

Urban Green Space Classification from Satellite Imagery: A Robustness Evaluation

Antonio Iyda Paganelli*, Patrick Nigri Happ†, Adriano Branco*, Luca Mottola‡

*Department of Informatics, Pontifical Catholic University of Rio de Janeiro, Brazil

Email: {apaganelli, abranco}@inf.puc-rio.br

†Department of Electrical Engineering, Pontifical Catholic University of Rio de Janeiro, Brazil

Email: patrick@puc-rio.br

‡Politecnico di Milano, Uppsala University, and RISE, Italy and Sweden

Email: luca.mottola@polimi.it

Abstract—We study the robustness of five model families spanning four classical supervised classifiers and one convolutional U-Net, tested with two backbones in identifying urban green spaces from satellite imagery. We test these techniques against a *real-world* application case across variations in locations, seasons, and years. We train the models once on autumn imagery from a single reference city in Lombardy, Italy, and investigate the robustness of a technique without retraining across shift scenarios. Our results show that urban green space classification shows partial robustness across shifts, with dominant classes remaining stable while the minority bare soil class represents a consistent failure point and the source of algorithm-specific vulnerabilities. Two degradation mechanisms explain most failures: signal shift, which drives interannual and seasonal degradation, and class imbalance, which modulates spatial transfer performance. The latter severely degrades bare soil detection when it is underrepresented at the target site, while stabilizing or improving it when abundant. We also consistently find that across all models and shift types, bare soil F1 degrades earlier and more severely than overall accuracy; in winter, for instance, Random Forest misses bare soil entirely with an F1 score of zero, while macro F1 still reads 0.66. This insight suggests bare soil F1 as a candidate early warning indicator of model instability.

Index Terms—urban green spaces, satellite imagery, machine learning, land cover classification, robustness, natural shifts.

I. INTRODUCTION

More than half of the world’s population lives in cities. In Europe, this share exceeds 75% and continues to accelerate [1]. Green areas within cities, including parks, street trees, gardens, and other vegetated patches, deliver benefits that range from temperature regulation and flood mitigation to pollution reduction [2], [3]. Taken together, this interconnected network of green elements, often called green infrastructure (GI), is increasingly seen as a core strategy for urban climate-change adaptation [3]. Managing it effectively, however, requires up-to-date information on where these areas are and how they evolve. Obtaining such information traditionally relies on field surveys, ground-based sensors, and manual photo-interpretation [4], [5]. These approaches are laborious, expensive, and difficult to replicate at the frequency and geographic extent that growing urban regions demand [6].

Satellite imagery. Satellite constellations offer a way to close this gap. Satellite sensors can capture imagery data of vast areas repeatedly, at low marginal cost, and under consistent

measurement conditions [7]. Programs such as Europe’s Sentinel-2 provide freely available imagery every five days [8], while commercial fleets like PlanetScope’s SuperDove cubesats deliver imagery with near-daily revisit [9]. Although satellite imagery still requires photo-interpretation, it represents an alternative for monitoring urban vegetation across metropolitan regions over extended periods [10], [11]. Doing so requires translating raw pixel measurements into land-cover maps, that is, assigning each location to a category such as green area, bare soil, or built-up surface [4], [12].

This translation is currently performed by supervised classification algorithms that learn each category’s spectral and spatial profile from labeled examples and generalize to unseen data [13]–[15]. These algorithms reach overall accuracy above 80% when trained and evaluated on data from the same source [12], [13]. These results, however, are typically obtained under controlled conditions: training and test images come from the same region, sensors, and period of the year [14], [16]. Whenever the statistical properties of new images differ from those seen during training, a phenomenon commonly referred to as *distribution shift* [17], [18], accuracy may sharply deteriorate. Handling distribution shifts in image classification is a well-documented challenge across machine learning and deep learning algorithms [19], [20].

For urban green space monitoring, such shifts are very common. Vegetation looks markedly different across seasons, and inter-annual climate variability adds further change from one year to the next [21], [22]. Moreover, a model trained on images from one location cannot be assumed to generalize to a nearby area with different plant species, urban density, or built environment [22], [23]. Yet, these are precisely the scenarios that make satellite monitoring valuable, namely the ability to *re-apply* a trained model across many locations and time periods at low additional cost [8]. Rebuilding or retraining a model for every new location, season, or year demands specialized expertise and substantial computational resources, making it impractical at scale [21].

Robustness to distribution shifts is therefore a key prerequisite for reusing models across space and time. Studies addressing this issue typically examine one source of variability at a time, treating temporal shifts, geographic transfers,

and sensor differences as separate problems [22], [24], [25]. This fragmentation makes it difficult to compare results across studies or to identify which type of shift is most damaging.

Crucially, however, the current state of affairs raises the question of how robustness performance, as reported in existing literature, carries over from controlled to *real-world conditions*, where distribution shifts occur along multiple dimensions in uncontrolled ways. The few existing results are somewhat contradictory, with some studies finding that methods evaluated on synthetically corrupted images fail to generalize to naturally occurring shifts [19], [26], while others report limited robustness depending on the shift type [20]. These gaps leave urban planners and designers with incomplete and often contradictory guidance on which algorithms to trust when conditions change in real deployments.

Contribution and results. To address this gap, we trained five classifiers on imagery from a single reference city and season, and evaluated them across three types of natural distributional shift: *spatial*, by testing on imagery from other cities; *interannual*, by testing on imagery from two other years in the same city and season; and *seasonal*, by testing on imagery from other seasons in the same city and year. Each type of shift is evaluated independently and without scenario-specific retraining. Beyond shift types, we treat *class imbalance*, an inherent property of urban land cover [27], [28], as a natural feature of the data rather than a problem to be corrected, with only a lightweight post-processing step applied to the minority class. This setup allows the effect of each shift type to be attributed with greater confidence and compared across algorithms, under a common controlled protocol.

Our results show that no single technique is robust across all shift types, and that each model has a distinct vulnerability profile associated with the shift type. Class imbalance and signal shift emerge as two degradation mechanisms, with class imbalance modulating spatial transfer performance and signal shift driving interannual and seasonal degradation. We also find that bare soil F1, the minority class in naturally imbalanced urban scenes, consistently degrades earlier and more severely than macro-level metrics. In winter, for instance, a Random Forest model misses bare soil entirely resulting in an F1 score of 0, while the macro F1 score still reads 0.66. This suggests that minority-class performance may act as an early and accessible indicator of broader model instability.

II. RELATED WORK

The robustness of classifiers when applied to satellite data works depends on four interconnected threads: *i)* the imagery and features available for classification, *ii)* the algorithms that process them, *iii)* the conditions under which those algorithms degrade, and *iv)* the challenge posed by minority classes.

A. Satellite Imagery for Urban Surface Mapping

What a satellite measures, in practice, is a snapshot of how each ground patch reflects sunlight across multiple wavelength bands, including ranges invisible to the human eye.

Vegetation, bare soil, and paved surfaces have distinct reflectance profiles, particularly in the near-infrared range, where healthy green cover reflects strongly while bare soil and built surfaces absorb. Derived scores exist, such as the Normalized Difference Vegetation Index (NDVI), that quantify the density and health of green cover [29]. These scores can compress the features of imagery data into signals that classification algorithms can use directly.

Sensor choice determines how much urban detail is preserved: Sentinel-2, a widely used free-access satellite constellation, provides 10-meter pixels [8], which in dense urban areas can mix heterogeneous surfaces (tree canopy, vehicles, pavement) into a single measurement, whereas PlanetScope SuperDove offers 3-meter pixels across eight spectral bands and near-daily revisit, enabling finer discrimination of urban green features [9], [30]–[32]. Even at high resolution, mapping remains difficult because spectrally similar surface types, such as dry grass and bare soil, can be nearly indistinguishable in many band combinations [10], [11]. The corresponding signatures are not fixed but shift with season, location, and time, manifesting across years as inter-annual change.

B. Classification Algorithms

Supervised classifiers learn from manually labeled pixels and assign land-cover labels to the remaining grid cells. These include, for example, logistic regression [33], support vector machines [34], [35], tree ensembles such as random forest [36], and XGBoost, as well as convolutional architectures such as U-Net [37], [38]. The first three classify each pixel from its spectral values alone, differing in how they draw class boundaries and in their tolerance to noisy inputs, whereas U-Net analyzes each pixel together with its spatial neighbors inside a moving window or tile, learning shape and texture patterns rather than isolated spectral values. However, this usually requires more labeled training data.

A comprehensive review reports overall accuracy for machine learning methods spanning 73–97% and deep learning architectures such as U-Net reaching 96–99% [6], [38]. These ranges, however, aggregate results from studies using entirely different sensors, tasks, class schemes, geographic regions, and data preparation methods, making it difficult to understand whether one technique would outperform another on any specific setup [13], [39], [40].

C. Performance Degradation Under Shifting Conditions

The core issue behind accuracy degradation is that spectral signatures are not fixed properties of material. They change with the state of the surface being observed. Vegetation reflects light differently depending on its phenological stage, the point in the plant’s annual growth cycle from budding in spring to leaf drop in autumn, so the same tree canopy can produce NDVI values close to zero in winter and above 0.7 at peak summer [41], [42]. Urban surface materials, dominant tree species, and local atmospheric conditions also vary from city to city, meaning a classifier trained on October imagery from

one municipality encounters a statistically different input when applied to another location or season.

Accuracy losses due to shifts are well documented in the literature. In crop mapping, for instance, models transferred across regions show systematic per-class accuracy drops [23], and existing reviews report overall accuracy (OA) for machine learning methods spanning 24 percentage points across studies with different sensors and geographic contexts, a spread too large to be explained by algorithm differences alone [6]. It remains contested whether training with synthetic perturbations, that is, artificially constructed corruptions meant to mimic real conditions, improves robustness to natural shifts. Some studies suggest that natural shifts are more harmful than synthetic perturbations predict [19], while others report that synthetic training only offers partial protection [20]. Domain adaptation methods, which realign the training and deployment distributions, serve as a mitigation strategy [25], but they are evaluated against a single shift type at a time. This leaves open the question of how different shift types affect different algorithm families under controlled, comparable conditions.

D. Class Imbalance in Urban Land-Cover Classification

Urban land cover is rarely distributed evenly across classes. Paved surfaces typically dominate city scenes, while bare soil covers only a small fraction of any urban image [27], [28], [38]. This imbalance creates a subtle trap. Overall accuracy, which counts the percentage of correctly labeled pixels across all classes, can appear high even when a model has learned to ignore a minority class entirely [14]: a classifier that labels every pixel as vegetation in a scene where bare soil covers 5% of the area still achieves 95% overall accuracy.

Per-class metrics reveal failures that overall accuracy masks [14]. For example, producers' accuracy measures the fraction of actual bare soil pixels the model correctly identifies. Yet, users' accuracy measures the fraction of bare soil predictions that were correct, and the F1-score combines both into a single high value only when the model is good at both. F1-macro then averages this score across all classes, treating each class equally, regardless of how frequently it appears in the image. Standard remedies include oversampling minority classes with synthetic examples before training [27] and penalizing minority-class errors more heavily through modified loss functions [43]. Bare soil remains particularly difficult anyway, as its reflectance overlaps with dry vegetation and bright impervious surfaces, while its spatial extent varies with construction activity and season [28], [38].

E. This Work

Despite the existing literature, it is difficult to identify a study that tests multiple shift types on the same set of algorithms within a common experimental setup. Because existing work addresses each shift in isolation, it is problematic to tell which degradation patterns are intrinsic to each model family and which arise from differences in study design. In particular, while individual studies hint at specific algorithm

vulnerabilities, it is unclear whether different algorithm families exhibit distinct degradation patterns across shift types when trained and evaluated under the same conditions, and how minority-class imbalance interacts with that degradation. In contrast, this work presents such a systematic evaluation.

III. METHODOLOGY AND SETUP

We use satellite imagery from small municipalities in Lombardy, Italy, to train and test five supervised classifiers under a common protocol covering study design, feature extraction, model training, and evaluation metrics.

A. Study Design and Scenarios

Our study is designed to evaluate a concrete deployment-oriented question: whether a classifier trained once on a controlled reference domain can be reused under the natural distributional shifts that arise in urban green-space monitoring. Following the Vegetation-Impervious-Surface standard [44], we consider three land-cover classes: *i*) green vegetation, *ii*) built-up areas, and *iii*) bare soil. This keeps the scheme simple enough for controlled cross-scenario comparison.

All experiments use PlanetScope SuperDove imagery across eight spectral bands acquired over Lombardy, Italy. The town of Codogno, south of Milano, serves as the reference domain, providing cloud-free autumn imagery and manually labeled polygons drawn in QGIS and cross-checked against high-resolution Google and Esri satellite basemaps. Models are trained on Codogno autumn data only and then evaluated without scenario-specific retraining in four settings:

- 1) a *baseline* test on unseen Codogno autumn data;
- 2) *spatial transfer* to two nearby municipalities, using imagery from the same season and year;
- 3) *interannual transfer* to past-year acquisitions at the same location, keeping season fixed;
- 4) *seasonal transfer* to winter, spring, and summer imagery from the same year and location.

This design varies one condition at a time, while keeping the others fixed, enabling direct comparison of natural distributional shifts across model families.

B. Feature Extraction and Data Preparation

All methods use the same input representation to ensure comparability across models. This design follows three principles: preserving the sensor-native spectral information, incorporating domain-informed vegetation cues, and adding limited local context. We retain the eight spectral bands and compute three metrics: NDVI as introduced earlier, the Normalized Difference Water Index (NDWI) [45], which contrasts the green and near-infrared bands to highlight sensitivity to vegetation water changes [46], and the red-edge ratio, which captures chlorophyll-related contrasts that NDVI, which uses only the red and NIR bands, does not [47]. For each normalized spectral band, we additionally compute the local mean and standard deviation over a 7×7 pixel neighborhood to encode small-scale spatial context. Altogether, the 8 spectral bands, 3 spectral indices, and 16 local statistical features

yield a 27-dimensional representation. We normalize channels independently for each scene using robust percentile scaling (2nd–98th percentiles) to reduce sensitivity to outliers and scene-specific radiometric extremes.

Scenes are divided into overlapping 128×128 tiles. Pixels flagged as cloud- or shadow-contaminated by the PlanetScope UDM2 mask are excluded from both training and evaluation. We use two disjoint polygon sets: 294 polygons for model fitting and validation, and 24 independently annotated held-out polygons over unseen regions of the same scene, on which all reported metrics are computed. The spatial, interannual, and seasonal transfer scenarios are evaluated against their own independently annotated polygons. During training, we apply standard data augmentation through random geometric transformations and mild radiometric perturbations.

C. Models and Training Protocol

We evaluate five families of supervised models with different algorithmic principles: Logistic Regression (LogReg), Support Vector Machine with radial basis function kernel (SVM), Random Forest (RF), XGBoost (XGB), and a lightweight U-Net architecture (U-Net) tested with two backbones, ResNet-18 (RNet18) and MobileNet-V2 (MNetV2). These models span linear, kernel-based, tree-ensemble, boosting, and convolutional approaches, allowing robustness to be assessed across distinct algorithmic principles rather than within a single model family. All models use the same data volume, input representation, and the same train/validation split.

For the machine learning models, we train on random samples of active pixels extracted from the training tiles, with at most 200,000 pixels for RF and XGB and 50,000 for SVM, to keep optimization tractable and consistent across repeated runs. We choose hyperparameters based on preliminary baseline-oriented trials on the Codogno autumn dataset. Our goal is not exhaustive model tuning, but to define reasonable, stable configurations that can be reused unchanged across all evaluation scenarios. RF uses 200 trees with unrestricted depth and out-of-bag estimation; SVM uses an RBF kernel with $C = 1.0$ and $\gamma = \text{scale}$; and XGB uses 200 boosting rounds, maximum depth 6, learning rate 0.1, and subsampling ratios of 0.8 for both rows and features.

We apply the same strategy for the two U-Net variants we test. The two encoder variants, RNet18 and MNetV2, are trained from scratch with AdamW, a learning rate of 10^{-3} , and mini-batches of 8 tiles, for up to 80 epochs with early stopping (patience 15) on the validation loss. We minimize a combined loss equal to $0.7 \text{CrossEntropy} + 0.3 \text{Dice}$, weighting the cross-entropy by class frequencies to handle imbalance.

Standard softmax thresholds used in classifier algorithms tend to underperform on minority classes because the model is implicitly biased toward more frequent classes during training. To improve the robustness of the minority bare soil class, we refine predictions with a lightweight probability-based post-processing step using a class-specific threshold τ_2 and a confidence margin for ambiguous cases. We calibrate these

TABLE I
BASELINE PERFORMANCE ON CODOGNO (MEAN $\pm$ STD OVER 20 RUNS).

Model	$F1_{\text{macro}}$	Urban F1	Green F1	Bare soil F1
SVM	0.971 ± 0.001	0.995 ± 0.001	0.963 ± 0.001	0.956 ± 0.001
RF	0.967 ± 0.008	0.992 ± 0.006	0.957 ± 0.003	0.953 ± 0.018
XGBoost	0.965 ± 0.004	0.992 ± 0.002	0.950 ± 0.006	0.953 ± 0.006
LogReg	0.921 ± 0.165	0.973 ± 0.064	0.918 ± 0.216	0.871 ± 0.226
MNetV2	0.951 ± 0.014	0.991 ± 0.006	0.943 ± 0.020	0.920 ± 0.024
RNet18	0.932 ± 0.051	0.979 ± 0.027	0.933 ± 0.023	0.885 ± 0.123

parameters once on the Codogno autumn validation split and then fix them for all scenarios.

D. Evaluation and Statistical Analysis

Across all scenarios, all models use the same input features, training data, and evaluation protocol, with no scenario-specific retraining, recalibration, or data adjustment. This ensures that differences in results can be attributed to the algorithms and shift types rather than to confounding factors in the experimental setup.

Under this common protocol, we evaluate performance at the pixel level using Overall Accuracy (OA), the macro-average F1-score ($F1_{\text{macro}}$), and per-class F1-Scores. OA is reported for comparison with prior work and to provide an overall summary of map accuracy, although, as discussed in Sec. II, it can obscure minority-class failures under class imbalance. For this reason, $F1_{\text{macro}}$ and per-class metrics are used to support the interpretation of robustness across scenarios. To account for the stochastic nature of training and sampling, we repeat every model-scenario combination 20 times, each with a different random seed. Results are reported as mean $\pm$ standard deviation. We perform pairwise comparisons using the Wilcoxon signed-rank test on per-run scores, with Holm–Bonferroni correction at $\alpha = 0.05$.

IV. RESULTS

We organize the section around three natural distributional shifts, spatial, interannual, and seasonal, each evaluated against a strong in-domain baseline. Beyond reporting performance for each shift, we introduce the degradation mechanisms that emerge from the data and identify how each family of algorithms responds to them.

A. Baseline Performance (In-domain Reference)

All models achieve strong in-domain performance with $F1_{\text{macro}} \geq 0.92$, as seen in Table I, with SVM attaining the highest score and minimal variability across runs. Per-class results reveal a different situation, though. The urban class is nearly trivial scoring above 0.97 for all models. Green vegetation, the target class of this study, is also consistently well classified, with $F1_{\text{macro}} \geq 0.93$ for every model except LogReg, and a tight run-to-run spread. Bare soil is the most challenging class: SVM, RF, and XGBoost still exceed 0.95, but LogReg and both UNet variants score lower and display substantially higher run-to-run variability.

Pairwise differences in $F1_{\text{macro}}$ are significant across all model pairs. This establishes a strong in-domain reference, ensuring that subsequent degradation under shift reflects genuine robustness differences rather than insufficient model capacity.

B. Spatial transfer

Spatial transfer performance is strongly dependent on bare soil prevalence at the target site, as seen in Table II. In Bresso, where bare soil covers only 0.98% of labeled pixels, $F1_{\text{macro}}$ performance drops sharply across all models despite near-perfect OA, with bare soil F1 losses ranging from 0.38 (SVM) to 0.62 (MNetV2). F1 performance for the urban and green classes remains high for every model (≥ 0.94), confirming that the $F1_{\text{macro}}$ collapse is driven entirely by failure on the bare soil class rather than a generalized loss of transfer capability. In Cusano Milanino, where bare soil covers 35.99%, most models maintain or improve performance.

F1 for Urban and Green classes is also stable in Cusano Milanino (≥ 0.97) for all models except RF, whose Urban class F1 falls to 0.91 despite favorable class prevalence, diverging from XGBoost, which improves under the same conditions. Bare soil class imbalance dominates the modulation of results, causing a severe degradation in $F1_{\text{macro}}$ when extremely underrepresented at the target site, while stabilizing or improving it relative to baseline performance when overrepresented.

C. Interannual transfer

Interannual transfer, shown in Table III, reveals year-dependent degradation that cannot be explained by class distribution alone. Bare soil prevalence is higher in 2020 (25.8%) than in 2021 (13.4%), yet degradation is substantially more severe in 2020, particularly for tree-based ensembles: RF collapses to $F1_{\text{macro}} = 0.62$ and Bare Soil F1 = 0.16, while XGBoost follows a similar pattern.

The 2020 shift also impacts the Urban class, whose F1 drops to 0.72 for RF and 0.79 for XGBoost, while F1 for the Green class remains stable for all models. Both classes recover in 2021. SVM maintains comparatively stable performance across both years. This dissociation between class prevalence and degradation severity points to signal shift as the primary driver of interannual instability.

Moreover, the shift, propagating into the abundant Urban class while sparing the Green class, distinguishes this scenario from spatial transfer, in which only the bare soil class was affected. Urban F1 exposes a clear failure mode. In 2020, RF Urban F1 drops by 0.27, XGBoost by 0.20, and MNetV2 by 0.17, despite urban remaining the dominant class. Signal shift might reflect land-use change, such as new construction or soil disturbance between years, which alters the spectral profile, especially for bare soil, independently of its spatial extent [28].

D. Seasonal Transfer

Seasonal transfer reveals that winter is the most challenging condition across all models, while spring yields the highest stability, as observed in Table IV. This pattern cannot be attributed solely to class distribution: winter and spring have

similar bare soil prevalence (11.2% and 12.6%, respectively), yet winter induces substantially greater degradation. Tree-based ensembles are most vulnerable, with RF collapsing to $F1_{\text{macro}} = 0.66$ and Bare Soil F1 = 0.00 in winter. UNet variants show higher winter resilience, with RNet18 $F1_{\text{macro}} = 0.88$ and especially Bare Soil F1 = 0.67. Summer exhibits an intermediate behavior, with additional Bare Soil degradation partly explained by lower class prevalence (7.6%). Phenological cycles alter surface reflectance substantially across the year [41], [42]. In winter, most models collapse spectral separability regardless of class prevalence. In these conditions, UNet architectures may be more resilient based on their spatial recognition capabilities [37], [38].

Urban and Green class F1 remain at or above 0.94 across all models and seasons, confirming that seasonal drift concentrates almost entirely in the bare soil class. Bare soil is spectrally unstable by nature, overlapping with dry vegetation and bright impervious surfaces across seasons and locations [28]. In winter, despite moderate bare soil prevalence (11.2%), XGBoost and SVM retain $F1_{\text{macro}}$ above 0.74 while Bare Soil F1 falls to 0.26 and 0.32, respectively. Even UNet, the most seasonally resilient model, drops to Bare Soil F1 = 0.67 while maintaining $F1_{\text{macro}} = 0.88$.

E. Degradation Mechanisms

Our results reveal two distinct degradation mechanisms that operate independently across shift types. Class imbalance, where the target domain contains a substantially different proportion of bare soil than the training data, modulates spatial transfer, consistent with evidence from crop mapping showing systematic per-class accuracy drops when class prevalence differs across regions [23]. Signal shift, where the spectral signature of the same surface type changes, dominates inter-annual and seasonal transfer. These two mechanisms do not occur systematically together; their relative contribution varies by shift type, which helps explain why robustness studies examining one shift type in isolation produce results that are difficult to generalize [19], [22].

The independence of these two mechanisms is illustrated by a direct comparison. When bare soil is nearly absent at the target site (Bresso, 0.98%), all models degrade on that class, including SVM, which drops to 0.38 in Bare Soil F1, showing that class imbalance alone is sufficient to cause failure. When bare soil is abundant (2020, 25.8%), class distribution cannot explain degradation, yet RF collapses to $\Delta F1_{\text{macro}} = -0.35$; $\Delta \text{BareSoilF1} = -0.79$, while SVM degrades only modestly with $\Delta F1_{\text{macro}} = -0.08$, showing that signal shift acts independently and affects models differently.

The relative contribution of each mechanism varies by shift type, and Bare Soil F1 consistently degrades earlier and more severely than macro-level metrics, while overall accuracy remains misleadingly stable [14]. Tracking Bare Soil F1 alongside OA during deployment could therefore provide an accessible early warning signal before broader map quality deteriorates [27], [48], making it the most sensitive indicator of model instability across all scenarios.

TABLE II

SPATIAL TRANSFER PERFORMANCE AND PENALTY RELATIVE TO THE CODOGNO UNSEEN BASELINE, ORGANIZED BY TARGET SITE. **BOLD**: BEST VALUE FOR COLUMN WITHIN SITE; *italic*: WORST VALUE.

Model	OA	$F1_{\text{macro}}$	$\Delta F1_{\text{macro}}$	Urban F1	Green F1	Bare Soil F1	Δ Bare Soil F1
<i>Target: Bresso (bare soil $\approx 0.98\%$)</i>							
LogReg	0.95 ± 0.12	0.75 ± 0.14	-0.17	0.96 ± 0.08	0.95 ± 0.22	0.35 ± 0.14	-0.52
RF	0.99 ± 0.00	0.79 ± 0.01	-0.18	0.99 ± 0.00	1.00 ± 0.00	0.37 ± 0.03	-0.59
SVM	0.99 ± 0.00	0.86 ± 0.04	-0.12	1.00 ± 0.00	1.00 ± 0.00	0.57 ± 0.10	-0.38
XGBoost	0.99 ± 0.00	0.83 ± 0.04	-0.14	0.99 ± 0.00	1.00 ± 0.00	0.50 ± 0.12	-0.46
MNetV2	0.95 ± 0.02	0.74 ± 0.05	-0.21	0.95 ± 0.02	0.99 ± 0.01	0.30 ± 0.14	-0.62
RNet18	0.95 ± 0.02	0.73 ± 0.04	-0.20	0.94 ± 0.03	0.99 ± 0.01	0.26 ± 0.09	-0.62
<i>Target: Cusano Milanino (bare soil $\approx 35.99\%$)</i>							
LogReg	0.92 ± 0.11	0.91 ± 0.15	-0.02	0.97 ± 0.09	0.89 ± 0.17	0.86 ± 0.20	-0.01
RF	0.90 ± 0.05	0.90 ± 0.07	-0.07	0.91 ± 0.04	0.98 ± 0.03	0.80 ± 0.15	-0.16
SVM	0.99 ± 0.00	0.99 ± 0.01	+0.02	0.99 ± 0.00	0.99 ± 0.01	0.98 ± 0.01	+0.03
XGBoost	0.99 ± 0.02	0.99 ± 0.02	+0.02	0.99 ± 0.02	0.99 ± 0.01	0.98 ± 0.04	+0.03
MNetV2	0.97 ± 0.02	0.97 ± 0.02	+0.02	0.98 ± 0.01	0.98 ± 0.03	0.95 ± 0.04	+0.03
RNet18	0.98 ± 0.02	0.98 ± 0.02	+0.05	0.99 ± 0.01	0.98 ± 0.03	0.97 ± 0.03	+0.08

TABLE III

INTERANNUAL PERFORMANCE (2020–2021); VALUES ARE MEAN $\pm$ STD.
BARE SOIL PREVALENCE: 2020 = 25.8%, 2021 = 13.4%.

Model	Year	$F1_{\text{macro}}$	Urban F1	Green F1	Bare Soil F1
LogReg	2020	0.92 ± 0.18	0.94 ± 0.10	0.91 ± 0.22	0.89 ± 0.25
LogReg	2021	0.90 ± 0.16	0.98 ± 0.06	0.93 ± 0.22	0.79 ± 0.21
RF	2020	0.62 ± 0.06	0.72 ± 0.02	0.98 ± 0.00	0.16 ± 0.15
RF	2021	0.90 ± 0.09	0.99 ± 0.02	0.98 ± 0.00	0.74 ± 0.26
SVM	2020	0.89 ± 0.04	0.88 ± 0.04	0.98 ± 0.00	0.82 ± 0.09
SVM	2021	0.95 ± 0.00	1.00 ± 0.00	0.98 ± 0.00	0.87 ± 0.01
XGB	2020	0.75 ± 0.13	0.79 ± 0.08	0.98 ± 0.00	0.48 ± 0.32
XGB	2021	0.94 ± 0.01	0.99 ± 0.00	0.97 ± 0.00	0.85 ± 0.02
MNetV2	2020	0.80 ± 0.13	0.83 ± 0.09	0.97 ± 0.04	0.61 ± 0.30
MNetV2	2021	0.90 ± 0.03	0.98 ± 0.02	0.96 ± 0.02	0.76 ± 0.07
RNet18	2020	0.91 ± 0.07	0.90 ± 0.07	0.97 ± 0.01	0.85 ± 0.15
RNet18	2021	0.90 ± 0.05	0.98 ± 0.02	0.96 ± 0.02	0.77 ± 0.12

F. Algorithms' Vulnerabilities

The differences in vulnerability reflect how an algorithm processes satellite imagery. SVM identifies boundaries between classes using the most contrasting signal combinations, which helps it remain stable when signals shift moderately [34], [35]: even in winter, the most challenging season, it retains the highest $F1_{\text{macro}} = 0.77$ among classical models. LogReg fits linear boundaries directly on per-pixel spectral values, without kernelization, ensembling, or spatial context [33]. Its mean performance remains competitive across shifts, but its run-to-run variance is higher than other classical models. RF builds many independent decision trees and averages their answers, making it sensitive when the signal of a class changes substantially across time or season [36]: it collapses in winter with Bare Soil $F1 = 0.00$ and in 2020 with $F1_{\text{macro}} = 0.62$, yet performs well in spring with $F1_{\text{macro}} = 0.95$ when signals are stable. XGBoost learns sequentially, focusing on the most informative signal combinations, which appears to confer greater stability than RF under signal shift. In 2020, XGBoost achieved $F1_{\text{macro}} = 0.75$ compared to RF's 0.62, despite similar baseline performance.

UNet variants incorporate the spatial arrangement of pixels around each location, which helps when seasonal changes reduce spectral contrast [37], [38]. In winter, UNet achieves

$F1_{\text{macro}} = 0.88$ and Bare Soil $F1 = 0.67$, substantially above RF's 0.66 and 0.00, respectively. This advantage does not transfer to new cities, where spatial patterns also differ from those in training. Worth noting is that UNet variants are trained on the same data volume as machine learning models, despite their greater data requirements [37], [38]. Their spatial transfer vulnerability may therefore reflect data constraints as much as architectural characteristics, and their performance could improve with larger or more diverse training sets.

Signal shift dominates interannual and seasonal transfer. Degradation is driven by changes in spectral signatures rather than class prevalence, as evidenced by RF and XGBoost collapsing in 2020 despite abundant bare soil, and by winter inducing greater degradation than spring with similar bare soil fractions. No single model dominates across shifts. SVM is robust, tree-based ensembles are most vulnerable to signal shift, and UNet variants are resilient to seasonal shift but more vulnerable to spatial transfer. Across all scenarios, Bare soil $F1$ degrades earlier and more severely than macro-level metrics.

V. DISCUSSION

We reveal that distributional shift affects the three classes asymmetrically. The green class is unaffected by the shifts we test. The urban class is robust in most settings, with modest drops in spatial transfer and a sharp drop in the 2020 interannual shift, whereas the bare soil class is most consistently affected. These findings carry direct implications for practitioners deploying satellite-based urban green space monitoring. Extending this evaluation to other climatic regions and urban morphologies remains an important next step.

Before deploying a trained model to a new city, year, or season, practitioners should *verify that both the class prevalence and the spectral signatures of each class at the target site resemble those of the training data*. Class-prevalence mismatch predicts degradation of rare classes, as observed for bare soil in our experiments. Even dominant classes can fail when their spectral appearance drifts, as shown by the 2020 interannual transfer, which occurred without any corresponding change in class prevalence, though the specific cause of this signal shift is a natural follow-on investigation. Moreover, *model*

TABLE IV
SEASONAL TRANSFER ON CODOGNO (AUTUMN IS PRESENTED AT TABLE I).

Metric	Season	LogReg	RF	SVM	XGB	RNet18	MNetV2
OA	winter	0.97 ± 0.07	0.98 ± 0.00	0.98 ± 0.00	0.98 ± 0.00	0.99 ± 0.01	0.98 ± 0.01
	spring	0.97 ± 0.11	0.99 ± 0.00	0.99 ± 0.00	1.00 ± 0.00	0.99 ± 0.01	0.99 ± 0.01
	summer	0.96 ± 0.11	0.96 ± 0.01	0.98 ± 0.00	0.97 ± 0.00	0.98 ± 0.01	0.98 ± 0.01
Macro F1	winter	0.76 ± 0.15	0.66 ± 0.00	0.77 ± 0.05	0.74 ± 0.05	0.88 ± 0.09	0.85 ± 0.09
	spring	0.91 ± 0.16	0.95 ± 0.01	0.95 ± 0.00	0.96 ± 0.01	0.94 ± 0.03	0.94 ± 0.02
	summer	0.93 ± 0.15	0.75 ± 0.08	0.90 ± 0.01	0.85 ± 0.01	0.90 ± 0.06	0.92 ± 0.04
Urban F1	winter	0.99 ± 0.04	0.99 ± 0.00	0.99 ± 0.00	0.99 ± 0.00	0.99 ± 0.01	0.99 ± 0.01
	spring	0.98 ± 0.07	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.01	0.99 ± 0.01
	summer	0.97 ± 0.08	0.97 ± 0.01	0.99 ± 0.00	0.99 ± 0.00	0.98 ± 0.01	0.99 ± 0.01
Green F1	winter	0.94 ± 0.22	0.98 ± 0.01	0.99 ± 0.00	0.98 ± 0.00	0.98 ± 0.01	0.99 ± 0.01
	spring	0.94 ± 0.22	0.99 ± 0.00	0.99 ± 0.00	1.00 ± 0.00	0.99 ± 0.00	0.99 ± 0.01
	summer	0.95 ± 0.18	0.98 ± 0.00	0.98 ± 0.00	0.97 ± 0.00	0.99 ± 0.01	0.99 ± 0.01
Bare F1	winter	0.35 ± 0.31	0.00 ± 0.00	0.32 ± 0.16	0.26 ± 0.16	0.67 ± 0.27	0.56 ± 0.25
	spring	0.81 ± 0.20	0.86 ± 0.03	0.86 ± 0.01	0.90 ± 0.02	0.83 ± 0.09	0.85 ± 0.07
	summer	0.86 ± 0.20	0.30 ± 0.23	0.74 ± 0.03	0.59 ± 0.04	0.72 ± 0.16	0.78 ± 0.11

choice should be guided by the expected shift type. SVM and XGBoost are more robust to signal shifts, while UNet variants are preferable when seasonal variation is the primary concern.

Retraining a model from scratch for every new location or season can be impractical at scale because it requires manually labeled reference data for each deployment context [21]. Targeted strategies worth exploring include enriching training data with bare soil samples from multiple seasons or cities, incorporating bare soil-specific spectral indices [49], or investigating whether multi-temporal imagery in training could improve transfer without full retraining. Domain adaptation methods, which re-align training and deployment distributions, offer a more systematic mitigation path, but evidence of their performance across multiple shift types is lacking [24], [25]; a limitation we explicitly address.

Label-efficient paradigms such as active, semi-supervised, and self-supervised learning may also reduce the annotation costs [21]. Finally, incorporating bare soil F1 as a routine monitoring metric during deployment, rather than relying solely on overall accuracy, provides practitioners with a model-agnostic and computationally inexpensive tool to determine when a trained classifier needs to be updated or replaced.

VI. CONCLUSION

Our findings point to a broader conclusion: a distributional shift degrades urban green space classifiers in ways that are class-specific. Green class classification remains stable. Urban class classification may degrade due to spatial and interannual shifts, and bare soil class classification degrades across all scenarios. Class imbalance modulates spatial transfer in the sites examined while signal shift drives interannual and seasonal degradation, and no single algorithm dominates across all conditions. Bare soil F1 consistently anticipates broader model instability across all shift types and models, offering a simple, model-agnostic deployment diagnostic that overall accuracy alone cannot provide. Robustness evaluation for urban monitoring applications should move beyond single-shift assessments toward unified frameworks that expose the complete vulnerability profile of each algorithm.

REFERENCES

- [1] United Nations, Department of Economic and Social Affairs, Population Division, "World urbanization prospects: The 2025 revision," United Nations, New York, Tech. Rep., 2025. [Online]. Available: <https://population.un.org/wup/>
- [2] M. C. Pastore, C. I. M. Parenti, and C. Patetta, "Measuring accessibility of green spaces for the health and wellbeing of inhabitants of the milan metropolitan area," *Land*, vol. 14, no. 1, p. 97, Jan. 2025. [Online]. Available: <http://dx.doi.org/10.3390/land14010097>
- [3] R. Ramyar, A. Ackerman, and D. M. Johnston, "Adapting cities for climate change through urban green infrastructure planning," *Cities*, vol. 117, p. 103316, Oct. 2021.
- [4] D. García-Álvarez, M. T. Camacho Olmedo, J.-F. Mas, and M. Paegelow, *Land Use Cover Mapping, Modelling and Validation. A Background*. Springer International Publishing, 2022, p. 21–33.
- [5] Y. Sun *et al.*, "Indoor drone localization and tracking based on acoustic inertial measurement," *IEEE Transactions on Mobile Computing*, 2023.
- [6] D. Dobrinčić, M. Miler, and D. Medak, "Mapping the green urban: a comprehensive review of materials and learning methods for green infrastructure mapping," *Sensors*, vol. 25, no. 2, p. 464, 2025.
- [7] P. Griffiths, P. Hostert, O. Gruebner, and S. v. der Linden, "Mapping megacity growth with multi-sensor data," *Remote Sensing of Environment*, vol. 114, no. 2, p. 426–439, Feb. 2010. [Online]. Available: <http://dx.doi.org/10.1016/j.rse.2009.09.012>
- [8] D. Phiri, M. Simwanda, S. Salekin, V. Nyirenda, Y. Murayama, and M. Ranagalage, "Sentinel-2 data for land cover/use mapping: A review," *Remote Sensing*, vol. 12, no. 14, p. 2291, Jul. 2020. [Online]. Available: <http://dx.doi.org/10.3390/rs12142291>
- [9] A. G. P. Silva, L. S. Galvão, L. G. Ferreira Júnior, N. M. Teles, V. V. Mesquita, and I. Haddad, "Discrimination of degraded pastures in the brazilian cerrado using the planetscope superdove satellite constellation," *Remote Sensing*, vol. 16, no. 13, p. 2256, Jun. 2024. [Online]. Available: <http://dx.doi.org/10.3390/rs16132256>
- [10] J. Degerickx, M. Hermy, and B. Somers, "Mapping functional urban green types using high resolution remote sensing data," *Sustainability*, vol. 12, no. 6, p. 2144, 2020.
- [11] E. Honeck, A. Sanguet, M. A. Schlaepfer, N. Wyler, and A. Lehmann, "Methods for identifying green infrastructure," *SN Applied Sciences*, vol. 2, no. 11, Oct. 2020. [Online]. Available: <http://dx.doi.org/10.1007/s42452-020-03575-4>
- [12] S. Talukdar, P. Singha, S. Mahato, Shahfahad, S. Pal, Y.-A. Liou, and A. Rahman, "Land-use land-cover classification by machine learning classifiers for satellite observations—a review," *Remote Sensing*, vol. 12, no. 7, p. 1135, Apr. 2020. [Online]. Available: <http://dx.doi.org/10.3390/rs12071135>
- [13] L. Ma, M. Li, X. Ma, L. Cheng, P. Du, and Y. Liu, "A review of supervised object-based land-cover image classification," *ISPRS Journal of Photogrammetry and Remote Sensing*, 2017. [Online]. Available: <http://dx.doi.org/10.1016/j.isprsjprs.2017.06.001>
- [14] G. M. Foody, "Status of land cover classification accuracy assessment," *Remote Sensing of Environment*, vol. 80, no. 1, pp. 185–201, Apr. 2002. [Online]. Available: [http://dx.doi.org/10.1016/S0034-4257\(01\)00295-4](http://dx.doi.org/10.1016/S0034-4257(01)00295-4)

- [15] F. Österlind *et al.*, “Strawman: resolving collisions in bursty low-power wireless networks,” in *Proceedings of the 11th international conference on Information Processing in Sensor Networks*, 2012.
- [16] L. Mottola *et al.*, “Pervasive games in a mote-enabled virtual world using tuple space middleware,” in *Proceedings of 5th ACM SIGCOMM workshop on Network and system support for games*, 2006.
- [17] A. Subbaswamy, R. Adams, and S. Saria, “Evaluating model robustness and stability to dataset shift,” in *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, ser. Proceedings of Machine Learning Research, vol. 130. PMLR, 2021.
- [18] C. Heinze-Deml and N. Meinshausen, “Conditional variance penalties and domain shift robustness,” *Machine Learning*, vol. 110, no. 2, p. 303–348, Nov. 2020. [Online]. Available: <http://dx.doi.org/10.1007/s10994-020-05924-1>
- [19] R. Taori, A. Dave, V. Shankar, N. Carlini, B. Recht, and L. Schmidt, “Measuring robustness to natural distribution shifts in image classification,” in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 18 583–18 599.
- [20] D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Durando, R. Desai, T. Zhu, S. Parajuli, M. Guo, D. Song, J. Steinhardt, and J. Gilmer, “The many faces of robustness: A critical analysis of out-of-distribution generalization,” in *Proceedings of the IEEE/CVF Intl. Conference on Computer Vision (ICCV)*, October 2021, pp. 8340–8349.
- [21] A. Xiao, W. Xuan, J. Wang, J. Huang, D. Tao, S. Lu, and N. Yokoya, “Foundation models for remote sensing and earth observation: A survey,” *IEEE Geoscience and Remote Sensing Magazine*, vol. 13, no. 4, p. 297–324, Dec. 2025. [Online]. Available: <http://dx.doi.org/10.1109/MGRS.2025.3576766>
- [22] D. Tuia, C. Persello, and L. Bruzzone, “Domain adaptation for the classification of remote sensing data: An overview of recent advances,” *IEEE Geoscience and Remote Sensing Magazine*, vol. 4, no. 2, p. 41–57, Jun. 2016. [Online]. Available: <http://dx.doi.org/10.1109/MGRS.2016.2548504>
- [23] D. M. Kluger, S. Wang, and D. B. Lobell, “Two shifts for crop mapping: Leveraging aggregate crop statistics to improve satellite-based maps in new regions,” *Remote Sensing of Environment*, vol. 262, p. 112488, Sep. 2021. [Online]. Available: <http://dx.doi.org/10.1016/j.rse.2021.112488>
- [24] J. Peng, Y. Huang, W. Sun, N. Chen, Y. Ning, and Q. Du, “Domain adaptation in remote sensing image classification: A survey,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 9842–9859, 2022.
- [25] S. Lyu, Q. Zhao, Z. Zhou, M. Li, Y. Zhou, D. Yao, G. Cheng, H. Zhou, and Z. Shi, “Deep learning based domain adaptation methods in remote sensing: A comprehensive survey,” *arXiv:2510.15615*, 2025.
- [26] S. Mei, J. Lian, X. Wang, Y. Su, M. Ma, and L.-P. Chau, “A comprehensive study on the robustness of deep learning-based image classification and object detection in remote sensing: Surveying and benchmarking,” *Journal of Remote Sensing*, vol. 4, Jan. 2024. [Online]. Available: <http://dx.doi.org/10.34133/remotesensing.0219>
- [27] G. Douzas, F. Bacao, J. Fonseca, and M. Khudinyan, “Imbalanced learning in land cover classification: Improving minority classes’ prediction accuracy using the geometric smote algorithm,” *Remote Sensing*, vol. 11, no. 24, p. 3040, Dec. 2019. [Online]. Available: <http://dx.doi.org/10.3390/rs11243040>
- [28] B. Delaney, K. Tansey, and M. Whelan, “Satellite remote sensing techniques and limitations for identifying bare soil,” *Remote Sensing*, vol. 17, no. 4, p. 630, Feb. 2025. [Online]. Available: <http://dx.doi.org/10.3390/rs17040630>
- [29] C. J. Tucker, “Red and photographic infrared linear combinations for monitoring vegetation,” *Remote Sensing of Environment*, vol. 8, no. 2, pp. 127–150, May 1979. [Online]. Available: [http://dx.doi.org/10.1016/0034-4257\(79\)90013-0](http://dx.doi.org/10.1016/0034-4257(79)90013-0)
- [30] R. Neyns, K. Efthymiadis, P. Libin, and F. Canters, “Fusion of multi-temporal planetscope data and very high-resolution aerial imagery for urban tree species mapping,” *Urban Forestry and Urban Greening*, vol. 99, p. 128410, Sep. 2024. [Online]. Available: <http://dx.doi.org/10.1016/j.ufug.2024.128410>
- [31] M. Gašparović, D. Dobričić, and I. Pilaš, “Mapping of allergenic tree species in highly urbanized area using planetscope imagery—a case study of zagreb, croatia,” *Forests*, vol. 14, no. 6, p. 1193, Jun. 2023. [Online]. Available: <http://dx.doi.org/10.3390/f14061193>
- [32] C. A. Boano *et al.*, “Exploiting the lqi variance for rapid channel quality assessment,” in *International Conference on Information Processing in Sensor Networks*, 2009.
- [33] G. Mallinis and N. Koutsias, “Spectral and spatial-based classification for broad-scale land cover mapping based on logistic regression,” *Sensors*, vol. 8, no. 12, pp. 8067–8085, Dec. 2008. [Online]. Available: <http://dx.doi.org/10.3390/s8128067>
- [34] M. Pal and P. M. Mather, “Support vector machines for classification in remote sensing,” *International Journal of Remote Sensing*, vol. 26, no. 5, pp. 1007–1011, Mar. 2005. [Online]. Available: <http://dx.doi.org/10.1080/01431160512331314083>
- [35] G. Mountrakis, J. Im, and C. Ogole, “Support vector machines in remote sensing: A review,” *ISPRS journal of photogrammetry and remote sensing*, vol. 66, no. 3, pp. 247–259, 2011.
- [36] M. Belgiu and L. Drăguț, “Random forest in remote sensing: A review of applications and future directions,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 114, p. 24–31, Apr. 2016. [Online]. Available: <http://dx.doi.org/10.1016/j.isprsjprs.2016.01.011>
- [37] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, “Convolutional neural networks for large-scale remote-sensing image classification,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 2, pp. 645–657, 2017.
- [38] B. Neupane, T. Horanont, and J. Aryal, “Deep learning-based semantic segmentation of urban features in satellite images: A review and meta-analysis,” *Remote Sensing*, vol. 13, no. 4, p. 808, Feb. 2021. [Online]. Available: <http://dx.doi.org/10.3390/rs13040808>
- [39] R. R. Vatsavai, E. Bright, C. Varun, B. Budhendra, A. Cheriyaad, and J. Grasser, “Machine learning approaches for high-resolution urban land cover classification: a comparative study,” in *Proceedings of the 2nd International Conference on Computing for Geospatial Research & Applications (COM.Geo '11)*. ACM Press, 2011, pp. 1–10. [Online]. Available: <http://dx.doi.org/10.1145/1999320.1999331>
- [40] V. Chaturvedi and W. T. de Vries, “Machine learning algorithms for urban land use planning: A review,” *Urban Science*, 2021. [Online]. Available: <http://dx.doi.org/10.3390/urbansci5030068>
- [41] C. Y. S. Wong, “Plant optics: underlying mechanisms in remotely sensed signals for phenotyping applications,” *AoB PLANTS*, vol. 15, no. 4, Jul. 2023. [Online]. Available: <http://dx.doi.org/10.1093/aobpla/plad039>
- [42] R. De Jong, J. Verbesselt, A. Zeileis, and M. Schaepman, “Shifts in global vegetation activity trends,” *Remote Sensing*, vol. 5, 2013. [Online]. Available: <http://dx.doi.org/10.3390/rs5031117>
- [43] N. Kumari and S. Minz, “Focal-hinge loss based deep hybrid framework for imbalanced remote sensing scene classification,” *INFOCOMP Journal of Computer Science*, vol. 23, no. 1, Jul. 2024. [Online]. Available: <http://177.105.60.18/index.php/infocomp/article/view/3320>
- [44] M. K. Ridd, “Exploring a vis (vegetation-impervious surface-soil) model for urban ecosystem analysis through remote sensing: comparative anatomy for cities,” *International journal of remote sensing*, vol. 16, no. 12, pp. 2165–2185, 1995.
- [45] S. K. McFeeters, “The use of the normalized difference water index (ndwi) in the delineation of open water features,” *International Journal of Remote Sensing*, vol. 17, no. 7, pp. 1425–1432, May 1996. [Online]. Available: <http://dx.doi.org/10.1080/01431169608948714>
- [46] Y. Zeng, D. Hao, A. Huete, B. Dechant, J. Berry, J. M. Chen, J. Joiner, C. Frankenberg, B. Bond-Lamberty, Y. Ryu, J. Xiao, G. R. Asrar, and M. Chen, “Optical vegetation indices for monitoring terrestrial ecosystems globally,” *Nature Reviews Earth & Environment*, vol. 3, no. 7, pp. 477–493, May 2022. [Online]. Available: <http://dx.doi.org/10.1038/s43017-022-00298-5>
- [47] W. J. Frampton, J. Dash, G. Watmough, and E. J. Milton, “Evaluating the capabilities of sentinel-2 for quantitative estimation of biophysical variables in vegetation,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 82, pp. 83–92, 2013.
- [48] Y. Deng, G. Chen, B. Tang, X. Duan, L. Zuo, and H. Zhao, “Study on class imbalance in land use classification for soil erosion in dry-hot valley regions,” *Remote Sensing*, vol. 17, no. 9, p. 1628, May 2025. [Online]. Available: <http://dx.doi.org/10.3390/rs17091628>
- [49] C. T. Nguyen, A. Chidthaisong, P. Kieu Diem, and L.-Z. Huo, “A modified bare soil index to identify bare land features during agricultural fallow-period in southeast asia using landsat 8,” *Land*, vol. 10, no. 3, p. 231, Feb. 2021. [Online]. Available: <http://dx.doi.org/10.3390/land10030231>