WDTS Training Dataset — User Guide

Data file: `WDTS_training_dataset.csv` — 3.0 MB · 1,083 data rows × 268 columns
Companion files: `BadBand.csv` — 213 rows × 3 columns, the band mask (see section 6.1a); and `quarry_spec_10nm.csv` — 6 rows × 214 columns, the per-year quarry reference spectra used for cross-year normalization (see section 5.1).
Distribute all three files together.
Format: UTF-8, comma-delimited, one header row
Guide generated: 2026-09-10

---

1. Overview

This dataset pairs **canopy-level imaging-spectroscopy reflectance** with **co-located foliar functional-trait measurements**, assembled as a training set for empirical (e.g. PLSR) spectra-to-trait models. Each row is one plot–year–scene observation: a single reflectance spectrum plus whatever traits were measured for that plot. It merges four source campaigns spanning the continental United States, 2008–2018:

`Dataset`	Rows	Years	Plots	Region	Lat range	Lon range
`FFT`	465	2008–2011	192	Upper Midwest / Northeast	39.53 – 47.84	−92.83 – −73.74
`CA2013`	197	2013	75	California	33.52 – 38.05	−121.77 – −116.16
`CA2018`	49	2018	47	California	34.00 – 37.17	−119.82 – −117.70
`NEON_avs`	372	2017–2018	4 sites	Sierra Nevada, CA	36.95 – 37.08	−119.32 – −118.97
Total	**1,083**	**2008–2018**	**318**		**33.52 – 47.84**	**−121.77 – −73.74**

Overall bounding box (WGS84):

```
W  -121.765687
S    33.517759
```

```
E   -73.741331
N    47.838408
```

The four campaigns differ substantially in which traits were measured, in ancillary metadata, and in spectral quality. **Read section 6 before using the data.**

---

2. File structure

268 columns in three blocks:

Block	Columns	Count	Contents
A. Identifiers & location	1–10	10	IDs, coordinates, plot and species composition
B. Traits & uncertainty	11–55	45	21 traits + 21 SD columns + `needle_f`, `Dataset`, `Date`
C. Reflectance spectrum	56–268	213	Numeric wavelength headers `380` … `2500`

Rows are unique: no duplicate `plotyear-scene`, and no duplicate `(lat, lon, Year)` triples.

The distribution also includes two companion files, both keyed to the same 213 spectral bands:

File	Shape	Purpose
`BadBand.csv`	213 × 3	Band mask and FWHM. **Required** to use block C correctly — see section 6.1a.
`quarry_spec_10nm.csv`	6 × 214	Per-year quarry reference spectra used to normalize across acquisition years — see section 5.1.

---

3. Block A — identifiers and location

Column	Type	Non-null	Description
`plotyear-scene`	string	1083	**Primary key.** `<plotyear>-<scene>` for airborne-scene records; for `NEON_avs` there is no scene suffix and the value equals `plotyear`.
`lat`	float	1083	Plot centroid latitude, decimal degrees, **WGS84 assumed [TBD — confirm datum]**.

Column	Type	Non-null	Description
`lon`	float	1083	Plot centroid longitude, decimal degrees, negative = west.
`plotyear`	string	1083	`<Plot>_<Year>`. 715 unique values.
`Plot`	string	1083	Field plot identifier. 318 unique. Naming differs by campaign — see below.
`Year`	int	1083	Calendar year of acquisition/sampling. 2008–2018.
`spp1`	string	465	Dominant species, 4-letter USDA-style code. `FFT` **only.**
`rba1`	float	465	Relative basal area of `spp1`, fraction 0.18–1.00.
`spp2`	string	462	Second species code. `FFT` only.
`rba2`	float	462	Relative basal area of `spp2`, fraction 0.01–0.42.

Plot naming by campaign:

- `FFT` — site code + plot number, e.g. `DC03`, `AK14B`, `BH03`
- `CA2013` — `T_###`, e.g. `T_007`
- `CA2018` — `H3_####`, e.g. `H3_0010`
- `NEON_avs` — `Plot` holds the **site–year** (`SOAP_2017`, `SOAP_2018`, `TEAK_2017`, `TEAK_2018`); the individual sample id is in `plotyear`, e.g. `SOAP_2017_0099`

Scene identifiers. The suffix of `plotyear-scene` is an AVIRIS-style flight-line ID, e.g. `f130626t01p00r06` (`f` + `YYMMDD` + run/pass/reflectance tags).
100 unique scenes are represented.

---

4. Block B — traits

4.1 Trait columns and summary statistics

Each trait `X` has a paired `X_sd` column.

Trait	Inferred unit	n	Min	Median	Mean	Max	SD
`LMA`	g m^{-2}	1081	37.03	163.25	174.09	436.51	86.96
`Carbon`	mg g^{-1}	1083	410.56	507.80	513.19	582.95	26.00
`Nitrogen`	mg g^{-1}	1082	5.70	18.63	19.16	47.07	6.19
`Cellulose`	mg g^{-1}	886	41.30	152.73	153.64	266.73	29.51
`Lignin`	mg g^{-1}	886	80.44	204.55	205.59	367.60	38.68
`Fiber`	mg g^{-1}	885	10.02	336.50	340.78	530.40	74.09

Trait	Inferred unit	n	Min	Median	Mean	Max	SD
`Starch`	mg g^{-1}	884	11.03	41.68	41.78	77.11	9.61
`Sugar`	mg g^{-1}	884	108.26	189.97	192.14	313.52	35.13
`NSC`	mg g^{-1}	884	126.32	248.89	252.26	396.44	53.11
`Phenolics`	mg g^{-1}	878	21.88	71.22	73.71	173.72	22.65
`Flavonoids`	mg g^{-1}	884	5.06	77.22	76.21	141.54	24.09
`Phosphorus`	mg g^{-1}	883	0.63	1.48	1.53	2.89	0.42
`Potassium`	mg g^{-1}	882	0.73	7.56	7.68	19.64	2.70
`Calcium`	mg g^{-1}	881	0.83	9.30	9.95	30.07	3.95
`Sulfur`	mg g^{-1}	884	0.30	1.27	1.32	2.61	0.41
`d13C`	‰ vs VPDB	884	−31.72	−26.21	−26.25	−21.29	1.63
`d15N`	‰ vs AIR	886	−7.98	−2.61	−2.88	1.38	1.55
`Chlorophylls_area`	µg cm^{-2}	837	19.16	46.00	45.96	70.43	7.39
`Carotenoids_area`	µg cm^{-2}	837	4.55	10.09	10.16	15.75	1.54
`Chlorophylls_mass`	mg g^{-1}	836	0.41	4.54	5.24	10.50	2.20
`Carotenoids_mass`	mg g^{-1}	837	0.23	0.97	1.07	1.93	0.34

`NSC` = non-structural carbohydrates.

4.2 `*_sd` uncertainty columns

Every `X_sd` column is populated for **465 rows only — exactly the `FFT` subset** (463 for a few traits). They represent the standard deviation across model permutations; see Singh et al. (2015) for details.

4.3 Other Block B columns

Column	Non-null	Description
`needle_f`	465 (`FFT` only)	**Fraction of conifers in the plot, by basal area**, 0–1 — the same basis as `rba1` / `rba2`. 162 rows = 0 (no conifers), 11 rows = 1 (entirely conifer), 292 rows mixed. Useful for stratifying broadleaf vs. needleleaf canopies, whose spectra differ markedly.
`Dataset`	1083	Source campaign: `FFT`, `CA2013`, `CA2018`, `NEON_avs`. **Use this to stratify.**
`Date`	49 (`CA2018` only)	Acquisition date as integer `YYYYMMDD`: 20180701 (10), 20180702 (6), 20180710 (21), 20180711 (12). For all other rows, recover the date from the scene ID.

---

5. Block C — reflectance spectra

- **213 bands**, 380–2500 nm, **uniform 10 nm spacing**, and a constant **10 nm FWHM** (from `fwhm` in `BadBand.csv` — every band)
- Column headers are **bare wavelength integers in nm** (`380`, `390`, … `2500`).
- **Units: fractional reflectance (0–1), not scaled integers.** Median values are ~0.034 in the red (680 nm) and ~0.27 in the NIR plateau (780–1080 nm), consistent with vegetated canopy reflectance.
- **No missing values** — 0 NaN cells across all 230,679 spectral cells.
- Spectra have been resampled to a common 10 nm grid from campaign-specific sensors, see Shafron et al., (2025) for details.
- **The bands are not all usable.** 66 of the 213 are excluded by the default mask in `BadBand.csv` — see section 6.1a.
- **Spectra are normalized across acquisition years** against a quarry reference target, supplied in `quarry_spec_10nm.csv` — see section 5.1.

Reference percentiles at selected wavelengths (all 1,083 rows):

nm	min	p05	median	p95	max
480	0.0028	0.0183	0.0343	0.0494	0.0749
680	0.0062	0.0120	0.0320	0.0653	0.1205
780	0.1192	0.1822	0.2839	0.4916	0.6353
1080	0.1495	0.2070	0.3002	0.4937	0.6340
1680	0.0563	0.0940	0.1530	0.2095	0.2900
2180	0.0254	0.0411	0.0680	0.1142	0.1842

5.1 Companion file: `quarry_spec_10nm.csv`

The spectra in this dataset were acquired across several years, and sensor and atmospheric conditions drift between campaigns. To make them comparable, they are normalized against a **quarry** — a bright, spectrally stable, pseudo-invariant ground target imaged in each year's AVIRIS-Classic flights. This file holds that reference spectrum, one row per acquisition year.

The normalization method itself is documented in the WDTS AVIRIS-C foliar trait user guide — Zheng et al. (2025), ORNL DAAC, https://daac.ornl.gov/WDTS/guides/WDTS_AVIRIS-C_foliar_trait_15m.html. Consult it before applying or reversing the normalization; this section describes only the file's contents.

Structure. 6 rows × 214 columns.

Column	Description
`year`	Acquisition year. One row each for 2013–2018.
`380` … `2500`	213 reflectance values, **byte-identical headers** to the main file's spectral block and to `WVL(nm)` in `BadBand.csv` — same order, same 10 nm grid.

Contents by year. Statistics below are over the 147 retained bands only:

`year`	Min	Median	Max
2013	0.1762	0.2086	0.2299
2014	0.1763	0.2141	0.2327
2015	0.1754	0.2029	0.2260
2016	0.1627	0.1951	0.2265
2017	0.1587	0.1901	0.2095
2018	0.1436	0.1761	0.1997

The quarry is spectrally flat and bright, as expected for a bare rock surface. Its apparent brightness declines steadily from 2013 to 2018 — the median falls from 0.209 to 0.176, about 16% — which is the cross-year drift the normalization is there to remove.

> ⚠ **`FFT` does not map to its own acquisition year**
>
> The quarry file covers **2013–2018**, but the dataset also contains **2008–2011** rows. A naive join on year therefore leaves **465 rows without a reference spectrum — exactly the `FFT` campaign**.
>
> **`FFT` was normalized using the 2017 quarry spectrum**, chosen for the spectral similarity between the two sets. When reproducing or reversing the normalization, map every `FFT` row to `year = 2017`, not to its own acquisition year. The remaining 618 rows use their own year directly.
>
> Conversely, 2014, 2015 and 2016 appear in the quarry file but match no dataset row; they are retained for completeness.

---

6. ⚠ Known data-quality issues

These are the most important things to understand before modelling.

6.1 Spectra are NOT pre-masked — bad bands are present with extreme values

The full 380–2500 nm range is retained, **including atmospheric water-vapor absorption windows and low-SNR detector edges**. Values there are physically meaningless and numerically extreme.

Affected regions:

Region	Cause	Severity
380–470 nm	Low SNR at the blue/UV detector edge	Moderate; mostly small negatives
1290–1420 nm	Water-vapour absorption	**Severe** — values to ±280
1790–1960 nm	Water-vapour absorption	**Severe** — values to ±292
2460–2500 nm	SWIR detector edge	Small negatives; systematic across all 465 `FFT` rows (≈ −0.002), consistent with a spectral-resampling edge effect

Use the supplied band mask. The companion file `BadBand.csv` carries the mask — see section 6.1a. Applying it removes every extreme value.

6.1a Companion file: `BadBand.csv`

The band mask ships as a separate 213-row CSV, one row per band, aligned **exactly** with the 213 spectral columns of the main file (verified: `WVL(nm)` matches the band headers element-for-element, in order). Join on wavelength, or rely on positional order.

Column	Description
`WVL(nm)`	Band centre wavelength, nm. 380–2500, 10 nm steps.
`fwhm`	Full width at half maximum, nm. Constant **10** for every band.
`BadBand`	The band mask. See the encoding note below.

> ⚠ **The encoding is inverted relative to the column name**
>
> **`1` = retain the band. `0` = exclude it.**
>
> Despite the name `BadBand`, a value of **1 does not flag a bad band** — it marks a band to **keep**. The water-vapor windows are coded `0`, and the vegetation-bearing regions are coded `1`.

Retained regions — 147 of 213 bands kept, 66 dropped:

Region	Bands	Covers
440–1240 nm	81	VNIR, through the 1.25 µm shoulder
1470–1760 nm	30	SWIR-1 window
2070–2420 nm	36	SWIR-2 window

The three windows are separated by the two atmospheric water-vapor absorption regions; the 380–430 nm and 2430–2500 nm detector edges are dropped as well.

6.2 Trait availability is severely unbalanced

n traits present	rows
2–3	197
11–20	62
21 (all)	**824**

`CA2013` is the key exception: **it carries only `LMA`, `Nitrogen` and `Carbon`** (≈197 rows each) and has zeros for all 18 other traits. `Chlorophylls_*` and `Carotenoids_*` are absent from **both** `CA2013` and `CA2018` — pigment models can only draw on `FFT` and `NEON_avs`.

Non-null counts by campaign:

Trait group	`CA2013`	`CA2018`	`FFT`	`NEON_avs`
LMA / Nitrogen / Carbon	195–197	49	465	371–372
Structural & chemical (18 traits)	**0**	48–49	463–465	366–372
Pigments (4 traits)	**0**	**0**	465	371–372