

# **S4 - Segment-based CLC Supervision and RF Classification on Sentinel Signatures**

Chris Reudenbach

2026-01-24

## **1 Purpose and position in the stack**

This module constructs a controlled scale transition from continental land-cover semantics (CLC, 100 m) to Sentinel-scale spatial resolution (10–20 m) using segment-level supervision.

The objective is not land-cover mapping as a cartographic product. The objective is scale stabilisation under semantic control. CLC provides semantic persistence across space and time. Sentinel provides spatial and spectral discrimination but no intrinsic semantic grounding. The workflow binds both without fabricating spatial precision.

In S-terms, this module occupies the transition from S3 to S4.

S3 provides geometry: spatial objects with stable topology and identifiers. S4 constructs deterministic numerical signatures and attaches supervision. Any learning component operates instrumentally in S4L. Decisions belong strictly to S5.

This separation is not stylistic. It prevents representational drift, hidden coupling between modelling and semantics, and uncontrolled feedback between optimisation and interpretation.

## **2 Input contracts**

The Sentinel predictor stack is a co-registered raster stack of continuous variables. It typically contains Sentinel-2 bands and derived indices such as EVI or kNDVI. All layers share grid geometry and projection. No categorical semantics are embedded.

The CLC raster contains true CLC3 integer codes in the range 111–523. Factor indices and colour tables are explicitly excluded. CLC is treated as a semantic reference frame, not as fine-scale truth.

Segments are polygon objects with stable segment\_id and valid geometry. The segmentation itself is upstream and immutable in this module.

These three inputs define the complete interface. No implicit data dependencies are tolerated.

### 3 Why CLC must be aggregated to objects

Resampling CLC to Sentinel resolution introduces false precision. Semantic resolution is silently mistaken for spatial resolution. Pixel-level mixing creates artefacts that cannot be interpreted or validated.

Therefore, CLC is aggregated strictly at object level.

For a segment  $s$  and a CLC class  $c$ , the area-weighted class share is defined as

$$\text{share} * s, c = \frac{\sum_i a * i, s, \mathbb{1}(x_i = c)}{\sum_i a_{i,s}}$$

where  $a_{i,s}$  is the fractional overlap between raster cell  $i$  and segment  $s$ , and  $x_i$  is the CLC code. The formulation enforces normalisation, determinism and scale consistency. Implementation uses exact area-weighted extraction.

From the class shares, the dominant class and its purity follow directly:

$$\text{class} * s = \arg \max_c (\text{share} * s, c), \quad \text{purity} * s = \max_c (\text{share} * s, c)$$

Purity is not cosmetic. It is an explicit label-quality control variable that limits semantic noise in training data.

CLC detail classes are mapped to a reduced functional ontology (urban\_sealed, cropland, grassland, forest subclasses, bare\_surface, others). This reduction sacrifices thematic detail in exchange for transferability, statistical stability and cross-region comparability.

### 4 Why segment statistics, not pixels

All modelling operates in segment space. A segment is a heterogeneous spatial object. Pixels are samples inside that object.

Segment signatures compress internal pixel distributions into controlled numerical descriptors: central tendency, dispersion and distribution shape. Typical statistics include mean, standard deviation, extrema, coefficient of variation and selected quantiles.

This representation suppresses pixel noise while preserving structural heterogeneity that remains relevant for classification and downstream modelling. It also stabilises learning under spatial autocorrelation and sampling imbalance.

The signature space is intentionally redundant. Redundancy increases representational capacity but must be controlled explicitly.

## 5 Feature contract and numerical stability

Predictor preprocessing follows the caret workflow for numerical stability and reproducibility (Kuhn and Johnson 2013).

Near-zero variance predictors are removed because they cannot support stable splits. Exact linear dependencies are eliminated to restore full column rank. Highly correlated predictors are filtered to reduce variance inflation and instability.

Correlation filtering uses the Pearson coefficient

$$r_{jk} = \frac{\text{Cov}(x_j, x_k)}{\sigma_{x_j} \sigma_{x_k}}$$

Predictors are removed if  $|r_{jk}|$  exceeds a defined threshold. Linear dependence is detected when a non-zero vector  $\mathbf{b}$  satisfies

$$\mathbf{X}\mathbf{b} = \mathbf{0}$$

The retained predictor set defines the only valid interface for both training and prediction. Any deviation violates the model contract.

## 6 Learning as an instrumental projection

Training data consist of segment-level pairs

$$(\mathbf{x} * s, y_s) * s = 1^n$$

where  $\mathbf{x}_s$  is the signature vector and  $y_s$  the dominant CLC-derived class.

Random Forest is used as an instrumental projector from signatures to class membership. An ensemble of trees  $h_t(\cdot)$  is fitted on bootstrap samples. Classification follows majority vote:

$$\hat{y}_s = \text{mode}h_1(\mathbf{x}_s), \dots, h_T(\mathbf{x}_s)$$

The method is robust to nonlinear boundaries, mixed predictor scales and moderate redundancy (Breiman 2001). The epistemic value lies in the representation and supervision, not in the classifier itself.

## 7 Why spatial validation is mandatory

Spatial samples violate the i.i.d. assumption due to spatial autocorrelation. Random cross-validation mixes neighbouring samples between training and testing and therefore produces optimistically biased performance estimates. This bias is well documented (Meyer et al. 2018; Roberts et al. 2017).

Spatially structured validation strategies address this leakage. Leave-location-out evaluates transfer to unseen locations. Leave-block-out enforces spatial independence by withholding entire spatial blocks. Both approximate real deployment conditions.

Extrapolation risk can be formalised using the Area of Applicability, which quantifies whether new samples fall inside the multivariate predictor domain supported by the training data (Meyer and Pebesma 2021). CAST operationalises these concepts in reproducible workflows (Meyer et al. 2025).

Hyperparameter tuning under spatial cross-validation behaves differently from random CV. Parameter settings that exploit local spatial structure are penalised. A coarse-to-fine tuning strategy balances robustness and computational cost.

Forward feature selection under spatial CV further suppresses locally overfitted predictors and improves transfer stability (Meyer et al. 2018, 2025).

## 8 Prediction contract

Prediction strictly reuses the same signature definitions and feature contract as training. No pixel-wise prediction is performed. Segments with incomplete signatures are flagged as unclassified.

This preserves semantic integrity and prevents silent contract drift.

## 9 Outputs

The module produces:

- a deterministically supervised segment layer (class, purity, n\_classes),
- a predicted segment layer (predicted class, optional probabilities),
- a persistent training table.

All artefacts are reproducible and auditable.

## 10 Key implications for modelling and teaching

Scale transitions are representation problems, not modelling problems. Semantic stability must precede spatial refinement. Object-level aggregation prevents false precision. Spatial validation is not optional. Models are instruments, not epistemic authorities.

## 11 Limitations

Semantic resolution is bounded by CLC. Rare classes remain unstable. Segmentation errors propagate. Purity thresholds trade bias against variance. These limitations are structural and must be managed explicitly.

## References

- Breiman, Leo. 2001. “Random Forests.” *Machine Learning* 45 (1): 5–32.
- Kuhn, Max, and Kjell Johnson. 2013. *Applied Predictive Modeling*. Springer.
- Meyer, Hanna, Carles Milà, Marvin Ludwig, Jan Linnenbrink, and Fabian Schumacher. 2025. *CAST: 'Caret' Applications for Spatial-Temporal Models*. <https://doi.org/10.32614/CRAN.package.CAST>.
- Meyer, Hanna, and Edzer Pebesma. 2021. “Predicting into Unknown Space? Estimating the Area of Applicability of Spatial Prediction Models.” *Methods in Ecology and Evolution* 12 (9): 1620–33.
- Meyer, Hanna, Christoph Reudenbach, Stefan Wöllauer, and Thomas Nauss. 2018. “Machine Learning-Based Modeling of Environmental Variables: Cross-Validation Strategies for Spatial Data.” *Ecological Modelling* 372: 221–34.
- Roberts, David R., Volker Bahn, Simone Ciuti, et al. 2017. “Cross-Validation Strategies for Data with Temporal, Spatial, Hierarchical, or Phylogenetic Structure.” *Ecography* 40 (8): 913–29.