applicativo utilizziamo la mappa iperspettrale AVIRIS nella regione del Cuprite (Nevada, USA). Addestriamo dunque un Random Forest probabilistico sulla libreria spettrale USGS e definiamo lo spazio delle classi tramite una tassonomia mineralogica data-driven. Poiché non sono disponibili dati sulla reale composizione mineralogica pixel per pixel sull'intera scena, valutiamo i risultati del metodo usato mediante indicatori operativi (dimensione degli insiemi, frequenza di insiemi vuoti e distribuzione spaziale dell'incertezza). Per ridurre gli effetti della differenza di distribuzione tra spettri della libreria e i pixel di Cuprite, adottiamo una versione "weighted" della conformal prediction basata sulla costruzione di pesi per i dati di calibrazione.
I risultati mostrano che la standard conformal classification produce sulla mappa del Cuprite una percentuale non trascurabile di insiemi vuoti (22,9% dei pixel a un livello di copertura scelto del 95%), mentre la weighted conformal elimina gli insiemi vuoti allo stesso livello di copertura, a fronte di insiemi mediamente più ampi nelle aree più ambigue. Nel complesso, il lavoro fornisce un framework operativo per ottenere mappe mineralogiche iperspettrali con una rappresentazione esplicita dell'incertezza, anche in presenza di distribuzione diversa tra dati di addestramento e test.

Parole chiave: qui, le parole chiave, della tesi, in italiano

Acknowledgements

This work has been partially supported by ACCORDO Attuativo ASI- POLIMI "Attività di Ricerca e Innovazione" n. 2018-5-HH.0, collaboration agreement between the Italian Space Agency and Politecnico di Milano.