Spectral Provenance: A Labelled Corpus for Recovering Processing History from Reflectance Spectra

Satyajeeth Suresh Kannan
Innovation Academy · Mary's Lab
github.com/JeetuSK0808/maryslab

Datasheet · August 2026 · Mary's Lab dataset DS-01 · CC-BY-4.0

ABSTRACT

A datasheet for the `spectral-provenance` dataset: 4152 reflectance spectra on a fixed 38-sample grid, each labelled with its processing history. The positive class is real — 1752 published spectra from the public-domain USGS Spectral Library Version 7 — and the other four classes are those same spectra put through a recorded operation, plus analytic curves with no measurement history at all. The task is to recover the history from the numbers. It ships with a documented baseline failure rather than a solved leaderboard: a multinomial logistic regression on the four noise-floor statistics the field would reach for first scores 0.696 overall and 0 recall on the smoothed class — every smoothed spectrum in the test split is mistaken for a published one. Splits are grouped by source record so the test set is genuinely held out.

Keywords: dataset; data forensics; provenance; spectroscopy; reflectance; tabular classification

1. Motivation

Spectral libraries are cited as evidence, and the label "measured" on a spectrum is a claim about provenance that the numbers themselves do not carry. Files get smoothed for presentation, resampled onto new grids, and occasionally generated from models, and none of those operations changes the column headers. Mary's Lab built a provenance detector (TR-13) on the premise that measurement leaves a noise floor processing removes — and found, against a real published library, that the premise fails: published data is already processed, so the signature a naive detector looks for is gone.

That failure is the motivation. A task on which the obvious method demonstrably fails, with real positives and exactly recorded negative operations, is the kind of benchmark worth releasing — the gap between 0.696 and 1.0 is open and the full 38-sample spectrum is in every row, so shape-aware models have obvious headroom.

2. Composition

4152 rows, one spectrum each, in five classes. Every spectrum sits on the canonical grid: 380–750 nm at 10 nm, 38 samples, values in [0,1].

Table 1. Classes. The published class is as-distributed USGS data; every derived class records the exact operation applied.

label	rows	what it is
published	1752	as distributed by USGS (splib07a)
smoothed	600	two passes of a 3-tap moving average
interpolated	600	decimated 3× and linearly re-interpolated
noise_added	600	uniform ±0.01 added per sample
modeled	600	analytic Gaussian reflectance, never measured

Each row also carries the baseline detector's four scalar statistics (measured-likeness, high-frequency energy ratio against a Savitzky–Golay quadratic fit, roughness, distinct-value ratio) and its raised artifact flags, so the published baseline is reproducible from the data alone.

Splits

Train 3361, validation 390, test 401 — grouped by *source record*, not split at random. All variants derived from one spectrum land on the same side, because a random row split would let a model score by recognising the mineral rather than the processing.

3. Collection and labeling

Published spectra come from USGS Spectral Library Version 7 (Kokaly et al. 2017), a US federal work in the public domain, ingested through quality gates Q1–Q6 (provenance completeness, physicality, spectral coverage, finiteness, deduplication, diversity) and resampled to the canonical grid; snapshot id `usgs-splib07a-1`. Labels for derived classes are not annotations — they are the operations themselves, applied by the seeded generator, so label noise on the derived classes is zero by construction. The one label carrying an external claim is `published`, and it means exactly "as USGS distributed it", not "raw sensor output".

4. Uses and baseline

Intended uses: dataset-integrity screening (rank a library by processing evidence and inspect the tail), benchmarking provenance detectors against a documented failure, and teaching the difference between "measured" and "as published".

Table 2. Reference baseline: multinomial logistic regression on the four scalar features, trained on train, reported on test. Reproduce with node datasets/build/baseline.mjs.

quantity	value
accuracy	0.696
majority-class floor	0.424
recall, published	0.941
recall, noise_added	0.93
recall, interpolated	0.86
recall, modeled	0.283
recall, smoothed	0

The two failures are the interesting rows. Modeled collapses into published (43 of 60 test rows), which is TR-13's finding restated as a benchmark: noise-floor statistics cannot separate an analytic curve from a published spectrum. And smoothed is not detected at all — recall 0 — because smoothing an already-processed curve leaves nothing for four summary statistics to see. A model reading the full spectral shape is the obvious next attempt, and the dataset is built so that attempt is honest: grouped splits, recorded operations, no leakage.

5. Distribution, license, and maintenance

Distributed as JSONL splits with a dataset card carrying the same statistics as this datasheet. USGS source data is public domain (cite Kokaly et al. 2017); the derived rows and the compilation are CC-BY-4.0. The generator is `datasets/build/` in the repository and is seeded — the same commit reproduces the same bytes, and this datasheet's numbers come from the generator's own `stats.json`. Engine 0.4.1, snapshot usgs-splib07a-1. Maintained with the repository; corrections arrive as a new dataset version, never as a silent overwrite.

6. Limitations

This section states what the result does not establish. It is written to be read before the abstract is quoted.

- **"Published" is a pipeline, not ground truth about measurement.** USGS splices spectrometers, corrects against references, and resamples. A classifier separating published from modeled learns to recognise a particular processing pipeline — useful for integrity screening, not the same as detecting fabrication.
- **Derived operations are a sample, not a census.** Two smoothing passes, one decimation factor, one noise amplitude. A detector that wins here has seen these operations, not all operations.
- **One source library.** Every published spectrum is USGS splib07a. Generalisation to other libraries' pipelines is an open question the dataset cannot answer about itself.
- **Grid resolution bounds the claim.** 38 samples at 10 nm; structure below that scale is not represented. TR-13 checked the baseline's failure at native resolution (up to 4,595 channels) and it persists, but models trained here see only the 38-sample view.

7. Acknowledgements and disclosure

The instruments described here were implemented in Mary's Lab, an open-source computational color science project. Software design, implementation, and the preparation of this report were carried out with substantial assistance from a large language model (Anthropic Claude), under the author's direction and review. All numerical results were produced by executing the released code; none were generated by a language model. The author is responsible for the claims made here.

References

1. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., and Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12), 86–92.
2. Kokaly, R. F., et al. (2017). *USGS Spectral Library Version 7*. U.S. Geological Survey Data Series 1035. doi:10.3133/ds1035.
3. Savitzky, A. and Golay, M. J. E. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, 36(8), 1627–1639.

4. Kannan, S. S. (2026). *Too Smooth to Be Measured: What a Noise-Floor Provenance Detector Can and Cannot Tell You About a Published Spectral Library*. Mary’s Lab technical report TR-13.

Generator, statistics, and this datasheet's renderer: `github.com/JeetuSK0808/maryslab` under `datasets/build/`. The build is seeded — the same commit reproduces the same dataset byte for byte — and every number above resolves from the generator's own stats.json. Datasheet structure follows Gebru et al. (2021). Prepared with AI assistance (see Acknowledgements).