

SAMPLE · MANGO DRY MATTER

Entry 1: 600 fruit from three seasons, a fourth season held back

We want to predict dry matter in whole fruit from a scan. 600 fruit from three seasons build the model; 100 from a fourth season test it.

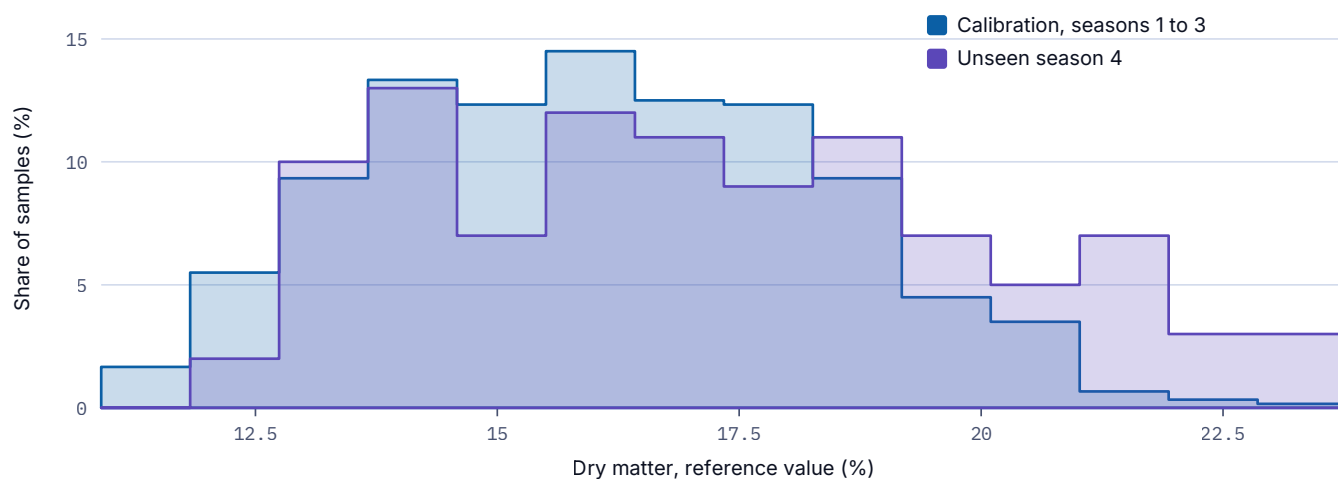
Entry 1 · Sample · 9 Oct 2026 · Lab engine run · page 1 of 8

WHAT WE MEASURED	Dry matter by laboratory, 10.9 to 22.9 % in calibration, 12.0 to 23.8 % in the unseen season.
WHAT WE FOUND	The two seasons overlap but the fourth runs higher on average.
WHAT TO DO	Check that your own test set covers a different season, lot or range.

REFERENCE

What the laboratory measured

Each bar is the share of fruit in a dry-matter range.



Source: Spectra Lab engine run, 9 Oct 2026, mango teaching subset (CC BY 4.0).

SOURCE

Where the data comes from

Anderson, Walsh, Subedi and Hayes (2020), Postharvest Biology and Technology 168; dataset licensed CC BY 4.0 (Mendeley Data, doi 10.17632/46htwnp833). Teaching subset: 600 calibration and 100 unseen-season spectra, 684 to 990 nm, 103 points.

SCAN · 684 TO 990 NM

Entry 2: every fruit gives a curve, and the curves look alike

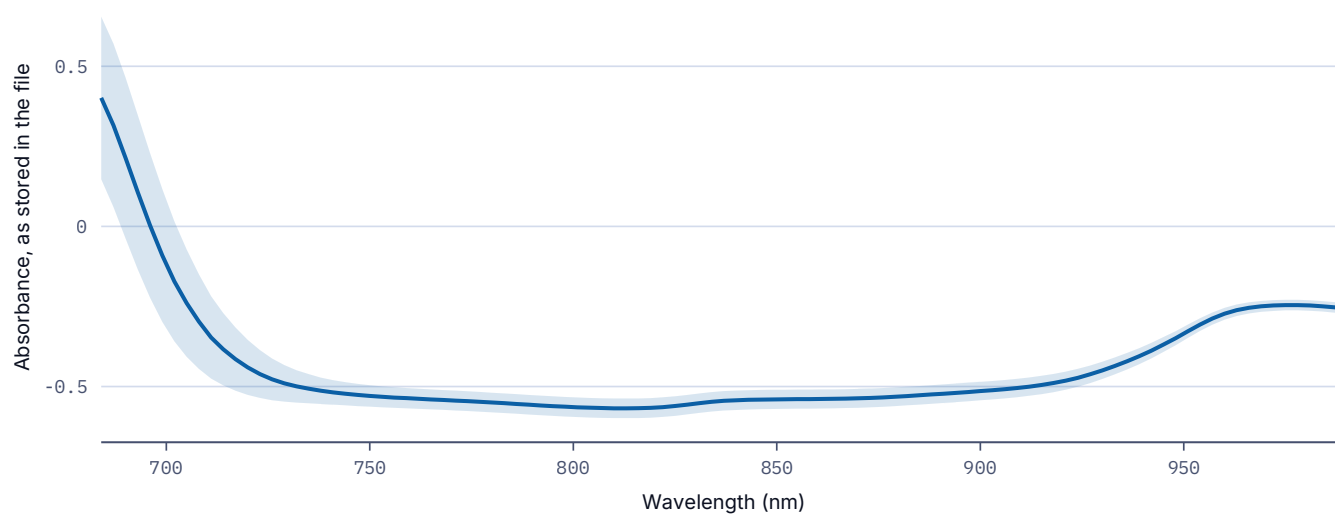
The instrument records absorbance at 103 wavelengths. The line is the mean; the band is one standard deviation across fruit.

Entry 2 · Scan · 9 Oct 2026 · Lab engine run · page 2 of 8

WHAT WE MEASURED	Absorbance at 103 wavelengths for each of 600 fruit.
WHAT WE FOUND	The mean curve falls steeply at the short end and rises again at the long end; the band shows how much fruit differs at each wavelength.
WHAT TO DO	Plot your own spectra before modelling. Odd curves here are easier to fix than odd predictions later.

SPECTRUM

What the instrument saw



Source: Spectra Lab engine run, 9 Oct 2026, mango teaching subset (CC BY 4.0).

PREPROCESSING

Entry 3: the same data gave 1.00 with a 1st derivative and 1.20 with SNV

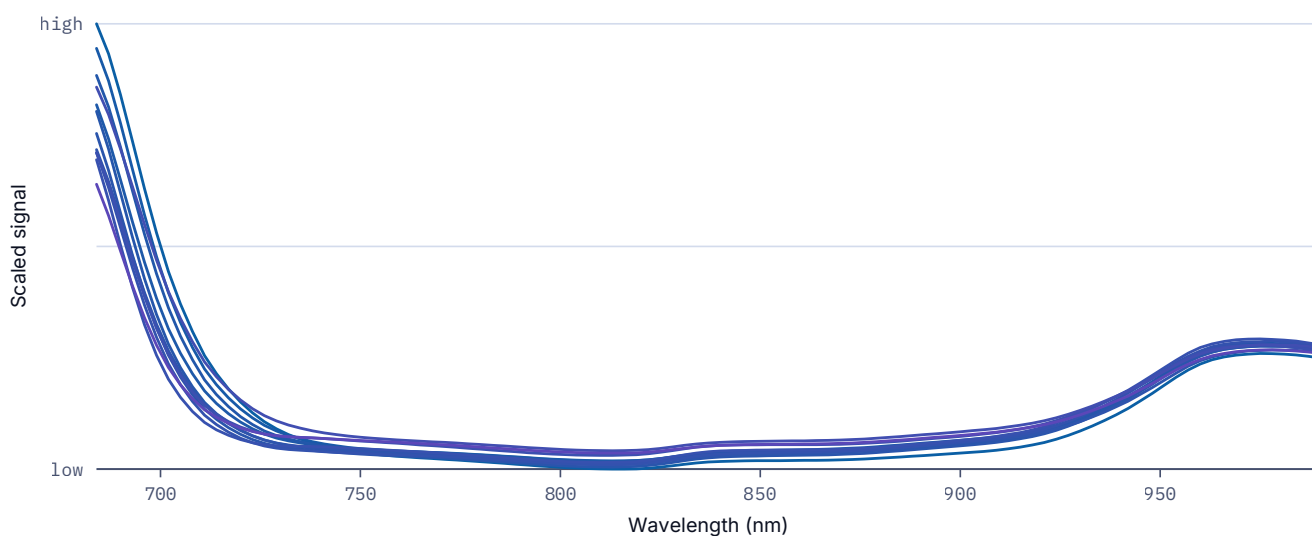
Four pipelines, one model type, one unseen season. SNV was about 20 % worse than the 1st derivative here; on other data the order can change.

Entry 3 · Preprocessing · 9 Oct 2026 · Lab engine run · page 3 of 8

WHAT WE MEASURED	Raw spectra and three corrections, each fitted on the calibration fruit only.
WHAT WE FOUND	Derivatives remove the height and slope differences. The lowest unseen-season error came from the 1st derivative.
WHAT TO DO	Compare on your own data. Do not carry a choice over from someone else's.

BEFORE AND AFTER

Drag from raw to corrected

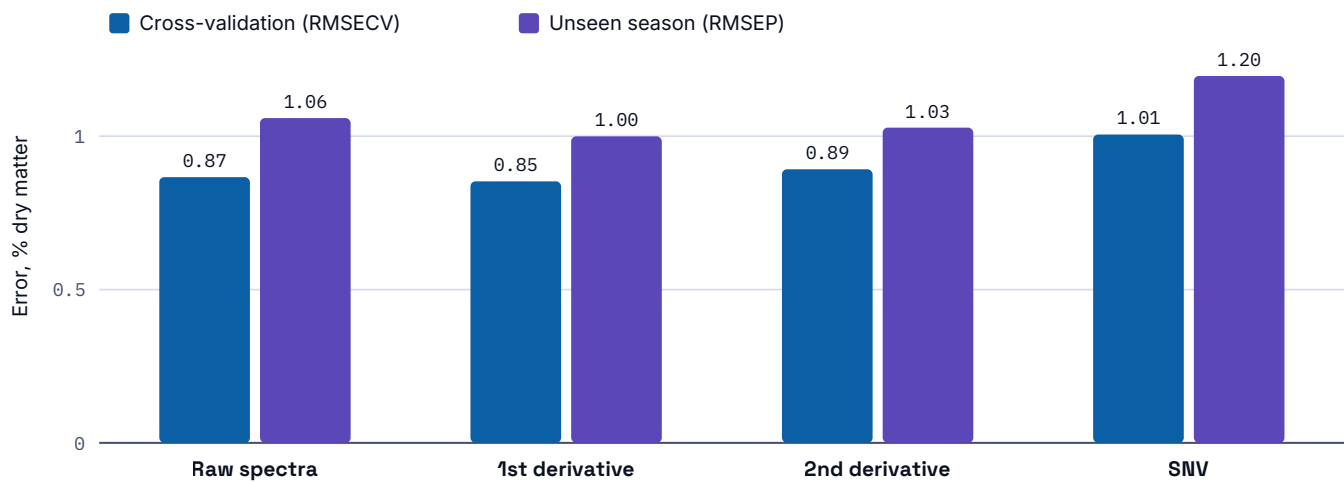


Raw spectra: the curves differ mostly by height and slope.

Source: Spectra Lab engine run, 9 Oct 2026, mango teaching subset (CC BY 4.0). Twelve of the 600 calibration fruit, coloured from low to high dry matter.

COMPARE

Error by preprocessing



Source: Spectra Lab engine run, 9 Oct 2026, mango teaching subset (CC BY 4.0).

NUMBERS

All the numbers

Bias is predicted minus reference; SEP is the error after removing the bias.

PIPELINE	COMPONENTS	RMSECV	RMSEP	BIAS	SEP	RPD
Raw spectra	11	0.867	1.059	-0.206	1.044	2.78
1st derivative	10	0.853	0.999	-0.187	0.987	2.94
2nd derivative	9	0.892	1.028	-0.028	1.033	2.81
SNV	10	1.005	1.197	+0.038	1.202	2.42

MODEL

Entry 4: the error stops falling around 10 components, and the smoothing window barely matters

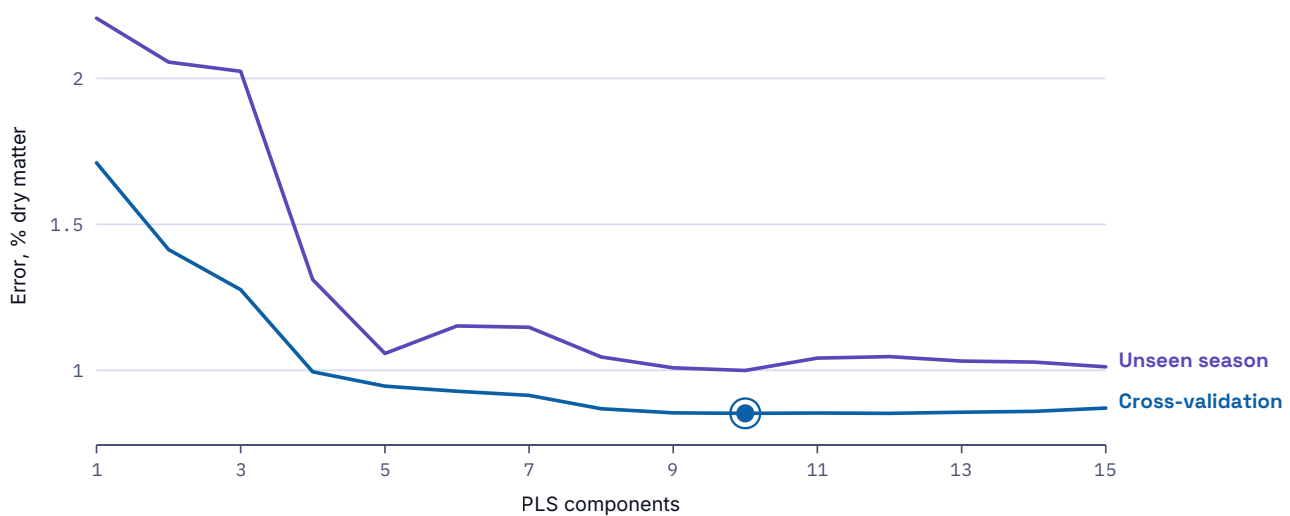
The Lab fits 1 to 15 PLS components and picks the lowest cross-validated error. The slider shows what each number does to the error on the unseen season.

Entry 4 · Model · 9 Oct 2026 · Lab engine run · page 4 of 8

WHAT WE MEASURED	PLS with 1 to 15 components on each pipeline, and seven smoothing windows from 5 to 31 points.
WHAT WE FOUND	Both errors are flat from about 9 components. The window changes the error by about 0.04.
WHAT TO DO	Stop where the curve goes flat. A simpler model transfers better.

COMPONENTS

Drag the number of components

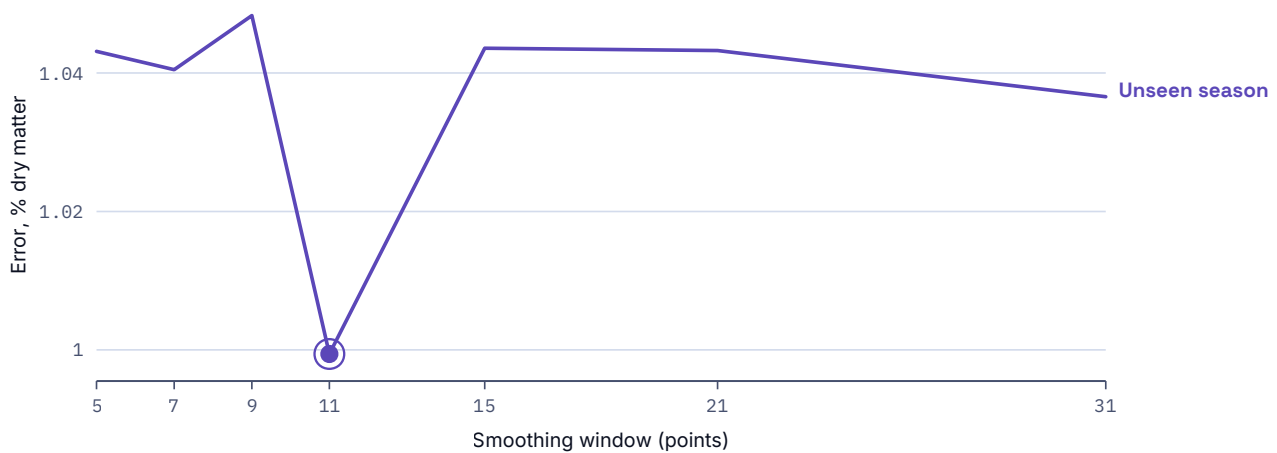


10 components with 1st derivative.

Source: Spectra Lab engine run, 9 Oct 2026, mango teaching subset (CC BY 4.0).

SMOOTHING

Drag the smoothing window



Window 11 points: the Lab default.

Source: Spectra Lab engine run, 9 Oct 2026, mango teaching subset (CC BY 4.0). 1st derivative, engine Savitzky-Golay with a 2nd-order polynomial.

VALIDATION

Entry 5: on a season the model never saw, the typical error is 1.00 % and the model under-predicts by 0.19

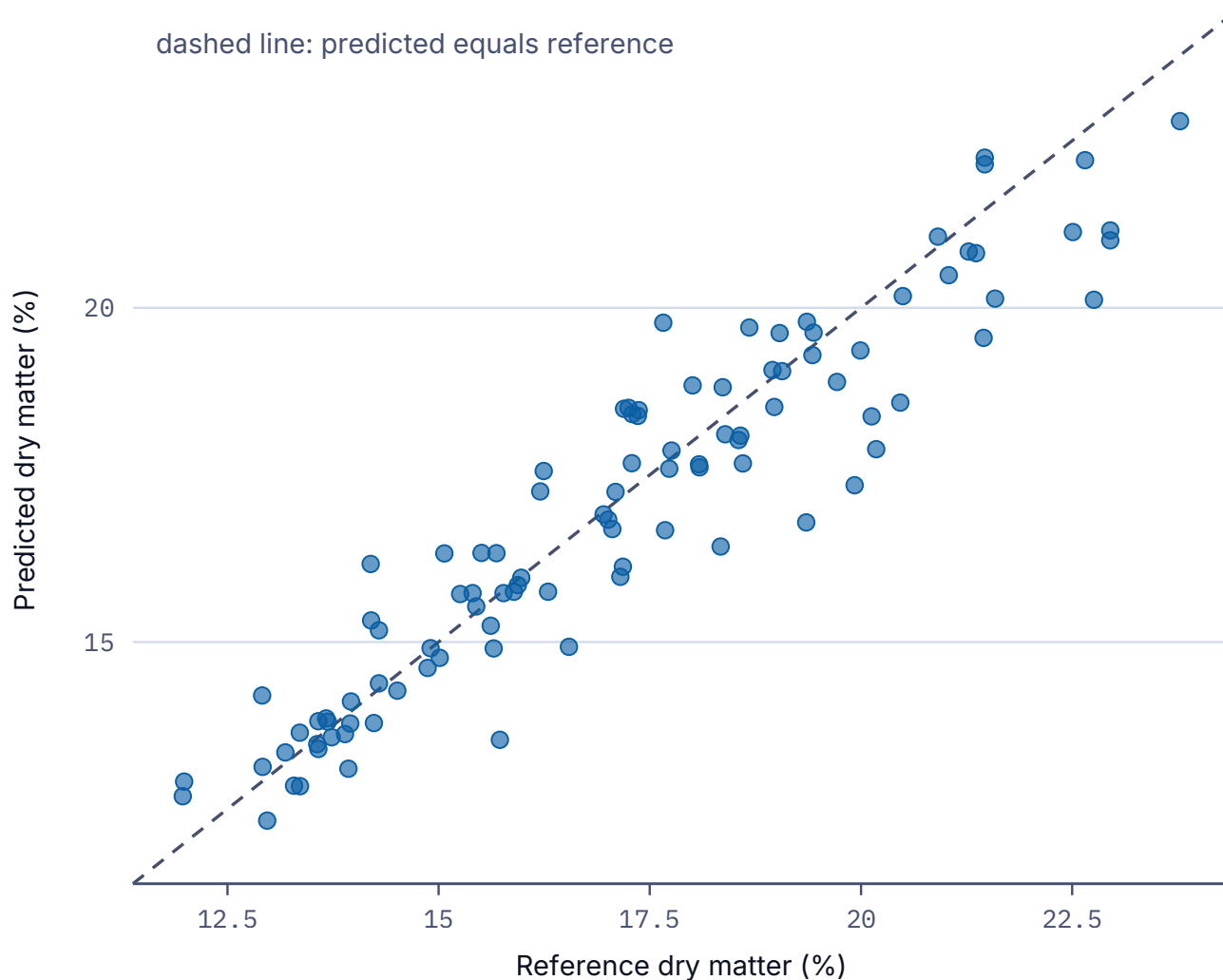
100 fruit from a season the model never saw: predicted dry matter against the laboratory value.

Entry 5 · Validation · 9 Oct 2026 · Lab engine run · page 5 of 8

WHAT WE MEASURED	Predictions from the model for 100 fruit in the fourth season, against the laboratory values.
WHAT WE FOUND	Most points sit near the dashed line. High values run low, which is why the slope is below 1.
WHAT TO DO	Report the number from data the model never saw, with its bias.

CHART

Predicted against reference



1st derivative: typical error **1.00** % dry matter, bias **-0.19**, error after removing bias **0.99**, RPD **2.94**, which the Lab bands as **rough quantitative**.

Source: Spectra Lab engine run, 9 Oct 2026, mango teaching subset (CC BY 4.0). 100 fruit from a season the model never saw. RPD here is the spread of the reference values divided by the error after removing bias.

STATISTICS

What the numbers say

STATISTIC	VALUE	WHAT IT TELLS YOU
RMSEP	0.999 %	typical error on new samples
Bias	-0.187 %	average predicted minus reference
SEP	0.987 %	error after removing the bias
RPD	2.94	spread of reference divided by SEP
Slope	0.855	1.000 would mean predictions scale perfectly with reference
R ²	0.886	squared correlation, reference and predicted

VERDICT

Entry 6: rough quantitative, RPD 2.94

The Lab sorts models by RPD into four bands. Whether the band is enough depends on what you need it for.

Entry 6 · Verdict · 9 Oct 2026 · Lab engine run · page 6 of 8

WHAT WE MEASURED	RPD on the unseen season, with the Lab's bands: below 1 no predictive value, 1 to 2 screening only, 2 to 3 rough quantitative, 3 or more quantitative.
WHAT WE FOUND	The model is 2.94, just under the quantitative line.
WHAT TO DO	Put your own limit in and read the verdict.

VERDICT

Where the model sits



Source: Spectra Lab engine run, 9 Oct 2026, mango teaching subset (CC BY 4.0).

YOUR LIMIT

Is that good enough for you?

Your limit is 1.00; 1st derivative gives a typical error of 1.00 on a season it never saw, and an RPD of 2.94 (rough quantitative).

Source: Spectra Lab engine run, 9 Oct 2026, mango teaching subset (CC BY 4.0). Your limit is your own requirement; the Lab does not set it.

OUTLIERS

Entry 7: 33 of 100 fruit exceed the T² limit; 4 exceed the Q limit

T² says a sample sits far from the calibration centre; Q says the model cannot rebuild its spectrum. A new season shifts spectra, so flags are information about the season.

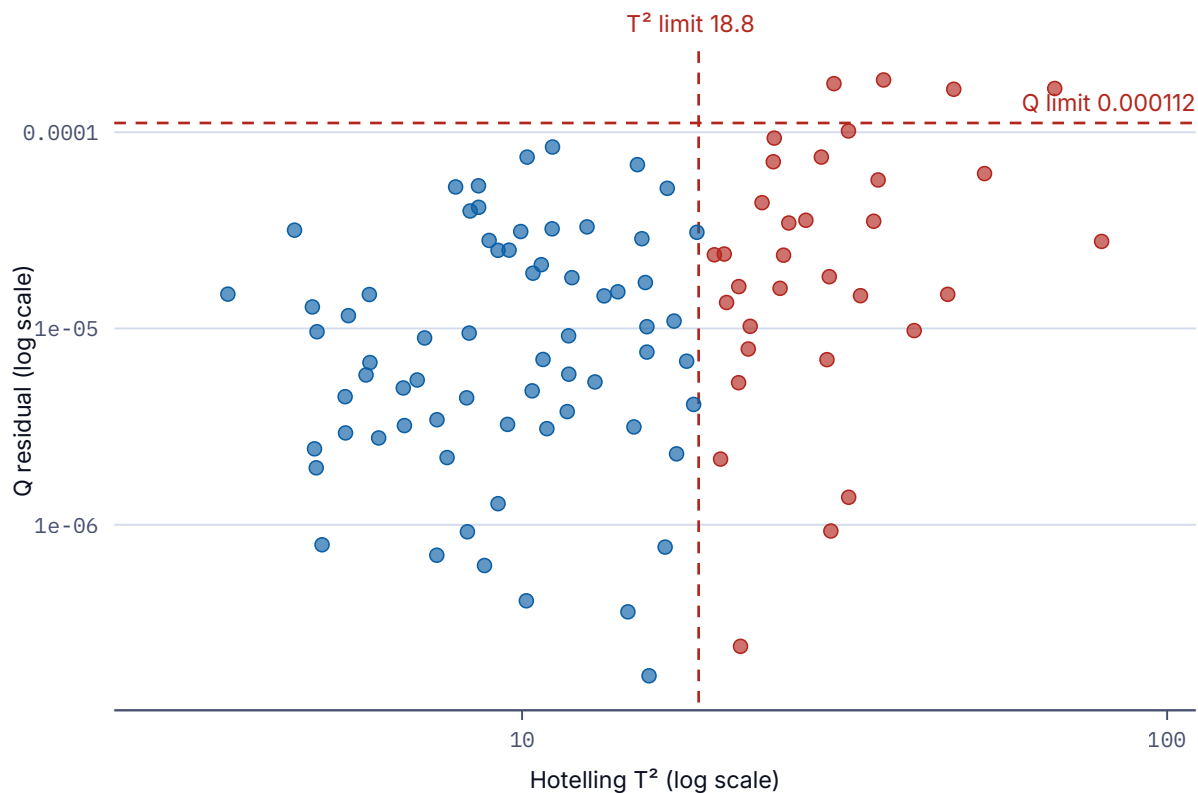
Entry 7 · Outliers · 9 Oct 2026 · Lab engine run · page 7 of 8

WHAT WE MEASURED	T ² and Q for each unseen-season fruit against the limits set on the calibration fruit.
WHAT WE FOUND	A third of the fruit are flagged on T ² . That tells us the season differs, not that the scans are bad.
WHAT TO DO	Inspect the flagged samples before dropping any. We are still refining how the Lab explains this.

CHART

Each point is one unseen-season fruit

Dashed lines are the limits; red points are beyond at least one.



Source: Spectra Lab engine run, 9 Oct 2026, mango teaching subset (CC BY 4.0).

LIMITS AND SOURCES

Entry 8: what this report does not cover, and where the data came from

One dataset, one split, one model family.

Entry 8 · Limits · 9 Oct 2026 · Lab engine run · page 8 of 8

LIMITS

Six limits

- **One dataset, one split.** Mango fruit, seasons 1 to 3 against season 4. Grain, dairy or feed would behave differently.
- **One model family.** Partial least squares, 5-fold cross-validation, 1 to 15 components.
- **Not an accredited method.** The report supports decisions; validate against your own requirements and standards before relying on a calibration.
- **Teaching subset.** 600 + 100 spectra from the larger archive of 11,691.
- **Reproducible.** The same files through the same engine give the same numbers; we re-ran it to check.
- **RPD here** is the spread of the unseen-season reference values divided by the error after removing bias; the Lab's model overview computes RPD from cross-validation instead.

SOURCES

Attribution

Anderson, Walsh, Subedi and Hayes (2020), Postharvest Biology and Technology 168; dataset licensed CC BY 4.0 (Mendeley Data, doi 10.17632/46htwnp833). Teaching subset: 600 calibration and 100 unseen-season spectra, 684 to 990 nm, 103 points.

Engine: the Spectra Lab modelling engine (scikit-learn PLS). The published 1.01 comparison is the figure from the dataset paper for PLS on the full data.

Spectra Lab · Real run on a public dataset · not a customer result

Figures carry their sources; assumptions are labelled as assumptions.