Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Botanical blind spots in a megadiverse archipelago: Geographic, taxonomic, temporal, and participation gaps in community documentation of Philippine plants

Adriane Tobias

atobias@ateneo.edu

Ateneo de Manila University

Research Article

Keywords: biodiversity informatics, citizen science, iNaturalist, sampling bias, Wallacean shortfall, conservation decision support

Posted Date: September 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10679230/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Community-generated biodiversity records can enlarge the evidence base for tropical conservation, but raw record density conflates biological diversity with accessibility, identification capacity, temporal organization, and participation. This study analyzed 280,194 deduplicated plant observations contributed by 15,029 observers to 17 regional projects of Talahalamanan ng Pilipinas. Current Philippine Standard Geographic Code units, reconciled 2024 population catchments, administrative land-area denominators, 5- and 10-km equal-area grids, observer-cohort retention, taxonomic-resolution indicators, Good coverage and Chao diagnostics, and multivariate province profiles were integrated for 82 provinces and the National Capital Region. Sixty-four observations lacking a defensible current-unit assignment were retained nationally but excluded from province analyses. Documentation was highly concentrated: one province supplied 41.5% and the leading five units 69.2% of assigned observations (Gini = 0.825). Observation volume was strongly associated with observer richness (Spearman rho = 0.929) and occupied 10-km cells (rho = 0.777), but not with species-resolution rate. Spatial conclusions were robust to grid choice: area-normalized 5- and 10-km footprints had rho = 0.975 and 80.5% quartile agreement. Median Research Grade, Good coverage, observer Gini, and 365-day retention were 24.7%, 55.2%, 0.676, and 31.7%, respectively. Three April peaks contained 44.2% of dated records, and 12 units remained documentation frontiers. Weak cluster separation supported continuous diagnostic profiles over a province league table. The framework assigns action-specific priorities such as field documentation, geographic expansion, identifier mobilization, community diversification, seasonal follow-up, or stratified repetition without treating sparse records as evidence of low biodiversity.

Introduction

The Philippines combines exceptional plant richness and endemism with intense habitat pressure across a fragmented archipelago. Its global biodiversity-hotspot status gives botanical knowledge immediate conservation relevance, yet available evidence remains uneven among islands, habitats, and administrative areas (Myers et al., 2000; Posa et al., 2008). Similar imbalances occur globally: plant-occurrence databases contain pronounced spatial, taxonomic, and temporal biases, while the places most in need of biodiversity information often have the weakest coverage (Amano et al., 2016; Meyer et al., 2016). A sparsely documented Philippine province may be genuinely species-poor, logistically difficult to reach, weakly connected to institutions, poorly served by taxonomic expertise, or simply absent from organized documentation. Identical record totals can therefore arise from very different biological and knowledge-production processes.

Community science can reduce some of these gaps. Platforms that enable georeferenced photographic observations expand observation networks, provide recent occurrence evidence, support environmental education, and connect public participation with research and management (Bonney et al., 2009; Chandler et al., 2017; Dickinson et al., 2010; McKinley et al., 2017). iNaturalist is particularly relevant because it combines broad taxonomic scope, public georeferencing, community identification, and an explicit data-quality assessment. Its use in biodiversity research spans species distributions, range

dynamics, conservation, invasion biology, and taxonomic discovery (Mason et al., 2025). Yet more records do not automatically mean less bias. Opportunistic observations remain shaped by roads, cities, institutions, protected-area access, participant preferences, taxonomic conspicuousness, and uneven expertise (Beck et al., 2014; Bird et al., 2014; Geldmann et al., 2016; Troudet et al., 2017).

Participation structure is part of both data quality and system resilience. A project with thousands of observations may represent a broad community of recurring contributors or depend on one exceptionally active observer. These situations have different implications for continuity, geographic reach, and vulnerability to participant turnover. Citizen-science contributions commonly follow long-tailed distributions in which a small fraction of participants supplies a large fraction of the data (Sauermann & Franzoni, 2015). iNaturalist participants also differ markedly in activity, taxonomic breadth, spatial behavior, and duration of engagement (Di Cecco et al., 2021). Observation totals should therefore be interpreted together with contribution concentration, effective observer number, recruitment, and retention.

Taxonomic maturity is a separate dimension. Photographic records can remain useful before species-level identification, but a large unresolved backlog constrains analyses of composition, range, nativity, and conservation status. Research Grade is an iNaturalist platform category rather than a universal validation standard, and species-level resolution is not equivalent to Research Grade. Conversely, Needs ID observations may remain informative at genus or family level, while zero exported identification agreements do not necessarily indicate an incorrect record. Diagnosis requires multiple indicators of taxonomic depth, agreement, and backlog rather than a single quality label (iNaturalist, 2026; Kosmala et al., 2016).

Sampling completeness also requires caution. Equal record counts can represent different degrees of repeated encounter with the assemblage sampled by the observation process, and raw observed richness is not comparable when effort varies. Coverage-based methods standardize comparisons by sample completeness rather than sample size (Chao & Jost, 2012; Chao et al., 2014). Opportunistic observations are not standardized ecological samples, however. Good coverage and nonparametric estimators can diagnose documentation saturation, but they cannot be interpreted as the percentage of a province's true flora known. Spatial footprint likewise requires area standardization and sensitivity to grid size.

Talahalamanan ng Pilipinas was conceived, established, and coordinated by the author as a national network of region-specific iNaturalist projects for community documentation of Philippine plants. The initiative permits a national diagnosis of where botanical knowledge is geographically sparse, taxonomically immature, temporally discontinuous, socially concentrated, or repeatedly incomplete. Its central distinction is between biological priority and documentation priority: few records do not demonstrate low diversity, while many records do not demonstrate adequate documentation.

This study quantified six complementary dimensions, namely – effort, spatial spread, taxonomic maturity, participation resilience, temporal continuity, and documentation completeness, across current

Philippine provinces and the National Capital Region. It tested whether observation volume tracked participation and spatial footprint more strongly than taxonomic maturity; evaluated the robustness of spatial conclusions to 5- versus 10-km grids; incorporated 90- and 365-day observer retention; examined whether provinces formed defensible discrete types; developed an exploratory Philippine Floristic Documentation Index (PFDI) with weight-sensitivity analysis; fitted transparent documentation-process models; and translated the diagnostic profiles into province-specific interventions. The principal expectation was that documentation would be strongly unequal and multidimensional, making component profiles more defensible than a single ranking.

Methods

Initiative provenance and source snapshot

Talahalamanan ng Pilipinas was conceived, established, and coordinated by the author. The underlying observations and identifications were contributed by individual iNaturalist users and constitute the community-generated evidentiary basis of the initiative.

The analytical snapshot was compiled in August 2026 from 17 regional project exports. Exported fields included observation identifiers and dates, observer accounts, quality grades, identification agreement and disagreement counts, public coordinates, positional uncertainty, administrative labels, image availability, captive/cultivated status, and exported taxonomic hierarchy. Observation dates ranged from 1 May 1907 to 1 August 2026; historical dates can occur when users upload older observations.

The regional exports contained 280,471 observation-project membership rows. National deduplication used the unique iNaturalist observation identifier while preserving a separate many-to-many project-membership table. The procedure produced 280,194 unique observations and retained 277 observations represented in more than one regional project. Validation required unique non-null observation keys, unique observation-project pairs, internally consistent national and province totals, valid date and grid fields, and reproducible table manifests.

Analytical populations and data-quality rules

Analytical populations were matched to each question rather than filtered through one universal quality rule. All 280,194 observations characterized community activity. Records not marked captive/cultivated (n = 185,223) were treated as candidate wild-flora observations, without interpreting the absence of a captive marker as proof of wild origin. Species-resolved candidate-wild records supported species-name frequency and documentation-completeness diagnostics. Research Grade was treated as a platform data-quality state based on verifiability and community identification, not as voucher-confirmed taxonomy (iNaturalist, 2026).

Taxonomic maturity was summarized using the proportions resolved to species or below, Research Grade and Needs ID proportions, and the proportion with zero exported identification agreements. Platform species-name values were retained as operational taxonomic units for this documentation-process paper. No analysis of floristic composition, nativity, endemism, or threat status depended on those names, thereby avoiding unsupported conclusions before accepted-name and conservation-status review.

Public coordinates were available for nearly all observations and images for 99.7%. Public positional uncertainty was reported for most records. Only public coordinates including locations already displaced by iNaturalist geoprivacy were used. Exact coordinates, usernames, URLs, and free-text localities were omitted from public-safe outputs. Fine-scale interpretation was avoided for obscured or low-accuracy records.

Current geographic units and denominator reconciliation

Province was the principal comparison unit. The reference table used the Philippine Standard Geographic Code as of 30 June 2026, representing 82 provinces and treating the National Capital Region as a province-equivalent analytical unit (83 units total; Philippine Statistics Authority, 2026). Legacy labels were reconciled to current units. Sixty-four observations lacked a defensible current-unit assignment and were excluded from province analyses, leaving 280,130 assigned observations; the complete national total retained all 280,194 records.

Population denominators began with 2024 POPCEN province totals. Because the observation assignments represented geographic catchments rather than administrative supervision, populations of geographically embedded highly urbanized cities, Isabela City, and the Special Geographic Area were added to their surrounding catchments where the PSA province total excluded them. This reconciliation summed exactly to the national 2024 population of 112,727,776. Population-normalized measures were observations per 1,000 catchment residents and observers per 100,000 catchment residents.

Land-area denominators used the province and city areas reported in the Philippine Statistical Yearbook and associated Land Management Bureau-derived administrative tables. Areas of geographically embedded independent cities were added where province tabulations treated them separately. Because detailed cadastral areas are administrative denominator values and may not sum to the 300,000-km2 national constant, they were used only for within-study relative rates. The principal area-standardized measures were observations and occupied grid cells per 1,000 km2; no exact percentage of provincial land sampled was claimed.

Spatial footprint and sensitivity

Public coordinates were aggregated to national equal-area grids in an Albers conic equal-area projection centered on the Philippines. The 10-km grid was primary and the 5-km grid was retained for sensitivity

analysis. Province metrics included occupied cells, occupied cells per 1,000 km2, the share of observations in the most sampled 10-km cell, a spatial Herfindahl concentration index, and the effective number of occupied cells (the inverse of the Herfindahl index; Jost, 2006).

Robustness to spatial resolution was assessed among the 82 positive-data units using Spearman correlation between 5- and 10-km occupied-cell counts and between area-normalized occupied-cell densities. Units were also assigned to quartiles at both resolutions, and quartile agreement was reported. This approach tests whether the substantive prioritization changes with grid size without pretending that a centroid-based occupied-cell count is an exact polygon-edge coverage percentage.

Participation, retention, and temporal continuity

Participation metrics included distinct pseudonymous observers per unit, observations per observer, the shares contributed by the largest and five largest observers, the Gini coefficient and Herfindahl index of observer contributions, effective observer number, and observers per 100,000 catchment residents. Province observer counts are not additive nationally because one person may contribute in several provinces. Concentration measures describe dependence and resilience rather than the correctness of observations.

Observer retention was calculated from new-observer cohorts within each geographic unit. An observer was eligible for the 90-day or 365-day measure only when the interval between first activity and the snapshot date was long enough to observe a return. Retention was the proportion of eligible observers who contributed again at or after the specified interval. This design addresses right-censoring but measures return within the analytical project system, not all subsequent activity on iNaturalist.

Temporal continuity was measured from sampling days, active year-months, active years, the share of observations in the most active month, a monthly Herfindahl index, and effective active months. Large temporal pulses were described only as high-intensity months. Event or campaign labels were not inferred from temporal spikes without organizer records.

Documentation completeness

For candidate-wild, species-resolved observations, the analysis calculated observed species-name values, abundance singletons and doubletons, and day-incidence uniques and duplicates. Good's abundance coverage proxy was 1 - f1/N, where f1 is the number of singleton species-name values and N is the number of species-resolved candidate-wild observations (Good, 1953). Bias-corrected Chao1 and day-incidence Chao2 diagnostics were calculated from singleton/doubleton and unique/duplicate-day frequencies (Chao, 1984). Observed-to-estimated ratios summarized documentation saturation.

An approximate 95% uncertainty interval for Good coverage was derived by applying a Wilson interval to the singleton proportion and reversing its limits. This is an uncertainty diagnostic for the current record

pool, not a confidence interval for the proportion of the true provincial flora documented. The analysis deliberately avoids equating observed species-name totals or Chao diagnostics with true flora size.

Statistical analysis and documentation-process models

Descriptive analyses reported medians, interquartile ranges, ranges, Gini coefficients, and shares of national totals contributed by the leading one and five units. Ten preplanned monotonic relationships were evaluated among positive-data units with Spearman rank correlation. The Holm procedure controlled family-wise error (Holm, 1979).

Three count-process models and one taxonomic-resolution model were fitted. Negative-binomial models examined observation volume as a function of standardized log catchment population, log catchment land area, active years, top-month share, and 365-day retention; and occupied 10-km cells as a function of log observations, log effective observers, and top-cell share with log equivalent 10-km land cells as an offset (Lindén & Mäntyniemi, 2011). A province-level ordinary least-squares model with HC3 robust standard errors used a continuity-corrected logit of the species-resolution rate to avoid treating opportunistic observations as independent Bernoulli trials. Effects are reported as exponentiated standardized coefficients. Models diagnose the documentation process and are not causal models of biodiversity.

All analyses were performed in Python 3.13 using pandas 2.2.3, NumPy 2.3.5, SciPy 1.17.0, statsmodels 0.14.6, and scikit-learn 1.8.0. Reproducible scripts, input hashes, derived tables, and validation reports accompany the submission package.

Table 1
National analytical snapshot and completed denominator framework.

Element	Value	Interpretation
Regional source projects	17	Region-organized iNaturalist exports
Deduplicated observations	280,194	All national activity
Assigned province-level observations	280,130	64 unresolved observations excluded only from province analyses
Unique observers	15,029	National unique accounts
Exported species-name values	6,599	Platform name units; not accepted-name flora
Candidate-wild observations	185,223	Not marked captive/cultivated
Research Grade observations	45,470 (16.2%)	Platform DQA state
Species-or-below observations	142,136 (50.7%)	Taxonomic-resolution indicator
Analytical geographic units	83	82 provinces plus NCR
2024 catchment population	112,727,776	Reconciled to the national POPCEN total
Primary spatial resolution	10 km	5-km sensitivity retained

Results

National scale, inequality, and documentation frontiers

Across the 83 current geographic units, 82 contained at least one assigned observation and Batanes contained none in the analytical snapshot. Davao del Norte contributed 116,254 observations (41.5% of assigned records), followed by Davao del Sur (26,660), the National Capital Region (24,183), Rizal (15,691), and Davao Oriental (10,962). The leading five units supplied 69.2% of all assigned observations. Observation totals had a Gini coefficient of 0.825, compared with 0.687 for province observer counts, 0.583 for sampling days, and 0.360 for occupied 10-km cells. Thus, documentation volume was much more geographically concentrated than spatial footprint.

Twelve units had fewer than 100 assigned observations: Batanes (0), Basilan (4), Sulu (16), Catanduanes (28), Tawi-Tawi (35), Abra (40), Apayao (46), Sarangani (58), Biliran (70), Davao Occidental (74), Masbate (89), and Zamboanga Sibugay (97). These units were treated as documentation frontiers rather than assigned unstable multivariate scores.

Table 2
Distributions of principal documentation indicators among positive-data geographic units.

Indicator	n	Median	Interquartile range
Total observations	82	667.5	227.2-1,485.2
Observations per 1,000 km2	82	179.47	70.04-485.81
Observers	82	79.0	35.8-214.5
Observers per 100,000 catchment population	82	9.82	6.25–18.67
Occupied 10-km cells	82	31.5	19.0-46.8
Occupied 10-km cells per 1,000 km2	82	9.92	6.11–14.98
Species-resolution rate	82	60.5%	55.6%-66.7%
Research Grade rate	82	24.7%	20.6%-27.7%
Needs ID rate	82	54.9%	48.1%-61.8%
Good's coverage proxy	82	55.2%	36.5%-67.6%
Observer Gini	82	0.676	0.568–0.740
365-day retention	79	31.7%	25.3%-38.2%
Top-month share	82	22.6%	14.4%-33.3%

Spatial footprint and grid-scale robustness

Palawan had the broadest raw footprint (196 occupied 10-km cells), followed by Quezon (91), Cebu (90), Negros Occidental (85), and Davao Oriental (83). Area standardization changed the ordering: Siquijor, the National Capital Region, Camiguin, Rizal, and Laguna had the highest occupied-cell densities per 1,000 km2. Observation volume was strongly related to occupied 10-km cells (rho = 0.777, Holm-adjusted p < .001), but the association with area-normalized occupied-cell density was more moderate (rho = 0.495, adjusted p < .001). High record volume therefore reflected both a broader footprint and repeated recording within already occupied areas.

Spatial conclusions were robust to grid choice. Raw 5- and 10-km occupied-cell counts had rho = 0.963; area-normalized densities had rho = 0.975. Quartile classifications agreed for 80.5% of positive-data units. The 10-km grid was retained for primary profiles because it reduced sensitivity to public-coordinate displacement and positional uncertainty, while the 5-km analysis demonstrated that the major spatial contrasts did not depend on a single resolution.

Taxonomic maturity and documentation completeness

Across positive-data units, the median species-resolution rate was 60.5% (interquartile range 55.6–66.7%), median Research Grade rate was 24.7% (20.6–27.7%), median Needs ID rate was 54.9% (48.1–61.8%), and median zero-agreement rate was 55.4% (52.2–59.5%). Observation volume was not associated with species-resolution rate (rho = -0.042, adjusted p = .708) or Research Grade rate (rho = -0.216, adjusted p = .152). More observations therefore did not automatically produce a more taxonomically mature province dataset.

Median Good coverage was 55.2% (36.5–67.6%). Median observed-to-Chao1 and observed-to-Chao2 ratios were 42.1% and 38.5%, respectively. Good coverage was strongly associated with observation volume (rho = 0.912, adjusted p < .001), reflecting increased repeat encounter in larger record pools. Nevertheless, high-volume provinces still differed in identification maturity and singleton structure. These diagnostics describe saturation of the community observation process, not completeness of the provincial flora.

Participation concentration and retention

The median positive-data unit had 79 observers, but the median effective number was only 17.0. Median observer Gini was 0.676, and the leading observer supplied a median 16.2% of provincial records. In the most concentrated units the leading observer contributed more than half of all observations. Observation volume increased with observer richness (rho = 0.929, adjusted p < .001) and with observer inequality (rho = 0.483, adjusted p < .001). Large datasets were therefore often broad in headcount but still supported by strongly unequal contribution structures.

Among the 79 units with eligible cohorts, median 90-day retention was 29.2% and median 365-day retention was 31.7%. Retention was not significantly associated with leading-observer share after Holm correction (rho = -0.147, adjusted p = .392). Retention and contribution concentration therefore capture different aspects of participation resilience. Raw observer totals alone would miss both.

Temporal concentration

Of 275,929 dated observations represented in the province-month panel, 64,820 occurred in April 2026, 32,751 in April 2025, and 24,509 in April 2024. These three months accounted for 44.2% of dated records. Across positive-data units, the median top-month share was 22.6%, and the median effective number of active months was 10.6. Observation volume was negatively associated with top-month share in the bivariate analysis (rho = -0.342, adjusted p = .007), although the process model identified a positive partial association after population, land area, active years, and retention were considered.

The timing and geographic composition of the three April peaks showed that national growth was not temporally even. The analysis does not attribute these peaks to named campaigns, because event calendars and upload histories were not part of the analytical snapshot. Their conservation relevance lies in whether concentrated activity is followed by repeated observation, geographic broadening, contributor return, and identification follow-through.

Multivariate structure, PFDI sensitivity, and intervention profiles

The first three principal components explained 61.4% of standardized variation (PC1 26.6%, PC2 20.3%, PC3 14.5%). No tested Ward solution produced strong separation: the highest silhouette coefficient was 0.182 at k = 5. Province documentation was therefore predominantly continuous and multidimensional rather than partitioned into well-separated natural classes.

The PFDI retained six separate components and used the composite only as an exploratory screening summary. The median rank under 20,000 random weighting schemes was almost perfectly correlated with the unweighted gap rank (rho = 0.999), indicating that the extremes were robust to plausible weight changes. Component profiles remained essential, however, because similar composite scores arose from different deficits.

Rule-based descriptions among the 71 stable units yielded 16 mature-documentation profiles, 14 completeness frontiers, five expert-dominated profiles, four community-rich but identification-poor profiles, two spatially concentrated profiles, and 30 mixed or transitional profiles. The rules are interpretive labels rather than empirical clusters. The weakest components assigned 17 units to identifier mobilization, 14 to geographic expansion, 14 to stratified repeated surveys, 13 to community diversification and retention, seven to new field documentation, and six to repeated seasonal follow-up. The 12 low-volume frontiers were assigned basic field documentation and geographic verification before finer ranking.

Documentation-process models

In the denominator-and-duration negative-binomial model, standardized log catchment population (incidence-rate ratio [IRR] 1.71, 95% CI 1.25–2.34, p < .001) and active years (IRR 4.49, 2.15–9.38, p < .001) were positively associated with observation volume, whereas catchment area was not. In the process-augmented model, population (IRR 2.19, 1.68–2.84), active years (IRR 3.49, 2.17–5.63), top-month share (IRR 1.48, 1.04–2.10), and 365-day retention (IRR 1.85, 1.07–3.20) were significant; catchment area remained nonsignificant. The model explained a modest portion of process variation (pseudo-R2 = 0.087).

With equivalent land cells as an offset, occupied 10-km cells increased with observation volume (IRR 1.59, 1.29–1.97, p < .001), whereas effective observers and top-cell share were not significant. In the province-level logit model, species-resolution rate was positively associated only with leading-observer share (odds ratio 1.17, 1.04–1.32, p = .008); observation volume, cultivated share, top-month share, and retention were nonsignificant. These associations characterize the knowledge-production process and do not identify causes of biodiversity patterns.

Table 3

Preplanned Spearman relationships among positive-data geographic units; p values are Holm-adjusted.

Relationship	n	rho	Adjusted p	Significant
Observation volume vs observer richness	82	0.929	< .001	Yes
Observation volume vs occupied 10-km cells	82	0.777	< .001	Yes
Observation volume vs area-normalized occupied-cell density	82	0.495	< .001	Yes
Observation volume vs species-resolution rate	82	-0.042	= .708	No
Observation volume vs Research Grade rate	82	-0.216	= .152	No
Observation volume vs Good coverage proxy	82	0.912	< .001	Yes
Observation volume vs top-month share	82	-0.342	= .007	Yes
Observation volume vs observer Gini	82	0.483	< .001	Yes
Effective observers vs occupied 10-km cells	82	0.639	< .001	Yes
365-day retention vs leading-observer share	79	-0.147	= .392	No

Table 4

Documentation-process model effects. Predictors were standardized; values are incidence-rate ratios for count models and exponentiated coefficients for the species-resolution logit model.

Model	Predictor	Effect (95% CI)	p
Observation volume baseline	Catchment population	1.71 (1.25–2.34)	< .001
Observation volume baseline	Catchment area	0.95 (0.68–1.31)	= .737
Observation volume baseline	Active years	4.49 (2.15–9.38)	< .001
Observation volume augmented	Catchment population	2.19 (1.68–2.84)	< .001
Observation volume augmented	Catchment area	0.81 (0.63–1.04)	= .094
Observation volume augmented	Active years	3.49 (2.17–5.63)	< .001
Observation volume augmented	Top-month share	1.48 (1.04–2.10)	= .031
Observation volume augmented	365-day retention	1.85 (1.07–3.20)	= .028
Occupied 10-km cells	Observation volume	1.59 (1.29–1.97)	< .001
Occupied 10-km cells	Effective observers	0.87 (0.71–1.06)	= .164
Occupied 10-km cells	Top-cell share	1.01 (0.89–1.14)	= .882
Species resolution	Observation volume	0.95 (0.82–1.11)	= .534
Species resolution	Captive/cultivated share	1.09 (0.96–1.23)	= .186
Species resolution	Leading-observer share	1.17 (1.04–1.32)	= .008
Species resolution	Top-month share	0.98 (0.87–1.11)	= .795
Species resolution	365-day retention	0.95 (0.85–1.05)	= .315

Table 5
Diagnostic conditions and principal interventions. Counts for the first six rows refer to 71 stable-profile units; the final row contains the 12 documentation frontiers.

Dimension/status	Diagnostic condition	Principal intervention	Units
Effort	Low total and area-standardized effort	New field documentation	7
Spatial spread	Low cell density/effective cells or high concentration	Geographic expansion	14
Taxonomic maturity	Low resolution/Research Grade or large Needs ID backlog	Identifier mobilization	17
Participation resilience	Low effective participation, high inequality, or weak retention	Community diversification and retention	13
Temporal continuity	Few effective months or high temporal concentration	Repeated seasonal follow-up	6
Documentation completeness	Low Good/Chao saturation	Stratified repeated surveys	14
Frontier status	Fewer than 100 observations	Basic field documentation and geographic verification	12

Discussion

The national analysis shows that botanical blind spots are not a single deficit. Observation volume, spatial spread, taxonomic maturity, participation resilience, temporal continuity, and documentation completeness varied partly independently. More records generally brought more observers, more occupied cells, and greater repeat encounter, but they did not guarantee species-level identification, Research Grade, distributed contribution, or temporal continuity. Treating raw observation totals as biodiversity or documentation quality would therefore obscure the intervention each province requires.

The concentration of 69.2% of assigned observations in five units is striking, but spatial inequality was substantially lower than volume inequality. This distinction matters: a province can occupy many cells with relatively few observations per cell, while another can accumulate enormous volume within a narrow footprint. Area normalization also changed the interpretation of small, intensively documented units. The high agreement between 5- and 10-km rankings supports the robustness of the principal spatial contrasts, although occupied-cell density remains an observation-process measure rather than a complete land-coverage percentage.

Taxonomic processing did not scale automatically with upload volume. The absence of a positive volume-resolution relationship and the large median Needs ID and zero-agreement shares indicate that identification capacity must be treated as infrastructure. Mobilizing identifiers, organizing taxon-focused identification events, improving photographic evidence, and connecting difficult groups to specialists may yield more analytical value than simply adding records in provinces whose principal deficit is

taxonomic maturity. Research Grade should nevertheless remain one indicator among several, because its platform rules and the photographic medium differ from voucher-based determination.

Good coverage and Chao diagnostics added a dimension that raw richness lacks. Their strong relation to volume was expected: larger record pools repeatedly encounter common species-name units. Yet median observed-to-estimated ratios remained below one-half, and low completeness sometimes persisted where volume was not minimal. Repeated, stratified observation across seasons, habitats, and localities can therefore be a distinct priority. These diagnostics should guide the design of further documentation, not be recast as estimates of the proportion of true flora known.

Participation profiles revealed both scale and fragility. A median effective observer number of 17 contrasted with a median raw observer count of 79, and observer inequality remained high. Retention added information not captured by concentration. Provinces dependent on one or two contributors need succession planning and broader local recruitment even when their observations are high quality; provinces with many first-time contributors but limited return need follow-up activities, feedback, and visible pathways from observation to identification and conservation use.

Temporal pulses can rapidly enlarge a national dataset, but three months accounting for 44.2% of dated records also expose dependence on concentrated activity. The process model’s partial association between top-month share and volume is consistent with high-intensity periods generating large datasets, while retention’s positive association suggests that durable participation also matters. Without verified event metadata, these are process associations rather than campaign effects. Future program evaluation should link events to post-event retention, spatial broadening, and identification backlog resolution, not merely upload totals.

The weak cluster structure is substantively useful. Provinces did not fall into sharply separated natural types, so the PFDI should not become a league table. Component profiles preserve what a rank discards. Cebu, for example, combined strong overall documentation with participation resilience as its weakest dimension; Palawan’s broad footprint did not remove its need for taxonomic support. Conversely, low-volume frontiers require basic documentation before fine distinctions among their rates are meaningful. The intervention table converts diagnosis into operational choices while keeping biological value separate from documentation status.

The documentation-process models support a cautious explanatory narrative. Population and years of activity were consistently associated with observation volume, whereas land area was not. Once area was offset, observation volume was the clearest correlate of occupied-cell spread. Species resolution showed a positive association with leading-observer share, which may indicate the role of highly engaged, knowledgeable contributors, but it also signals potential dependence. The modest model fit and associational design prohibit claims that these variables cause the observed gaps. The models are intended to make the remaining process variation explicit and testable.

Reducing documentation uncertainty is conservation infrastructure. Contemporary photographic records can provide recent evidence of presence, phenology, introduced species, range extensions, and places requiring formal survey. They can complement herbarium records and direct scarce expert effort. The framework is transferable to other biodiversity-rich countries because its minimum requirements are a dated observation table, versioned administrative units, spatial aggregation, participant identifiers, taxonomic resolution, and reproducible diagnostics. The central safeguard should remain unchanged: a documentation gap is uncertainty in a knowledge system, not evidence of low biodiversity.

Limitations

The study is complete for its stated documentation-process scope, but several boundaries of inference remain scientifically unavoidable. First, the dataset is opportunistic and project-bounded. Observation count is not abundance, observed species-name count is not total flora, a missing record is not evidence of absence, and a sampling day is not a standardized survey. Good coverage and Chao diagnostics are sensitive to uneven effort and future identification changes.

Second, the spatial analysis uses public or platform-obscured coordinates aggregated to cell centroids. The 5- and 10-km sensitivity analysis limits dependence on one resolution, and area-standardized densities correct for gross differences in unit size, but they do not represent exact polygon-intersection percentages or environmental representativeness. The 64 unresolved province assignments were excluded transparently and are too few to alter national conclusions, but they should be resolved in any fine-scale operational map.

Third, platform species-name units were not converted into an accepted-name floristic checklist. This is a deliberate scope boundary rather than a blank analytical module: no composition, nativity, endemism, or threat-status claim is made in this paper. A separate conservation-status analysis requires licensed source integration, synonym review, expert verification, and safeguards for sensitive taxa. Similarly, the present process models do not include roads, forest cover, protected areas, institutions, or travel time and therefore do not claim to explain accessibility or biological richness.

Fourth, observer retention is based on pseudonymous account activity within the project system. It does not measure offline participation, identifying activity by people who did not upload observations, or return under a different account. Exported agreement counts also do not reconstruct the identifier network or time to identification. Finally, historical observation dates and bulk uploads complicate temporal interpretation; named campaign effects require verified event and upload-time metadata.

Conclusions

Talahalamanan ng Pilipinas demonstrates both the scale of community capacity and the unevenness of the botanical knowledge it produces. More observations were associated with more observers, broader occupied-cell footprints, and greater repeat encounter, but they did not guarantee taxonomic maturity,

temporal continuity, or distributed participation. A few provinces and high-intensity months dominated the national record pool, while 12 units remained documentation frontiers.

A defensible priority framework asks what is missing and which intervention can supply it. Some provinces need basic field documentation; others need geographic expansion, identifier support, participant diversification, repeated sampling, or continuity between high-intensity periods. Current administrative and population reconciliation, area-standardized spatial diagnostics, grid-scale sensitivity, retention, completeness uncertainty, multivariate profiles, and explicit process models make those recommendations reproducible. They reduce documentation uncertainty without confusing data scarcity with biological scarcity.

Declarations

Acknowledgments

Talahalamanan ng Pilipinas was conceived, established, and coordinated by the author. Its national dataset exists because thousands of people documented and identified plants through iNaturalist. The author gratefully acknowledges these observers and identifiers, as well as the educators, facilitators, partner institutions, and local communities that supported participation. The author also acknowledges the developers and maintainers of iNaturalist and members of the Philippine botanical and biodiversity-informatics communities who sustain the wider knowledge infrastructure on which this work depends.

Funding: No funds, grants, or other financial support were received for conducting this study or preparing the manuscript.

Competing interests: The author declares no relevant financial or non-financial interests.

Author contributions: Adriane Tobias conceived, established, and coordinated Talahalamanan ng Pilipinas; designed the study and analytical framework; curated and integrated the data; performed the formal analyses and visualization; administered the initiative; drafted and revised the manuscript; and approved the final version.

Ethics approval: This study analyzed biodiversity observation records and did not involve intervention with human participants, animals, or plant material. Observer-level analyses used pseudonymous identifiers and were reported only in aggregate. Exact coordinates and direct account identifiers were excluded from public outputs.

Consent to participate and consent for publication: Not applicable.

Data availability: The de-identified analytical tables, province-level derived outputs, spatially aggregated figure data, analysis code, graphical abstract, and validation reports accompany this submission package. The underlying observations and identifications remain attributable to individual iNaturalist contributors and subject to their source licenses. Exact coordinates, observer account names,

observation and image URLs, and free-text localities are withheld pending privacy, licensing, and sensitive-species review. A repository DOI should be added before acceptance.

Code availability: Reproducible Python scripts and nonsensitive derived analytical outputs accompany this submission package.

References

1. Amano T, Lamming JDL, Sutherland WJ (2016) Spatial gaps in global biodiversity information and the role of citizen science. Bioscience 66(5):393–400. https://doi.org/10.1093/biosci/biw022
2. Barcelona JF, Nickrent DL, LaFrankie JV, Callado JRC, Pelser PB (2013) Co's Digital Flora of the Philippines: Plant identification and conservation through cybertaxonomy. Philippine J Sci 142(Special Issue):57–67
3. Beck J, Böller M, Erhardt A, Schwanghart W (2014) Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. Ecol Inf 19:10–15. https://doi.org/10.1016/j.ecoinf.2013.11.002
4. Bird TJ, Bates AE, Lefcheck JS, Hill NA, Thomson RJ, Edgar GJ, Stuart-Smith RD, Wotherspoon S, Krkošek M, Stuart-Smith JF, Pecl GT, Barrett N, Frusher S (2014) Statistical solutions for error and bias in global citizen science datasets. Biol Conserv 173:144–154. https://doi.org/10.1016/j.biocon.2013.07.037
5. Boakes EH, McGowan PJK, Fuller RA, Chang-qing D, Clark NE, O'Connor K, Mace GM (2010) Distorted views of biodiversity: Spatial and temporal bias in species occurrence data. PLoS Biol 8(6):e1000385. https://doi.org/10.1371/journal.pbio.1000385
6. Bonney R, Cooper CB, Dickinson J, Kelling S, Phillips T, Rosenberg KV, Shirk J (2009) Citizen science: A developing tool for expanding science knowledge and scientific literacy. Bioscience 59(11):977–984. https://doi.org/10.1525/bio.2009.59.11.9
7. Callaghan CT, Poore AGB, Mesaglio T, Moles AT, Nakagawa S (2021) Bioscience 71(1):55–63. https://doi.org/10.1093/biosci/biaa131. Three frontiers for the future of biodiversity research using citizen science data
8. Chandler M, See L, Copas K, Bonde AMZ, López BC, Danielsen F, Legind JK, Masinde S, Miller-Rushing AJ, Newman G, Rosemartin A, Turak E (2017) Contribution of citizen science towards international biodiversity monitoring. Biol Conserv 213:280–294. https://doi.org/10.1016/j.biocon.2016.09.004
9. Chao A (1984) Nonparametric estimation of the number of classes in a population. Scand J Stat 11(4):265–270
10. Chao A, Gotelli NJ, Hsieh TC, Sander EL, Ma KH, Colwell RK, Ellison AM (2014) Rarefaction and extrapolation with Hill numbers: A framework for sampling and estimation in species diversity studies. Ecol Monogr 84(1):45–67. https://doi.org/10.1890/13-0133.1

11. Chao A, Jost L (2012) Coverage-based rarefaction and extrapolation: Standardizing samples by completeness rather than size. Ecology 93(12):2533–2547. https://doi.org/10.1890/11-1952.1
12. Daru BH, Park DS, Primack RB, Willis CG, Barrington DS, Whitfeld TJS, Seidler TG, Sweeney PW, Foster DR, Ellison AM, Davis CC (2018) Widespread sampling biases in herbaria revealed from large-scale digitization. New Phytol 217(2):939–955. https://doi.org/10.1111/nph.14855
13. Di Cecco GJ, Barve V, Belitz MW, Stucky BJ, Guralnick RP, Hurlbert AH (2021) Observing the observers: How participants contribute data to iNaturalist and implications for biodiversity science. Bioscience 71(11):1179–1188. https://doi.org/10.1093/biosci/biab093
14. Dickinson JL, Zuckerberg B, Bonter DN (2010) Citizen science as an ecological research tool: Challenges and benefits. Annu Rev Ecol Evol Syst 41:149–172. https://doi.org/10.1146/annurev-ecolsys-102209-144636
15. Geldmann J, Heilmann-Clausen J, Holm TE, Levinsky I, Markussen B, Olsen K, Rahbek C, Tøttrup AP (2016) What determines spatial bias in citizen science? Exploring four recording schemes with different proficiency requirements. Divers Distrib 22(11):1139–1149. https://doi.org/10.1111/ddi.12477
16. Good IJ (1953) The population frequencies of species and the estimation of population parameters. Biometrika 40(3–4):237–264. https://doi.org/10.2307/2333344
17. Holm S (1979) A simple sequentially rejective multiple test procedure. Scand J Stat 6(2):65–70
18. Hsieh TC, Ma KH, Chao A (2016) iNEXT: An R package for rarefaction and extrapolation of species diversity (Hill numbers). Methods Ecol Evol 7(12):1451–1456. https://doi.org/10.1111/2041-210X.12613
19. iNaturalist (2026) What is the Data Quality Assessment (DQA) and how do observations qualify to become Research Grade? iNaturalist Help. https://help.inaturalist.org/en/support/solutions/articles/151000169936
20. Isaac NJB, van Strien AJ, August TA, de Zeeuw MP, Roy DB (2014) Statistics for citizen science: Extracting signals of change from noisy ecological data. Methods Ecol Evol 5(10):1052–1060. https://doi.org/10.1111/2041-210X.12254
21. Jost L (2006) Entropy and diversity. Oikos 113(2):363–375. https://doi.org/10.1111/j.2006.0030-1299.14714.x
22. Kosmala M, Wiggins A, Swanson A, Simmons B (2016) Assessing data quality in citizen science. Front Ecol Environ 14(10):551–560. https://doi.org/10.1002/fee.1436
23. Lindén A, Mäntyniemi S (2011) Using the negative binomial distribution to model overdispersion in ecological count data. Ecology 92(7):1414–1421. https://doi.org/10.1890/10-1831.1
24. Mason BM, Mesaglio T, Heitmann JB, Chandler M, Chowdhury S, Gorta SBZ, Grattarola F, Groom Q, Hitchcock C, Hoskins L, Lowe SK, Marquis M, Pernat N, Shirey V, Baasanmunkh S, Callaghan CT (2025) iNaturalist accelerates biodiversity research. Bioscience 75(11):953–965. https://doi.org/10.1093/biosci/biaf104

25. McKinley DC, Miller-Rushing AJ, Ballard HL, Bonney R, Brown H, Cook-Patton SC, Evans DM, French RA, Parrish JK, Phillips TB, Ryan SF, Shanley LA, Shirk JL, Stepenuck KF, Weltzin JF, Wiggins A, Boyle OD, Briggs RD, Chapin FS, Soukup III, M. A (2017) Citizen science can improve conservation science, natural resource management, and environmental protection. Biol Conserv 208:15–28. https://doi.org/10.1016/j.biocon.2016.05.015
26. Meyer C, Weigelt P, Kreft H (2016) Multidimensional biases, gaps and uncertainties in global plant occurrence information. Ecol Lett 19(8):992–1006. https://doi.org/10.1111/ele.12624
27. Myers N, Mittermeier RA, Mittermeier CG, da Fonseca GAB, Kent J (2000) Biodiversity hotspots for conservation priorities. Nature 403:853–858. https://doi.org/10.1038/35002501
28. Pelser PB, Barcelona JF, Nickrent DL (eds) (2011) -present Co's Digital Flora of the Philippines. https://www.philippineplants.org/
29. Philippine Statistics Authority (2026) Philippine Standard Geographic Code as of 30 June 2026. https://psa.gov.ph/classification/psgc/
30. Philippine Statistics Authority (2025) Philippine Statistical Yearbook 2025: Table 1.1, Population, land area, and density by region and province, census years 2000 to 2024. https://psa.gov.ph/philippine-statistical-yearbook
31. Posa MRC, Diesmos AC, Sodhi NS, Brooks TM (2008) Hope for threatened tropical biodiversity: Lessons from the Philippines. Bioscience 58(3):231–240. https://doi.org/10.1641/B580309
32. Rousseeuw PJ (1987) Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 20:53–65. https://doi.org/10.1016/0377-0427(87)90125-7
33. Sauermann H, Franzoni C (2015) Crowd science user contribution patterns and their implications. Proc Natl Acad Sci USA 112(3):679–684. https://doi.org/10.1073/pnas.1408907112
34. Theobald EJ, Ettinger AK, Burgess HK, DeBey LB, Schmidt NR, Froehlich HE, Wagner C, HilleRisLambers J, Tewksbury JJ, Harsch MA, Parrish JK (2015) Global change and local solutions: Tapping the unrealized potential of citizen science for biodiversity research. Biol Conserv 181:236–244. https://doi.org/10.1016/j.biocon.2014.10.021
35. Troudet J, Grandcolas P, Blin A, Vignes-Lebbe R, Legendre F (2017) Taxonomic bias in biodiversity data and societal preferences. Sci Rep 7:9132. https://doi.org/10.1038/s41598-017-09084-6

Figures

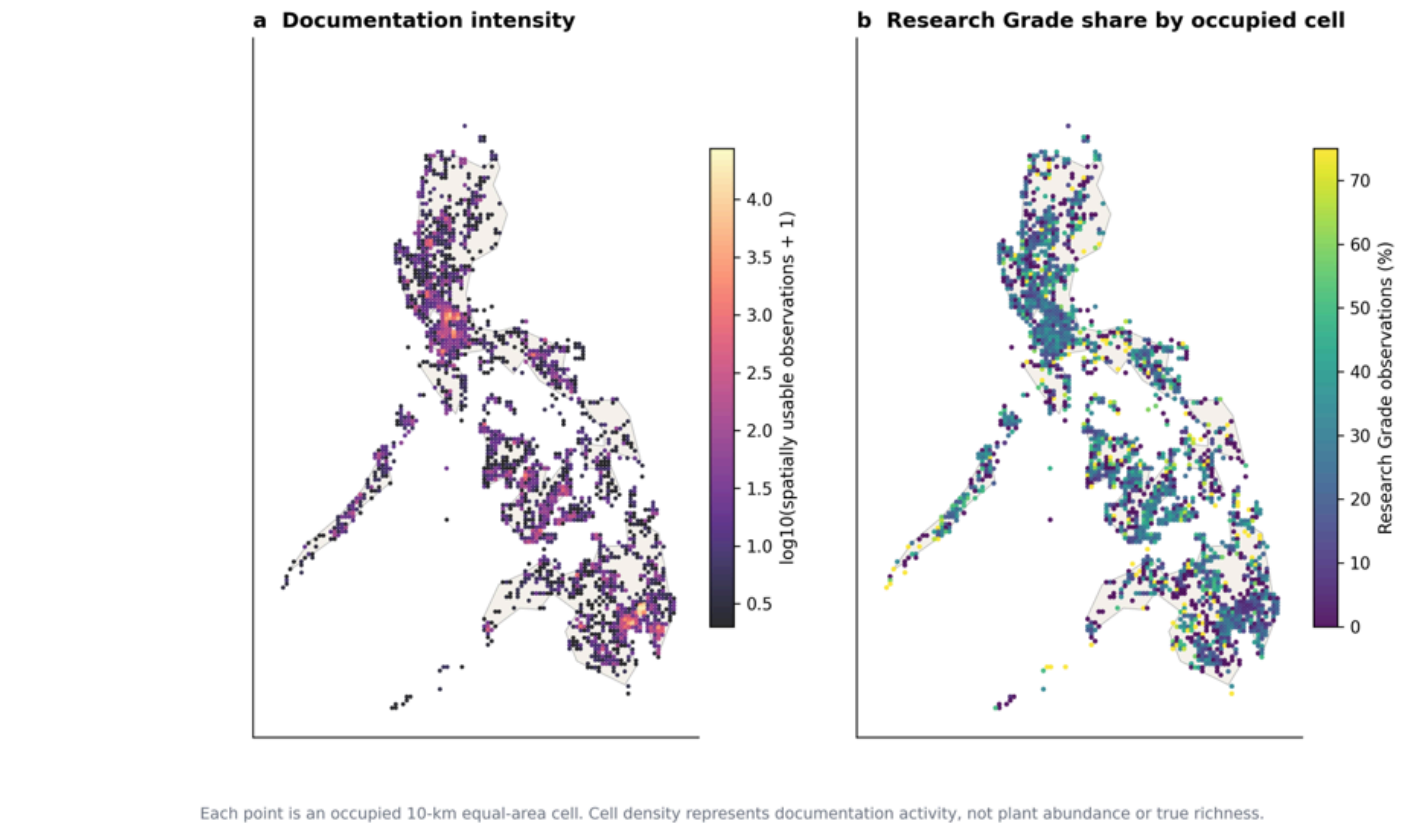


Figure 1

National footprint and taxonomic maturity of community plant documentation. (a) Log-scaled observation intensity in occupied 10-km equal-area grid cells. (b) Research Grade share within occupied cells. Maps portray documentation, not plant abundance or true richness.

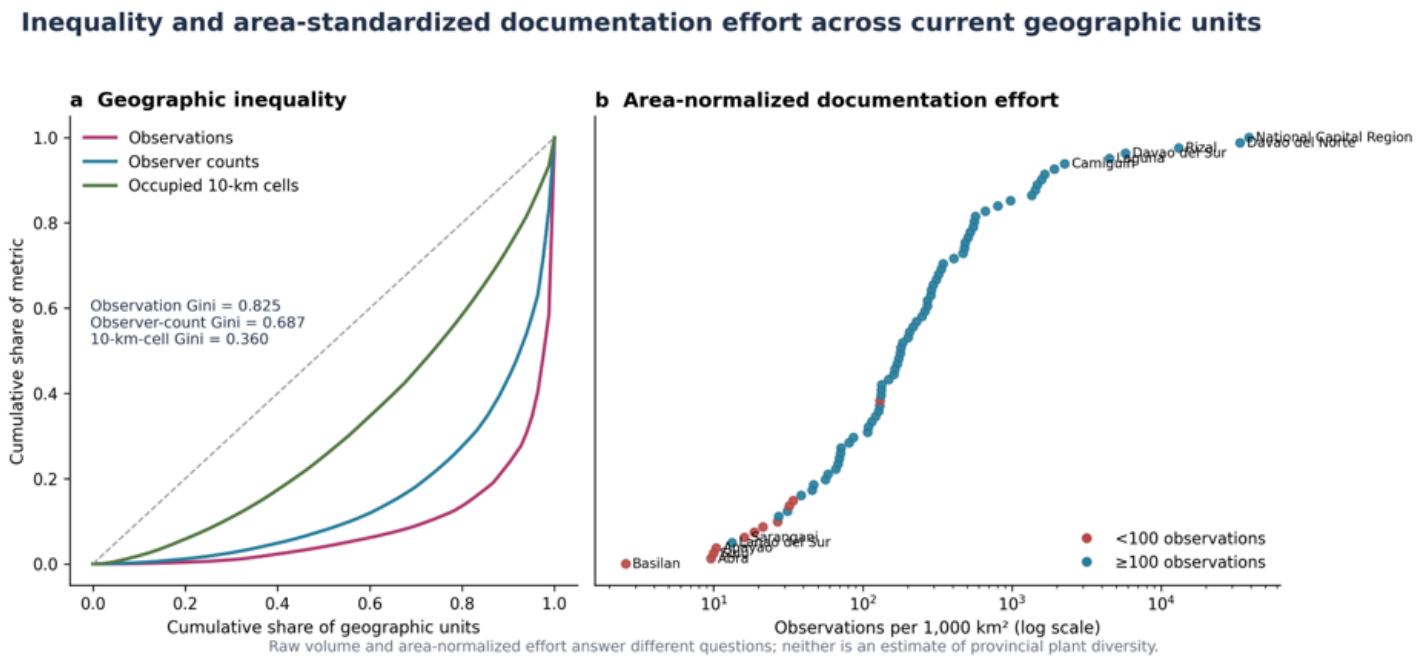


Figure 2

Inequality and area-standardized documentation across current geographic units. (a) Lorenz curves compare the concentration of assigned observations, observer counts, and occupied 10-km cells. (b) Ranked observation effort per 1,000 km2 distinguishes raw volume from area-standardized intensity.

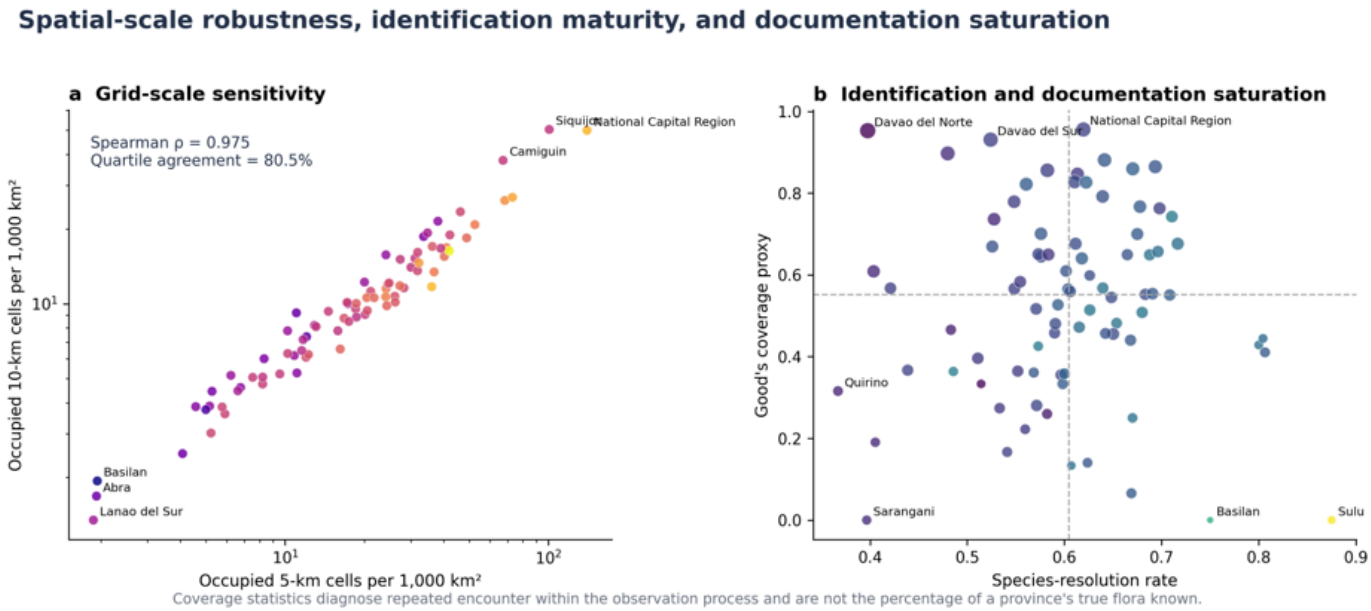


Figure 3

Spatial-scale robustness, identification maturity, and documentation saturation. (a) Area-normalized 5- and 10-km occupied-cell densities. (b) Species-resolution rate. (c) Good coverage proxy. Low-volume frontiers are shown separately because proportion estimates are unstable.

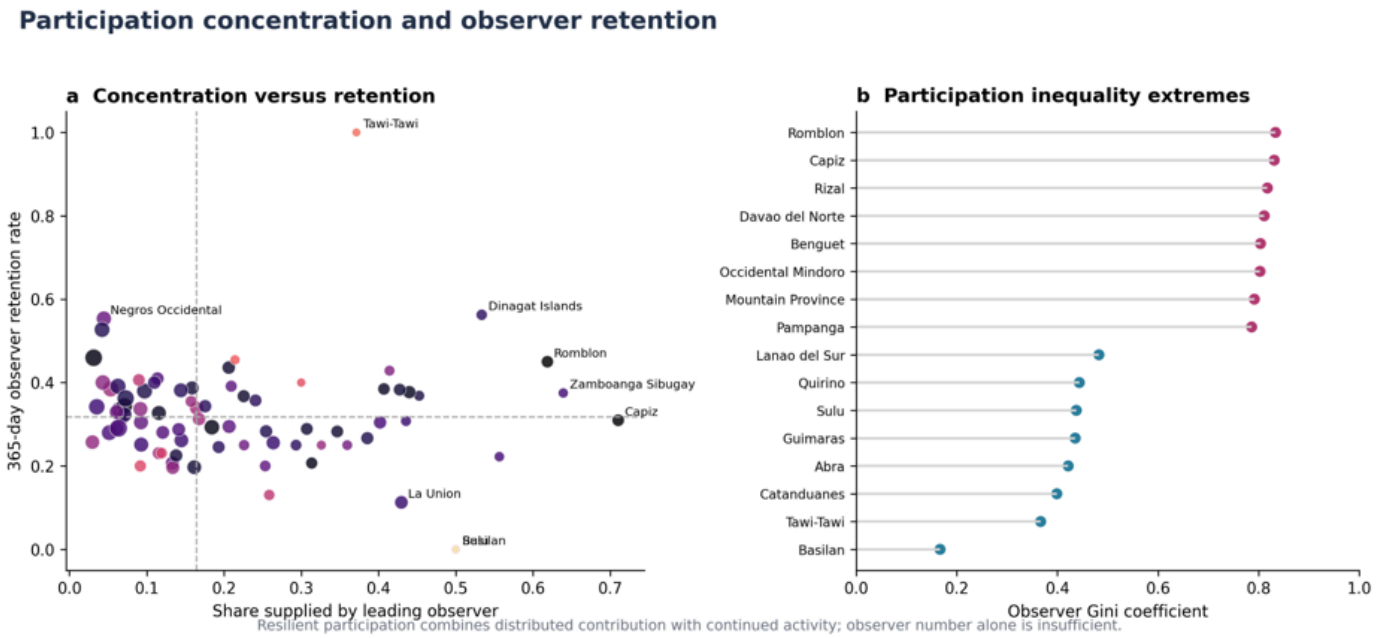


Figure 4

Participation concentration and observer retention. Province profiles compare leading-observer share, observer Gini, and 365-day cohort retention; low-volume units are interpreted cautiously.

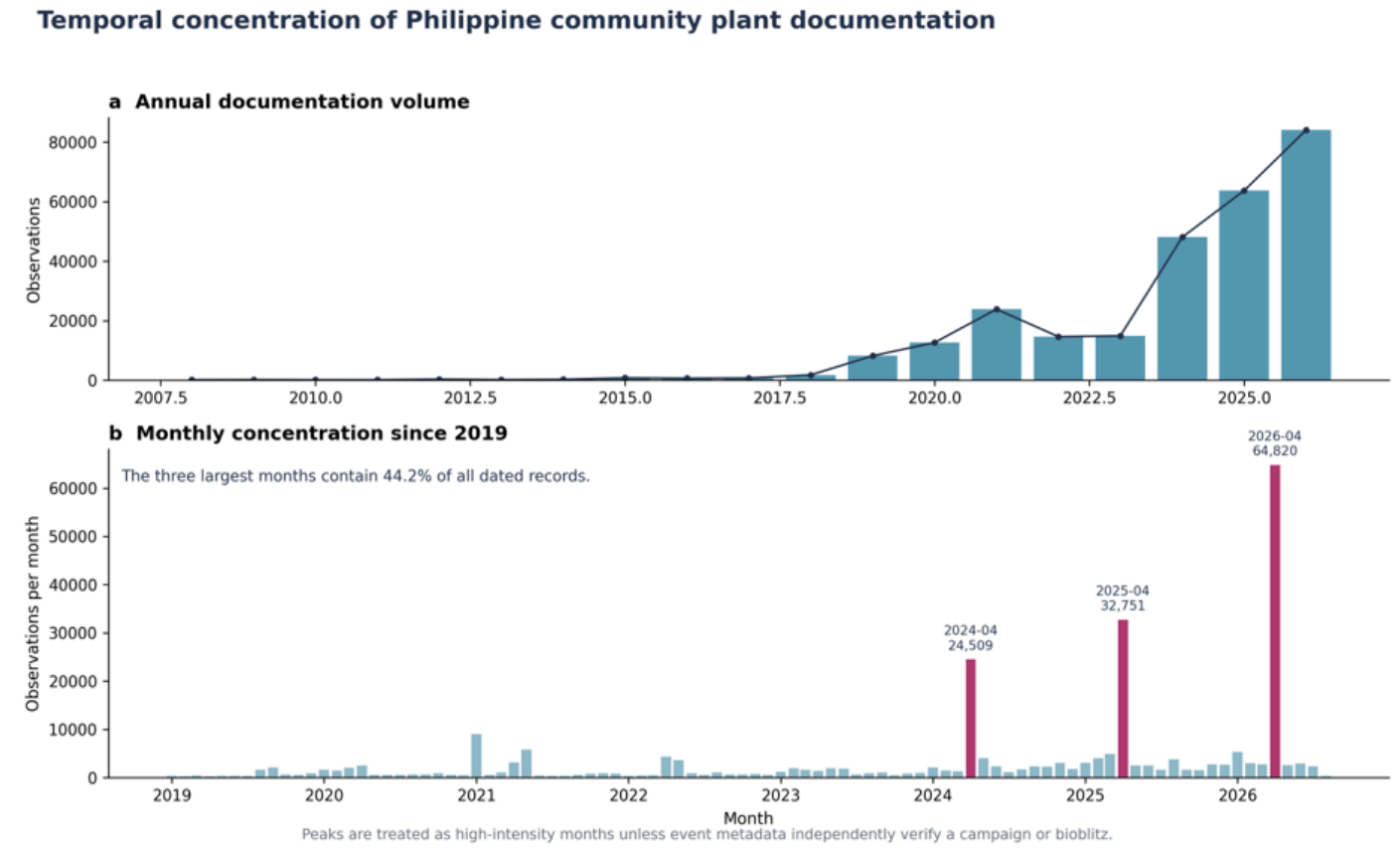


Figure 5

Temporal concentration of Philippine community plant documentation. Annual and monthly series show long-term growth and pronounced high-intensity periods in April 2024, April 2025, and April 2026.

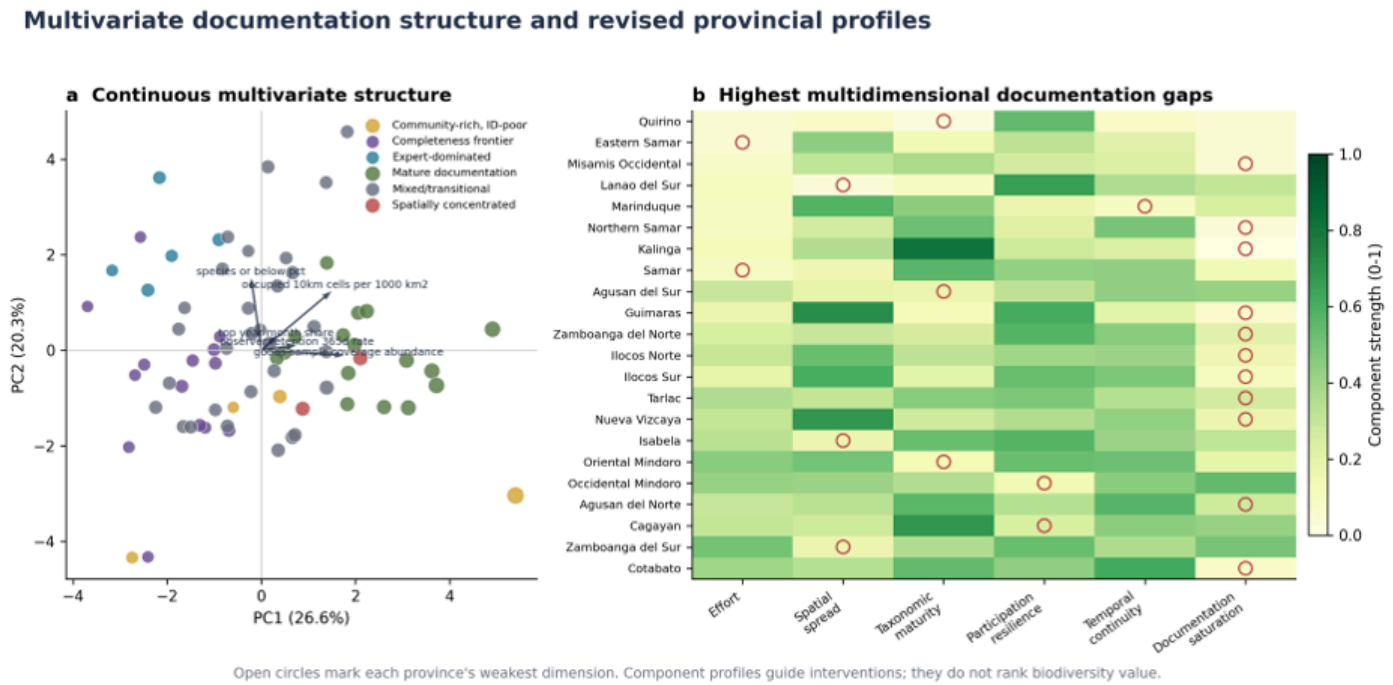


Figure 6

Multivariate documentation structure and provincial profiles. (a) PCA scores show continuous variation rather than discrete groups. (b) A component heat map identifies the weakest dimension among high-gap provinces; open circles mark each weakest component.

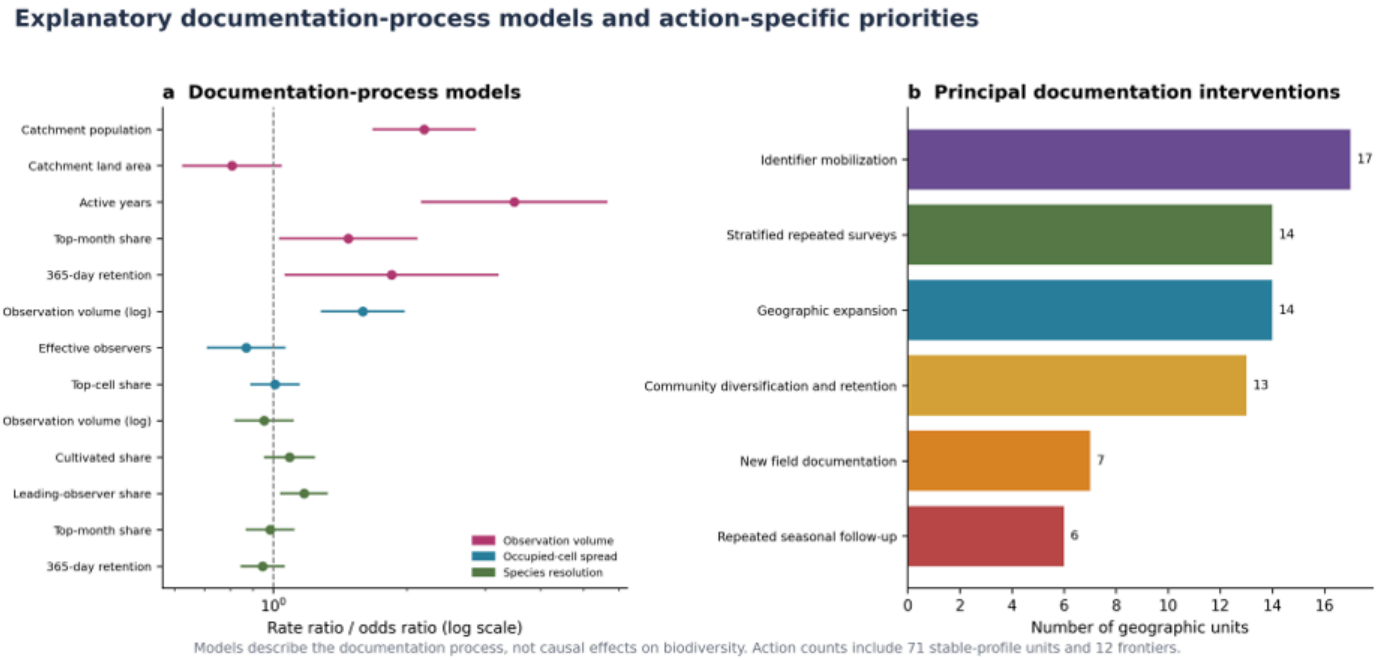


Figure 7

Documentation-process models and action-specific priorities. (a) Exponentiated standardized coefficients with 95% confidence intervals. (b) Number of stable-profile units assigned to each principal intervention.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- TalahalamananPaper1SupplementaryTablesComplete.xlsx