Scientific Reports

https://doi.org/10.1038/s41598-026-37043-7

Article in Press

Unsupervised mapping of urban tree diversity using spatially-aware visual clustering

Received: 18 November 2025

Accepted: 19 January 2026

Published online: 25 February 2026

Cite this article as: Abuhani D.A., Seccaroni M., Mazzarello M. *et al.* Unsupervised mapping of urban tree diversity using spatially-aware visual clustering. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-37043-7

Diaa Addeen Abuhani, Marco Seccaroni, Martina Mazzarello, Imran Zualkernan, Fabio Duarte & Carlo Ratti

Unsupervised Mapping of Urban Tree Diversity using Spatially-aware Visual Clustering

Diaa Addeen Abuhani[1,2*], Marco Seccaroni[1,3], Martina Mazzarello[1*], Imran Zualkernan[2], Fabio Duarte[1], Carlo Ratti[1,3]

[1*]Department of Urban Studies and Planning, Senseable City Laboratory, Massachusetts Institute of Technology, 77 Massachusetts Ave, Cambridge, MA 02139, Massachusetts, USA.
[2]Department of Computer Science and Engineering, American University of Sharjah, 127 Street, University City, 26666, Sharjah, UAE.
[3]Department of Architecture, Built Environment, and Construction Engineering, Politecnico di Milano, Piazza Leonardo da Vinci, 32, Milano, 20133, Milan, Italy.

*Corresponding author(s). E-mail(s): diaa@mit.edu; mmazz@mit.edu;
Contributing authors: marco.seccaroni@polimi.it; izualkernan@aus.edu; fduarte@mit.edu; ratti@mit.edu;

Abstract

Urban tree biodiversity is critical for climate resilience, ecological stability, and livability in cities, yet most municipalities lack detailed knowledge of their canopies. Field-based inventories provide reliable estimates of Shannon and Simpson diversity but are costly and time-consuming, while supervised AI methods require labeled data that often fail to generalize across regions. We introduce an unsupervised clustering framework that integrates visual embeddings from street-level imagery with spatial planting patterns to estimate biodiversity without labels. Applied to eight North American cities, the method recovers genus-level diversity patterns with high fidelity, achieving low Wasserstein distances to ground truth for Shannon and Simpson indices and preserving spatial autocorrelation. This scalable, fine-grained approach enables biodiversity mapping in cities lacking detailed inventories and offers a pathway for continuous, low-cost monitoring to support equitable access to greenery and adaptive management of urban ecosystems.

Keywords: Biodiversity, Unsupervised Learning, Clustering, Streetviews

1 Introduction

Quantifying biodiversity in urban forests is essential to assess ecosystem health and to guide evidence-based management and policy decisions for urban green spaces. Widely used metrics like Shannon entropy and Simpson's index integrate both richness (number of taxa) and evenness (distribution of abundances) into a single value, providing interpretable indicators for comparing ecological conditions across locations and over time [1, 2]. However, accurate estimation of these indices at the genus or species level typically requires detailed and geo-referenced taxonomic inventories, most often collected through arborist-led field surveys [3]. Although these surveys are reliable, they are resource intensive. Urban street-tree programs may spend up to $65 per tree annually on inventory, monitoring, and management, which has limited large-scale implementation to only a few well-funded cities [4, 5].

In the past 50 years, adapting to climate change has become a priority for cities around the world. Seminal and contemporary studies [6–9] highlight that anthropogenic changes in the past few decades have caused a rapid loss of urban biodiversity. In addition to supporting biodiversity and urban ecosystems [10], trees play a critical role in maintaining livability by contributing to equitable access to greenery [11–13], improved air quality [14, 15], and thermal comfort [16, 17]. Yet, tree planting strategies are often guided by aesthetic preferences rather than ecological function [18], resulting in reduced diversity and increased vulnerability to pests, diseases, and climate stressors [13]. A more diverse and ecologically informed urban canopy can buffer against these risks and adapt over time [19–21].

Open access data sources, such as Street View Imagery (SVI), offer a scalable alternative for biodiversity monitoring due to their extensive spatial coverage [22]. However, SVI datasets are unlabeled, meaning the images do not specify the tree species present and require substantial pre-processing to extract ecological information suitable for diversity estimations. Identifying individual taxa from imagery is particularly challenging: trees of the same species may look very different depending on season, age, or management, while distinct species can appear deceptively similar. These difficulties make automated classification highly sensitive to subtle morphological cues. Current supervised ecological AI models [23, 24], while effective within their training domains, often under-perform when applied to regions or taxa not well represented in training data, largely due to geographic domain shift [25].

Self-supervised learning offers a promising alternative by eliminating the need for labeled data [26]. However, when applied to limited or homogeneous datasets, these methods often under-perform due to overfitting and poor generalization, as robust representations typically require large and diverse training samples. Few clustering approaches have been explicitly designed to preserve abundance distributions in a way that enables accurate Shannon and Simpson estimation from unlabeled imagery. This

capability is essential for producing ecologically meaningful biodiversity metrics from large-scale, unlabeled datasets.

In this work, we address the challenge of large-scale, label-free biodiversity monitoring by developing an unsupervised clustering approach that estimates genus-level urban tree diversity directly from street-level imagery. We apply the framework across eight North American cities with contrasting forest compositions and imbalance, and evaluate its ability to recover biodiversity metrics such as Shannon and Simpson indices as well as spatial distribution patterns. While our evaluation is necessarily constrained to cities with available ground-truth inventories, the results demonstrate that the approach generalizes across diverse urban forest contexts and provides a scalable pathway for integrating biodiversity insights into urban resilience planning.

2 Results

We evaluated our framework in eight North American cities spanning a wide range of forest compositions and genera imbalances (see Section 4.1). All biodiversity metrics are computed at the level of spatial grid cells, obtained by partitioning each city into regular 500-m grids, with each cell representing a local urban tree community. Accordingly, richness, Shannon, and Simpson are all treated as local (α-diversity) measures, and $RMSE_\alpha$ quantifies error in cell-level richness rather than city-wide (γ) diversity. The method consistently recovered meaningful biodiversity patterns, demonstrating robustness even in cities dominated by a few genera. Table 1 reports results for the best-performing configuration, selected through hyperparameter tuning (see Appendix A) and ablation study (see Appendix B), which achieved a balanced trade-off between richness accuracy, clustering quality, and diversity preservation.

Table 1: Clustering performance for our clustering framework across cities. Metrics include richness error ($RMSE_\alpha \downarrow$), clustering quality (V-score $\uparrow$), and distributional agreement with ground-truth Shannon (scaled) and Simpson diversity measured using the 1-Wasserstein distance ($W_1 \downarrow$). Values are reported as mean [95% CI] through bootstrapping [27].

City	$RMSE_\alpha \downarrow$	V-score $\uparrow$	$W_{1,\text{Shannon}} \downarrow$	$W_{1,\text{Simpson}} \downarrow$
Calgary	4.39 [4.20–4.59]	0.397 [0.387–0.406]	0.215 [0.208–0.222]	0.178 [0.169–0.186]
New York	3.42 [3.34–3.51]	0.476 [0.471–0.480]	0.073 [0.070–0.077]	0.064 [0.059–0.068]
Columbus	5.99 [5.76–6.22]	0.430 [0.422–0.437]	0.177 [0.172–0.183]	0.140 [0.134–0.147]
Denver	5.39 [5.15–5.62]	0.487 [0.480–0.494]	0.069 [0.065–0.073]	0.040 [0.036–0.045]
Los Angeles	3.74 [3.63–3.84]	0.477 [0.473–0.481]	0.070 [0.067–0.073]	0.067 [0.064–0.071]
Seattle	9.94 [9.56–10.3]	0.487 [0.481–0.493]	0.155 [0.150–0.160]	0.095 [0.089–0.101]
Washington	15.6 [14.7–16.3]	0.498 [0.491–0.505]	0.133 [0.127–0.140]	0.033 [0.029–0.039]
San Francisco	10.0 [9.17–10.8]	0.493 [0.485–0.501]	0.034 [0.028–0.041]	0.028 [0.023–0.034]

Richness and cluster quality results.

As shown in Table 1, genera imbalance strongly influenced richness estimation error ($RMSE_\alpha$), though not uniformly across cities. In highly uneven forests such as Washington and Seattle, $RMSE_\alpha$ exceeded within-city variability, underscoring the difficulty of estimating richness under strong dominance patterns. By contrast, in more balanced systems such as Los Angeles, Denver, and New York, errors were smaller than or comparable to observed standard deviations, and in several cases outperformed a naïve mean-predictor baseline (see Appendix E). Despite these differences, the V-score remained consistent across all datasets (0.397–0.498), indicating that clustering preserved the overall partitioning structure even when richness estimates deviated. These findings suggest that richness is not the primary strength of our approach, but performance is competitive in balanced forests and improves over statistical baselines. Richness estimation is inherently more sensitive to long-tail genus distributions, as rare taxa contribute little to the dominant visual and spatial signal regardless of learning paradigm. This limitation is therefore not specific to unsupervised clustering, and could be mitigated in future work through targeted sampling [28], weak priors on rarity [29], or emerging long-tail–aware generative augmentation strategies [30], which remain challenging in ecological settings.

Diversity results.

For Shannon and Simpson diversity, we assessed agreement with ground truth using the Wasserstein distance [31]. This metric quantifies the minimum "cost" of transforming one distribution into another, accounting not only for absolute differences but also for their positions across the value range. Unlike RMSE, which captures pointwise error, the Wasserstein distance is sensitive to systematic shifts in distributional shape and spread, making it particularly suitable for ecological indices. We focus on Shannon and Simpson entropies due to their long-standing theoretical foundation, widespread adoption, and interpretability in ecological research [1]. These classical metrics capture complementary aspects of community structure such as balanced weighting of common and uncommon taxa (Shannon) and dominance-driven structure (Simpson)[32].

In our datasets, Simpson diversity is bounded between 0 and 1, while Shannon diversity can reach values of about 3.5 (see Section 4.1). To enable direct comparison across indices and cities, Shannon values were normalized by their relative maximum, placing both metrics on a common [0,1] scale. In this context, Wasserstein distances below 0.05 indicate very strong alignment with ground truth distributions, while values near 0.1 represent moderate but still ecologically meaningful agreement [33]. Cities such as Washington (0.133 for Shannon, 0.033 for Simpson) and San Francisco (0.034 for Shannon, 0.028 for Simpson) demonstrated strong alignment for both indices, with proportionally smaller errors for Simpson. Calgary, by contrast, exhibited the weakest alignment across metrics. We hypothesize that this may be linked to more aggressive visual obfuscation apparent in Canadian Street View imagery compared to the United States, where stronger privacy safeguards and stricter blurring protocols are enforced [34]. Fig. 1 illustrates this by comparing predicted and observed Shannon values in Los Angeles (Fig. 1a) and Simpson values in New York (Fig. 1b). A full set of city-level results, including baseline comparisons and variability relative to observed standard

deviations, is provided in Appendix E, which contextualizes how performance varies across urban forest structures.

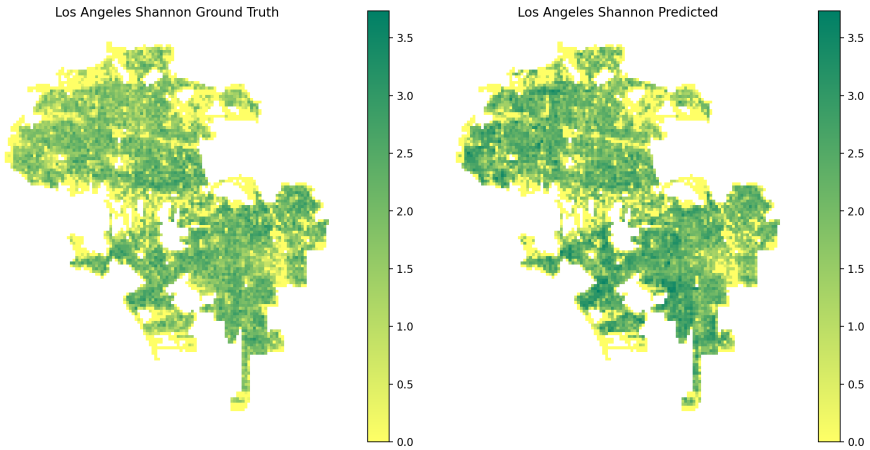
Diagnosis of possible entropy inflation.

One potential concern with unsupervised clustering is that lower distributional errors may arise not from genuine recovery of community structure but from artificial entropy inflation, e.g., through oversplitting or the proliferation of singleton clusters. To assess this, we performed a dedicated diagnostic analysis (Appendix D) that examined oversplitting factors, purity, evenness shifts, and rank–abundance correlations between predicted and observed communities against a supervised model using linear probing on BioCLIP [23] model. Across all cities, the singleton ratio was zero, oversplitting factors remained close to one, and evenness shifts were modest, with New York even showing a slight decrease. At the same time, rank–abundance correlations were consistently high ($\rho \approx 0.9$), indicating that predicted clusters preserved the overall shape of community distributions. These findings confirm that the improved alignment of Shannon and Simpson diversity under the unsupervised approach is not an artifact of inflated entropy, but rather reflects ecologically meaningful recovery of biodiversity structure.

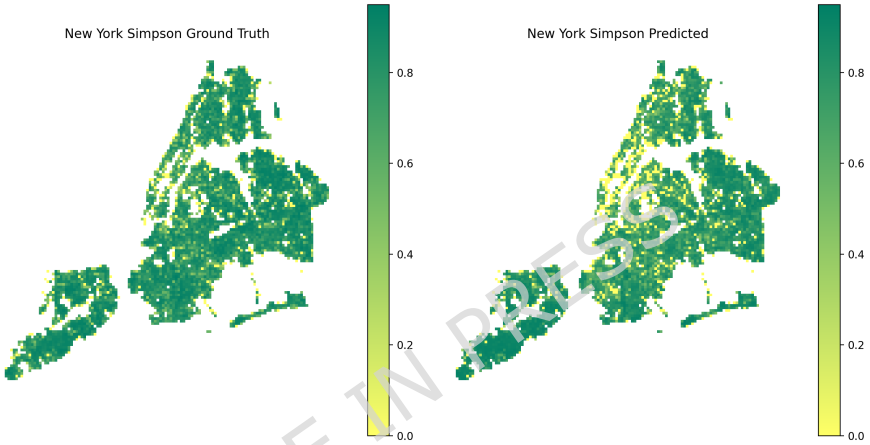
Overall, the method achieved low Wasserstein distances for both Shannon and Simpson diversity across diverse urban forest conditions, including in highly imbalanced contexts. High Simpson accuracy suggests the framework effectively captured dominant genera distributions, while strong Shannon performance indicates robustness to abundance evenness patterns. Taken together, these results demonstrate that our approach can accurately recover both abundance distributions and spatial diversity patterns with strong spatial fidelity (see Appendix C) from entirely unlabeled street-level imagery, directly addressing the lack of scalable, label-free biodiversity mapping methods highlighted in the Introduction.

3 Discussion

Our results demonstrate that an unsupervised and spatially aware clustering framework can recover both magnitude and spatially weighted biodiversity measures, where diversity is computed locally and evaluated across spatial grid cells, from unlabeled street-level imagery with high fidelity, even under markedly imbalanced genus distributions. Across diverse urban contexts, the method achieved low Wasserstein distances for Shannon and Simpson diversity, while preserving spatial autocorrelation patterns with minimal deviation from ground truth. These findings are consistent with previous reports that dominant taxa patterns are easier to recover than local evenness structure driven by the relative abundances of non-dominant genera [1, 22], but extend prior work by showing that evenness-sensitive indices such as Shannon can also be estimated reliably when spatial priors are incorporated. Crucially, by enabling robust, label-free biodiversity mapping, this approach offers a scalable tool that could be integrated into municipal monitoring systems, guide nature-based policy, and support adaptive planning in cities with limited ecological data infrastructure.



(a) Los Angeles's Shannon Entropy Ground Truth and Predicted Results



(b) New York's Simpson Entropy Ground Truth and Predicted Results

Fig. 1: A comparison between ground truth and predicted values of Shannon entropy in Los Angeles and Simpson entropy in New York.

Our ablation analysis (Appendix B) showed that the joint use of outlier elimination and cluster merging provided the best balance between clustering quality, richness accuracy, and preservation of diversity structure. This result indicates that addressing noise (through elimination) and fragmentation (through merging) in tandem is particularly effective for recovering meaningful ecological patterns from unlabeled imagery. More broadly, it suggests that unsupervised clustering in environmental applications benefits from strategies that explicitly mitigate both spurious assignments and oversegmentation. While prior studies have applied label-free clustering to infer urban

landforms from very high-resolution satellite data [35] or crop types from multispectral time series [36], such approaches typically rely on costly or proprietary datasets. By contrast, our framework demonstrates that freely available street-level imagery can support scalable, label-free biodiversity assessment in urban forests, extending applicability to regions where high-resolution remote sensing is not accessible.

From an applied perspective, the framework offers a scalable alternative to field-based inventories for monitoring urban tree diversity and can be integrated into urban forestry programs. Spatially explicit Shannon and Simpson maps at the city scale can guide planting strategies, highlight areas at risk of monocultures, and support temporal monitoring, while reducing the effort and cost of repeated ground surveys. Importantly, this directly addresses the central gap identified in the Introduction in regards to the inability of existing methods to recover abundance distributions and spatial diversity patterns from unlabeled imagery. By overcoming this limitation, our approach provides a pathway for continuous, cost-effective biodiversity monitoring that can inform urban resilience planning. Here, "city-scale" refers to the aggregation and analysis of spatially explicit grid-level diversity patterns, rather than to the direct estimation of total city-wide (γ) richness. City-level insights are obtained by summarizing the distribution and spatial organization of grid-level Shannon and Simpson diversity values, enabling identification of dominance hotspots, monoculture risk, and spatial inequities without explicitly inferring total genus counts.

Several constraints remain. Street-level imagery is updated at irregular intervals, sometimes years apart, which complicates the interpretation of temporal change. In practice, this means the framework is better suited to structural baselines and cross-sectional monitoring than to precise year-to-year trend detection. Coverage is uneven globally, with gaps in rural areas, private developments, and in certain countries; thus, results are most applicable in dense urban cores of cities with established GSV infrastructure, and less transferable to places where coverage is patchy. Privacy-driven blurring can obscure morphological features needed for genus discrimination, potentially biasing genus detection toward larger or more distinctive trees, and under-representing genera with finer-grained leaf or bark traits. Seasonal foliage changes, variable lighting, and transient occlusions (e.g. vehicles or pedestrians) introduce visual noise, which can reduce consistency across time and space and necessitate larger sample sizes to stabilize estimates. In such cases, phenological variation may cause visually distinct seasonal appearances of the same genus to be assigned to separate pseudo-taxa, potentially affecting cluster purity or inflating effective richness. Additionally, the method only captures the visible streetscape; off-street vegetation is excluded, which can underestimate diversity in areas where backyards, courtyards, or parks hold a significant share of urban trees.

Looking ahead, performance could be enhanced by incorporating vision–language models (VLMs) that leverage fine-grained textual descriptors of morphological traits [37], improving discrimination between visually similar trees. As model efficiency advances, large-scale VLM integration may become feasible. Further gains may come from foundational models trained on ecologically diverse, high-resolution vegetation datasets, coupled with domain-adaptive fine tuning to improve generalization to under-represented regions. In parallel, prior work shows that satellite RGB imagery alone

is insufficient for reliable urban tree assessment [38, 39]; integrating complementary modalities such as airborne LiDAR for structural detail and high-resolution multispectral data or recent foundation models [40] for spectral cues would enrich representation and enable a more holistic view of urban forest diversity. Beyond technical gains, these enhancements would directly support cross-city generalization, allowing models to adapt to ecological and urban form variability across different contexts. Moreover, the fusion of tree diversity indicators with other environmental datasets such as air quality, temperature, or land use would situate this work within the broader trend of multimodal urban environmental monitoring, creating synergies for applications in climate resilience, ecosystem services evaluation, and nature-based urban planning.

In sum, this work establishes that unsupervised clustering of spatial and visual embeddings can be a practical, transferable, and data-efficient approach to quantifying urban tree diversity. By recovering both the magnitude and spatial organization of Shannon and Simpson diversity indices, the framework offers a pathway for continuous, city-scale biodiversity monitoring and supports the development of more resilient, ecologically informed urban forests. By enabling the recovery of both abundance distributions and spatial diversity patterns from unlabeled imagery, our approach opens the door to scalable, globally comparable biodiversity mapping for urban environments.

4 Methods

We approach urban tree biodiversity with an unsupervised clustering framework driven by spatial priors, using neighborhood structure and planting patterns as the primary signal. Rather than relying on datasets that label each tree by species or genus and then training a supervised model on those labels, our method operates without labels and instead groups trees into taxonomically coherent clusters using visual embeddings extracted from street-level imagery. Each cluster functions as a pseudo-taxon, allowing the computation of Shannon, Simpson, and related indices without the need for explicit species identification. The end-to-end pipeline consists of five stages: data description, spatial and visual embedding extraction, spatial–visual clustering, cluster refinement, and biodiversity metric computation. An overview of the methodology is shown in Fig. 2 and detailed later in this section.

4.1 Data Description

This study relies on the Auto Arborist dataset [38], a large-scale benchmark of urban street trees derived from dense sampling of Google Street View (GSV) imagery. While the dataset spans numerous cities, we selected eight cities that span a wide range of urban form, genera composition, and spatial distribution, providing a robust testbed for cross-city evaluation of the framework. Across these cities, the dataset contains hundreds of thousands of geotagged tree instances annotated at the genus level.

To move from city-level coverage to local-scale analysis, each city is discretized into regular 500-m grid cells, which serve as the fundamental spatial units of analysis. All biodiversity metrics are computed at the grid-cell level, with each cell representing a

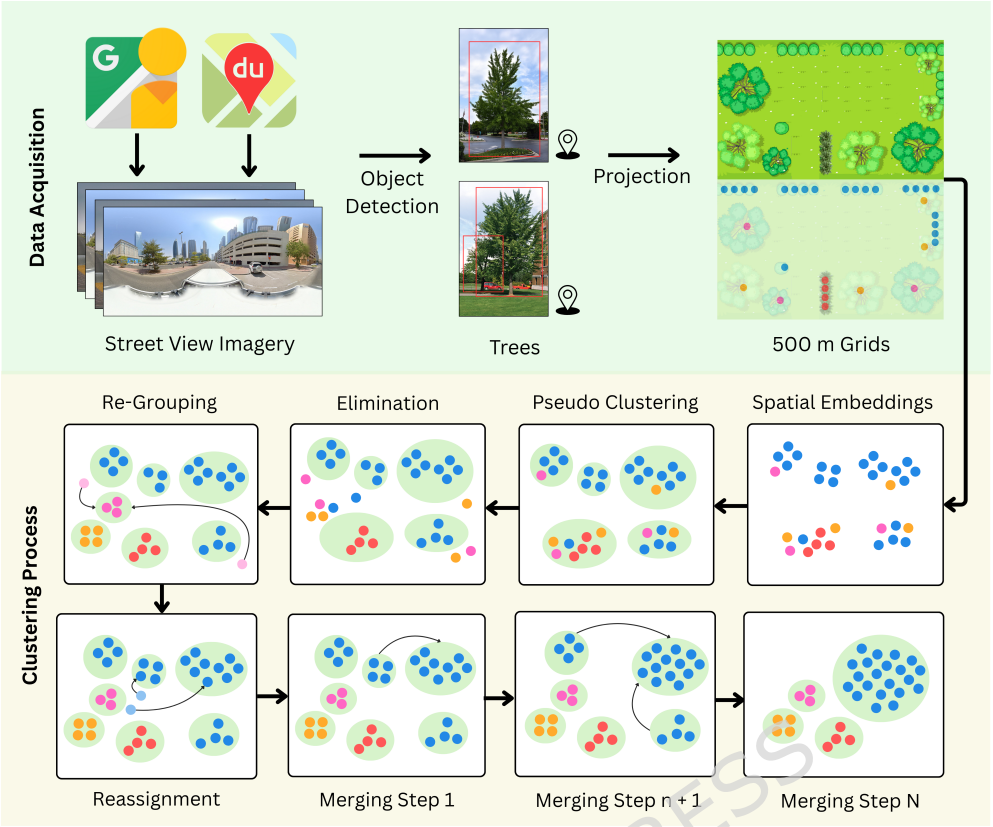


Fig. 2: Overview of the methodology. Street-level imagery (see Section 4.1) is processed using an object-detection model from prior work to extract tree instances, which are geo-located and mapped into 500 m grid cells (top row). The bottom rows illustrate the iterative clustering process applied to spatial embeddings: from left to right, the procedure consists of pseudo-clustering, elimination, and re-grouping (top sequence), followed by reassignment and successive merging steps (bottom sequence), culminating in stable genus-level clusters. Together, these stages yield coherent genus-level tree distributions across space.

local urban tree community. The dataset is no longer publicly hosted due to maintenance constraints, but remains available from the authors upon request. Full details of the acquisition and preprocessing pipeline are provided in [38].

All analyses are conducted at the genus level, which represents a well-established and pragmatic taxonomic resolution in large-scale urban tree studies. At this level, taxonomic groupings remain ecologically meaningful while avoiding the substantial ambiguity and visual overlap that often limit reliable species-level discrimination from street-level imagery. Many functional traits, environmental tolerances, and management considerations are conserved within genera, making genus-level aggregation

suitable for biodiversity assessment and urban forestry applications [41–43]. This resolution also provides a stable basis for comparative analysis across cities while remaining compatible with the structure of widely used urban tree inventories. To ensure reproducibility and enable further research, we will release the spatial and visual embeddings of some cities generated in this study, providing a standardized, high-quality resource for consistent and cross-city evaluation.

Table 2: Summary of urban forest characteristics for the eight North American cities in the Auto Arborist dataset. Reported metrics include urban area, total tree counts, number of genera, genera imbalance ($\mu_{IR} \pm \sigma_{IR}$), and the maximum grid-cell–level Shannon diversity ($H'_{\max}$), shown as a reference for local α-diversity rather than city-wide entropy.

City	Urban Area (km^2)	#Trees	#Genera	$\mu_{\text{IR}} \pm \sigma_{\text{IR}}$	$H'_{\max}$
Calgary, AB	621.7	64,476	35	11.28 ± 18.02	2.47
New York, NY	1,223.6	560,069	68	11.33 ± 14.67	3.12
Columbus, OH	1,336.5	114,536	81	17.96 ± 24.95	3.18
Denver, CO	1,760.0	175,438	97	18.40 ± 20.25	3.27
Los Angeles, CA	5,910.0	391,788	202	20.98 ± 27.90	3.50
Seattle, WA	2,546.0	150,983	142	34.06 ± 37.07	3.44
Washington, DC	3,351.0	152,983	71	43.13 ± 43.38	2.89
San Francisco, CA	1,330.0	154,698	195	57.96 ± 77.78	3.48
Average	2,384.5	220,746	111	26.43 ± 32.00	3.17

The dataset used encompasses eight major North American cities (Table 2), capturing substantial variation in urban extent, tree abundance, taxonomic richness, and genera imbalance. Urban area spans nearly an order of magnitude, from 621.7 km^2 in Calgary to 5,910 km^2 in Los Angeles. Total tree counts range from roughly 64,000 in Calgary to over 560,000 in New York, reflecting differences in both city size and canopy coverage. Taxonomic richness, measured as the number of genera, peaks in Los Angeles (202) and San Francisco (195), indicating broad genera variety, and is lowest in Calgary (35), suggesting a more homogeneous composition.

The genera imbalance factor (IR), defined as the mean ± standard deviation of the maximum-to-minimum (MM) genus count ratio within city grids, highlights marked contrasts in dominance patterns. New York exhibits the most even composition (11.33 ± 14.67), whereas San Francisco (57.96 ± 77.78), Washington (43.13 ± 43.38), and Seattle (34.06 ± 37.07) display strong skew toward a few dominant genera, because richness is driven by the accurate detection of rare genera, which often appear as low-abundance or singleton occurrences and are therefore more susceptible to misassignment or merging errors in clustering-based approaches. Such high imbalance factors present greater challenges for biodiversity estimation, particularly for richness metrics. Importantly, IR should be interpreted as a measure of dominance rather than a direct indicator of overall biodiversity: cities with high IR can still host many

genera, but those genera are unevenly distributed, with a few disproportionately represented. In our analysis, we found that while extreme IR values often coincided with higher richness errors, they did not consistently predict lower Shannon or Simpson diversity, since these indices emphasize evenness and relative abundance rather than absolute richness. In particular, Simpson diversity and, to a lesser extent, Shannon entropy largely reflect dominant and common genera, whereas rare genera contribute disproportionately to richness but minimally to these entropy-based measures [32].

Shannon diversity, which in this dataset can reach values of approximately 3.5, further distinguishes the cities. Larger and compositionally diverse cities such as Los Angeles, San Francisco, and Seattle attain the highest maxima, reflecting both high richness and evenness in certain localities. By contrast, cities with fewer genera or pronounced dominance, such as Calgary and Washington, tend toward lower Shannon maxima, consistent with reduced local evenness.

Collectively, these differences position the dataset used as a diverse and challenging benchmark to test the generalizability of unsupervised clustering methods in urban biodiversity assessment. A detailed pseudo-code of the pipeline is provided in Appendix B (Algorithm **??**), and the complete implementation, including hyperparameter tuning and preprocessing scripts, is available on Github.

4.2 Clustering Framework

Clustering urban tree data from street-level imagery requires methods that can accommodate both fine-grained visual differences between genera and spatial patterns imposed by urban design. Our framework combines visual and spatial embeddings in a multi-stage pipeline that progressively refines initial groupings, removing visual outliers, merging highly similar clusters, and reassigning ambiguous samples. This design balances taxonomic accuracy with spatial representativeness, ensuring that diversity metrics capture both the magnitude and spatial structure of urban forest biodiversity.

4.2.1 Pseudo Clustering

Urban trees are often arranged linearly along streets and boulevards to provide shade, enhance aesthetic appeal, and define pedestrian zones [44]. However, many clustering algorithms perform poorly on linear or anisotropic spatial patterns, as they often assume compact clusters with uniform densities. Clustering methods like Ordering Point to Identify the Clustering Structure (OPTICS) [45] or Fuzzy c-Means (FCM) [46] rely on parameters or assumptions that are not well suited for elongated or non-convex cluster geometries.

Spatial embeddings derived from TaxaBind [24], a multimodal foundation model that incorporates geographic location and environmental features globally, provide a powerful solution to the limitations of traditional clustering algorithms when applied to linearly arranged urban tree data. Unlike centroid- or density-based methods, which often fail to capture elongated or irregular structures [47], spatial embeddings transforms raw geographic coordinates into a learned representation that encodes relative spatial context and structural relationships. This allows linearly aligned trees to be represented as coherent patterns, even in the presence of spacing variability, gaps, or

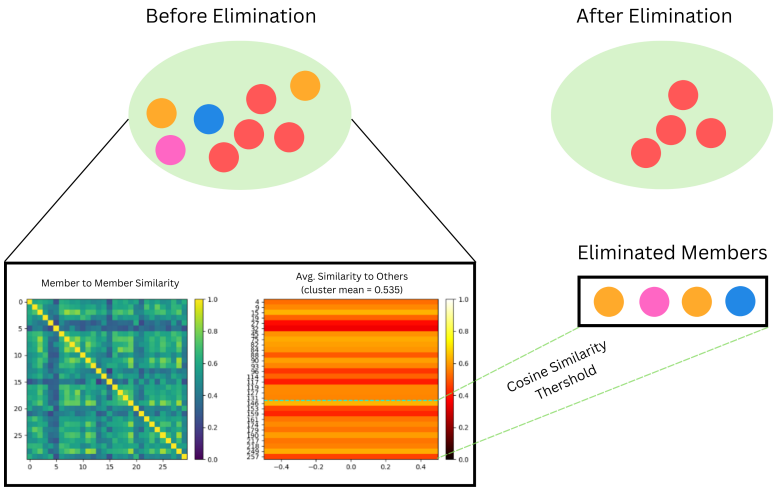


Fig. 3: Intra-cluster similarity filtering using a tuned threshold. Low-similarity members (highlighted) are excluded to increase cluster coherence.

missing data. By preserving and enhancing neighborhood relationships, spatial embeddings help group trees that share local spatial context, making clustering more robust to urban morphology.

For visual features, we use BioCLIP [23] to extract fine-grained, taxonomically relevant representations from street-level imagery. These visual embeddings complement the spatial representation by capturing leaf, bark, and crown characteristics that are essential for distinguishing between visually similar genera.

As an initial step, we group TaxaBind spatial embeddings into pseudo-clusters using the HDBSCAN algorithm [48], leveraging the tendency of urban trees to be planted in structured, often linear arrangements. This approach typically results in taxonomically pure clusters due to the spatial regularity introduced by common urban planting strategies such as row planting along streets or uniform spacing in public parks.

4.2.2 Outlier Elimination

Outlier elimination based on intra-cluster visual similarities aims to improve the coherence and reliability of clustering by removing samples that visually deviate from the dominant patterns within their assigned group. After forming initial clusters based on spatial proximity, we compute pairwise cosine similarities [49] between all visual embeddings within each cluster. Cosine similarity is particularly suitable here because it measures the angular closeness of vectors rather than their raw magnitudes, which is advantageous in high-dimensional embedding spaces where differences in vector length are less informative than differences in orientation. This makes it effective for capturing whether samples share a common visual representation, even when the embedding space is large and sparse. Samples with consistently low similarity to other members are considered visually inconsistent and are eliminated as outliers.

Because visual differences among tree genera can be subtle and fine-grained, especially in urban settings with morphologically similar genera, directly detecting outliers is challenging. To address this, we adopt a conservative strategy in which outliers are identified as those samples whose average similarity to other cluster members falls below a threshold determined through hyperparameter tuning (see Appendix A). This thresholded filtering prioritizes consistency while avoiding premature exclusion of valid but visually nuanced samples, resulting in more stable and interpretable clusters in subsequent steps. This process is further illustrated in Fig. 3.

4.2.3 Re-grouping and Re-assignment

The re-grouping and re-assignment process targets samples eliminated during the outlier detection stage, which were initially clustered based primarily on spatial proximity rather than visual coherence. Since such pseudo-clusters may group together visually dissimilar trees due to their physical arrangement, the eliminated members are revisited to assess whether they form more semantically consistent groupings.

Visual embeddings of all eliminated samples are compared, and if a subset exhibits mutual similarity above the tuned cosine similarity threshold (Appendix A), eliminated samples are grouped into a new cluster, capturing visual structure that spatial proximity alone may have missed. Remaining samples are evaluated for reassignment to the most visually similar existing cluster, again using the tuned threshold. If no such reassignment is possible, the sample is treated as a singleton cluster.

This step enhances visual coherence in the final clustering while improving the model's ability to capture the long-tail distribution of urban tree genera where many rare or underrepresented classes naturally appear as singletons.

4.2.4 Merging Process

Within each post-processing iteration, cluster merging serves as an adaptive consolidation step, reducing over-fragmentation as the clustering structure evolves. After outlier elimination, some clusters become more visually coherent, and new clusters may emerge from outlier grouping. These updates can reveal opportunities to combine clusters that now exhibit strong visual similarity.

As shown in Fig. 4, we compute the average cosine similarity between the embedding centroids of all cluster pairs. If the similarity between two clusters exceeds a threshold determined through hyperparameter tuning (see Appendix A), they are merged into a single group. Performing this merging iteratively, rather than in a single final pass, ensures that decisions are made on the most up-to-date cluster compositions, accounting for changes introduced by outlier removal, grouping, and reassignment.

This incremental approach prevents premature large-scale merges that risk collapsing meaningful fine-grained distinctions, while still allowing the framework to consolidate redundant clusters as they appear. Over successive iterations, the result is a set of visually coherent, non-redundant clusters that better reflect the ecological structure of the dataset. By reducing spurious fragmentation and redundancy, these refinements directly improve the accuracy of abundance distributions, which in turn leads to more reliable Shannon and Simpson diversity estimates.

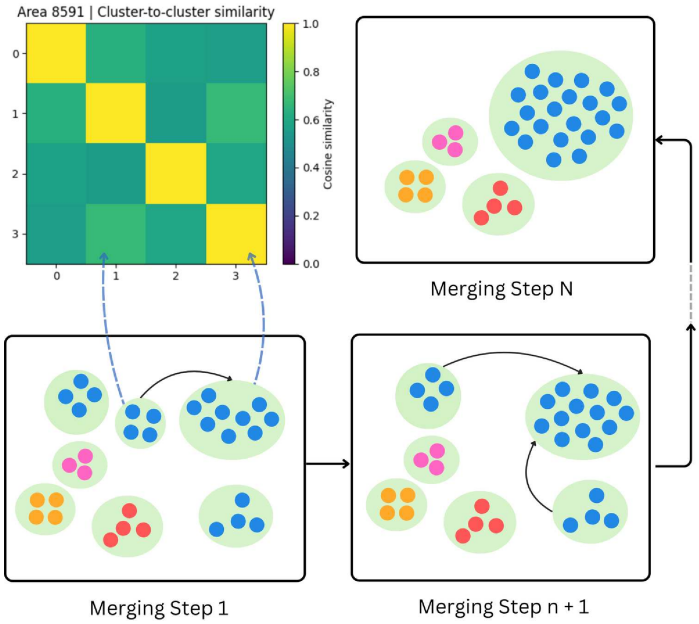


Fig. 4: Iterative merging of visually similar clusters, guided by centroid similarity, to consolidate over-fragmented groups.

5 Conclusion

This study introduced an unsupervised clustering framework for quantifying urban forest diversity directly from street level imagery, capturing both evenness and dominance based diversity indices. Extensive hyperparameter tuning and ablation experiments revealed that combining outlier elimination with cluster merging provided the best trade off between clustering quality, richness accuracy, and preservation of diversity structure, while additional reassignment steps offered only marginal benefits. Across diverse urban contexts, the method achieved low Wasserstein distances for both Shannon and Simpson diversity, indicating strong agreement with ground truth even in cities with highly imbalanced genera distributions.

Importantly, the framework preserved the spatial organization of diversity, as reflected by minimal differences in Moran's I values for Shannon and Simpson indices across cities. This ability to recover both the magnitude and spatial structure of key biodiversity metrics underscores the method's potential as a scalable monitoring tool for urban tree communities. While challenges remain including uneven imagery availability, occlusion from privacy blurring, and environmental variability, these results demonstrate the promise of computer vision based approaches for large scale, fine grained biodiversity assessment and their potential integration into urban forestry policy and planning.

Acknowledgements. The authors thank Dubai Future Foundation, UnipolTech, Consiglio per la Ricerca in Agricoltura e l'Analisi dell'Economia Agraria, Volkswagen Group America, FAE Technology, Samoo Architects & Engineers, Shell, GoAigua, ENEL Foundation, Kyoto University, Weizmann Institute of Science, Universidad Autónoma de Occidente, Instituto Politecnico Nacional, Imperial College London, Universitá di Pisa, KTH Royal Institute of Technology, AMS Institute and all the members of the MIT Senseable City Lab Consortium for supporting this research.

Declarations

5.1 Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

5.2 Conflict of interest

The authors declare no competing interests.

5.3 Ethics approval and consent to participate

Not applicable.

5.4 Consent for publication

Not applicable.

Data availability

All data supporting the findings of this study are available within the paper and its Supplementary Information. The code used for analysis is available at https://github.com/Diaa340/Releaf/.

5.5 Materials availability

Not applicable.

5.6 Author contribution

M.M., F.D., I.Z., M.S., and D.A.A. conceptualized the study. D.A.A. and M.S. performed the methodology development and data curation, conducted the formal analysis, and wrote the original draft. M.M., F.D., and I.Z. supervised the research, C.R. provided resources, and handled funding acquisition. All authors reviewed and edited the manuscript.

References

[1] Magurran, A.E.: Diversity indices and species abundance models. In: Magurran, A.E. (ed.) Ecological Diversity and Its Measurement, pp. 7–45. Springer, Dordrecht (1988). https://doi.org/10.1007/978-94-015-7358-0_2

[2] Magurran, A.E.: Choosing and interpreting diversity measures. In: Magurran, A.E. (ed.) Ecological Diversity and Its Measurement, pp. 61–80. Springer, Dordrecht (1988). https://doi.org/10.1007/978-94-015-7358-0_4

[3] Nielsen, A.B., Östberg, J., Delshammar, T.: Review of urban tree inventory methods used to collect data at single-tree level. Arboriculture & Urban Forestry (AUF) **40**(2), 96–111 (2014)

[4] Davies, Z.G., Edmondson, J.L., Heinemeyer, A., Leake, J.R., Gaston, K.J.: Mapping an urban ecosystem service: quantifying above-ground carbon storage at a city-wide scale. Journal of applied ecology **48**(5), 1125–1134 (2011)

[5] Escobedo, F., Seitz, J.: Costs of Managing an Urban Forest. EDIS **2009**(7) (2009) https://doi.org/10.32473/edis-fr279-2009

[6] Boonman, C.C.F., Serra-Diaz, J.M., Hoeks, S., Guo, W.-Y., Enquist, B.J., Maitner, B., Malhi, Y., Merow, C., Buitenwerf, R., Svenning, J.-C.: More than 17,000 tree species are at risk from rapid global change. Nature Communications **15**(1), 166 (2024) https://doi.org/10.1038/s41467-023-44321-9 . Publisher: Nature Publishing Group

[7] Cardinale, B.J., Duffy, J.E., Gonzalez, A., Hooper, D.U., Perrings, C., Venail, P., Narwani, A., Mace, G.M., Tilman, D., Wardle, D.A., Kinzig, A.P., Daily, G.C., Loreau, M., Grace, J.B., Larigauderie, A., Srivastava, D.S., Naeem, S.: Biodiversity loss and its impact on humanity. Nature **486**(7401), 59–67 (2012) https://doi.org/10.1038/nature11148 . Publisher: Nature Publishing Group

[8] Paul, B.: The scale of the biodiversity crisis laid bare. Nature **635**, 31 (2024)

[9] Aronson, M.F.J., La Sorte, F.A., Nilon, C.H., Katti, M., Goddard, M.A., Lepczyk, C.A., Warren, P.S., Williams, N.S.G., Cilliers, S., Clarkson, B., Dobbs, C., Dolan, R., Hedblom, M., Klotz, S., Kooijmans, J.L., Kühn, I., MacGregor-Fors, I., McDonnell, M., Mörtberg, U., Pyšek, P., Siebert, S., Sushinsky, J., Werner, P., Winter, M.: A global analysis of the impacts of urbanization on bird and plant diversity reveals key anthropogenic drivers. Proceedings of the Royal Society B: Biological Sciences **281**(1780), 20133330 (2014) https://doi.org/10.1098/rspb.2013.3330 . Publisher: Royal Society

[10] Helden, A.J., Stamp, G.C., Leather, S.R.: Urban biodiversity: comparison of insect assemblages on native and non-native trees. Urban Ecosystems **15**(3), 611–624 (2012) https://doi.org/10.1007/s11252-012-0231-x

[11] Gunnarsson, B., Knez, I., Hedblom, M., Sang, A.O.: Effects of biodiversity and environment-related attitude on perception of urban green space. Urban Ecosystems **20**(1), 37–49 (2017) https://doi.org/10.1007/s11252-016-0581-x

[12] Lin, B.B., Fuller, R.A., Bush, R., Gaston, K.J., Shanahan, D.F.: Opportunity or Orientation? Who Uses Urban Parks and Why. PLOS ONE **9**(1), 87422 (2014) https://doi.org/10.1371/journal.pone.0087422 . Publisher: Public Library of Science

[13] Liu, J., Slik, F.: Are street trees friendly to biodiversity? Landscape and Urban Planning **218**, 104304 (2022) https://doi.org/10.1016/j.landurbplan.2021.104304

[14] Hewitt, C.N., Ashworth, K., MacKenzie, A.R.: Using green infrastructure to improve urban air quality (GI4AQ). Ambio **49**(1), 62–73 (2020) https://doi.org/10.1007/s13280-019-01164-3

[15] McPherson, E.G., Doorn, N., Goede, J.: Structure, function and value of street trees in California, USA. Urban Forestry & Urban Greening **17**, 104–115 (2016) https://doi.org/10.1016/j.ufug.2016.03.013

[16] Wang, X., Dallimer, M., Scott, C.E., Shi, W., Gao, J.: Tree species richness and diversity predicts the magnitude of urban heat island mitigation effects of greenspaces. Science of The Total Environment **770**, 145211 (2021) https://doi.org/10.1016/j.scitotenv.2021.145211

[17] Ishimatsu, K., Ito, K., Mitani, Y.: Developing Urban Green Spaces and Effective Use of Rooftop Spaces for Cooling and Urban Biodiversity. In: Ito, K. (ed.) Urban Biodiversity and Ecological Design for Sustainable Cities, pp. 217–240. Springer, Tokyo (2021). https://doi.org/10.1007/978-4-431-56856-8_10

[18] Zhao, J., Xu, W., Li, R.: Visual preference of trees: The effects of tree attributes and seasons. Urban Forestry & Urban Greening **25**, 19–25 (2017) https://doi.org/10.1016/j.ufug.2017.04.015

[19] Muluneh, M.G., Worku, B.B.: Contributions of urban green spaces for climate change mitigation and biodiversity conservation in Dessie city, Northeastern Ethiopia. Urban Climate **46**, 101294 (2022) https://doi.org/10.1016/j.uclim.2022.101294

[20] Esperon-Rodriguez, M., Tjoelker, M.G., Lenoir, J., Baumgartner, J.B., Beaumont, L.J., Nipperess, D.A., Power, S.A., Richard, B., Rymer, P.D., Gallagher, R.V.: Climate change increases global risk to urban forests. Nature Climate Change **12**(10), 950–955 (2022) https://doi.org/10.1038/s41558-022-01465-8 . Publisher: Nature Publishing Group

[21] Jenerette, G.D., Clarke, L.W., Avolio, M.L., Pataki, D.E., Gillespie, T.W., Pincetl, S., Nowak, D.J., Hutyra, L.R., McHale, M., McFadden, J.P., Alonzo,

M.: Climate tolerances and trait choices shape continental patterns of urban tree biodiversity. Global Ecology and Biogeography **25**(11), 1367–1376 (2016) https://doi.org/10.1111/geb.12499 . Publisher: John Wiley & Sons, Ltd

[22] Beery, S.M.: Where the Wild Things Are: Computer Vision for Global-Scale Biodiversity Monitoring. Ph.D., California Institute of Technology, United States – California (2023). https://doi.org/10.7907/m4mt-2q51 . ISBN: 9798380870627. https://www.proquest.com/docview/2898870216/abstract/D69C70D751F84DC2PQ/1

[23] Stevens, S., Wu, J., Thompson, M.J., Campolongo, E.G., Song, C.H., Carlyn, D.E., Dong, L., Dahdul, W.M., Stewart, C., Berger-Wolf, T., *et al.*: Bioclip: A vision foundation model for the tree of life. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19412–19424 (2024)

[24] Sastry, S., Khanal, S., Dhakal, A., Ahmad, A., Jacobs, N.: Taxabind: A unified embedding space for ecological applications. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 1765–1774 (2025). IEEE

[25] Sierra, E., Gillespie, L.E., Soltani, S., Exposito-Alonso, M., Kattenborn, T.: Divshift: Exploring domain-specific distribution shifts in large-scale, volunteer-collected biodiversity datasets. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, pp. 28386–28396 (2025)

[26] Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning, pp. 1597–1607 (2020). PmLR

[27] Haukoos, J.S., Lewis, R.J.: Advanced statistics: bootstrapping confidence intervals for statistics with "difficult" distributions. Academic emergency medicine **12**(4), 360–365 (2005)

[28] Li, T., Cao, P., Yuan, Y., Fan, L., Yang, Y., Feris, R.S., Indyk, P., Katabi, D.: Targeted supervised contrastive learning for long-tailed recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6918–6928 (2022)

[29] Bhat, S.D., More, A., Soni, M., Agrawal, S.: Prior2posterior: Model prior correction for long-tailed learning. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 1289–1298 (2025). IEEE

[30] Liu, J., Sun, Y., Han, C., Dou, Z., Li, W.: Deep representation learning on long-tailed data: A learnable embedding augmentation perspective. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2970–2979 (2020)

[31] Panaretos, V.M., Zemel, Y.: Statistical aspects of wasserstein distances. Annual Review of Statistics and Its Application **6**(1), 405–431 (2019) https://doi.org/10.1146/annurev-statistics-030718-104938

[32] Jost, L.: Entropy and diversity. Oikos **113**(2), 363–375 (2006)

[33] Panaretos, V.M., Zemel, Y.: An Invitation to Statistics in Wasserstein Space, 1st edn. SpringerBriefs in Probability and Mathematical Statistics. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-38438-8

[34] Murphy, H.S.: Mapping privacy protection in the digital world: study of the privacy implications of street-level imaging applications (2011)

[35] Metzler, A.B., Nathvani, R., Sharmanska, V., Bai, W., Moulds, S., Owoo, N.S., Fynn, I.E.M., Muller, E., Dufitimana, E., Akara, G.K., *et al.*: Unsupervised deep clustering of high-resolution satellite imagery reveals phenotypes of urban development in sub-saharan africa. Science of The Total Environment **988**, 179739 (2025)

[36] Wang, S., Azzari, G., Lobell, D.B.: Crop type mapping without field-level labels: Random forest transfer and unsupervised clustering techniques. Remote Sensing of Environment **222**, 303–317 (2019) https://doi.org/10.1016/j.rse.2018.12.026

[37] Saha, O., Van Horn, G., Maji, S.: Improved zero-shot classification by adapting vlms with text descriptions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 17542–17552 (2024)

[38] Beery, S., Wu, G., Edwards, T., Pavetic, F., Majewski, B., Mukherjee, S., Chan, S., Morgan, J., Rathod, V., Huang, J.: The Auto Arborist Dataset: A Large-Scale Benchmark for Multiview Urban Forest Monitoring Under Domain Shift. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 21262–21275. IEEE, New Orleans, LA, USA (2022). https://doi.org/10.1109/CVPR52688.2022.02061 . https://ieeexplore.ieee.org/document/9878589/

[39] Wegner, J.D., Branson, S., Hall, D., Schindler, K., Perona, P.: Cataloging Public Objects Using Aerial and Street-Level Images — Urban Trees. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6014–6023. IEEE, Las Vegas, NV, USA (2016). https://doi.org/10.1109/CVPR.2016.647 . http://ieeexplore.ieee.org/document/7781016/

[40] Brown, C.F., Kazmierski, M.R., Pasquarella, V.J., Rucklidge, W.J., Samsikova, M., Zhang, C., Shelhamer, E., Lahera, E., Wiles, O., Ilyushchenko, S., Gorelick, N., Zhang, L.L., Alj, S., Schechter, E., Askay, S., Guinan, O., Moore, R., Boukouvalas, A., Kohli, P.: AlphaEarth Foundations: An embedding field model for accurate and efficient global mapping from sparse label data (2025). https://arxiv.org/abs/2507.22291

[41] Barbier, S., Chevalier, R., Loussot, P., Bergès, L., Gosselin, F.: Improving biodiversity indicators of sustainable forest management: Tree genus abundance rather than tree genus richness and dominance for understory vegetation in french lowland oak hornbeam forests. Forest ecology and management **258**, 176–186 (2009)

[42] Sjöman, H., Östberg, J., Bühler, O.: Diversity and distribution of the urban tree population in ten major nordic cities. Urban Forestry & Urban Greening **11**(1), 31–39 (2012)

[43] Nowak, D.J., Greenfield, E.J., Hoehn, R.E., Lapoint, E.: Carbon storage and sequestration by trees in urban and community areas of the united states. Environmental pollution **178**, 229–236 (2013)

[44] Mattocks, R.H.: Street Trees: Their Selection, Planting and After-Care: Illustrated. Town Planning Review **11**(1), 21 (1924) https://doi.org/10.3828/tpr.11.1.u56t446671022x19

[45] Ankerst, M., Breunig, M., Kriegel, H.-P., Ng, R., Sander, J.: Ordering points to identify the clustering structure. In: Proc. Acm Sigmod, vol. 99 (2008)

[46] Bezdek, J.C., Ehrlich, R., Full, W.: FCM: The fuzzy c-means clustering algorithm. Computers & geosciences **10**(2-3), 191–203 (1984). ISBN: 0098-3004 Publisher: Elsevier

[47] Bhattacharjee, P., Mitra, P.: A survey of density based clustering algorithms. Frontiers of Computer Science **15**(1), 151308 (2021)

[48] McInnes, L., Healy, J., Astels, S., *et al.*: hdbscan: Hierarchical density based clustering. J. Open Source Softw. **2**(11), 205 (2017)

[49] Xia, P., Zhang, L., Li, F.: Learning similarity with cosine similarity ensemble. Information sciences **307**, 39–52 (2015)