Evaluating the Transferability of Machine-Learning Models for Pre-Emergence Bark Beetle Detection Using Multispectral and Hyperspectral UAV Data

Geng Wang[1], Aurora Bozzini[2], Luiz Henrique Elias Cosimo[1], Emma Turkulainen[3], Salma Bijou[4], Raquel Alves de Oliveira[3], Juha Suomalainen[3], Erika Vennervirta[3], Niko Koivumäki[3], Roope Näsi[3], Lucie Kupková[4], Massimo Faccoli[2], Eija Honkavaara[3], Langning Huo[1]

[1] Department of Forest Resource Management, Swedish University of Agricultural Sciences, 90 183, Umeå, Sweden - langning.huo@slu.se, luiz.cosimo@slu.se, geng.wang@slu.se

[2] Department of Agronomy Food Natural Resources Animals and Environment, University of Padua, 35020, Legnaro (Padova), Italy - aurora.bozzini@phd.unipd.it, massimo.faccoli@unipd.it

[3] Department of Remote Sensing and Photogrammetry, Finnish Geospatial Research Institute FGI, 00521 Helsinki, Finland - raquel.alvesdeoliveira@maanmittauslaitos.fi, juha.suomalainen@nls.fi, erika.vennervirta@maanmittauslaitos.fi, niko.koivumaki@maanmittauslaitos.fi, roope.nasi@maanmittauslaitos.fi, eija.honkavaara@nls.fi

[4] Department of Applied Geoinformatics and Cartography, Faculty of Science, Charles University, Albertov 6, Prague 2, Czech Republic - bijous@natur.cuni.cz, lucie.kupkova@natur.cuni.cz

Keywords: UAV, hyperspectral imaging, multispectral imaging, European spruce bark beetle, forest health monitoring

Abstract

UAV-based bark beetle detection often remains site-specific. We assessed transferability using seven DroneNet4Beetles UAV crown-reflectance datasets (multispectral and hyperspectral VNIR) from Sweden, Finland, Italy, and Czechia. We compared Random Forest (RF) and generalized linear model (GLM) using (i) multivariable spectra (multispectral) or multivariable VI sets, and (ii) single-VI GLMs, and evaluated performance along a four-tier ladder (within-plot, cross-plot within stand, cross-stand, cross-country) using Cohen’s kappa over weekly time series. RF and multivariable models generally performed best within-domain, but transferability dropped sharply at the cross-stand level and became most variable across countries. In contrast, the Green-shoulder Curvature Ratio Index (GSCR1, updated formulation in 2026) showed the most consistent transfer between stands and countries, with strong SE↔CZ and CZ↔FI transfer but weak SE↔FI, indicating domain distance—not geographic distance—drives generalization. Transfer involving different hyperspectral sensors was not systematically worse, but multispectral cross-country transfer was highly sensitive, consistent with red-edge band-centre mismatch reducing VI generalizability.

1. Introduction

Outbreaks of the European spruce bark beetle (*Ips typographus*) have intensified across Central and Northern Europe, driven by droughts, storms, and other extreme climatic events (Seidl et al., 2017; Hlásny et al., 2021). The resulting mortality of Norway spruce has reduced growing stock and impaired forest carbon uptake, highlighting an urgent need for rapid and operational tools for early detection. Pre-emergence detection (i.e., identifying infested trees before brood emergence) is a crucial management strategy, yet field surveys are too costly and slow for large-scale implementation (Bárta et al., 2021). UAV-based optical remote sensing provides high-resolution, flexible, and timely mapping at the single-tree level, enabling detailed monitoring of spectral changes soon after attack (Näsi et al., 2015; Huo et al., 2021; Bozzini et al., 2025; Bijou et al., 2025).

However, despite a rapidly growing body of UAV-based studies, it remains unclear how early infestations can be reliably detected and how well-developed methods transfer among regions. Heterogeneity in sensor types (multispectral vs. hyperspectral), band configurations—particularly in the red-edge and green-shoulder regions—and analysis methods often yields inconsistent and sometimes contradictory results (Huo et al., 2024; Cosimo et al., 2025). Many studies use machine learning or deep learning within single sites, but often lack standardized accuracy metrics and cross-site validation, limiting our understanding of model robustness and generalizability under varying ecological and climatic conditions (Huo et al., 2023a).

To address these gaps, we assembled six UAV datasets from four major outbreak regions—southern Sweden, southern Finland, the southeast Alps in Italy, and Czechia—encompassing both multispectral and hyperspectral campaigns at the single-tree level. Using this comprehensive dataset, we develop and compare machine-learning models for classifying tree health status (e.g., healthy vs. infested) based on spectral features and vegetation indices. We systematically evaluate how model performance depends on sensor characteristics, spectral regions, and mono- vs. bi-temporal inputs.

A key focus of this study is transferability. We define a domain as a specific data distribution shaped by distinct spatial boundaries, ecological conditions, and sensor characteristics. To quantify how well machine-learning approaches generalize under varying degrees of domain shift, we evaluate models across nested validation tiers ranging from local infestation plots to entire countries. By combining multi-country datasets with a unified evaluation framework, we identify spectral features and modelling strategies that remain effective across domains, thereby providing practical recommendations for operational early-warning systems.

2. Materials and methods

2.1. Study area and datasets

We used seven UAV crown-reflectance time-series datasets compiled in the DroneNet4Beetles workflow, covering Norway spruce stands across Sweden (SE), Finland (FI), Czechia (CZ), and Italy (IT). Each dataset contains (i) a tree-level ground-truth table with infestation status and plot/stand identifiers, and (ii) weekly crown reflectance matrices extracted for the same set of delineated tree crowns (Table 1). Selected vegetation indices (VIs, Table 2) were calculated. More details on the data can be found in Huo et al. (2026).

2.2. Machine learning and statistical models

We evaluated three types of machine-learning and statistical models for classification:
(1) Random forest (RF) (Breiman, 2001) with all bands normalized by NIR for multispectral datasets and all VIs for hyperspectral datasets.
(2) Generalized linear model (GLM) (McCullagh, 2019) using all indices, with a binomial distribution with logit link.
(3) Generalized linear models using a single index (named by each VI), with a binomial distribution with logit link.

When applicable, we randomly partitioned 60% of the samples for training and 40% for testing. This process was repeated using 50 random splits, and the average accuracy was reported. Classification performance was measured using Cohen’s kappa (Cohen, 1960), which is particularly suitable for imbalanced sample sizes between classes. Kappa is reported for each available time point, presenting the means over repeated runs and boxplots over split replicates where applicable.

2.3. Transferability test

(1) Within-plot level (infestation clusters). Among the 83 plots from all datasets, 19 plots containing >10 infested trees were selected for within-plot transferability testing, ensuring sufficient sample sizes for a 60/40 training-to-testing split within each plot. Training and testing samples were randomly split 50 times, and the average kappa coefficients for RF, GLM, and individual VIs were presented to represent within-plot transferability.

(2) Cross-plot within-stand level. SE2021-MS (6 stands × 4 plots), SE2023-MS (4 stands × 6 plots), IT2023-MS (2 stands × 4 plots), and SE2023-HS (4 stands × 6 plots) were utilized for this test, given their hierarchical stand-plot structures. Within each stand, models were trained on a subset of plots and tested on held-out plots per week. The plot-combination schemes depended on the stand’s plot count (explicit combinations for 4-plot stands; combinatorial splits for 6-plot stands). All training-testing combinations of plots were executed, yielding a distribution of kappa values across stand×split replicates.

(3) Cross-stand within-country level. SE2021-MS, SE2023-MS, and SE2023-HS were used for this evaluation. For each week, models were trained on a subset of stands and tested on held-out stands within the same dataset to assess cross-stand transferability.

(4) Cross-country and cross-sensor level.
Multispectral: SE2023-MS and FI2023-MS from matched acquisition weeks (week 26 and weeks 33/34) were used to evaluate multispectral model transferability. Models were first trained and tested on 60% and 40% of the data from one country using 50 random splits to establish a baseline average kappa. Subsequently, the trained models were applied directly to the multispectral data of the other country to assess cross-country generalization.
Hyperspectral: SE2023-HS, CZ2022-HS, and FI2021-HS were used to test the transferability of VI-based models across countries and sensors. Two matched phenological time points were selected: T1 representing approximately 6 weeks post-infestation (week 26 for SE2023-HS and CZ2022-HS; week 34 for FI2021-HS), and T2 representing a later stage (weeks 28/29 for SE2023-HS and CZ2022-HS; week 36 for FI2021-HS). Similar to the multispectral design, models were trained using data from one country with an internal holdout and then evaluated on the remaining two countries.

Table 1. Summary of the seven remote sensing datasets used in this study

Dataset	SE2021-MS	SE2023-MS	FI2023-MS	IT2023-MS	SE2023-HS	CZ2022-HS	FI2021-HS
Region	Sweden	Sweden	Finland	Italy	Sweden	Czech Republic	Finland
Forest type	Managed forest	Managed forest	Nature-served forest	Nature-served forest	Managed forest	Managed forest in national park	Nature-served forest
Elevation (masl)	140	140	25	1250-1425	140	1030	25
Year	2021	2023	2023	2023	2023	2022	2021
Time of attacks	Week 20–23	Week 20–23	-	Week 21	Week 20–23	Week 22	Week 27-29 (second swarm)
Number of stands	6	4	1	4	4	1	1
Number of plots	24	24	-	8	24	-	-
Total number of trees	977	1018	133	374	1018	47	76
Number of healthy trees	770	909	101	223	909	11	47
Number of attacked trees	151	101	32	151	101	11	29
Sensor Type	Multispectral	Multispectral	Multispectral	Multispectral	Hyperspectral	Hyperspectral	Hyperspectral
Sensor name	MAIA S2	MAIA S2	MicaSense Altum-PT	DJI Phantom 4	SPECIM AFX10	Headwall Nano	SPECIM AFX10
Number of bands	9	9	5	5	224	269	224
Number of acquisition times	4	9	8	5	7	9	4
Acquisition weeks	19, 25, 33, 42	17, 20, 22, 24, 26, 28, 33, 35, 47	21, 24, 26, 27, 30, 32, 34, 39	24, 25, 26, 27, 28/29	16, 18, 23, 26, 28, 30, 32	20, 22, 23,24, 25, 26, 29, 31	30, 32, 34, 36

https://doi.org/10.5194/isprs-archives-XLIX-B3-2026-1163-2026

Table 2. Details of the vegetation indices (VIs) used.

Index	Definition	Reference
MRDSWI1	$\frac{R_{705} \times R_{705}}{R_{550} \times R_{783}}$	Huo et al. (2023a)
MRDSWI2	$\frac{R_{705} \times R_{705} \times R_{865}}{R_{550} \times R_{783} \times R_{783}}$	Huo et al. (2023a)
REIPguyot	$700 + 40 \times \frac{\frac{R_{670} + R_{780}}{2} - R_{700}}{R_{740} - R_{700}}$	Guyot et al. (1988)
Chlrededge	$\frac{NIR}{\mathrm{RedEdge710}} - 1$	(Gitelson et al., 2003)
CIrededge	$\frac{NIR}{\mathrm{RedEdge730}} - 1$	(Gitelson et al., 2003)
NGRDI	$\frac{R_{550} - R_{665}}{R_{550} + R_{665}}$	Tucker (1979)
NDVI	$\frac{R_{865} - R_{665}}{R_{865} + R_{665}}$	Rouse et al. (1973)
GSCR1-2026	$\ln\left(\left\lvert\frac{(R550 - R530 + 0.15)^2}{[R530 - 0.5 \times (R490 + R550) + 0.15] \times (R530 - R490)}\right\rvert\right)$	Huo et al. (2024)

Note: R_λ denotes the reflectance at the specific center wavelength (nm). For multispectral sensors lacking these exact theoretical narrow bands (e.g., DJI Phantom 4, MicaSense Altum-PT), the nearest available broadband corresponding to the respective spectral region (e.g., Red, Red-edge, NIR) was utilized to compute the indices.

3. Results

3.1 Plot-level transferability across weeks

Across datasets, plot-level performance showed a strong dependence on the acquisition week, with kappa values often close to zero at early acquisition weeks and increasing later in the season. In SE2021-MS and SE2023-MS, RF and VI-based models exhibited a clear step increase around the mid-to-late season (e.g., after ~week 33), reaching high agreement in later weeks, while GLM generally yielded lower kappa values compared to RF when using the full multispectral predictors. In FI2023-MS, kappa increased more gradually across the available weeks, and the relative ranking between indices and models was less consistent than in Sweden. In contrast, IT2023-MS remained close to chance level across all weeks for all tested models and indices, indicating limited separability under those specific acquisition and site conditions.

For hyperspectral datasets, performance increases were similarly concentrated in later weeks. SE2023-HS and CZ2022-HS showed marked improvement from early to late weeks, with several VI-based models approaching strong agreement in the late season. Among the tested hyperspectral indices, the updated green-shoulder formulation (GSCR1-2026) frequently ranked among the strongest or most stable curves in late weeks, whereas red-edge-based indices showed dataset-dependent behaviour. FI2021-HS displayed more modest kappa levels and weaker separation between methods across its available weeks.

3.2 Cross-plot within-stand transferability

When models were transferred to unseen plots within the same stand, performance became more variable, as reflected by the wider distributions of kappa values. Nevertheless, the seasonal pattern observed at plot level largely persisted: SE2021-MS maintained high late-season kappa for RF and several VIs, while earlier weeks remained low. SE2023-MS showed broader dispersion across plot splits, suggesting stronger within-stand heterogeneity and/or sensitivity to plot-level differences compared with SE2021-MS. IT2023-MS again showed limited performance across all methods. For SE2023-HS, kappa distributions expanded substantially at early weeks and tightened at later weeks, indicating that cross-plot transfer became feasible mainly in the late-season window, with GSCR1-2026 standing out as one of the more transferable VI options.

3.3 Cross-stand-within-country transferability

Transferring models across stands within the same dataset further increased variability and reduced performance relative to within-plot and within-stand tests. SE2021-MS retained strong late-season transferability across stands for RF and multiple VI-based models, whereas SE2023-MS showed lower and less stable kappa across weeks, consistent with stronger stand-level domain shift. In the hyperspectral case (SE2023-HS), cross-stand transferability improved primarily in later weeks, and again the best-performing VI differed between early and late season, highlighting the importance of phenological timing when extrapolating across stands.

3.4 Cross-country transferability

Cross-country experiments revealed a clear domain-shift penalty. For the multispectral Sweden–Finland transfer (Fig. 4a), within-country training/testing provided a baseline that was consistently higher than out-of-country testing, while cross-country performance often dropped toward a low kappa, especially at the earlier matched week. The relative ranking between RF, GLM, and single VIs also depended on transfer direction, indicating that country-specific acquisition/stand conditions can dominate the learned decision boundary.

For hyperspectral cross-country transfer among Sweden, Czechia, and Finland (Fig. 4b–d), transferability was highly pair-dependent and time-point dependent (T1 vs T2). Some train–test pairs retained moderate agreement and others collapsed toward near-zero kappa, even when within-country performance was acceptable. Across several pairings, GSCR1-2026 and selected red-edge indices were among the more competitive VI-based options, but no single model/VI dominated across all country pairs and both time points, underlining the need for multi-domain calibration and standardized acquisition protocols when aiming for generalizable HS detection.

The VI time-series plots help explain the cross-country results by revealing systematic baseline shifts and dataset-specific seasonal trajectories (Fig. 5). For MRDSWI1/2 and CIrededge, the absolute value ranges differed between SE2023-HS, CZ2022-HS, and FI2021-HS, implying that a fixed decision threshold learned in one country may not transfer directly to another. REIPguyot exhibited a narrower dynamic range but still showed site-dependent offsets. Importantly, the healthy–infested separation was not uniform across countries: some indices showed clearer separation in SE2023-HS and CZ2022-HS than in FI2021-HS, consistent with the heterogeneous cross-country kappa patterns in Fig. 4b–d.

3.5 Where transferability breaks across validation tiers

To identify the spatial level at which transferability is lost, we tracked Cohen’s kappa across the four nested validation tiers (Figs. 1–4). We interpreted a “loss of transferability” as kappa distributions either collapsing toward near-zero agreement or becoming highly unstable across splits.

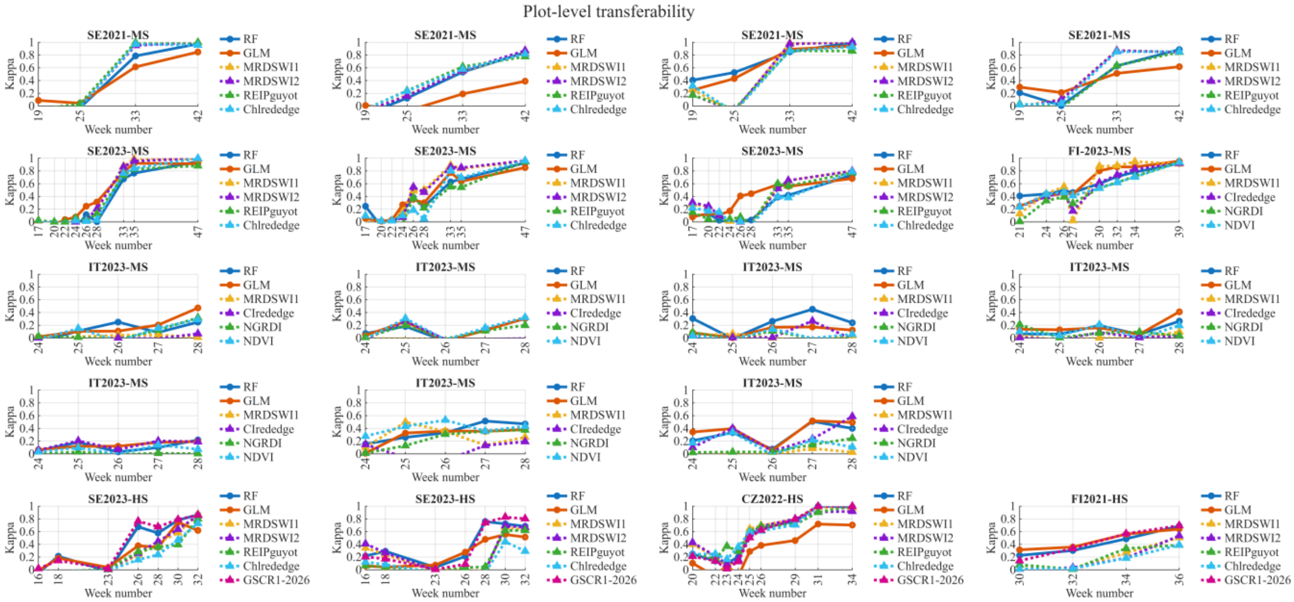

Figure 1. Within-plot (tree holdout) performance across acquisition weeks for each plot. Lines show mean Cohen’s kappa over 50 random splits for RF and GLM and selected VI-based GLMs/RF where applicable.

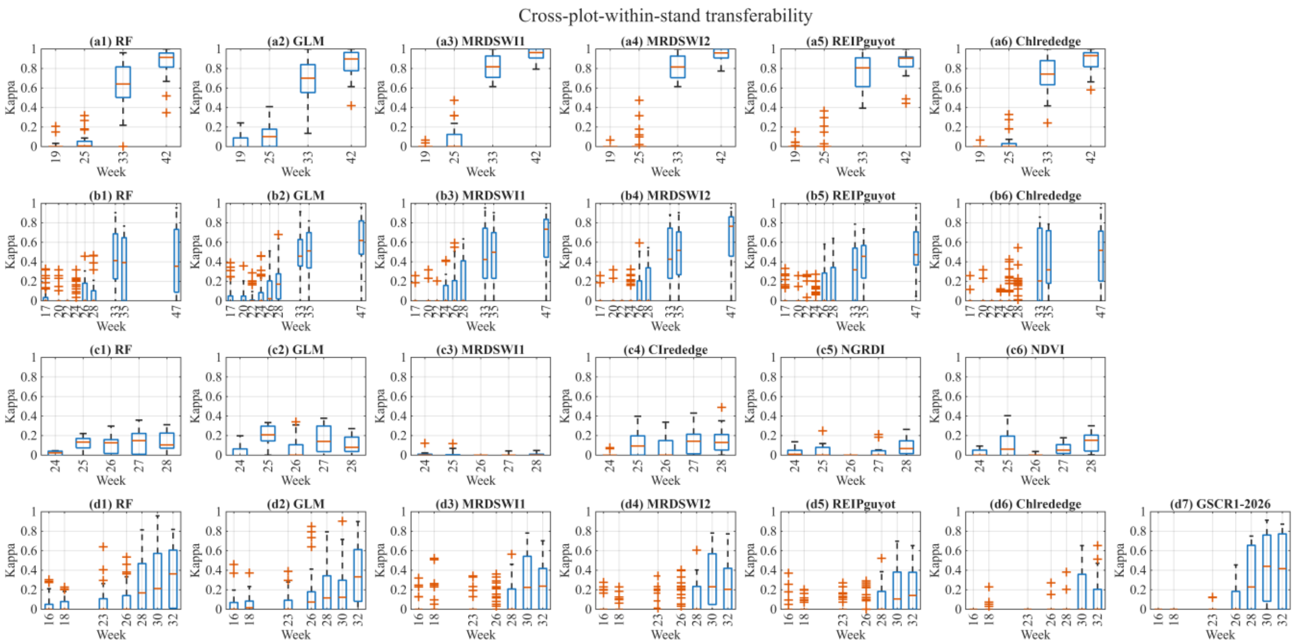

Figure 2. Cross-plot-within-stand transferability on SE2021-MS (a), SE2023-MS (b), IT2023-MS (c), and SE2023-HS (d). Boxplots show kappa distributions across plot-split replicates for each week, comparing RF/GLM and selected VIs.

3.5.1 Breakpoint by tier

Across datasets, most methods showed their first substantial degradation when moving from within-plot to cross-plot within stand (Figs. 1–2), visible as a widening of kappa distributions even when the within-plot trajectories were strong. The more decisive breakpoint occurred at cross-stand within country (Fig. 3), where several methods that performed well within plots/stands exhibited reduced and less consistent kappa, indicating that stand-level domain shift (structure, illumination, phenology, and label heterogeneity) is a major driver of transfer failure. The strongest and most frequent loss of transferability was observed under cross-country transfer (Fig. 4), where many train–test directions dropped to low kappa despite acceptable within-country baselines—demonstrating that country-to-country differences dominate the learned decision boundary unless explicitly modelled or cross-calibrated.

3.5.2 Models vs VIs: Multivariate predictors break earlier while some single VIs exhibit greater stability

Within-domain (Figs. 1–3), RF was typically among the best performers and often showed steeper late-season gains than GLM, consistent with its ability to capture nonlinear separability. However, under stronger domain shift (especially cross-country; Fig. 4), multivariate RF/GLM models frequently degraded sharply or inconsistently, suggesting sensitivity to baseline shifts and interaction patterns learned from the training domain. In contrast, single-VI GLMs often showed lower peak performance within a site but, for some indices, retained moderate agreement more consistently across domains (Fig. 4b–d). This behaviour is consistent with single indices acting as constrained, physiologically interpretable features that are less prone to overfitting site-specific covariance structures.

3.5.3 Which VIs retain transferability best

In the hyperspectral cross-country experiments (Fig. 4b–d), the updated green-shoulder index GSCR1-2026 frequently increased from T1 to T2 and remained competitive across multiple train–test pairs, indicating comparatively stable transfer in late-season conditions. Red-edge-related indices (MRDSWI1/2, Chlrededge, REIPguyot) showed pair-dependent transferability, performing well in some directions but not others. The VI distribution plots (Fig. 5) support this: several indices exhibit systematic country-specific offsets and different seasonal trajectories, implying that a model trained in one country can fail in another due to baseline shift even if the healthy–infested separation exists locally.

3.5.4 Multispectral cross-country transfer is highly sensitive

In the multispectral Sweden–Finland transfer (Fig. 4a), out-of-country performance was generally low and strongly dependent on direction and week, with occasional isolated cases of higher kappa for a specific VI in one week but not sustained across time. This indicates that cross-country generalization with multispectral features is particularly sensitive to acquisition/phenology mismatch and sensor/processing differences, and requires explicit calibration or domain adaptation beyond standard within-site validation.

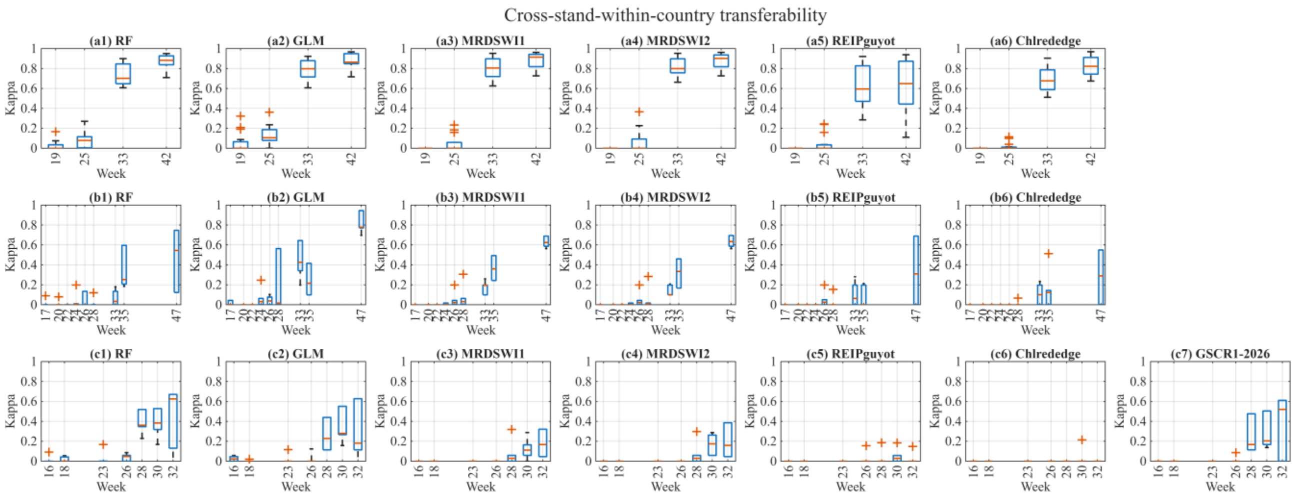

Figure 3. Cross-stand-within-country transferability on SE2021-MS (a), SE2023-MS (b), and SE2023-HS (c). Boxplots show kappa distributions across stand-split replicates for each week.

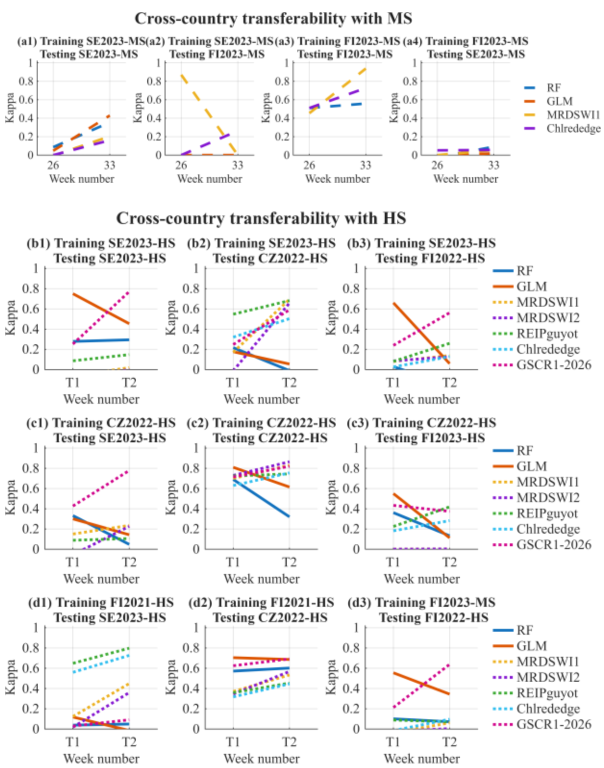

Figure 4. Cross-country transferability of UAV-based infestation detection using multispectral and hyperspectral data: (a) multispectral Sweden–Finland; (b–d) hyperspectral Sweden–Czechia–Finland. Kappa for internal holdout (training domain) and external test (other country) at matched weeks, comparing RF, GLM, and selected VIs.

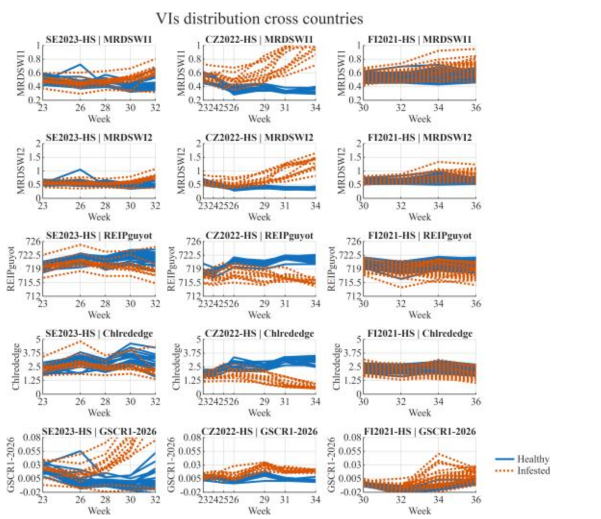

Figure 5. Hyperspectral VI time-series distributions across countries. Healthy vs infested crown trajectories across weeks for selected VIs.

https://doi.org/10.5194/isprs-archives-XLIX-B3-2026-1163-2026

4. Discussion

4.1 Transferability breakpoints across validation tiers

The four-tier evaluation (within-plot → cross-plot within stand → cross-stand within country → cross-country) makes the breakpoints of generalization explicit. Within-plot results (Fig. 1) can reach high kappa in favourable weeks, but this but this high performance is rarely maintained when transferring to unseen plots (Fig. 2) and unseen stands (Fig. 3). Cross-country evaluation (Fig. 4) is therefore the most stringent stress test in this study because it contains the largest failures and the strongest domain shifts (Wang et al., 2026). Importantly, however, transfer difficulty is not determined by geography alone; instead, it tracks the effective domain distance between training and test data (phenology, outbreak dynamics, radiometric baselines, and feature-space overlap).

4.2 Pair-dependent cross-country transferability in hyperspectral data

In the hyperspectral cross-country experiments (Fig. 4b–d), transferability was strongly pair-dependent: GSCR-based models transferred well between SE↔CZ and CZ↔FI, but showed weak transfer for SE↔FI. This indicates that “cross-country” difficulty is not determined by national boundaries per se, but by the effective domain distance between training and testing data. A plausible contributor is that FI2021-HS was acquired in late summer to autumn during the second swarming period, when both spruce phenology and the temporal trajectory of infestation symptoms can differ (Jönsson et al., 2007); in addition, weaker attack intensity may produce a slower or less consistent decline in crown condition, reducing separability relative to earlier-season outbreaks. Figure 5 supports this interpretation by showing baseline shifts and different seasonal trajectories for several indices between SE and FI, implying mis-calibration under direct transfer. Nevertheless, the GSCR range of healthy trees in FI and CZ remains broadly comparable (Fig. 5), consistent with GSCR being comparatively robust to site-specific baselines even when cross-domain classification performance varies.

4.3 Sensor effects: hyperspectral brand differences are not the dominant driver

Importantly, the observed pair-dependence does not support a simplistic assumption that differing sensors inevitably lead to poorer transferability. SE and FI were acquired with SPECIM hyperspectral systems, whereas CZ was acquired with a Headwall system, yet transferability involving CZ was not systematically lower. This suggests that, at least for the indices tested here (notably GSCR), hyperspectral sensor brand differences were not a dominant source of domain shift relative to phenological timing, outbreak dynamics, and site-specific baselines. In other words, cross-country generalization is primarily limited by when and under what outbreak/stand conditions the data are acquired, not merely by the hyperspectral manufacturer.

4.4 Multispectral cross-country failure and red-edge spectral mismatch

Cross-country failure is also evident for the multispectral transfer test (Fig. 4a), where performance drops more consistently than in several hyperspectral pairings. In this case, an additional source of domain shift is sensor spectral mismatch, particularly slightly different red-edge band centre wavelengths between the two multispectral sensors. Because several tested VIs depend directly on red-edge positioning (e.g., red-edge chlorophyll proxies), even small band-centre differences can alter index values and reduce transferability when models are trained in one sensor domain and applied to another (Wang et al., 2026). This highlights that, for multispectral systems, the same “index name” does not necessarily represent the same biophysical measurement across instruments.

4.5 Models versus VIs under domain shift

Within-domain, RF frequently achieves strong kappa and sharp late-season gains (Figs. 1–3), consistent with its ability to exploit nonlinear separation and feature interactions. Under stronger domain shift (especially cross-country; Fig. 4), multivariate models can degrade abruptly or inconsistently, likely because they leverage site-specific covariance structures that are not generalizable across sites (Huo et al., 2023b; Li et al., 2023). In contrast, single-VI GLMs often show lower peak performance but can be more stable across domains for selected indices. The combination of pair-dependent kappa in Fig. 4 and baseline diagnostics in Fig. 5 suggests that a substantial portion of cross-domain failure arises from mis-calibration due to baseline shift, rather than the complete absence of a spectral signal.

4.6 Implications for generalizable UAV early-detection workflows

Overall, cross-country validation remains the most stringent stress test (as it exhibited the most substantial performance drops), but it is not uniformly harder than cross-stand validation for every pairing; transferability is governed by phenological alignment, radiometric/processing consistency, sensor spectral compatibility, and coverage of the test-domain feature space in the training data, rather than geography alone. Practically, this motivates three steps for community benchmarking and deployment: (i) report performance across transfer tiers rather than only within-site results; (ii) match cross-site comparisons by phenophase (e.g., temperature sums) rather than week labels alone (Baier et al., 2007); and (iii) screen for baseline shifts using plots like Fig. 5 and apply simple calibration (e.g., healthy-baseline alignment or anomaly features) prior to transferring decision thresholds.

4.7 Limitations and next steps

This study intentionally uses transparent, reproducible models (RF and GLM) and a compact VI set to isolate transfer effects. Remaining limitations include: (i) cross-country tests are sensitive to the availability of matched time points and may still contain phenological mismatch; (ii) multispectral VI comparability is constrained by band-centre differences; and (iii) cross-domain calibration was not explicitly applied. Next steps are to (a) align cross-site comparisons by phenophase (e.g., growing degree days) (Jönsson et al., 2007), (b) report band-centre/FWHM explicitly for all sensors and VI implementations, and (c) test lightweight calibration schemes (baseline alignment, standardization, or domain-adaptive thresholds) to separate “transfer failure due to baseline shift” from “transfer failure due to missing separability.”

5. Conclusions

- **Within-domain accuracy vs. transferability**: RF and multivariable models (spectra or multi-VI sets) generally achieved the highest within-domain accuracy, but they also showed lower and less stable transferability between sites/stands, implying that strong local performance can reflect site-specific covariance/baselines and should not be interpreted as operational capability without cross-domain validation.

- **Optimal transferable features:** Across the transfer tests, GSCR (especially GSCR1-2026) was among the most consistently transferable VI options in the hyperspectral experiments, and often remained competitive when multivariate models degraded under domain shift.
- **Critical transferability breakpoint:** Cross-stand transfer caused a clear drop and larger variability relative to within-plot/within-stand tests, highlighting stand-level heterogeneity as a major obstacle to generalizable UAV detection.
- **Cross-country is pair-dependent, not geo-distance driven**: In hyperspectral cross-country transfer, GSCR-based models performed well for SE↔CZ and CZ↔FI but weakly for SE↔FI, showing that transferability depends on domain distance (phenology/baselines/outbreak context) rather than geographic distance alone.
- **Hardware variations are not the dominant driver in hyperspectral transfer:** Good transfer involving CZ (Headwall) versus SE/FI (SPECIM) indicates that hyperspectral manufacturer differences were not a primary source of domain shift for the tested indices.
- **Multispectral transferability is highly sensitive to spectral mismatch:** Cross-country transfer failures in multispectral data are consistent with spectral band mismatch. Particularly, slight shifts in red-edge band centre wavelengths directly alter red-edge VIs and significantly reduce model generalizability.

Acknowledgements

The datasets used in this study were assembled within the DroneNet4Beetles initiative, an ongoing collaborative effort to develop and harmonise multi-site drone datasets for the early detection of spruce bark beetle infestation. Another study using the same data as this study is Huo et al. (2026), which provides more details on the data description. We sincerely thank all partners, collaborators, and field teams for their contributions to data collection, harmonisation, and sharing. Efforts are underway to make these datasets publicly accessible, and updates will be provided on the website above. We also welcome contributions from peers interested in sharing data and joining the DroneNet4Beetles initiative; interested researchers may contact langning.huo@slu.se. This work was funded by the Swedish Research Council for Sustainable Development Formas (project grant 2024-00404 to LH) and the SLU Forest Damage Centre.

References

Baier, P., Pennerstorfer, J., & Schopf, A. (2007). PHENIPS—A comprehensive phenology model of Ips typographus (L.) (Col., Scolytinae) as a tool for hazard rating of bark beetle infestation. Forest Ecology and Management, 249(3), 171-186.

Bárta, V., Lukeš, P., & Homolová, L. (2021). Early detection of bark beetle infestation in Norway spruce forests of Central Europe using Sentinel-2. International Journal of Applied Earth Observation and Geoinformation, 100, 102335.

Bijou, S., Kupková, L., Červená, L., & Lysák, J. (2025). Assessing early bark beetle infestation in Norway spruce using multitemporal hyperspectral UAV imagery: Is detection during the green stage always timely? *Ecological Indicators*, 178, 113869.

Bozzini, A., Huo, L., Brugnaro, S., Morgante, G., Persson, H.J., Finozzi, V., Battisti, A., & Faccoli, M. (2025). Multispectral drone images for the early detection of bark beetle infestations: assessment over large forest areas in the Italian South-Eastern Alps. Frontiers in Forests and Global Change, 8, 1532954.

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1), 37-46.

Cosimo, L.H.E., Lindberg, E., Persson, H., Bohlin, J., & Huo, L. (2025). Hyperspectral drone images indicate that green shoulder indices are robust in pre-emergence detection of spruce bark beetle infestation across spatial and temporal scales. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLVIII-2/W11-2025, 73–79.

Gitelson, A.A., Gritz, Y., & Merzlyak, M.N. (2003). Relationships between leaf chlorophyll content and spectral reflectance and algorithms for non-destructive chlorophyll assessment in higher plant leaves. Journal of Plant Physiology, 160, 271-282.

Guyot, G., Baret, F. 1988. “Utilisation de la Haute Resolution Spectrale pour Suivre L’etat des Couverts Vegetaux“. In Spectral Signatures of Objects in Remote Sensing, Proceedings of the conference held 18-22 January, 1988 in Aussois (Modane), Francex, edited by Guyenne, T.D., and J.J. Hunt. European Space Agency. 279.

Hlásny, T., Zimová, S., Merganičová, K., Štěpánek, P., Modlinger, R., & Turčáni, M. (2021). Devastating outbreaks of bark beetles in the Czech Republic: Drivers, impacts, and management implications. Forest Ecology and Management, 490, 119075.

Huo, L., Cosimo, L.H.E., Bozzini, A., Bijou, S., Oliveira, R.A. de, Suomalainen, J., Vennervirta, E., Koivumäki, N., Näsi, R., Kupková, L., Faccoli, M., & Honkavaara, E. (2026). A cross-European assessment on the pre-emergence detection of trees attacked by spruce bark beetle using UAV imagery. Remote Sensing of Environment, 342, 115445.

Huo, L., Koivumäki, N., Oliveira, R.A., Hakala, T., Markelin, L., Näsi, R., Suomalainen, J., Polvivaara, A., Junttila, S., & Honkavaara, E. (2024). Bark beetle pre-emergence detection using multi-temporal hyperspectral drone images: Green shoulder indices can indicate subtle tree vitality decline. ISPRS Journal of Photogrammetry and Remote Sensing, 216, 200–216.

Huo, L., Lindberg, E., Bohlin, J., & Persson, H.J. (2023). Assessing the detectability of European spruce bark beetle green attack in multispectral drone images with high spatial- and temporal resolutions. Remote Sensing of Environment, 287, 113484.

Huo, L., Persson, H.J., Bohlin, J., & Lindberg, E. (2023). Green Attack or Overfitting? Comparing Machine-Learning- and Vegetation-Index-Based Methods to Early Detect European Spruce Bark Beetle Attacks Using Multispectral Drone Images. In IGARSS 2023 IEEE International Geoscience and Remote Sensing Symposium, Pasadena, CA, USA (pp. 546-549).

Huo, L., Persson, H.J., & Lindberg, E. (2021). Early detection of forest stress from European spruce bark beetle attack, and a new vegetation index: Normalized distance red & SWIR (NDRS). Remote Sensing of Environment, 255, 112240.

Jönsson, A.M., Harding, S., Bärring, L., & Ravn, H.P. (2007). Impact of climate change on the population dynamics of Ips typographus in southern Sweden. Agricultural and Forest Meteorology, 146(1-2): 70-81.

Li, N., Huo, L., & Zhang, X. (2023). Exploring Common Hyperspectral Features of Early-Stage Pine Wilt Disease at Different Scales, for Different Pine Species, and at Different Regions. In 2023 IEEE International Geoscience and Remote Sensing Symposium IGARSS (pp. 2023: 7575-7578).

McCullagh, P. (2019). Generalized Linear Models. Routledge.

Näsi, R., Honkavaara, E., Lyytikäinen-Saarenmaa, P., Blomqvist, M., Litkey, P., Hakala, T., Viljanen, N., Kantola, T., Tanhuanpää, T., & Holopainen, M. (2015). Using UAV-based photogrammetry and hyperspectral imaging for mapping bark beetle damage at tree-level. Remote Sensing, 7(11), 15467-15493.

Rouse, J. W., Haas, R. H., Schell, J. A., & Deering, D. W. (1973). Monitoring vegetation systems in the Great Plains with ERTS.

Seidl, R., Thom, D., Kautz, M., Martin-Benito, D., Peltoniemi, M., Vacchiano, G., Wild, J., Ascoli, D., Petr, M., Honkaniemi, J., & Lexer, M.J. (2017). Forest disturbances under climate change. Nature Climate Change, 7(6), 395-402.

Tucker, C. J. (1979). Red and photographic infrared linear combinations for monitoring vegetation. Remote Sensing of Environment, 8(2), 127–150.

Wang, G., Chai, G., Aierken, N., Chen, L., Wang, J., Qian, Z., Lei, L., & Zhang, X. (2026). Monitoring forest damage caused by pine wilt disease with remote sensing technologies: A review. IEEE Geoscience and Remote Sensing Magazine, 14(1), 286-305.