Spatial intelligence in ethnobotany: A spatio-temporal framework for detecting biocultural knowledge systems

Muh Akbar Idris, Rainer W. Bussmann

Correspondence

Muh Akbar Idris[1,2*], Rainer W. Bussmann[3,4]

[1]Department of Statistics and Data Sciences, IPB University, Bogor, Indonesia.
[2]Faculty of Mathematics and Natural Sciences, Universitas Sulawesi Barat, Majene, Indonesia.
[3]Department of Ethnobotany, Institute of Botany and Bakuriani Alpine Botanical Garden, Ilia State University, Botanical Str. 1, 0105 Tbilisi, Georgia.
[4]Department of Botany and Mycology, State Museum of Natural History, Erbprinzenstr. 13, 76133 Karlsruhe, Germany

*Corresponding Author: muhakbaridris@gmail.com

Ethnobotany Research and Applications 36:50 (2026) - http://dx.doi.org/10.32859/era.36.50.1-22
Manuscript received: 17/09/2026 - Revised manuscript received: 23/11/2026 - Published: 24/11/2026

Research

Abstract

Background: Maps of ethnobotanical databases can conflate useful-plant diversity with documentation effort, plant distributions, and administrative scale. We tested whether spatial and network methods reveal robust structure while retaining these interpretive limits.

Methods: Taxon-level major uses in the Royal Botanic Gardens, Kew SEPASAL archive were linked to documented country distributions. The resulting country profiles were analyzed using effort adjustment, permutation-based Global and Local Moran statistics, ordination, clustering stability, Jaccard networks, documentation-time diagnostics, and spatially blocked prediction.

Results: The dataset contained 420,821 country–taxon–use associations for 6,428 taxa in 141 mapped countries. Effort-adjusted richness was spatially autocorrelated under symmetric eight-nearest-neighbor weights (Moran's I = 0.497; p = 0.0001). A balanced two-cluster K-means solution was stable under subsampling (median adjusted Rand index = 0.964). The similarity network contained 19 modules with modularity 0.636. Leave-continent-out prediction performed worse than random cross-validation, and spatial residual structure remained. ST-DBSCAN was feature-incoherent, while constrained K-means selected zero spatial and temporal penalties.

Conclusions: SEPASAL contains reproducible spatial structure, but the analysis does not locate community knowledge or cultural boundaries. Its strongest contribution is an auditable framework for distinguishing database organization from biocultural interpretation and for identifying comparative questions that require field validation. Country-level patterns should therefore guide carefully governed follow-up research rather than substitute for collaboration with knowledge holders.

Keywords: biocultural diversity; ethnobotanical database; knowledge networks; spatial autocorrelation; spatial cross-validation; useful plants

Background

Traditional ecological knowledge is relational: it emerges through sustained engagements among people, plants, landscapes, languages, livelihoods, and institutions. Ethnobotanical inventories capture parts of these relationships, yet analytical summaries commonly detach records from their spatial and documentary contexts. Counts of useful species or medicinal reports are frequently mapped as if they were direct and comparable measures of knowledge. That interpretation is rarely warranted because sampling effort, publication history, taxonomic coverage, political boundaries, and ecological availability all influence the resulting pattern. Quantitative ethnobotany is strongest when it makes those relationships and uncertainties explicit rather than treating a database as a neutral census (Albuquerque & Hanazaki 2009; Gaoue *et al.* 2017).

Biocultural diversity provides a useful conceptual frame because biological, linguistic, and cultural variation are often geographically associated without being reducible to one another (Maffi 2005; Loh & Harmon 2005; Gorenflo *et al.* 2012). Recent work has shown that much medicinal knowledge is linguistically unique and that knowledge networks can reveal vulnerability and complementarity that simple richness counts conceal (Cámara-Leret *et al.* 2019; Cámara-Leret & Bascompte 2021). The new Forest of Knowledge synthesis further demonstrates the scale of documented Amazonian plant-use diversity and the urgency of distinguishing cultural knowledge from the ecological distributions of useful species (Cámara-Leret *et al.* 2026). These studies motivate spatial inference, but they also warn against reading cultural processes directly from geographical coincidence.

Spatial statistics can test whether neighboring units resemble one another beyond descriptive appearance. Global Moran's I measures overall spatial dependence, while Local Indicators of Spatial Association identify local concentrations and spatial outliers (Anselin 1995). Clustering and network analysis ask different questions. Clustering summarizes multivariate profiles, whereas networks represent direct similarities or shared taxa and can reveal modules that cross administrative borders (Newman 2006; Blondel *et al.* 2008). Agreement among these representations is therefore an empirical result, not an assumption.

Spatio-temporal methods require particular caution in ethnobotany. Publication year is usually the date of documentation, not the date a practice began, changed, or disappeared. ST-DBSCAN identifies observations that are proximate in both geographic and temporal dimensions (Birant & Kut 2007). More recently, Idris *et al.* (2026) developed Spatio-Temporal Spatially Constrained K-Means (ST-SCKM), which extends centroid-based clustering by jointly incorporating spatial and temporal structure together with a K-nearest-neighbor soft constraint. Here, both methods are used diagnostically to test whether the spatial and bibliographic dimensions of the SEPASAL records contribute meaningful structure beyond conventional feature-based clustering. Bibliographic year is interpreted strictly as documentation time rather than evidence of temporal change in traditional ecological knowledge.

This study asks four questions. First, are country-level documented useful-plant repertoires spatially structured after a simple adjustment for documentation intensity? Second, do alternative clustering and network representations reveal stable regional structure? Third, does useful-plant similarity decline with geographic distance? Fourth, do spatio-temporal clustering and spatially blocked prediction add information beyond conventional summaries? The contribution is a transparent, reproducible workflow and an empirical assessment of where its inferences succeed or fail.

Recent syntheses show that ethnobotanical spatial research remains dominated by descriptive mapping and that the broader integration of statistics and data science is still uneven (Idris *et al.* 2026; Limba *et al.* 2026). Comparative work involving Bussmann further demonstrates why spatial proximity cannot be treated as a direct proxy for shared knowledge: documented plant repertoires may differ markedly between nearby markets or districts because of population histories, resource availability, and the scale of observation (Bussmann *et al.* 2018; Kunwar *et al.* 2018).

In this study, “detecting biocultural knowledge systems” therefore means detecting reproducible candidate structure in country-level documentation. It does not mean locating communities, assigning cultural boundaries, or validating knowledge independently of the people who hold it. Figure 1 makes this evidential boundary explicit before the analytical results are presented.



Figure 1. Evidential structure of the analysis. Database entities are converted into explicit analytical representations, while the lower boundary separates supported database-level inference from unsupported community or cultural claims.

Materials and Methods

Study design and data discovery

We conducted a structured search of Zenodo, Dryad, Figshare, institutional archives, and journal supplements for datasets containing plant names, use categories, cultural or locality identifiers, geography, time, licensing, and reproducible files. The candidate inventory is retained in metadata/source_inventory.csv. The Forest of Knowledge was the scientifically richest candidate but its approximately 11 GB archive was unnecessary once a smaller licensed dataset could support an end-to-end empirical run. Native American Ethnobotany contained stronger cultural identifiers but lacked defensible coordinates and had uncertain upstream licensing. Rizoma & Ewé offered valuable regional data but a narrower spatial basis (Souza & Hawkins 2020; Cordero *et al.* 2022).

We selected SEPASAL, a static Royal Botanic Gardens, Kew archive released under CC BY 4.0. The database collates useful wild and semi-domesticated plants of tropical and subtropical drylands and uses the TDWG Economic Botany Data Collection Standard. We downloaded the taxa, distribution, major-use, referenced specific-use, and bibliography tables and stored SHA-256 checksums. Natural Earth 1:110m country polygons supplied geographic context and representative points. No new human-participant data were collected.

Figure 2 summarizes the complete data-construction sequence from source acquisition to claim verification. The workflow identifies where TaxKey linkage, ISO-country harmonization, deduplication, representation building, and statistical validation alter the analytical object. Its audit points make the distinction between reproducible database processing and ethnobotanical interpretation visible before any model result is considered.

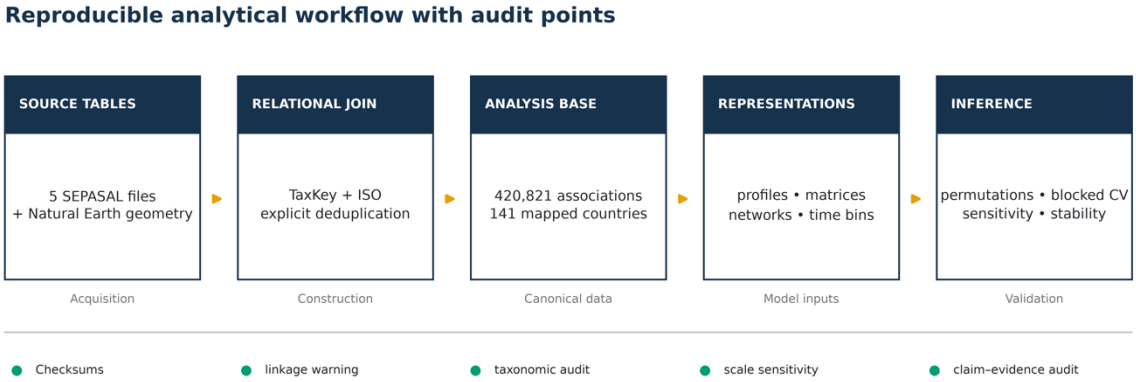


Figure 2. Reproducible data construction and analysis workflow. Audit points remain active from source acquisition through claim verification.

The sequence in Figure 2 makes clear that the analytical country profile is produced only after relational expansion and explicit deduplication. Checksums protect source identity, whereas linkage and taxonomic audits protect the meaning of the constructed associations. Sensitivity and claim–evidence audits then test whether numerical stability is being mistaken for ethnobotanical validity.

Analysis unit and relational interpretation

The country was the only spatial unit supportable without invented geocoding. We joined major-use and distribution tables by SEPASAL TaxKey, retained distribution rows with ISO country codes, and deduplicated each country–taxon–major-use–subuse combination. The resulting association means that a use is documented for a taxon whose SEPASAL distribution includes that country. It does not establish that the use was elicited in that country or practiced by every community within it. We therefore use the term documented useful-plant repertoire, not country knowledge inventory.

We decomposed the join as the sum of the number of retained ISO countries per distinct source TaxKey, major-use and subuse record. The distribution fields do not systematically identify a primary locality of use, and a botanical type locality would locate a specimen rather than a documented practice. We therefore repeated the country analysis after excluding taxa linked to more than 25 or 50 countries. Each test refitted the effort regression and two-cluster model and rebuilt the eight-neighbor graph and 90th-percentile Jaccard network for the countries with retained records. Partition agreement was evaluated against the original labels on those same countries.

This distinction changes the scientific target. Spatial pattern may arise from plant biogeography, the regional history of ethnobotanical documentation, shared use classifications, or genuine cultural continuity. The analysis cannot separate these mechanisms. Country units also combine diverse nations, languages, and ecosystems, so clusters are analytical summaries rather than biocultural territories.

Data cleaning and taxonomic handling

We normalized two-letter country codes, removed rows without country codes from spatial analysis, constructed binomial name strings from the supplied genus and specific epithet, simplified historical family labels, and retained SEPASAL TaxKey as the stable taxonomic identifier. Names linked internally to accepted TaxKeys were flagged, but uncertain names were not silently replaced. A complete crosswalk records the original name, accepted-link status, match method, and confidence. This conservative procedure avoids falsely claiming contemporary taxonomic resolution.

Ethnobotanical indicators and effort adjustment

For each country we calculated TaxKey richness, family richness, major-use richness, association counts, major-use category proportions, and Shannon diversity. We used residuals from ordinary least squares of log(1 + TaxKey richness) on log(1 + association count) as a common, interpretable archive-level benchmark. Association count is generated by the taxon-to-country join and serves as a proxy for record intensity. We compared the baseline with a cubic spline (four quantile knots), a common slope with continent intercepts, and an adjustment including the mean log number of countries per taxon in each repertoire. The sensitivity fits address functional form, mean continental shifts, and range breadth, respectively. Independent measures of collection effort and source coverage would be needed to assess residual documentation bias.

For Figure 16, we added 50 pseudo-observations distributed in proportion to the archive-wide major-use shares before computing each country profile and its log2 ratio to the archive-wide composition. We masked cells when a country had fewer than 100 associations, a cell had fewer than five associations, or the sign changed when the prior weight varied from 25 to 100. These display thresholds mark low support in the archive and do not estimate sampling uncertainty. We clipped displayed ratios at ±2.5 and ordered the shrunk profiles by average-linkage clustering within continents. These settings apply only to the display.

Spatial weights and autocorrelation

Country polygons were projected to Robinson coordinates and representative interior points were transformed back to longitude and latitude. Great-circle distances among 141 points were used to build symmetric k-nearest-neighbor graphs for $k = 4, 6, 8$, and 10. We selected $k = 8$ as the primary compromise between local connectivity and global smoothing. Global Moran's I was estimated for richness, use diversity, effort-adjusted richness, and documentation intensity. Two-sided permutation inference used 9,999 random permutations. Local Moran statistics were computed for effort-adjusted richness and p-values were corrected across countries using the Benjamini–Hochberg false-discovery-rate procedure.

Feature representation and baseline clustering

The clustering matrix combined standardized richness, diversity, effort-adjusted richness, reports per species, and proportions of 13 major-use categories. Principal component analysis supported ordination and sensitivity checking (Jolliffe and Cadima 2016). We compared K-means ($k = 2 - 8$), Ward hierarchical clustering, and feature-space DBSCAN. K-means used 50 initializations and deterministic seed 42, following the least-squares partitioning formulation of Lloyd (1982). Candidate models were evaluated with silhouette, Davies–Bouldin, Calinski–Harabasz, coverage, cluster balance, and spatial compactness. We rejected K-means solutions whose smallest-to-largest cluster-size ratio was below 0.10, then selected the best remaining multi-metric rank. Stability was assessed with 100 independent 80% subsamples using adjusted Rand index and normalized mutual information (Hubert & Arabie 1985).

For the extended K-means scan, we recorded within-cluster sum of squares for k = 1 to 12 and computed silhouette, cluster sizes, and stability for k = 2 to 12. Each candidate was compared with refits on the same 100 subsamples of 80% of countries. The size-ratio rule was held fixed to assess whether a finer solution remained admissible.

Documentation time and spatio-temporal clustering

We parsed the earliest four-digit year from bibliographic year strings, restricted years to 1500–2026, and aggregated taxon-use references into 25-year country bins. Country-period units with at least five linked records were retained. These units represent when information entered the documented bibliography and should not be interpreted as dates of knowledge emergence, transmission, or disappearance.

ST-DBSCAN was implemented using geographic and documentation-time distances, with haversine distance for the spatial component. We evaluated spatial thresholds of 500–4,000 km, temporal thresholds of 25–100 years, and minimum-sample values of 3, 5, and 8. Native noise classifications and coverage were retained for all configurations.

We used an augmented K-means diagnostic inspired by ST-SCKM (Idris et al. 2026). The archive implementation concatenates standardized log-transformed reference-link counts and taxon richness with standardized projected coordinates and median bibliography year, scaling the spatial and temporal columns by the square roots of their respective weights. The wildfire implementation uses a spatial neighbor penalty and profiles contextual variables after clustering. The two implementations therefore represent different objectives.

We scanned k = 3 to 6 and spatial and temporal weights of 0, 0.10, 0.25, 0.50 and 1. For the bin-width diagnostic, we rebuilt country-period units from the referenced records using 10-, 25- and 50-year bins and retained the same five-link inclusion rule. We compared ST-DBSCAN feature silhouettes at full coverage and the best feature silhouette with zero versus positive spatial or temporal weights. This comparison isolates feature separation from the compactness criteria used in the original scoring.

Similarity and network analysis

A binary country-by-TaxKey matrix supported pairwise Jaccard similarity. We related log geographic distance to repertoire dissimilarity using a Mantel-style permutation correlation with 4,999 permutations and a descriptive pairwise regression including same-continent status. Because pairwise observations are non-independent and Mantel tests can be limited, neither analysis was used alone to support mechanism. Cross-cultural similarity has substantive meaning only when availability and reporting are considered (Saslis-Lagoudakis *et al.* 2011; Kunwar *et al.* 2018).

We also estimated a descriptive pairwise model including the mean and absolute difference in log association count for each pair. We computed the partial Pearson correlation by residualizing both log distance and Jaccard dissimilarity against these effort terms and same-continent status. The same country appears in many dyads, so the adjusted coefficients and correlation are reported without ordinary regression tests. Independent plant-availability data would be required to condition on floristic overlap without reusing the response incidence matrix.

The bipartite network connected countries to the 250 most widespread useful taxa, permitting degree and approximate betweenness summaries. A separate country-similarity network connected pairs at or above the empirical 90th percentile of off-diagonal Jaccard similarity. Greedy modularity communities were compared with K-means using adjusted Rand index and normalized mutual information. Network modules were not equated with cultural groups.

For the similarity network, we varied the empirical cutoff across the 80th, 85th, 90th and 95th percentiles and compared weighted greedy modularity and Louvain at resolution 1. At the 90th percentile we also varied Louvain resolution from 0.5

to 2 and scored all partitions at the common resolution of 1 for comparison. Isolates were retained in module counts. In the bipartite network, we compared country degree ranks among the top 100, 250, 500 and all mapped taxa using Spearman correlations. The 250-taxon filter limits visual and computational complexity and explicitly selects high-prevalence plants.

GeoAI and spatial validation

A 500-tree random forest predicted effort-adjusted richness from country coordinates, documentation intensity, use diversity, and use-category proportions (Breiman 2001). Random five-fold predictions were compared with leave-continent-out predictions to expose geographic leakage, consistent with recommendations for structured cross-validation (Roberts *et al.* 2017; Valavi *et al.* 2019). We reported R^2, mean absolute error, root mean squared error, permutation importance, and Moran's I of spatial-CV residuals. Feature importance was interpreted descriptively, not causally. GeoAI was included only to test predictive transportability, not to produce a priority map.

We repeated random five-fold validation over 20 split seeds and report the observed range of pooled R^2. For the five continent-grouped folds, we report held-out fold sizes, R^2 relative to each fold mean, and mean absolute error. This measures sensitivity to region and split in the available archive (Mushagalusa et al. 2024). The target was defined by a single archive-wide effort regression prior to validation, so this exercise evaluates prediction of the constructed residual and does not validate the end-to-end workflow for new regions.

Reproducibility ethics and governance

All processing is scripted; raw files are checksummed; parameters are stored in config.yaml; random seed 42 is fixed; and manuscript values are read from outputs/statistical_report.json. These practices support FAIR principles of findability, accessibility, interoperability, and reuse (Wilkinson *et al.* 2016). FAIR does not determine whether data should be open. CARE principles require collective benefit, authority to control, responsibility, and ethics in the use of Indigenous data (Carroll *et al.* 2020). This study therefore avoids fine-scale mapping, does not infer sacred or restricted knowledge, and treats future community-level work as requiring governance and validation by knowledge holders.

For future community-governed data, the unit would be a permissioned observation linked to place, observation date, language or community identifier where authorized, and a record of consent and permitted uses (Pankararu et al. 2026). Authorized data stewards would decide which records could be combined and analyzed. The analysis would use use-report locations and survey effort, and would validate by site or community according to the stated target. Before release, stewards would review maps, model outputs and tables for disclosure through coordinates, small cells or combinations of identifiers and approve the level of aggregation or withhold the output. The research agreement would identify collective benefits and set procedures for access requests, correction and withdrawal.

Results

Dataset characteristics and geographic coverage

The processed dataset contained 420,821 country–taxon–use associations, 6,428 TaxKeys, 6,359 distinct supplied species strings, 208 families, and 13 major-use categories across 141 mapped countries. Median mapped-country richness was 212 taxa. The major-use CSV contains 28,219 data rows, excluding its header. Of these, 410 did not link to a retained ISO country. The 27,809 linked records generated a mean of 15.13 country associations each. Taxa recorded in at least 10 countries contributed 368,574 associations (87.6%), and those recorded in at least 25 contributed 265,424 (63.1%). The expansion is uneven across taxa and occurs before the restriction to mapped countries.

Table 1. Descriptive characteristics of the analytical dataset.

Metric	Value
Raw Taxa	17,853
Canonical Country Taxon Use Associations	420,821
Countries	141
Unique Taxa	6,428
Unique Species Strings	6,359
Families	208
Use Categories	13
Median Species Richness Per Country	212
Documentation Year Min	1,800
Documentation Year Max	2,024

Table 1 shows that the canonical analytical base is much larger than the raw taxon table because each documented taxon use can be propagated across multiple recorded distribution countries. The 6,428 unique TaxKeys and 6,359 supplied species strings also reveal a small but consequential distinction between database identifiers and textual names. These totals define the scope of the analysis and should not be interpreted as 420,821 independent field observations.

Figure 3 maps the geographic concentration of the 420,821 constructed country–taxon–use associations and identifies the countries contributing the largest totals. Panel A displays the global coverage gradient, whereas Panel B ranks the twelve highest country totals on a common scale. Reading the panels together shows whether large national counts form a broad regional pattern or arise as isolated database concentrations.

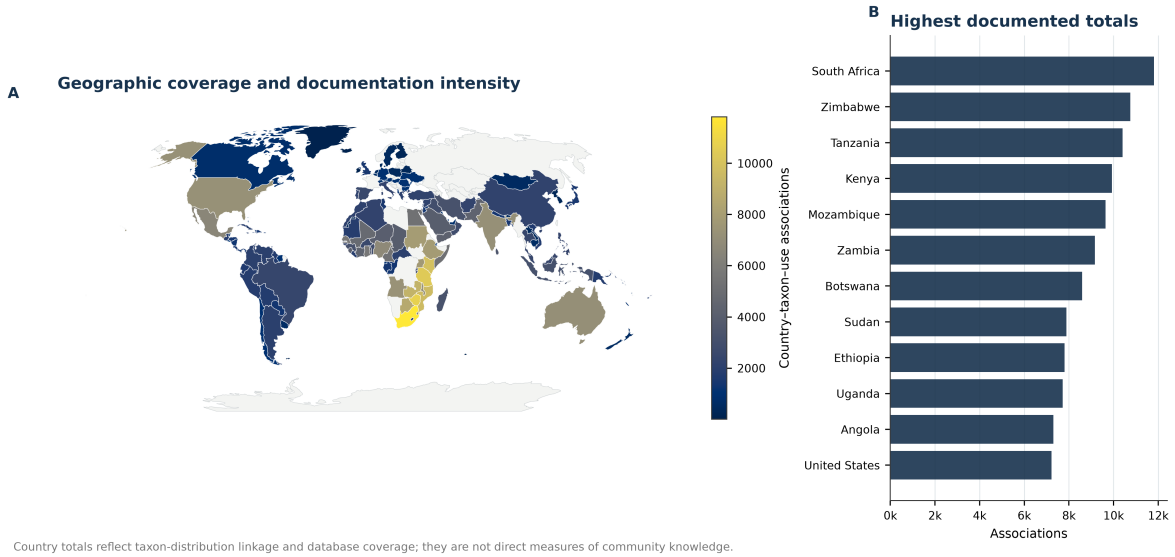


Figure 3. Geographic coverage and documentation intensity, with the countries having the largest constructed association totals.

The ranking is strongly concentrated in eastern and southern Africa: South Africa contributed 11,812 associations, followed by Zimbabwe with 10,748 and Tanzania with 10,407. Eleven of the twelve highest totals were African countries, with the United States as the sole exception in that group. This distribution most directly reflects the dryland focus, taxon-distribution expansion, and uneven database coverage, rather than a continental ranking of knowledge abundance.

Figure 4 separates the geography of effort-adjusted richness from the fitted relationship between documented species richness and association count. Panel A maps the residual component, so warm and cool colors indicate countries with respectively more or fewer documented taxa than expected from database intensity. Panel B exposes the fitted trend and the country-level departures that generate that residual map.

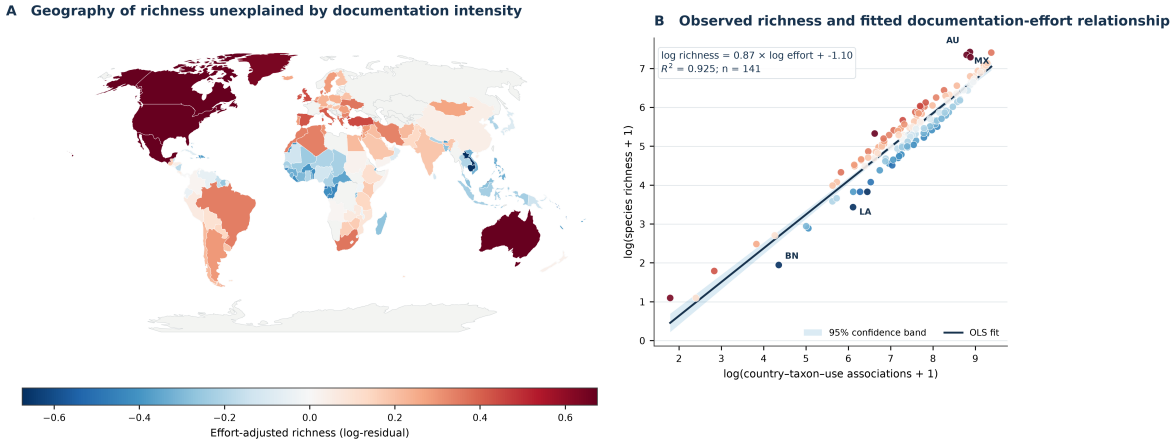


Figure 4. Effort-adjusted documented useful-plant richness and its fitted relationship with documentation intensity.

The largest positive effort-adjusted residuals occurred in Australia and Mexico, both above 0.80, followed by Canada and the United States at approximately 0.67. Laos, Brunei, and Kampuchea formed the most negative residual tail, indicating much lower documented richness than predicted by their association totals. Extreme residuals from very small totals, such as Greenland, require particular caution because a regression residual can be numerically large even when the underlying evidence base is sparse.

Global and local spatial dependence

Effort-adjusted richness showed strong positive global spatial autocorrelation under the primary symmetric 8-nearest-neighbour graph (I = 0.497; expected I = -0.007; permutation z = 13.657; p = 0.0001; 9,999 permutations). Moran's I remained positive at k = 4, 6, 8, and 10, declining gradually from 0.554 to 0.468 as neighborhoods broadened. Raw richness, use diversity, and documentation intensity were also positively autocorrelated. The concordance across raw and adjusted indicators confirms spatial structure in the database, but it does not identify whether biogeography or cultural continuity is the principal driver.

Global Moran's I remained positive after fitting the range-breadth covariate (0.554), spline (0.522), and continent-intercept adjustment (0.331). Each two-sided permutation p-value was 0.0001. The decline with continent intercepts indicates that broad regional means contribute to baseline spatial dependence. Residual correlations with the primary model were 0.762, 0.950 and 0.848, respectively. These checks address measured archive features and leave historical documentation effects unresolved.

After FDR correction, Local Moran analysis classified 11 countries as high-high, 17 as low-low, 1 as high-low, and 1 as low-high; 111 were not significant. High-high units were concentrated mainly in Europe, with smaller representation in Asia and North America. Low-low units were concentrated in Africa and Asia. These categories identify database-relative spatial association, not knowledge superiority or deficiency.

Capping distribution breadth at 25 and 50 countries retained 163,847 and 297,208 associations and 139 and 140 mapped units, respectively. Moran's I was 0.507 and 0.551, compared with 0.494 and 0.497 for the original residuals on the same two country sets. Two-cluster agreement with the original labels fell to ARI 0.444 and 0.486. The recalculated networks returned 14 and 18 modules (ARI 0.809 and 0.840 versus the primary network). The positive richness dependence persisted, while the membership of the centroid partition was sensitive to widespread taxa.

Table 2 reports the primary Global Moran estimates under symmetric eight-nearest-neighbor weights and provides the numerical basis for the spatial-dependence interpretation. Effort-adjusted richness produced the largest coefficient ($I = 0.497$), followed by association count ($I = 0.452$), species richness ($I = 0.263$), and use diversity ($I = 0.180$). All four permutation probabilities were 0.0003 or smaller, indicating that the observed positive spatial structure was unlikely under the corresponding randomization procedure.

Table 2. Global Moran results under the primary symmetric eight-nearest-neighbor weights.

Variable	Moran I	P Permutation	Interpretation
Species Richness	0.2628	0.0001	positive spatial dependence
Use Shannon	0.1802	0.0003	positive spatial dependence
Effort Adjusted Richness	0.4970	0.0001	positive spatial dependence
N Associations	0.4522	0.0001	positive spatial dependence

The ordering of the coefficients shows that adjusting richness for association count did not remove spatial structure; it strengthened the spatial signal relative to raw richness and use diversity. Positive dependence across all four indicators supports a regional organization of the assembled database, but the statistic does not identify whether floristic similarity, shared documentation practices, or knowledge circulation produced that organization. The multi-scale comparison in Figure 5 therefore tests whether this conclusion depends on the selected eight-neighbor graph.

Figure 5 compares global spatial-autocorrelation estimates across four outcomes and four neighborhood definitions. Panel A tests whether the direction and ordering of the coefficients persist as k increases from 4 to 10, while Panel B visualizes the

primary effort-adjusted result as the slope of the Moran scatterplot. This combination separates sensitivity to the spatial-weight definition from the country-level pattern underlying the selected coefficient.

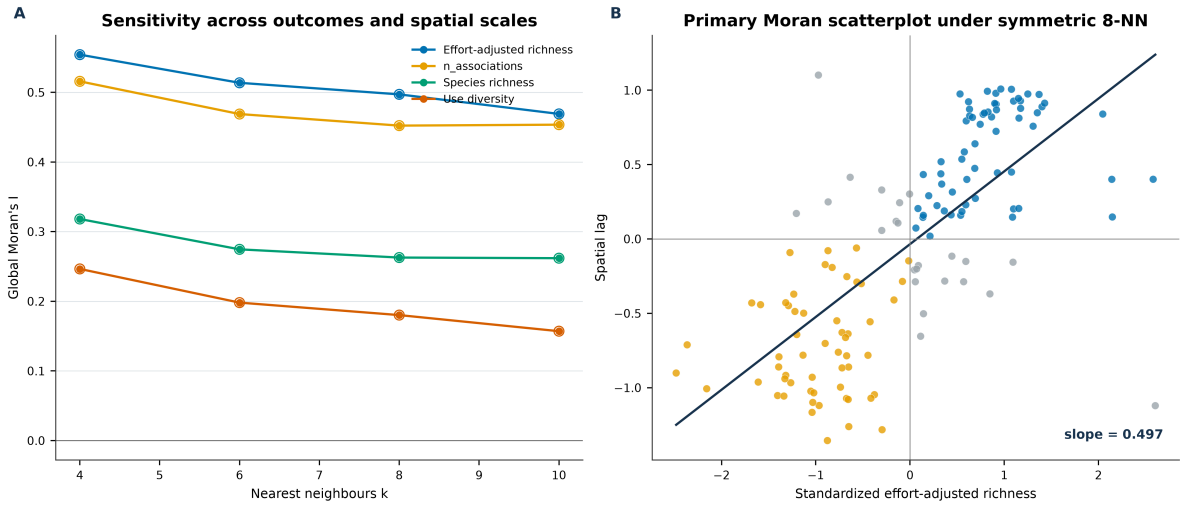


Figure 5. Global spatial autocorrelation across outcomes and neighborhood sizes, and the primary Moran scatterplot for effort-adjusted richness.

Effort-adjusted richness remained the most spatially structured outcome at every tested neighborhood size, declining from $I = 0.554$ at $k = 4$ to $I = 0.469$ at $k = 10$. Association count, species richness, and use diversity also remained positive, although their coefficients weakened or stabilized as broader neighborhoods smoothed local contrasts. The quadrant structure in Panel B shows that the global coefficient arises mainly from countries with concordant positive or negative values and neighboring lags, with a smaller set of spatially discordant observations.

Figure 6 localizes the statistically supported high–high, low–low, and spatial-outlier patterns after false-discovery-rate correction. Panel A shows where each Local Moran class occurs, and Panel B makes the strong imbalance between significant and non-significant countries explicit. The figure therefore identifies spatial concentrations that merit interpretation without visually overstating the sparse outlier classes.

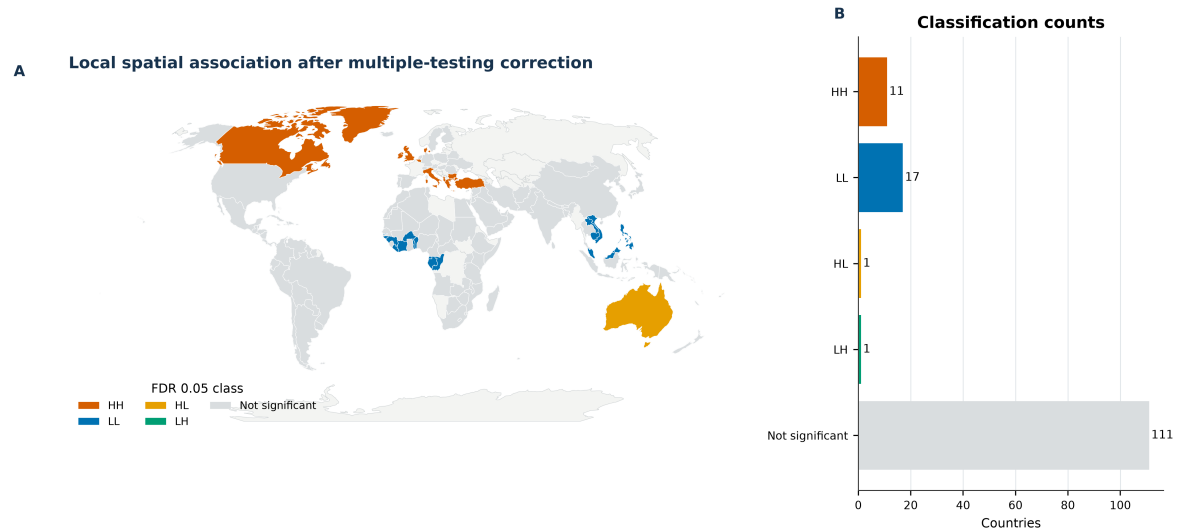


Figure 6. Local Moran classification after false-discovery-rate correction and the corresponding country counts.

After correction, 11 countries were classified high–high and 17 low–low, whereas only one high–low and one low–high outlier remained. Eight of the high–high countries were in Europe, while all low–low countries occurred in Africa or Asia.

Because 111 countries were not significant, the mapped classes should be read as localized database-relative concentrations rather than a complete typology of national ethnobotanical systems.

Ordination and baseline clustering

Principal-component ordination showed broad overlap among continents with a small number of extreme country profiles. Ward $k = 6$ achieved the highest composite rank but was severely imbalanced: its smallest cluster was approximately 1.5% of its largest. Applying the prespecified balance requirement retained K-means $k = 2$ as the interpretable baseline. The two clusters contained 87 and 54 countries. Their median taxon richness was 232 and 174, respectively, while median effort-adjusted richness had opposite signs (-0.206 and 0.276). The partition is therefore driven by composition and the relationship between richness and documentation, not by a simple high-versus-low knowledge ranking.

The retained partition was stable under repeated subsampling: median adjusted Rand index was 0.964, with interquartile range 0.027–1.000. Stability indicates numerical reproducibility under record removal but does not establish cultural reality. It supports retaining the partition as a descriptive baseline while requiring independent ethnobotanical evidence for any interpretation beyond database profiles.

Inertia for k = 1 to 12 was 2679, 2030, 1590, 1345, 1167, 1049, 901, 806, 734, 684, 642, 601. Median subsample ARI for k = 2 to 12 was 0.964, 1.000, 1.000, 0.981, 0.833, 0.714, 0.809, 0.800, 0.700, 0.744, 0.634. The inertia curve declined progressively and showed no unique break at two. For k = 3 to 12 the smallest class contained one or two countries, so none passed the 0.10 size-ratio rule. The two-cluster partition is a balanced summary at the chosen feature scale and can obscure regional subgroups.

Figure 7 displays the country ordination and the variable loadings that define its principal compositional gradient. Panel A shows the broad overlap among continental point clouds together with a small number of extreme profiles, whereas Panel B identifies the variables contributing most strongly to the first component. The figure is used to interpret multivariate structure, not to claim that continent labels constitute discrete cultural groups.

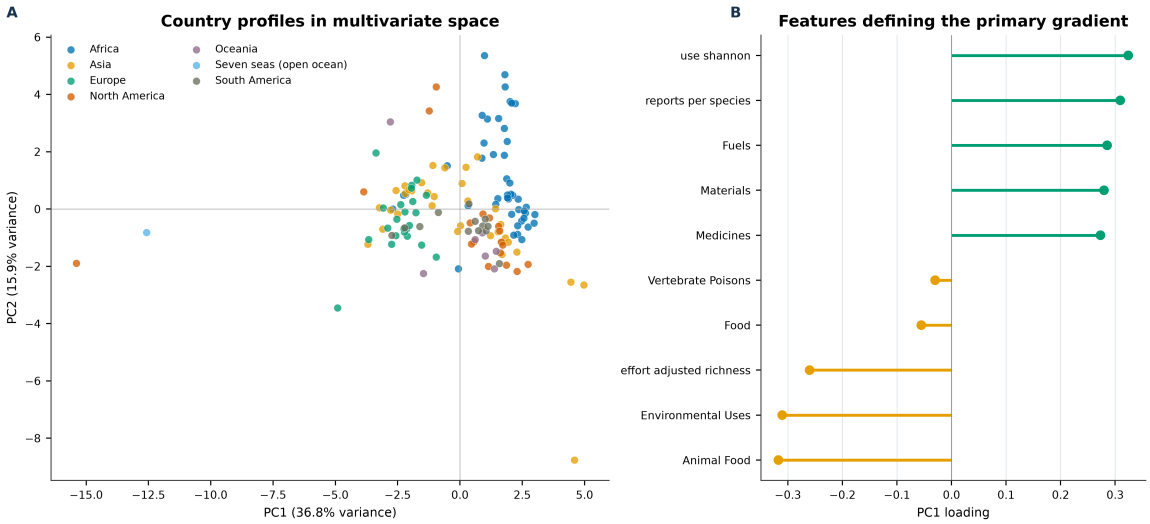


Figure 7. Principal-component ordination of country profiles and the variables defining the primary multivariate gradient.

The first component explained 36.8% of total profile variance and contrasted positive loadings for use diversity, reports per species, fuels, materials, and medicines with negative loadings for animal food, environmental uses, and effort-adjusted richness. Continental colors overlap across much of the ordination, showing that the principal compositional gradient is not reducible to continental membership. The isolated points at the margins warrant record-level inspection because leverage can reflect distinctive documentation profiles as well as genuine compositional differences.

Figure 8 maps the retained two-cluster solution and compares its standardized profiles. The map shows the geographic interspersion of the classes, while the profile panel reveals that their separation depends on combinations of richness, diversity, effort adjustment, and reporting intensity. These classes summarize database composition and must not be interpreted as cultural regions or ranked levels of ethnobotanical knowledge.

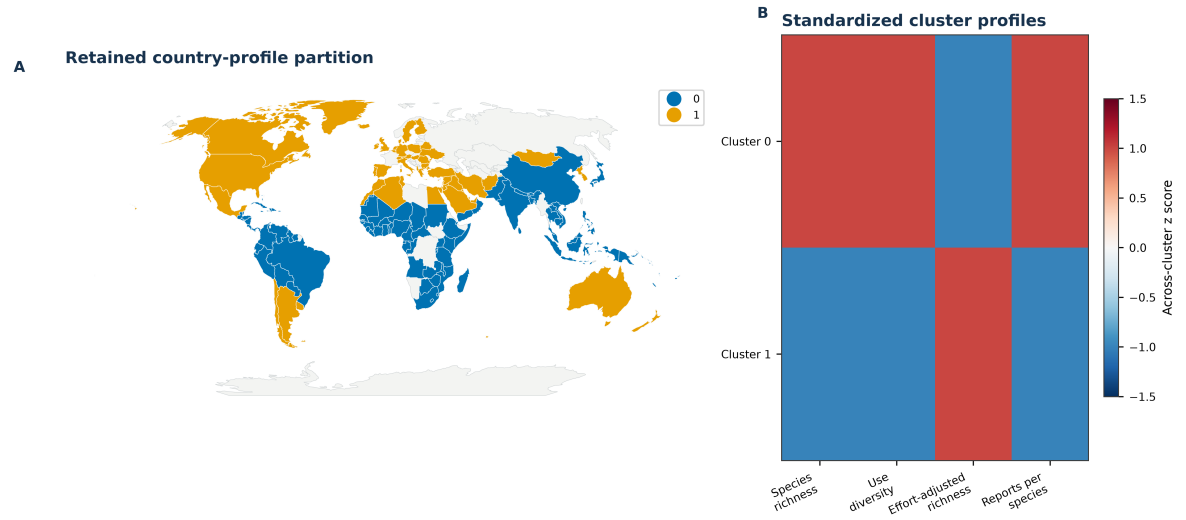


Figure 8. Retained two-cluster K-means partition and standardized cluster profiles. The classes are analytical summaries, not cultural regions.

The retained classes contained 87 and 54 countries and were numerically stable under repeated 80% subsampling, with a median adjusted Rand index of 0.964. Their standardized profiles differ in sign across effort-adjusted richness and the remaining summary variables, demonstrating that the partition is multivariate rather than a simple richness split. The geographic mixture within each color class reinforces the decision to treat the clusters as analytical summaries rather than mapped biocultural regions.

Spatio-temporal documentation diagnostics

The ST-DBSCAN grid did not reveal a defensible feature-coherent solution. The best complete-coverage configuration used a 4,000 km spatial threshold, 100-year temporal threshold, and minimum samples of 3, producing two clusters with silhouette -0.048 and Davies–Bouldin 6.876. The negative silhouette indicates that country-period documentation units were, on average, closer in feature space to another cluster than their assigned cluster. We therefore retain the output as a failed diagnostic rather than a substantive result.

The ST-SCKM parameter search selected $k = 3$, while the best-performing configuration assigned no additional spatial or documentation-time weighting. The corresponding feature-space silhouette was 0.511. Positive spatial or temporal regularization did not improve the ethnobotanical coherence of the partition and progressively reduced feature-space separation at larger weights. Thus, although ST-SCKM was originally developed to favor spatially and temporally coherent partitions (Idris *et al.* 2026), the present SEPASAL representation did not support additional constraint beyond the unconstrained feature structure.



Figure 9. Selected ST-DBSCAN assignments and the parameter trade-off showing negative feature-space coherence.

The best complete-coverage ST-DBSCAN configuration still produced a negative feature-space silhouette of -0.048 and a Davies–Bouldin value of 6.876. Panel B shows that configurations approaching complete coverage did not cross into positive silhouette space, so expanding spatial or temporal neighborhoods did not recover coherent documentation clusters. The result is retained as evidence against this representation and parameter family, not replaced by a more attractive unsupported partition.

Figure 10 illustrates why ST-SCKM does not provide additional spatial-temporal structure for this database representation. Panel A shows that the selected partition remains primarily differentiated along the documentation-feature gradient. Panel B reveals a systematic decline in silhouette as spatial and temporal regularization increases. Unlike the wildfire application in which spatial-temporal constraints improved contiguous risk zoning (Idris *et al.* 2026), the SEPASAL country-period data contain much weaker evidence of a comparable joint structure. This contrast is expected because bibliography years represent the history of documentation rather than direct observations of a continuously evolving spatial process.

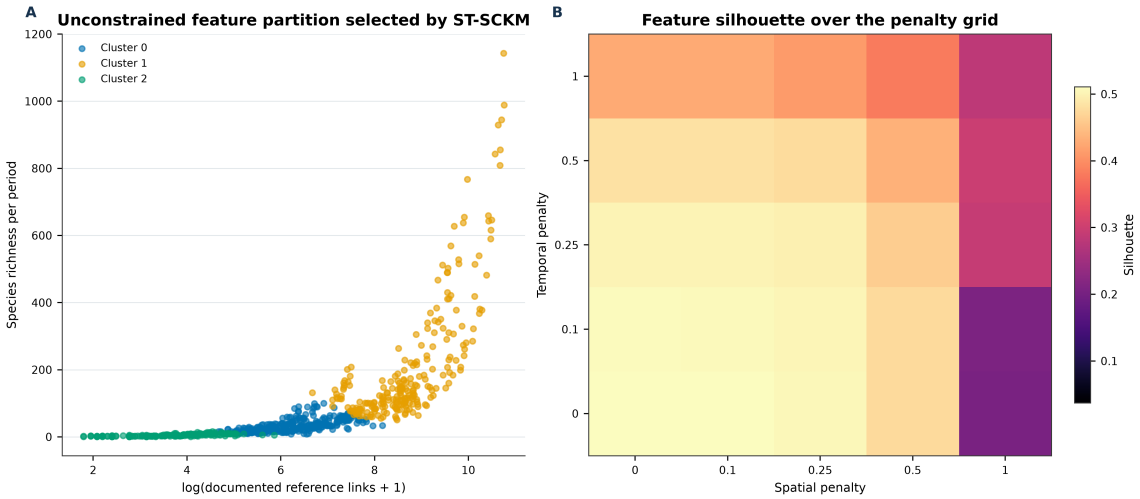


Figure 10. ST-SCKM partition and sensitivity to spatial-temporal regularization. (A) Feature-space representation of the partition associated with the selected ST-SCKM configuration. (B) Variation in feature-space silhouette across spatial and temporal penalty settings. The optimum occurred without additional spatial or documentation-time penalties, indicating that stronger constraints reduced rather than improved the coherence of documented useful-plant profiles.

The selected ST-SCKM solution used $k = 3$ and achieved a feature-space silhouette of 0.511 only when the additional spatial and temporal penalties were zero. Increasing either penalty progressively reduced separation, indicating that geographic proximity and bibliography time did not improve the country-period grouping. This negative comparison is ethnobotanically important because documentation year records scholarly entry into the archive rather than the time at which knowledge originated or circulated.

The 10-, 25- and 50-year bins retained 1,610, 772 and 526 country-period units. The respective best full-coverage ST-DBSCAN silhouettes were −0.060, −0.048 and −0.036. In the augmented K-means scan, the best zero-weight scores were 0.5193, 0.5110 and 0.5449. The best positive-weight scores were 0.5187, 0.5099 and 0.5454. A temporal weight of 0.10 gave the 50-year result, an improvement of less than 0.001. Bin width therefore affects the numerical scores but did not yield a material gain from adding spatial or temporal weights.

Similarity distance decay and network structure

Country useful-plant repertoire dissimilarity increased with log geographic distance in the supporting Mantel-style analysis ($r = 0.602$; permutation $p = 0.0002$; 4,999 permutations). A descriptive pairwise regression including same-continent status explained $R^2 = 0.379$; the same-continent coefficient was -0.042. The direction is consistent with distance decay, shared floras, and regional documentation, but the design cannot isolate cultural transmission from ecological availability. After including pairwise documentation intensity, the coefficient for log distance was 0.066 and the partial correlation was 0.461. The same-continent coefficient was -0.032 compared with −0.042 in the simpler model. On the unit Jaccard-dissimilarity scale this is about three hundredths of a point at the same distance and effort values, a modest difference among heterogeneous country profiles.

The country–plant bipartite network, restricted to the 250 most widespread useful taxa, contained 391 nodes and 11,957 edges. *Cynodon dactylon* had the highest plant degree (113 countries), followed by *Oxalis corniculata* (94) and *Cynosurus indicus* (91). India, South Africa, Kenya, and Tanzania had the highest country degrees within this deliberately prevalence-restricted network. Centrality reflects database breadth and distribution linkage, not cultural importance.

The 250 retained taxa accounted for 11,957 of 45,926 country-to-taxon edges in the full mapped network (26.0%). Country degree ranks for the 100-, 500- and all-taxon networks correlated 0.931, 0.972 and 0.935 with the 250-taxon rankings. These are country-rank checks. Plant degree and betweenness still depend on the prevalence screen and the sampled archive.

The country-similarity network used Jaccard threshold 0.217, corresponding to the 90th percentile of off-diagonal values. It contained 141 nodes, 987 edges, density 0.100, and 19 modules with modularity 0.636. Agreement between network modules and K-means was partial (adjusted Rand index = 0.346; normalized mutual information = 0.416). The mismatch is scientifically useful: thresholded pairwise similarity reveals regional substructure that a two-class global partition compresses. Figure 11 reveals the block structure and cross-continental exceptions in pairwise Jaccard similarity among documented country repertoires.

At the 80th, 85th, 90th and 95th percentile cutoffs, greedy modularity yielded 11, 12, 19 and 35 modules, including 5, 6, 11 and 22 isolates. Louvain at resolution 1 yielded 10, 13, 19 and 35 modules. At the primary cutoff Louvain increased modularity from 0.636 to 0.663, but agreement between algorithms was ARI 0.702. Louvain returned 18 to 21 modules over the tested resolutions. The stricter graph produces more isolates, so a higher modularity or module count does not demonstrate more distinct ethnobotanical groups.

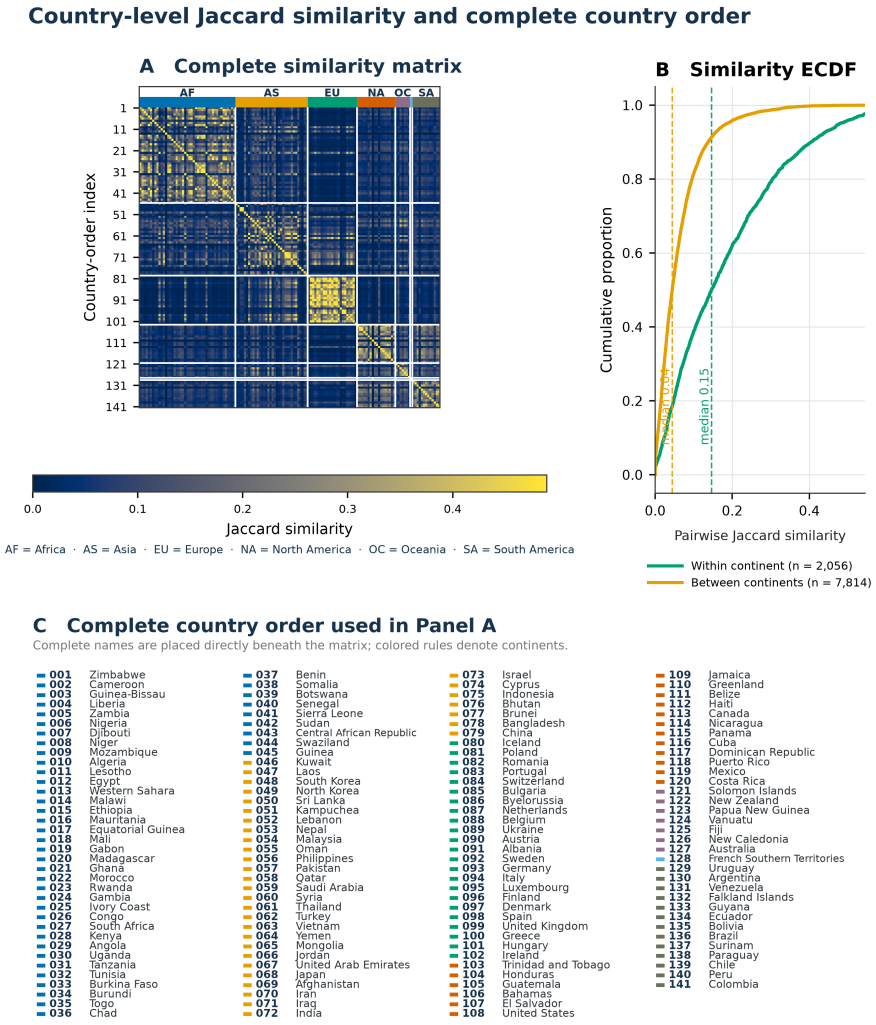


Figure 11. Pairwise Jaccard similarity of documented useful-plant repertoires, ordered within continents.

Country-level similarity of documented useful-plant repertoires. Panel A shows the complete 141-country Jaccard matrix ordered by continent, and Panel B compares empirical cumulative distributions for within-continent and between-continent pairs. Panel C is positioned immediately below the analytical panels and provides the complete, non-overlapping country order used by the numbered axes. All 141 country names are shown in full.

Continental summaries were themselves heterogeneous. Median within-continent similarity was highest in Europe (0.296), followed by North America (0.225), South America (0.204), and Africa (0.151), while Asia and Oceania were both approximately 0.087. These differences are compatible with variation in floristic overlap and documentation density, but they should not be read as rankings of cultural cohesion. The full country key in Figure 11 preserves the country-level units needed to inspect exceptions hidden by continental averages.

Figure 12 presents the bipartite country–plant network in a degree-ordered node-link layout. The display makes individual associations traceable, while the lower panel retains the complete network's degree structure and prevents the selected subnetwork from being mistaken for the full analysis. Reading the panels together distinguishes visually selected examples from the quantitative topology of the complete network.

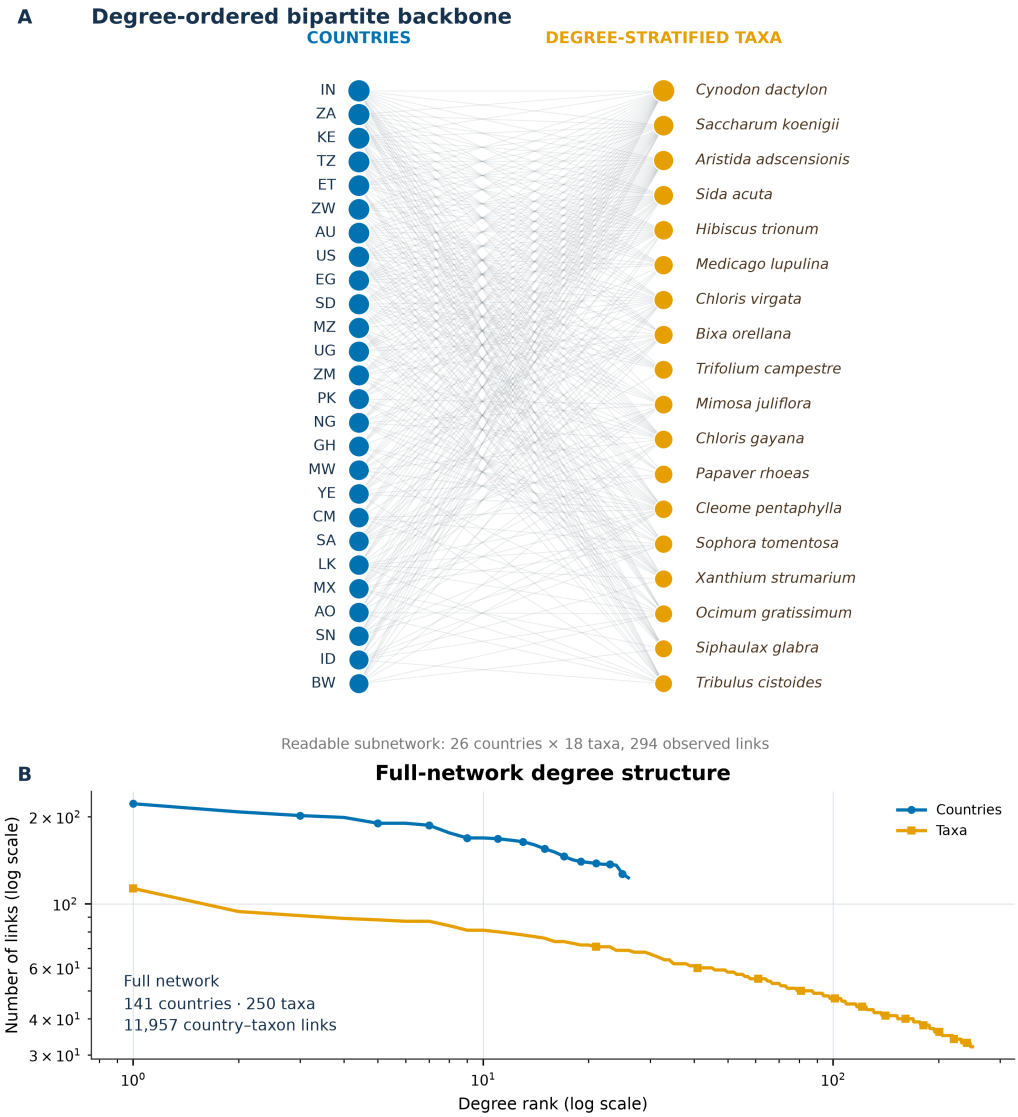


Figure 12. Country–plant bipartite network. Panel A shows a readable subnetwork of 26 high-degree countries and 18 taxa sampled across the degree gradient of the 250 retained widespread taxa; an edge denotes a documented country–taxon association. Panel B reports the complete degree-rank structure for all 141 countries, 250 taxa, and 11,957 links. Scientific plant names are italicized; node selection affects display only, not the network statistics.

Figure 13 places the strongest country-similarity ties in geographic context and summarizes the size distribution of the detected network modules. Panel A reveals where upper-tail Jaccard links form geographically concentrated or cross-border structures, and Panel B shows that three large modules dominate the partition. The joint display helps distinguish a globally modular network from the smaller isolates and dyads retained by thresholding.

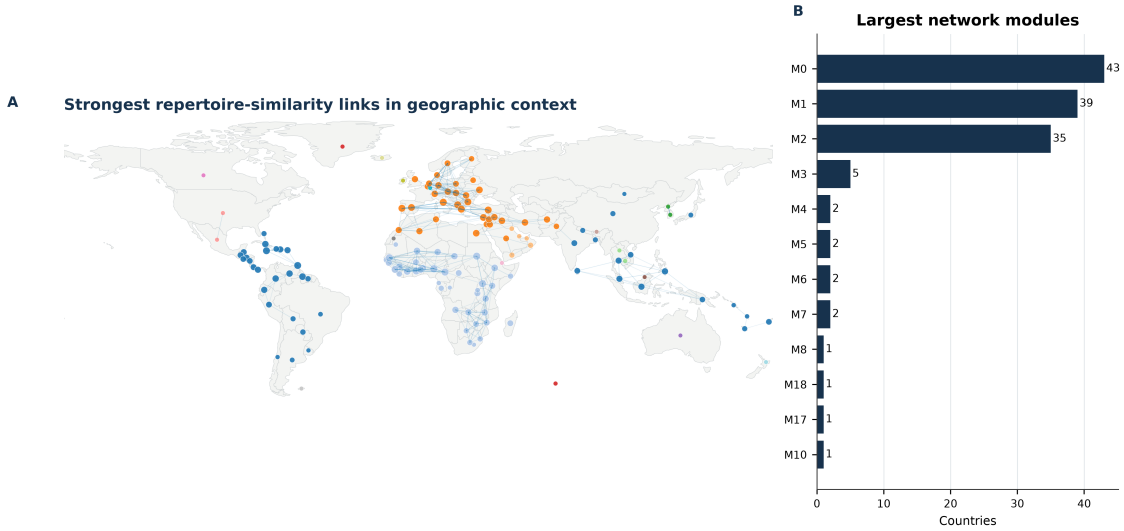


Figure 13. Strongest country-similarity links in geographic context and the sizes of the largest network modules.

The three largest network modules contained 43, 39, and 35 countries, whereas the remaining modules were much smaller. Geographic concentration within the large modules is compatible with shared flora, regional reporting histories, and cross-border circulation, but the map also retains long-distance and cross-continental similarities. Module membership is therefore most useful for selecting comparative cases and checking documentation pathways rather than assigning cultural identity.

GeoAI performance uncertainty and sensitivity

The original random five-fold split gave $R^2 = 0.685$, MAE = 0.111 and RMSE = 0.176. Continent-grouped validation gave pooled $R^2 = 0.587$, MAE = 0.136 and RMSE = 0.201. Across 20 random split seeds, pooled R^2 ranged from 0.639 to 0.709. The held-out fold sizes and R^2 were 45 and 0.596 for Africa, 34 and 0.411 for Asia, 23 and 0.083 for Europe, 19 and 0.635 for North America plus the open-ocean category, and 20 and 0.374 for Oceania plus South America. Fold MAE ranged from 0.111 to 0.172. These ranges describe variability across the observed splits. The original grouped-CV residuals remained positively autocorrelated (Moran's I = 0.198, permutation p = 0.0001).

Figure 14 brings the three principal sensitivity analyses together so robust findings can be distinguished from parameter-dependent ones. The panels compare ST-DBSCAN coverage, the ST-SCKM penalty response, and the neighborhood sensitivity of Global Moran's I using their relevant tuning axes. A result is treated as substantively reliable only when its interpretation survives these changes rather than depending on a single convenient parameter setting.

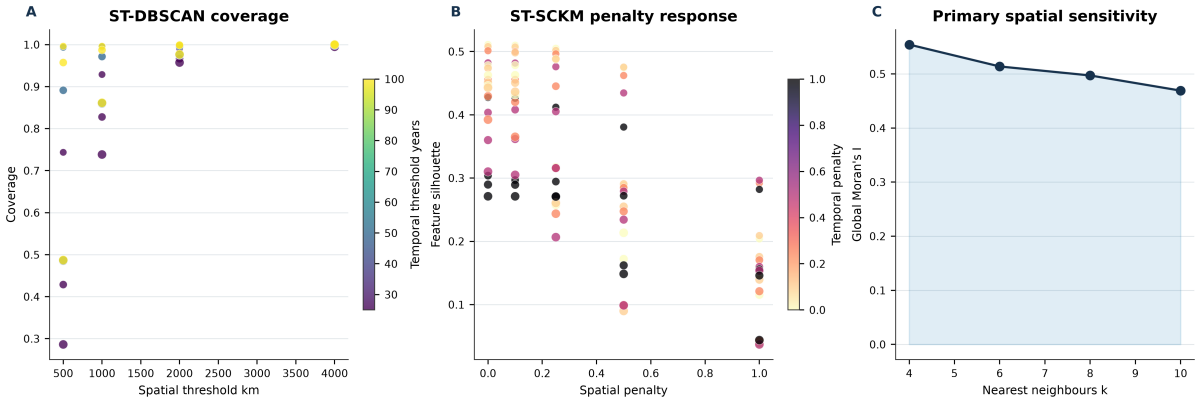


Figure 14. Joint sensitivity summary for ST-DBSCAN, constrained K-means, and the primary Global Moran result.

The sensitivity panels converge on two different conclusions: the Global Moran result remains positive as neighborhood size changes, whereas the spatio-temporal clustering results do not gain coherence under stronger constraints. ST-DBSCAN coverage can approach one while silhouette remains poor, and ST-SCKM separation declines as its penalties increase. Robust spatial dependence and failed spatio-temporal clustering can therefore coexist because they test different properties of the database.

Figure 15 compares effort-adjusted richness across combinations of K-means class and repertoire-distinctiveness tertile. The plot shows individual country values alongside the median and interquartile range in each group. Overlap between groups is visible directly.

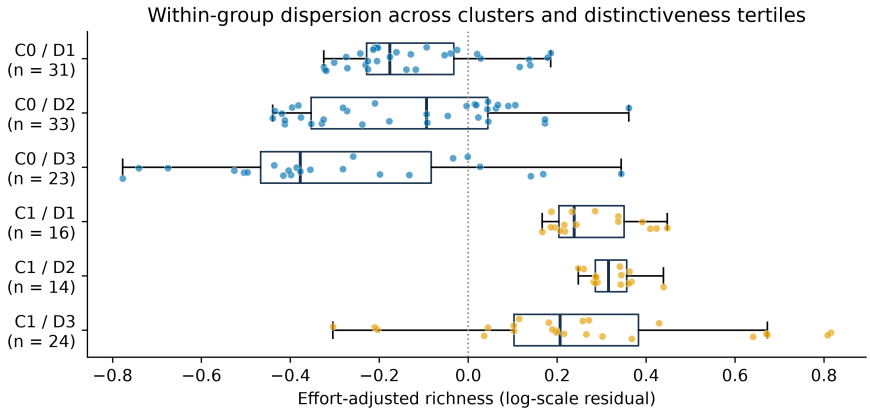


Figure 15. Distribution of effort-adjusted richness within K-means classes (C0 and C1) and repertoire-distinctiveness tertiles (D1 to D3). Points denote countries. Boxes span the interquartile range, with whiskers extending to the most extreme observed values within 1.5 interquartile ranges of the box. The dotted line marks zero.

Effort-adjusted richness varies within all six groups. Several distributions overlap even when their medians differ. Distinctiveness and the two-class partition describe different properties of the country profiles, and the displayed groups summarize variation within the archive.

Figure 16 decomposes the country profiles into 13 documented use categories and shows whether each category occupies a larger or smaller share of a country's recorded repertoire than in the dataset as a whole. The circular arrangement preserves continent-level context while exposing changes among neighboring country sectors and across category rings. It therefore complements the coarse cluster map by showing the compositional detail that a two-class solution necessarily compresses.

The support screen masked 221 of 1,833 cells, including all cells in six countries with fewer than 100 associations. Among displayed cells, European profiles had positive median ratios for bee plants and food additives and negative median ratios for fuels. Median fuel and medicine ratios were positive in Africa, and median animal and invertebrate food ratios were positive in Asia. These summaries describe the subset of cells with sufficient archive support under the stated display rule.

Main findings

Table 3 consolidates the main numerical results and states the evidential scope that should accompany each finding. The rows distinguish tested spatial dependence, descriptive profile clusters, thresholded network modules, and geographically blocked predictive performance. This separation prevents evidence generated for one analytical question from being reused as support for a stronger cultural or causal claim.

Taken together, the rows show that the strongest reproducible statement concerns spatial structure in documented country repertoires, not the direct detection of bounded knowledge communities. The contrast between a stable two-cluster summary and a 19-module similarity network demonstrates that inferred regionalization changes with representation and scale. The spatial-validation results further show that predictive performance must be reported with geographic blocking before any model is considered transportable.

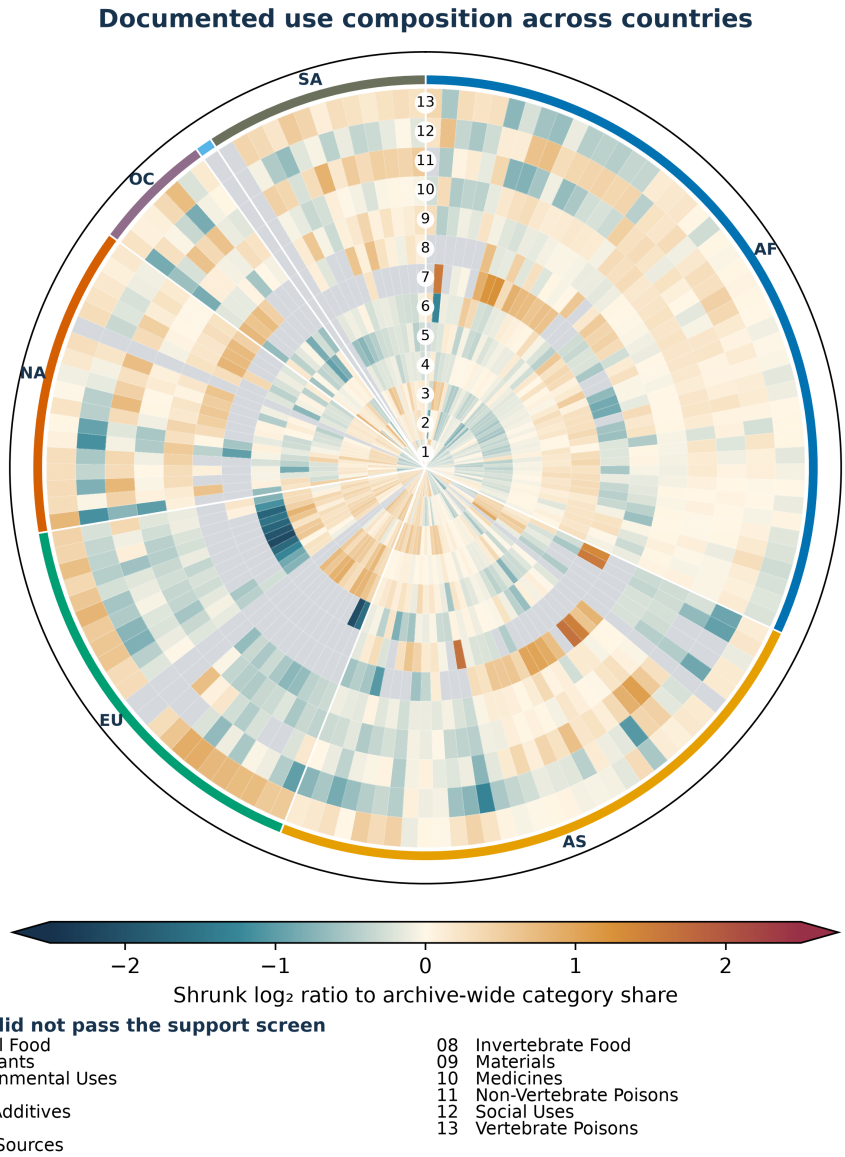


Figure 16. Circular heatmap of documented useful-plant composition. Sectors represent the 141 country profiles grouped by continent, and rings 1 to 13 represent use categories. Colors show shrunk log2 ratios to the archive-wide composition. Gray cells have low country totals, fewer than five cell associations, or an unstable sign across prior weights. The outer band marks continent groups.

Table 3. Main findings and their evidential scope.

Question	Finding
Spatial structure	Effort-adjusted richness Moran I = 0.497, permutation $p = 0.0001$ under symmetric 8-NN weights.
Biocultural regions	The selected baseline used $k = 2$; median 80% subsample ARI=0.964.
Knowledge similarity	The similarity network contained 19 modules with modularity 0.636; modules and K-means agreed weakly/moderately according to ARI 0.346.
GeoAI	Leave-continent-out $R^2 = 0.587$; random-fold R^2= 0.685.

Discussion

The wildfire application of ST-SCKM uses geographic locations and times associated with observed events, for which proximity and persistence can carry information about a physical process (Idris et al. 2026). In SEPASAL, different bibliography periods share the same country representative point, and one dated citation can contribute links across several distribution countries. Consecutive bins therefore reflect uneven publication and indexing histories alongside compositional differences. With those inputs, proximity constraints can join countries or periods that share location or date while differing in recorded repertoire features.

The strong differentiation reported among nearby Bogotá medicinal-plant markets by Bussmann *et al*. (2018) provides an important substantive parallel: repertoire dissimilarity can reflect mobility, provenance, access, and documentation context rather than geographic distance alone. Likewise, the participatory and ethical trajectory emphasized by Bussmann (2019) means that database-level spatial regularities should generate field questions, not replace collaboration with knowledge holders. These studies support interpreting the present country-level patterns as prompts for contextual comparison rather than as self-validating maps of culture.

Spatial organization of the recorded useful plant knowledge base

The clearest result is consistent positive spatial dependence in effort-adjusted richness across multiple neighborhood sizes. Nearby countries resemble one another more than expected under spatial permutation, even after accounting for the number of constructed taxon-use associations. This finding moves beyond a descriptive map because the spatial relationship is explicitly tested and remains stable across k. Yet the result belongs to the database as assembled. SEPASAL's dryland focus, plant distributions, source coverage, and national boundaries all contribute to the signal.

The Local Moran classes locate clusters of similar country residuals relative to the archive-wide count relationship. High-high units concentrated in Europe, while low-low units occurred in Africa and Asia. Continental intercepts attenuated global dependence, but the checks cannot rank the contributions of publication language, digitization, research history or plant availability.

Emergent regions are representation dependent

The stable two-cluster K-means solution summarizes a broad contrast, but the 19-module similarity network shows that finer regional structure remains. Partial ARI and NMI agreement demonstrates that neither representation is a privileged map of biocultural regions. K-means forces every country into a global partition based on multivariate centroids. The similarity network retains local pairwise links and permits isolated or weakly connected units. Their disagreement is therefore expected and informative.

Ward sometimes achieved strong separation by isolating very small groups, while DBSCAN often sacrificed coverage. A similar trade-off appeared as k increased in K-means. The retained classes support comparisons of broad profile contrasts, and pairwise similarities can be used to identify more specific cases for qualitative comparison.

Spatial continuity versus useful plant similarity

The positive distance-decay relationship is consistent with several mechanisms. Neighboring countries share floras and ecological conditions; scholarly traditions and regional floras may also produce similar records; communities may exchange plants and knowledge across borders; and common ancestry may structure terminology and practice. The analysis cannot identify which mechanism dominates. A stronger test would combine community-level use observations, species availability, language phylogenies, historical movement, and repeated field research. The present result should be treated as a benchmark against which those richer models can be compared.

Network structure and central plants

The bipartite network highlights plants that connect many country repertoires, including *Cynodon dactylon*, *Oxalis corniculata*, and several widespread grasses. Their centrality is compatible with broad distributions, repeated documentation, and versatile uses. It is not equivalent to cultural importance, pharmacological efficacy, or priority for bioprospecting. Cross-cultural comparison can generate hypotheses, but ethical and ecological context should precede extraction-oriented interpretation (Saslis-Lagoudakis *et al.* 2011).

Networks can group related units when pairwise relationships vary across a system, as shown in studies of housing and commodity-price comovements (Xu & Zhang 2023, 2024a, 2024b). The analogy concerns the use of pairwise structure. In SEPASAL an edge represents shared recorded taxa. As ecological network sampling experiments show, uneven observation can change inferred network structure (Hervías-Parejo et al. 2025). The changing module counts and isolate frequencies in our threshold scan favor using modules to select comparative cases and reporting the parameters that define them.

Spatio-temporal documentation patterns

Finer bibliography bins captured more country-period units but produced no improvement in full-coverage ST-DBSCAN silhouette. The tiny positive-weight gain at 50 years was specific to that aggregation. Cumulative documentation curves could

characterize archive growth, although accumulating records would reinforce the distinction between publication intensity and the timing of community practice.

Implications for biocultural conservation

Spatially unusual repertoires can guide comparative case selection. Conservation decisions would require evidence on ecological exposure, community priorities, tenure and the permissions under which knowledge can be used. Figure 15 shows only the dispersion of archive-derived residuals within the six analytical groups.

The most useful conservation implication is procedural. National or global database maps should be treated as screening tools whose claims require community-led validation. Sensitive plant locations and restricted practices should not be exposed merely because a technical pipeline can map them. Collective benefit and authority to control are analytical constraints, not appendices to an otherwise completed model.

Methodological implications for ethnobotany

Spatial intelligence adds value when each method answers a distinct ethnobotanical question. Moran statistics test regional dependence. Clustering asks whether complete profiles can be summarized. Similarity networks identify cross-border connections. Spatial cross-validation tests whether models transport geographically. The methods become misleading when their outputs are treated as interchangeable evidence of culture. A reproducible workflow should preserve these distinctions and report negative results.

Random splitting gave higher predictive scores for this archive, while the continent-grouped folds showed substantial regional variation. The Europe fold illustrates how a restricted within-fold outcome range can yield a low R^2 even with a comparatively low absolute error. Residual spatial dependence is a diagnostic of structured prediction error. With 141 countries and five geographic folds, it cannot distinguish missing predictors from limited model capacity or geographic shifts in the feature-outcome relationship (Mushagalusa et al. 2024).

Limitations

The analysis has eight principal limitations. First, use records are linked to countries through plant distributions rather than locality-specific documentation. Second, countries are coarse administrative units containing many cultures and environments. Third, documentation effort is only partly represented by association count. Fourth, taxonomic names retain the historical SEPASAL backbone. Fifth, bibliography years do not measure change in living knowledge. Sixth, representative points simplify large and archipelagic countries. Seventh, similarity thresholds and cluster numbers remain analytical choices. Eighth, environmental exposure, linguistic relationships, and community-governance variables were unavailable. These constraints prevent submission-ready claims about biocultural regions or knowledge diffusion, even though the computational results themselves are reproducible.

Conclusion

Country-level documented useful-plant repertoires in SEPASAL are spatially structured. Effort-adjusted richness showed positive global autocorrelation across four neighborhood scales, and local tests identified regional concentrations and outliers. A balanced two-cluster summary was numerically stable, while a thresholded similarity network revealed finer modular structure. Geographic distance was associated with repertoire dissimilarity, but ecological availability and documentation history remain plausible alternative explanations.

Temporal and geographic validation narrowed the interpretation of the archive. Full-coverage ST-DBSCAN remained feature-incoherent at all three bin widths. The augmented K-means diagnostic found no material gain from adding space or bibliography year, while the random-forest results varied across held-out regions. These patterns set limits on what the available country and bibliography records can support.

The next scientific step is not a more elaborate algorithm. It is a governed community-level dataset that links use reports to places, peoples, observation dates, plant availability, and uncertainty with the authority of knowledge holders. Applied to such data, the workflow can help compare local distinctiveness, shared repertoires, and conservation questions without treating traditional ecological knowledge as decontextualized extractable information.

Declarations

List of abbreviations: ARI - adjusted Rand index; CARE - Collective Benefit, Authority to Control, Responsibility, and Ethics; FDR - false discovery rate; GeoAI - geospatial artificial intelligence; LISA - Local Indicators of Spatial Association; SEPASAL - Survey of Economic Plants for Arid and Semi-Arid Lands; ST-DBSCAN - spatial-temporal density-based spatial clustering of applications with noise.

Ethics approval and consent to participate: No new human-participant research was conducted. The analysis used openly available secondary botanical and bibliographic records. Community-level coordinates, sacred practices, and restricted knowledge were neither inferred nor published.

Consent for publication: Not applicable

Availability of data and materials: SEPASAL is available from Royal Botanic Gardens, Kew at https://sftp.kew.org/pub/data-repositories/sepasal/ under CC BY 4.0. Processed analytical data, code, configuration, checksums, figures, tables, and the claim–evidence audit accompany the reproducible research package.

Competing interests: Not applicable

Funding: Not applicable

Author contributions: M.A.I developed the analytical workflow, conducted the computational analyses, prepared the visualizations, and drafted the manuscript. R.W.B critically revised and edited the manuscript.

Acknowledgements

The authors acknowledge the Royal Botanic Gardens, Kew for maintaining and openly distributing the SEPASAL archive and the Natural Earth project for country geometries.

Literature cited

Anselin L. 1995. Local Indicators of Spatial Association—LISA. Geographical Analysis 27:93-115. doi: 10.1111/j.1538-4632.1995.tb00338.x.

Birant D, Kut A. 2007. ST-DBSCAN: An algorithm for clustering spatial–temporal data. Data & Knowledge Engineering 60:208-221. doi: 10.1016/j.datak.2006.01.013.

Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. 2008. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment 2008:P10008. doi: 10.1088/1742-5468/2008/10/p10008.

Breiman L. 2001. Random Forests. Machine Learning 45:5-32. doi: 10.1023/a:1010933404324.

Bussmann RW, Paniagua Zambrana NY, Romero C, Hart RE. 2018. Astonishing diversity—the medicinal plant markets of Bogotá, Colombia. Journal of Ethnobiology and Ethnomedicine 14:43. doi: 10.1186/s13002-018-0241-8.

Bussmann RW. 2019. Making friends in the field: How to become an ethnobotanist - A personal reflection. Ethnobotany Research and Applications 18:1-13. https://ethnobotanyjournal.org/index.php/era/article/view/1357

Cámara-Leret R, Bascompte J. 2021. Language extinction triggers the loss of unique medicinal knowledge. Proceedings of the National Academy of Sciences 118:e2103683118. doi: 10.1073/pnas.2103683118.

Cámara-Leret R, Fortuna MA, Bascompte J. 2019. Indigenous knowledge networks in the face of global change. Proceedings of the National Academy of Sciences 116:9913-9918. doi: 10.1073/pnas.1821843116.

Cámara-Leret R, Patrick R. Roehrdanz PR, Jordi Bascompte J. 2026. The forest of knowledge under global change. Nature 655:1212-1216. doi: 10.1038/s41586-026-10741-y.

Cámara-Leret R, Roehrdanz P, Bascompte J. 2026. The Forest of Knowledge. Zenodo dataset. doi: 10.5281/zenodo.19202485.

Cordero S, Abello L, Gálvez F. 2022. Rizom<<a: a new comprehensive database on traditional uses of Chilean native plants. Biodiversity Data Journal 10:e80002. doi: 10.3897/bdj.10.e80002.

do Nascimento Fernandes de Souza E, Julie A Hawkins JA. 2020. Ewé: a web-based ethnobotanical database for storing and analysing data. Database 2020:baz144. doi: 10.1093/database/baz144.

Gaoue OG, Coe MA, Bond M, Hart G, Seyler BC, McMillen H. 2017. Theories and Major Hypotheses in Ethnobotany. Economic Botany 71:269-287. doi: 10.1007/s12231-017-9389-8.

Gorenflo LJ, Romaine S, Mittermeier RA, Walker-Painemilla K. 2012. Co-occurrence of linguistic and biological diversity in biodiversity hotspots and high biodiversity wilderness areas. Proceedings of the National Academy of Sciences 109:8032-8037. doi: 10.1073/pnas.1117511109.

Hubert L, Arabie P. 1985. Comparing partitions. Journal of Classification 2:193-218. doi: 10.1007/bf01908075.

Hervías-Parejo S, Traveset A, Nogales M, Heleno R, Llewelyn J, Strona G. 2025. Sampling biases across interaction types affect the robustness of ecological multilayer networks. Ecological Informatics 89:103183. doi: 10.1016/j.ecoinf.2025.103183.

Idris MA, Aidi MN, Djuraidah A. 2026. Performance evaluation of ST-DBSCAN and spatio-temporal spatially constrained K-means (ST-SCKM) for wildfire risk zoning and resilience analysis. Journal of Safety Science and Resilience:100357. doi: 10.1016/j.jnlssr.2026.100357.

Idris MA, Yudanto AA, Putra MNHT, Andespa R, Syukri N, Fadila F, Safira I, Mutmainah Z, Resiloy UDA. 2026. From mapping to modeling: Spatial statistics in ethnobotanical research. Ethnobotany Research and Applications 34:1-10. https://ethnobotanyjournal.org/index.php/era/article/view/8467

Jolliffe IT, Cadima J. 2016. Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 374:20150202. doi: 10.1098/rsta.2015.0202.

Kor L, Fernández-Lucero M, Granados Flórez DA, Dawson TP, Mauricio Diazgranados M. 2024. Bridging local and scientific knowledge for area-based conservation of useful plants in Colombia. Ambio 53:309-323. doi: 10.1007/s13280-023-01921-5.

Kunwar RM, Fadiman M, Cameron M, Bussmann RW, -Magar KP, Rimal B, Sapkota P. 2018. Cross-cultural comparison of plant use knowledge in Baitadi and Darchula districts, Nepal Himalaya. Journal of Ethnobiology and Ethnomedicine 14:40. doi: 10.1186/s13002-018-0242-7.

Lloyd S. 1982. Least squares quantization in PCM. IEEE Transactions on Information Theory 28:129-137. doi: 10.1109/tit.1982.1056489.

Loh J, Harmon D. 2005. A global index of biocultural diversity. Ecological Indicators 5:231-241. doi: 10.1016/j.ecolind.2005.02.005.

Maffi L. 2005. Linguistic, cultural and biological diversityINGUISTIC, CULTURAL, AND BIOLOGICAL DIVERSITY. Annual Review of Anthropology 34:599-617. doi: 10.1146/annurev.anthro.34.081804.120437.

Mushagalusa CA, Fandohan AB, Glèlè Kakaï R. 2024. Random forest and spatial cross-validation performance in predicting species abundance distributions. Environmental Systems Research 13:23. doi: 10.1186/s40068-024-00352-9.

Newman MEJ. 2006. Modularity and community structure in networks. Proceedings of the National Academy of Sciences 103:8577-8582. doi: 10.1073/pnas.0601602103.

Pankararu CJ, Teixidor-Toneu I, Odonne G, et al. 2026. A global biodiversity use data infrastructure acknowledging indigenous and local knowledge. npj Biodiversity 5:7. doi: 10.1038/s44185-026-00121-0.

Paulino de Albuquerque U, Hanazaki N. 2009. Five Problems in Current Ethnobotanical Research—and Some Suggestions for Strengthening Them. Human Ecology 37:653-661. doi: 10.1007/s10745-009-9259-9.

Roberts DR, Bahn V, Ciuti S, Boyce MS, Elith J, Guillera-Arroita G, Hauenstein S, J. Lahoz-Monfort JJ, Schröder B, Thuiller W, Warton DI, Wintle BA, Hartig F, Dormann CF. 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography 40:913-929. doi: 10.1111/ecog.02881.

Royal Botanic Gardens, Kew. 2020. Survey of Economic Plants for Arid and Semi-Arid Lands SEPASAL database. Dataset. https://sftp.kew.org/pub/data-repositories/sepasal/README.txt

Russo Carroll S, Garba I, Figueroa-Rodríguez OL, Holbrook J, Lovett R, Materechera S, Parsons M, Raseroka K, Rodriguez-Lonebear D, Rowe R, Sara R, Walker JD, Anderson J, Hudson M. 2020. The CARE Principles for Indigenous Data Governance. Data Science Journal 19:43. doi: 10.5334/dsj-2020-043.

Saslis-Lagoudakis CH, Williamson EM, Savolainen V, Hawkins JA. 2011. Cross-cultural comparison of three medicinal floras and implications for bioprospecting strategies. Journal of Ethnopharmacology 135:476-487. doi: 10.1016/j.jep.2011.03.044.

Valavi R, Elith J, J. Lahoz-Monfort JJ, Guillera-Arroita. G 2019. block CV : An r package for generating spatially or environmentally separated folds for k -fold cross-validation of species distribution models. Methods in Ecology and Evolution 10:225-232. doi: 10.1111/2041-210x.13107.

Wang S, Huang X, Liu P, Zhang M, Biljecki F, Hu T, Fu X, Liu L, Liu X, Wang, R, Huang Y, Yan J, Jiang J, Chukwu M, Naghedi SR, Hemmati M, Shao Y, Jia N, Xiao Z, Tian T, Hu Y, Yu L, Yap W, Macatulad E, Chen Z, Cui Y,Ito K, Ye M, Fan Z, Lei B, Bao S. 2024. Mapping the landscape and roadmap of geospatial artificial intelligence (GeoAI) in quantitative human geography: An extensive systematic review. International Journal of Applied Earth Observation and Geoinformation 128:103734. doi: 10.1016/j.jag.2024.103734.

Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, Bonino da Silva Santos L, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJG, Groth P, Goble C, Grethe, JS, Heringa J, C 't Hoen PAC, Hooft R, Kuhn T, Kok R, Kok, J Lusher, SJ, Martone ME,

Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone SA, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B. 2016. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3:160018. doi: 10.1038/sdata.2016.18.

Xu X, Zhang Y. 2023. Spatial-Temporal Analysis of Residential Housing, Office Property, and Retail Property Price Index Correlations: Evidence from Ten Chinese Cities. International Journal of Real Estate Studies 17(2):1-13. doi: 10.11113/intrest.v17n2.274.

Xu X, Zhang Y. 2024a. Network analysis of comovements among newly-built residential house price indices of seventy Chinese cities. International Journal of Housing Markets and Analysis 17(3):726-749. doi: 10.1108/IJHMA-09-2022-0134.

Xu X, Zhang Y. 2024b. Network Analysis of Price Comovements Among Corn Futures and Cash Prices. Journal of Agricultural & Food Industrial Organization 22(1):53-81. doi: 10.1515/jafio-2022-0009.

Zignora Limba S, Awantara IG, Riansyah B, Idris MA. 2026. Integrating ethnobotany and quantitative methods: A bibliometric analysis of statistical, mathematical, and data science approaches (2016–2025). Ethnobotany Research and Applications 34:1-9. https://ethnobotanyjournal.org/index.php/era/article/view/8318