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.

An integrated bioinformatics and bioacoustics dataset for Pantanal fauna

Mountaga SY

mountagasy@muntigidi.ca

Research Article

Keywords:

Posted Date: July 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10323705/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

The Pantanal is one of the largest tropical wetlands and an important reservoir of Neotropical acoustic biodiversity. This Data Descriptor presents Pantanal Bioacoustics v2, a reproducible data infrastructure that combines a multi-source taxonomic reference, continental occurrence records, audio metadata, validated recordings and acoustic descriptors. The taxonomic master contains 10,088 entries, of which 6,657 are valid binomials or simple trinomials under the publication filter. Following multi-source discovery, file validation and SHA-256 duplicate quarantine, the downloadable corpus contains 11,785 validated recordings representing 4,476 species. Of these recordings, 415 are assigned to the strict Pantanal bounding box and 300 species have at least one in-box recording. The processing framework extracts 77 acoustic descriptors and supplies quality-audit reports, database fingerprints, command logs and machine-readable manifests. The resource is intended for passive acoustic monitoring, taxonomic coverage assessment, model development and targeted field acquisition. Source-level licence restrictions are retained, and Macaulay Library records are represented as metadata only.

Background and Summary

2.1 Scientific context

The Pantanal is a large Neotropical wetland shared by Brazil, Bolivia and Paraguay and covering approximately 150,000 km2. Its seasonal flood-drought dynamics support rich avian and acoustic communities. Wildfires, agricultural expansion and climate variability create an increasing need for repeatable, non-invasive monitoring. Passive acoustic monitoring can provide long-duration observations, but its regional use depends on a reference corpus that is taxonomically controlled, geographically interpretable and technically traceable.

The project began with a practical question: which birds, families and sound-producing taxa can be documented from available sources, and which of them are represented by usable audio? General regional syntheses often cite approximately 650 bird species for the Pantanal. That estimate is an ecological benchmark, whereas the 9,086 Aves entries in the merged master represent a much broader multi-source taxonomic domain before geographic and taxonomic filtering. The dataset therefore distinguishes listed taxa, valid scientific names, taxa with GBIF evidence in the study extent, taxa with validated audio, and taxa with strict in-box audio.

2.2 Objectives and scope

Consolidate GBIF, IUCN and Pantanal IBA taxonomic information into a traceable master reference.

Aggregate occurrence data for South and North America to support geographic and seasonal interpretation.

Build a validated, deduplicated audio corpus from complementary public sources.

Extract reproducible acoustic descriptors and document technical quality.

Provide data, code, audit outputs and manifests that enable independent reuse.

2.3 Headline dataset characteristics

Table 1
Headline characteristics of the post-quarantine dataset.

Indicator	Value
Taxonomic master	10,088 entries
Valid binomials or simple trinomials	6,657 (66.0%)
Valid Aves names	5,695 of 9,086
Species with GBIF presence in strict Pantanal extent	1,635
Validated recordings after duplicate quarantine	11,785
Species with at least one validated recording	4,476
Strict Pantanal-bbox recordings	415
Species with at least one strict-bbox recording	300
Acoustic descriptors per processed recording	77
Critical anomalies after final audit	0

Methods

3.1 Pipeline architecture

The workflow is organized into ten processing phases and a reporting and audit layer. It begins with taxonomic consolidation, continues through occurrence aggregation, Pantanal intersection, audio discovery, validation and descriptor extraction, and then adds ecological profiles, exploratory analyses, eBird cross-validation and large-scale multi-source expansion. Phase 10b produces metrics, figures, bilingual report sections, a 16-check audit, database fingerprints and a global manifest. Each phase uses pre-modification backups, dry-run defaults where appropriate, explicit application steps and append-only JSONL logs.

Table 2
Processing phases and principal outputs.

Phase	Purpose	Principal output
1	Taxonomic consolidation	master_v2.db
2–3	Continental GBIF aggregation and Pantanal intersection	Occurrence and geographic tables
4–5	Audio discovery and validation	Validated source-linked recordings
6	Acoustic descriptor extraction	77 descriptors per processed audio
7–9	Ecological profiles, PCA and eBird cross-validation	Derived ecological and validation tables
10	Large-scale multi-source audio expansion	Expanded audio_v2.db
10b	Metrics, figures, audit and manifests	Publication and reproducibility package

3.2 Data sources

The taxonomic layer combines the GBIF backbone, IUCN Red List information and Pantanal Important Bird and Biodiversity Area checklists. The occurrence layer uses two large GBIF downloads. The audio layer integrates Xeno-Canto, GBIF Media and iNaturalist files, with Macaulay Library represented in a dedicated metadata-only table because general redistribution is not permitted. eBird Status and Trends information is used for regional presence and seasonal cross-validation.

Table 3
Primary data sources, uses and constraints.

Source	Use in the dataset	Constraint
GBIF	Taxonomic backbone and occurrence downloads	Large external archives; source heterogeneity
IUCN Red List	Conservation status	Only a subset of master taxa is classified
Pantanal IBA resources	Regional checklist context	Mixed taxonomic formatting
Xeno-Canto	Principal downloadable audio source	Licences frequently restrict commercial reuse
GBIF Media	Audio links associated with occurrences	May re-index records from other platforms
iNaturalist	Citizen-science audio and coordinates	Sparse Pantanal audio participation
Macaulay Library	Coverage assessment through metadata	Audio files not redistributed
eBird	Regional presence and seasonal validation	Rate-limited access

3.3 Taxonomic filtering

The initial master contained 10,088 heterogeneous entries. Approximately 3,431 entries were not usable binomials or simple trinomials: 1,121 genus names with authority strings, 2,073 unresolved BOLD codes, 144 bare genera and 93 other patterns such as hybrids, sp. labels or higher ranks. A strict regular-expression filter retained 6,657 names. In Aves, 5,695 of 9,086 entries were retained. Reporting coverage against this filtered denominator prevents the non-canonical entries from artificially lowering estimated audio coverage.

3.4 GBIF occurrence downloads and geographic rules

The South American download contains 104,522,562 avian occurrences from 1,264 published datasets and has a compressed size of 15.4 GB. The North American download contains 1,348,301,479 avian occurrences from 677 datasets and has a compressed size of 178.1 GB. Both downloads use simple tab-separated values and filter for Aves, the relevant continent, valid coordinates and present occurrence status. The strict Pantanal extent is defined as latitude − 23 to -13 degrees and longitude − 61 to -52 degrees.

3.5 Audio discovery, download and validation

Phase 10 queried the 6,657 valid taxonomic names and identified 44,056 Xeno-Canto candidates and 15,649 GBIF Media candidates. After candidate aggregation and source-identifier deduplication, 58,048 candidates remained. Parallelised downloads were rate-limited by source. A total of 12,172 files were downloaded and 12,150 passed initial validation, corresponding to 99.8%. Twenty-two files failed because

of truncation, unexpected format or decoding errors. A recovery step reattached 183 files found on disk but absent from the database.

Validation used two stages: verification of the MP3, WAV, OGG or FLAC file signature, followed by effective decoding with checks for non-finite values, silence and a minimum duration of one second. Files passing these controls entered the acoustic-processing queue.

3.6 Acoustic descriptors

The 77 descriptors cover spectral centroid, bandwidth, rolloff, contrast and flux; MFCC 1–13 with first and second derivatives; onset rate and tempo; RMS energy, zero-crossing rate and dynamic range; a loudness proxy; and the ecological indices NDSI, ACI, BI, H, ADI and AEI. Phase 8b reuses the original Phase 6 feature extraction and joins it to corrected post-quarantine metadata. After excluding identifiers and non-feature fields, 69 numerical descriptors are available for the 2,666-recording exploratory subset.

3.7 Statistical and quality-control procedures

Coverage ratios use Wilson 95% confidence intervals. Acoustic niche partitioning compares within-family and between-family distances by 10,000 one-sided permutations with Benjamini-Hochberg correction at a 0.05 false-discovery rate. The audit suite contains seven critical checks, five major checks and four minor checks. Critical controls test missing files, suspiciously small files, SHA-256 content duplicates, geographic flag errors and records marked valid without a file path. Major controls examine duration, licence, URL, source identifier and discovery-score anomalies. Minor controls document missing records, date, licence or extension information.

Data Records

4.1 Taxonomic reference

Table 4
Taxonomic composition of the master reference.

Group	Total entries	Valid names	Valid (%)
Aves	9,086	5,695	62.7
Insecta	764	728	95.3
Reptilia	104	104	100.0
Mammalia	76	73	96.1
Amphibia	54	53	98.1
Total	10,088	6,657	66.0

The filtered reference contains 245 distinct families across all groups, including 157 valid Aves families distributed across 27 orders. The source compilation is strongly dominated by birds, which account for

90.1% of raw entries. Taxonomic pollution is also concentrated in Aves because of unresolved BOLD codes and authority-formatted checklist entries.

4.2 GBIF coverage of the Pantanal extent

Within the strict bounding box, 1,635 species have at least one GBIF occurrence. Coverage is uneven among the 264 raw families: 27 families have more than 80% of their species documented in the box, whereas 214 families have less than 20% coverage. These values describe the documentation available through GBIF rather than complete biological occupancy.

4.3 Post-quarantine audio corpus

Table 5
Expansion from the earlier validated corpus to the post-quarantine corpus.

Metric	Earlier corpus	Post-quarantine corpus	Change
Validated recordings	2,710	11,785	+ 9,075; x4.35
Strict-bbox recordings	22	415	+ 393; x18.9
Records flagged in Pantanal countries	1,361	5,604	+ 4,243; x4.1
Species with validated audio	931	4,476	+ 3,545; x4.8
Species with strict-bbox audio	Not reported	300	New field

Before duplicate quarantine, 12,150 recordings were marked valid. The SHA-256 resolver marked 365 records as duplicate content and reduced the publication corpus to 11,785 without materially changing species-level coverage. The pre-quarantine species_master flag identified 4,502 species with audio; the final post-quarantine count is 4,476.

4.4 Coverage by taxonomic group

Table 6
Audio coverage by taxonomic group using the filtered denominator.

Group	Valid species	With audio	Coverage and Wilson 95% CI	With strict-bbox audio
Amphibia	53	40	75.5% [62.4–85.1]	11
Aves	5,695	4,352	76.4% [75.3–77.5]	282
Insecta	728	26	3.6% [2.5–5.2]	0
Mammalia	73	43	58.9% [47.5–69.5]	7
Reptilia	104	11	10.6% [6.0-17.9]	0

The strong avian bias reflects the structure of the public source ecosystem, particularly Xeno-Canto. Insecta and Reptilia remain poorly represented and should be treated as acquisition gaps rather than

evidence of low biological diversity.

4.5 Conservation-status records

Table 7
Audio coverage by IUCN category in the valid-name subset.

IUCN status	Valid species	With audio	Coverage and Wilson 95% CI	With strict-bbox audio
EN	4	2	50.0% [15.0–85.0]	1
VU	6	4	66.7% [30.0-90.3]	2
NT	8	8	100.0% [67.6–100.0]	5
LC	162	131	80.9% [74.1–86.2]	41
DD	1	1	100.0%	0
Unassessed	6,476	4,356	67.3% [66.1–68.4]	262

Fourteen of the 18 threatened or near-threatened species have at least one validated recording and eight have a strict-bbox recording. Based on the category table, the four priority species without audio comprise two EN and two VU species. This corrects the inconsistent wording in the earlier narrative section of the source report.

4.6 Audio sources and licences

The corpus records source, source identifier, original URL, recordist, date, coordinates, licence and a derived licence class when available. Approximate post-quarantine source counts in the original report were about 9,300 Xeno-Canto records, 1,200 GBIF Media records and 580 iNaturalist records; these rounded counts were not intended to sum exactly to the final corpus and should be replaced by exact database queries in the archived data dictionary. Macaulay Library records are stored separately and do not contribute to the downloadable count.

The reported licence classes total 11,785 records: approximately 7,050 research-only or non-commercial records, 355 records under open licences such as CC0, CC-BY or CC-BY-SA, 1,740 restricted records and 2,640 records with unspecified licence information. Reuse must follow the licence attached to each source record.

Technical Validation

5.1 Quality-audit results

Table 8
Automated audit results before and after correction.

Category	Before correction	After correction	Action
Critical checks A1-A7	928 anomalies (7.64%)	0	Quarantine and field correction
SHA-256 duplicate content	729 records in 364 groups	0 active duplicate groups	365 records marked dup_content
Country false negatives	199	0	in_country corrected
Major checks B1-B5	2,123	2,000 documented issues	Retained for review or cosmetic normalisation
Corpus integrity score	92.36%	100.00%	No critical anomaly remaining

Most content duplicates resulted from GBIF Media re-indexing Xeno-Canto files or iNaturalist media. The resolver selected a winner by geographic relevance, source priority, identifier order and duration. The full audit report, anomaly CSV, duplicate-group CSV and executable quarantine SQL are part of the reproducibility package.

5.2 Acoustic descriptor distributions

In the processed corpus, 69.6% of recordings have an NDSI proxy above 0.5 and 16.9% have a value below zero. These values describe the signal-processing output and should not be treated as definitive ecological classifications. The source composition, recording conditions and incomplete geographic metadata influence the distribution.

5.3 Species-level PCA

A complementary PCA was calculated after averaging 11 standardised descriptors by species. It contains 924 species: 818 Aves, 42 Mammalia, 38 Amphibia, 19 Insecta and 7 Reptilia. PC1 explains 42.1% and is primarily associated with spectral brightness and timbre. PC2 explains 19.8% and is associated mainly with energy and intensity. The total for PC1 and PC2 is 61.97%. This result should not be compared directly with the per-recording PCA because aggregation reduces within-species variation.

5.4 Xeno-Canto quality metadata

The quality analysis covers 9,967 Xeno-Canto records. Quality A accounts for 8,579 records, or 86.1%. The taxonomic totals are 9,799 Aves, 91 Mammalia, 71 Amphibia and 6 Insecta. No validated Xeno-Canto reptile recording appears in this subset. Quality labels are contributor-provided metadata and are not a substitute for task-specific model evaluation.

5.5 Phase 8b per-recording PCA

Phase 8b joins the original Phase 6 descriptors to corrected post-quarantine metadata and the valid-name filter. The final exploratory matrix contains 2,666 recordings and 69 numerical descriptors. PC1 explains 8.22%, PC2 6.52%, PC3 6.16%, and PC1-PC10 47.79% cumulatively. The low variance captured by individual components is consistent with the diversity of taxa, sources and sound-production mechanisms.

5.6 Geographic completeness and family centroids

Of the 2,666 recordings in the Phase 8b feature subset, 282 have usable coordinates and 19 fall in the strict bounding box. This 10.58% coordinate rate applies only to the feature subset. It must not be presented as the coordinate-completeness rate of the full 11,785-recording database unless a separate database query confirms that value. In the same subset, 1,482 records are flagged as belonging to Brazil, Bolivia or Paraguay from country metadata, even when exact coordinates are unavailable.

Family centroids were calculated for 89 families represented by at least three recordings. They provide a compact summary of average acoustic position but do not establish biologically discrete clusters. Families with small sample sizes have greater uncertainty and should be interpreted with their dispersion statistics.

5.7 eBird cross-validation and migration fields

The regional eBird scan covered 1,743 codes and confirmed 1,707 species at least once. Twenty-four candidate species were identified for manual taxonomic review, but the master reference was not modified automatically. The Phase 7 summary reports 84 species with a migration signal, while the eBird comparison was applied to a subset of 81 candidates. Thirty-five of those 81 candidates were supported by regional seasonal information. The difference between the 84 total signals and the 81 cross-validated candidates is retained explicitly and should be documented in the released processing manifest.

5.8 Acoustic niche and biogeographic fields

Among 93 families with at least three recordings in the relevant analysis, 71 show statistically significant within-family acoustic proximity after Benjamini-Hochberg correction. This is a derived analysis field and not a claim that family identity can be inferred reliably from audio alone. The biogeographic classification assigns 4,893 valid names to a strict Neotropical category, 412 to a Nearctic-migrant category, 261 to a cosmopolitan category and 1,091 to an undetermined category. These labels reflect GBIF distribution patterns and do not establish endemism.

Usage Notes

6.1 Appropriate reuse

Use the valid-name flag when calculating coverage ratios or querying public APIs.

Separate the full post-quarantine database from the smaller Phase 8b feature subset in all analyses.

Use source identifiers and licence fields before redistributing any audio.

Treat absence of audio as an acquisition gap, not as biological absence.

Treat migration, NDSI and biogeographic classes as derived indicators requiring independent validation.

Use the audit manifests and SHA-256 files to confirm the exact data version used.

6.2 Known limitations

The dataset is dominated by avian sources. Insect, reptile, mammal and amphibian coverage is less complete and varies according to public-source participation. Macaulay Library audio is not part of the redistributed corpus. Coordinate availability is limited in the feature subset. Migration classification is heuristic, NDSI is an approximate proxy, and GBIF-based biogeographic classes are sensitive to reporting bias. The exploratory analyses therefore support quality assessment and dataset characterisation rather than tests of new ecological hypotheses.

6.3 Potential applications

The corpus can support passive acoustic monitoring pilots, priority-species reference libraries, source-aware model training, retrieval based on acoustic similarity and targeted field campaigns. Operational use requires field validation, especially for threatened species, migratory candidates and under-represented geographic zones. Hydrological or land-management decisions require independent environmental assessment and are not implied by the dataset alone.

Data Availability

7.1 External GBIF source data

Table 9
GBIF occurrence downloads used by the project.

Region	DOI	Created	Records	Published datasets	Compressed size	Format
North America	10.15468/dl.5wqvyd	28 March 2026, 21:48:12	1,348,301,479	677	178.1 GB	Simple TSV
South America	10.15468/dl.6t94dn	28 March 2026, 21:48:16	104,522,562	1,264	15.4 GB	Simple TSV

Both GBIF downloads filter for TaxonKey Aves, the relevant continent, HasCoordinate true and OccurrenceStatus present. The DOI landing pages identify the exact historical downloads used. These

external source archives do not need to be duplicated in the project repository, but their DOI, filter, creation date and role in the pipeline must remain in the manuscript and README.

7.2 Reproducing a GBIF request

A reviewer who wishes to create a new equivalent download must create a GBIF account and authenticate with the GBIF username and password. The asynchronous request returns a download key. When processing is complete, GBIF sends an email and the key can be used to retrieve the status, download link and DOI.The download key identifies a specific request and is not a permanent API credential. Very large requests may take several hours, and completion within 24 hours is not guaranteed.

7.3 Project-generated data and peer-review access

The data produced by the study should be supplied directly to editors and reviewers: master_v2.db, audio_v2.db, descriptor tables, metadata, data dictionaries, audit outputs, manifests, checksums, figures, scripts and a README. At first submission, an anonymous direct-download link can be used. Before publication, the files should be deposited in a recognised repository with a persistent identifier and direct HTTPS access. Open-licence audio may be archived separately. Restricted audio should be represented by source identifiers, URLs, licences, hashes and retrieval scripts rather than redistributed without permission.

Table 10
Recommended contents of the review and publication package.

Item	Directly supplied?	Recommended treatment
GBIF raw occurrence archives	No duplicate required	Cite persistent DOI and filters
Derived SQLite databases	Yes	Repository or anonymous review storage
Descriptor and summary tables	Yes	CSV, Parquet or structured database
Audit reports and manifests	Yes	JSON, CSV, Markdown and SHA-256 files
Code and environment information	Yes	Versioned archive with README
Open-licence audio	When included in the described release	Separate archive with licences
Restricted audio	No, unless permission exists	Identifiers, URLs, licences and retrieval scripts

Code Availability

8.1 Scripts and reproducibility artefacts

The codebase includes scripts 01_consolidation.py through 09_ebird_integration.py, 10_audio_expansion_v5.py, clean_master_binomials.py, phase10c_recover_orphans.py and the Phase 10b reporting suite. Critical reporting components include phase10b_run_all.py, phase10b_metrics.py, phase10b_figures.py, phase10b_report_sections.py, phase10b_quality_audit.py and phase10b_resolve_a3_dups.py. Each run can produce a command log, per-step output logs, a global manifest and SHA-256 fingerprints of the input databases.

8.2 Principal data artefacts

master_v2.db, approximately 4.7 MB, containing the consolidated taxonomy and reference fields.

audio_v2.db, approximately 402 MB after quarantine, containing candidates, validated records, metadata-only records and species progress.

gbif_world.db with the species_gbif_keys taxonomic-resolution table.

Phase 8b PCA metadata and figure manifests in JSON.

Quality-audit reports in JSON and Markdown, anomaly CSV files and quarantine SQL.

Publication figures in PNG, PDF and SVG; SVG or PDF should be retained as the archival vector format and PNG used for manuscript embedding.

8.3 Operational reproducibility

Database inserts use INSERT OR IGNORE or INSERT OR REPLACE with source and source_id deduplication keys. JSON cache writes are atomic, and SQLite databases use WAL mode with periodic commits. Timestamped backups are produced before schema migration, orphan recovery, quarantine and duplicate resolution. These controls reduce the risk of silent changes and allow the final publication figures to be traced to exact database states.

Declarations

9.1 Acknowledgements

The author acknowledges GBIF, Xeno-Canto, iNaturalist, eBird, the IUCN Red List, Pantanal IBA resources and Macaulay Library metadata, together with the observers, recordists, taxonomists and conservation researchers who created and maintained the underlying records.

9.2 Funding

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

9.3 Competing interests

The author declares no financial, commercial, professional or personal competing interests that could have influenced the work.

9.4 Author contributions

Mountaga Sy conceived and designed the study, developed the Pantanal Bioacoustics v2 pipeline, consolidated the taxonomic, occurrence, conservation and audio data, implemented audio discovery, validation, deduplication and audit procedures, performed the acoustic and statistical analyses, generated the figures and reproducibility outputs, and wrote and revised the manuscript.

References

1. Wilson EB (1927) Probable inference, the law of succession, and statistical inference. J Am Stat Assoc 22:209–212. https://doi.org/10.1080/01621459.1927.10502953
2. Brown LD, Cai TT, DasGupta A (2001) Interval estimation for a binomial proportion. Stat Sci 16:101–133. https://doi.org/10.1214/ss/1009213286
3. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Royal Stat Soc B 57:289–300
4. McFee B et al (2015) librosa: Audio and music signal analysis in Python. Proceedings of the 14th Python in Science Conference
5. Kahl S et al (2021) BirdNET: A deep learning solution for avian diversity monitoring. Ecol Inf 61:101236
6. Sueur J et al (2014) Acoustic indices for biodiversity assessment and landscape investigation. Acta Acustica united Acustica 100:772–781
7. Junk WJ et al (2006) Biodiversity and its conservation in the Pantanal of Mato Grosso, Brazil. Aquat Sci 68:278–309
8. Tubelis DP, Tomas WM (2003) Bird species of the Pantanal wetland, Brazil. Ararajuba 11:5–37
9. Sugai LSM et al (2019) Terrestrial passive acoustic monitoring: review and perspectives. Bioscience 69:15–25
10. Pieretti N, Farina A, Morri DA (2011) A new methodology to infer the singing activity of an avian community: The Acoustic Complexity Index. Ecol Ind 11:868–873. https://doi.org/10.1016/j.ecolind.2010.11.005
11. Sullivan BL et al (2009) eBird: A citizen-based bird observation network in the biological sciences. Biol Conserv 142:2282–2292. https://doi.org/10.1016/j.biocon.2009.05.006
12. Sueur J, Aubin T, Simonis C, Seewave (2008) A free modular tool for sound analysis and synthesis. Bioacoustics 18:213–226. https://doi.org/10.1080/09524622.2008.9753600

13. Villanueva-Rivera LJ, Pijanowski BC, Doucette J, Pekin B (2011) A primer of acoustic analysis for landscape ecologists. Landscape Ecol 26:1233–1246. https://doi.org/10.1007/s10980-011-9636-9
14. GBIF.org. GBIF Occurrence Download. https://doi.org/10.15468/dl.5wqvyd (2026)
15. GBIF.org. GBIF Occurrence Download. https://doi.org/10.15468/dl.6t94dn (2026)
16. BirdLife International. BirdLife DataZone: Important Bird and Biodiversity Areas (2026)
17. Xeno-Canto Foundation. Xeno-Canto: Sharing wildlife sounds from around the world (2026)
18. Cornell Lab of Ornithology (2026) Macaulay Library and eBird Status and Trends resources (2024&#8211
19. iNaturalist (2026) Research-grade observations. California Academy of Sciences and National Geographic Society

Figures

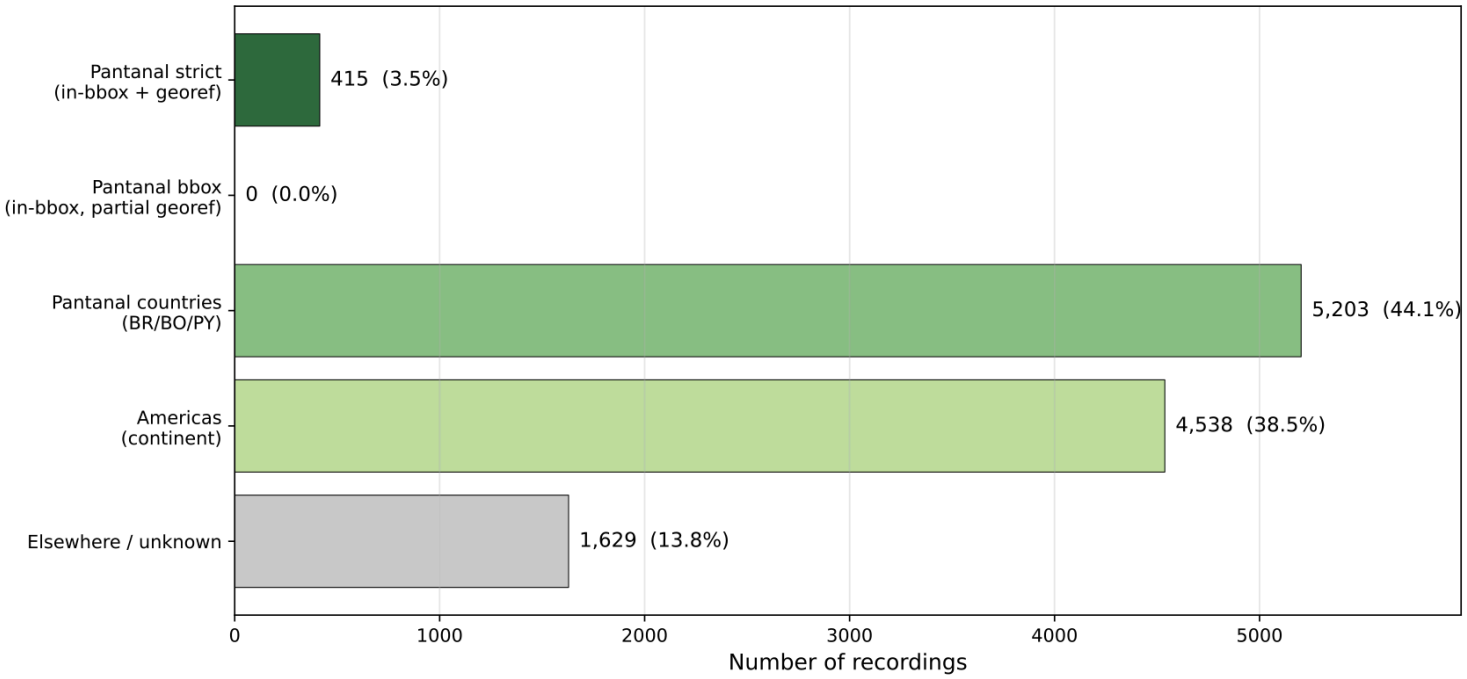


Figure 1

Geographic stratification of the 11,785 validated recordings. The plotted categories total 11,785 records. The 5,203 value is the country-only category used by this mutually exclusive figure and should not be confused with the separate database summary of 5,604 records flagged in Pantanal countries. The difference between those two derivations must remain documented in the repository metadata.

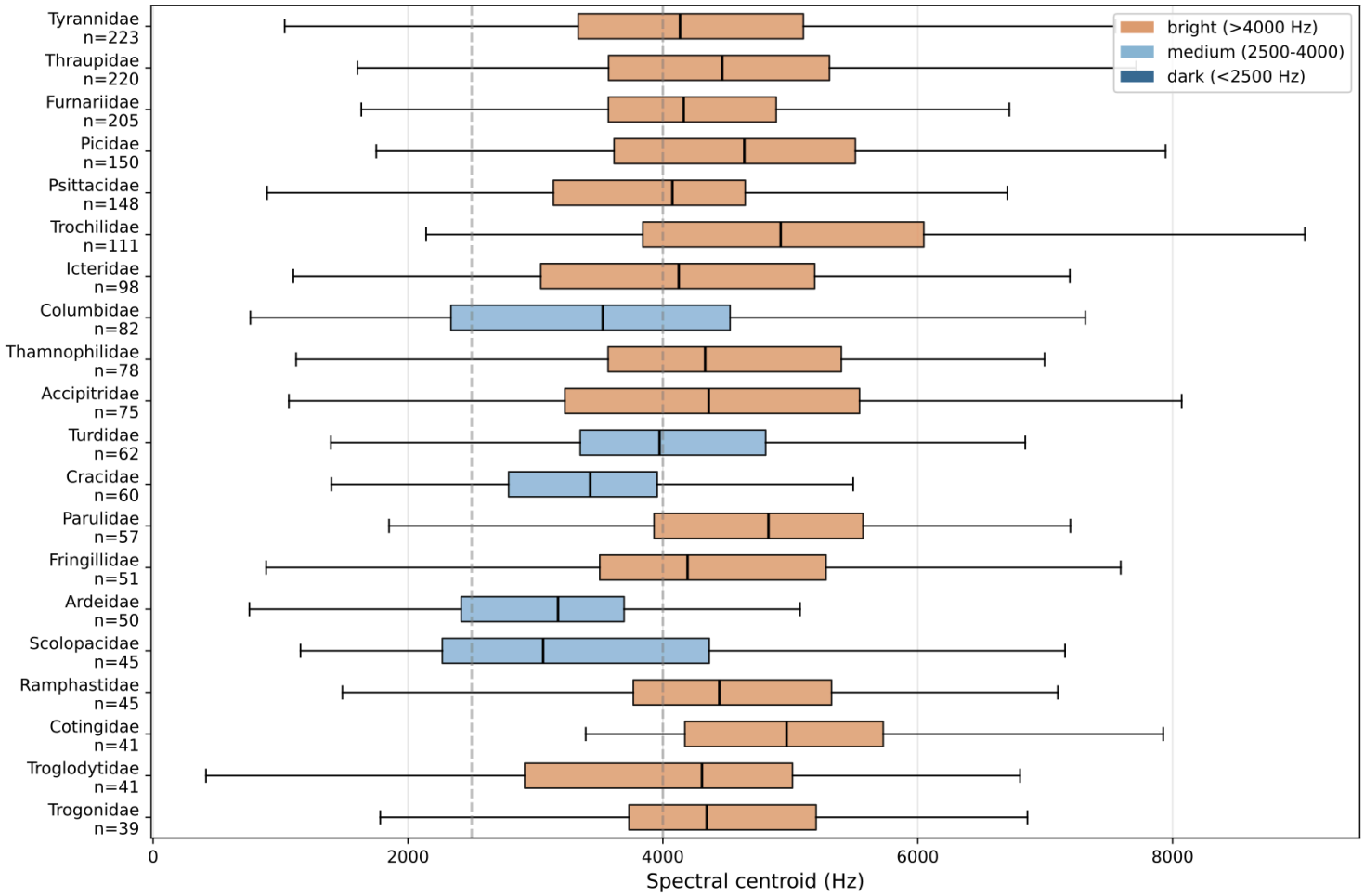


Figure 2

Spectral-centroid distributions for the 20 most represented Aves families in the 1,881-recording analysable subset. Dashed lines indicate 2,500 and 4,000 Hz. The figure describes the supplied feature subset and not the entire 11,785-recording corpus.

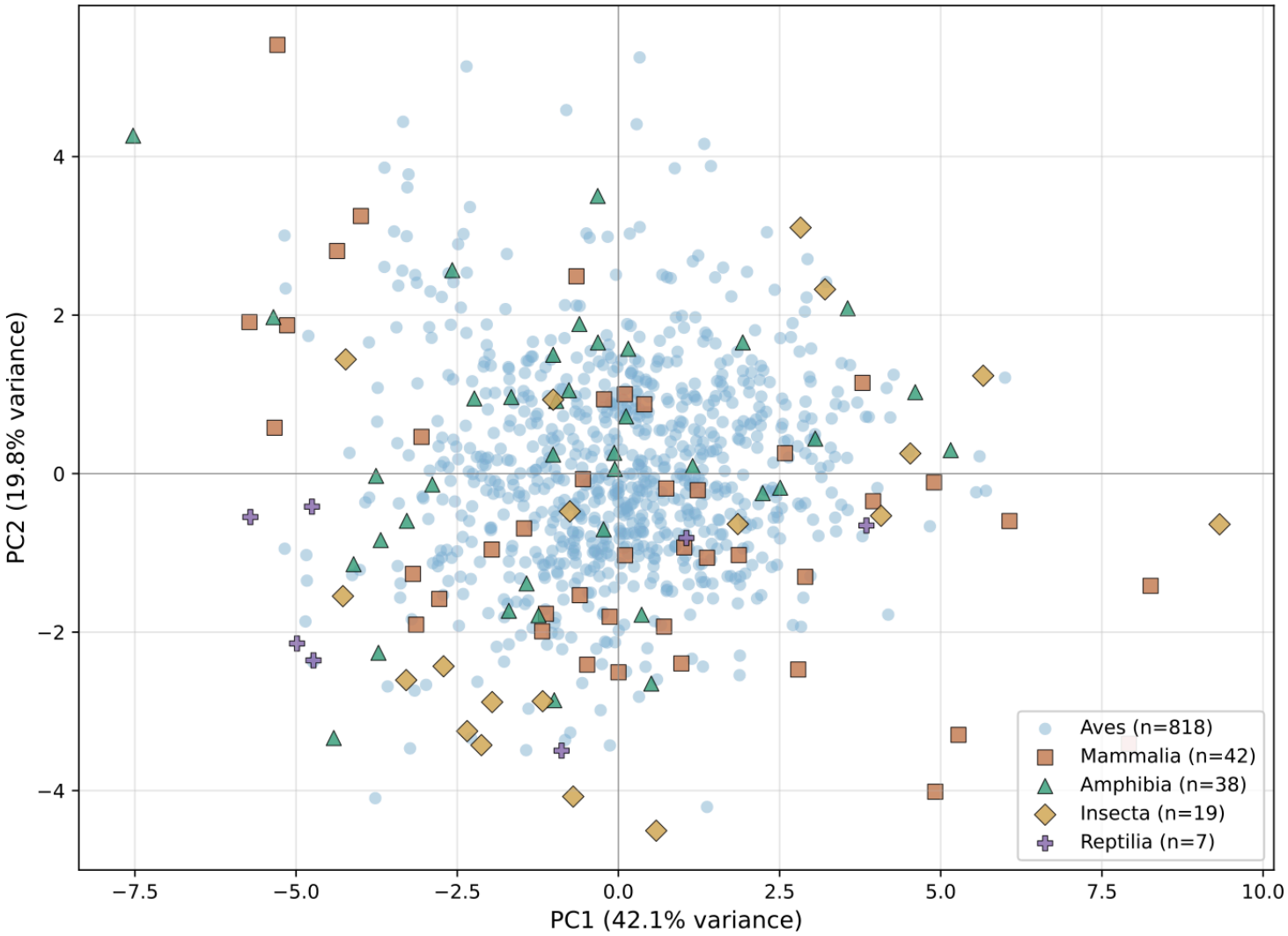


Figure 3

PCA of 11 standardised acoustic descriptors aggregated for 924 species. PC1 explains 42.1% and PC2 explains 19.8% of the variance.

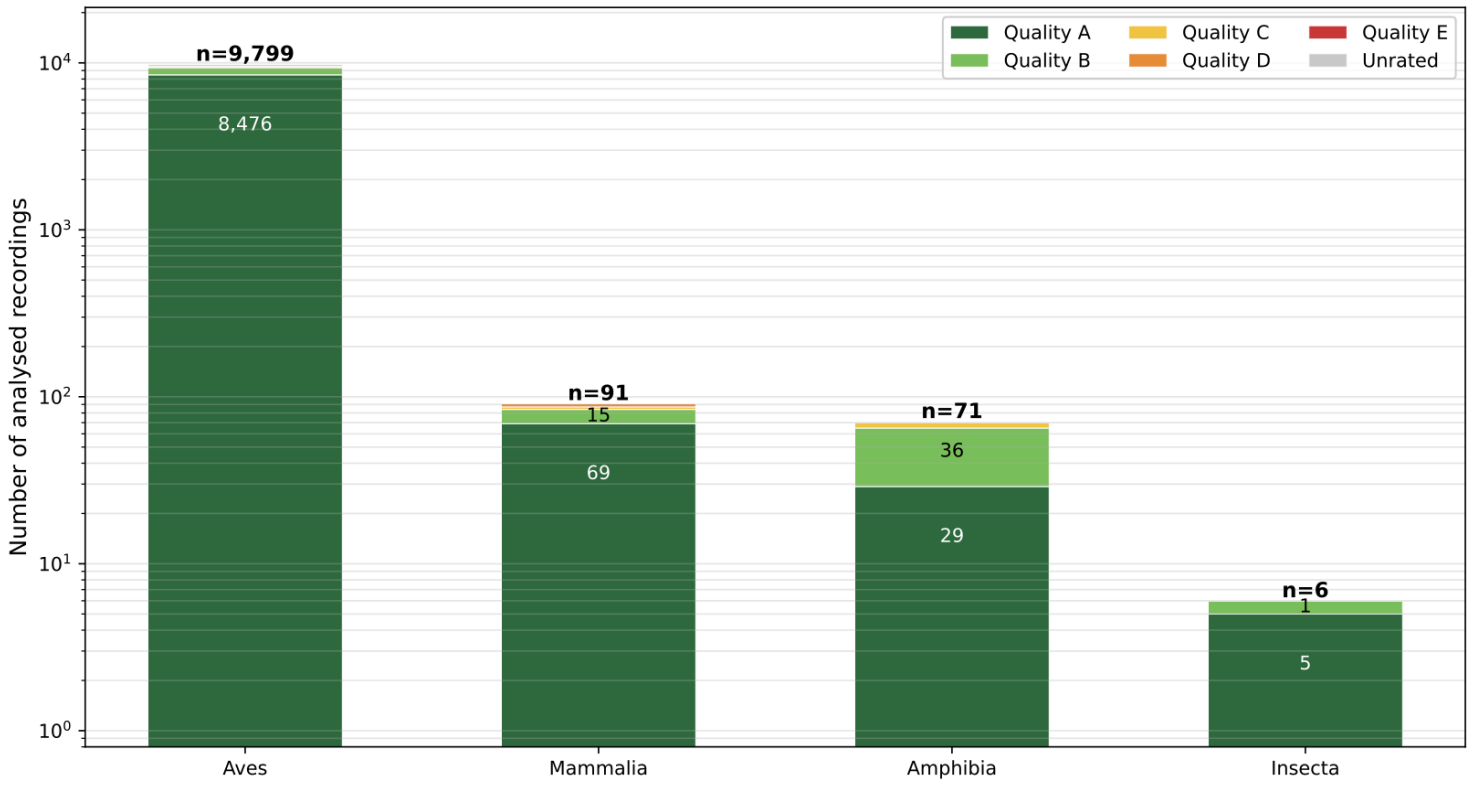


Figure 4

Xeno-Canto quality distribution by taxonomic group on a logarithmic scale. Quality A includes 8,476 Aves, 69 Mammalia, 29 Amphibia and 5 Insecta records, for a total of 8,579.

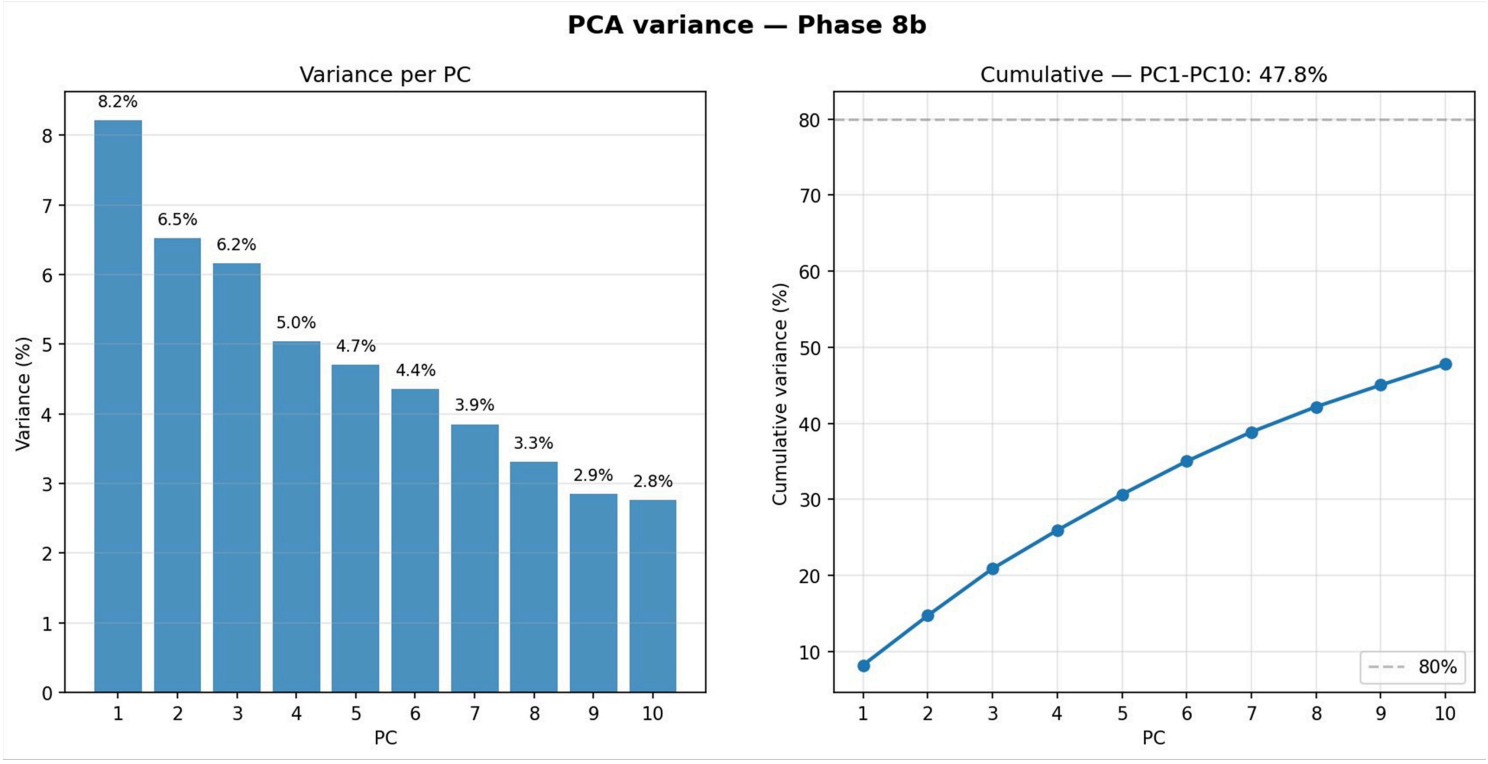


Figure 5

Variance explained by the first ten principal components in Phase 8b. PC1 and PC2 explain 14.74% together; PC1-PC10 explain 47.79%.

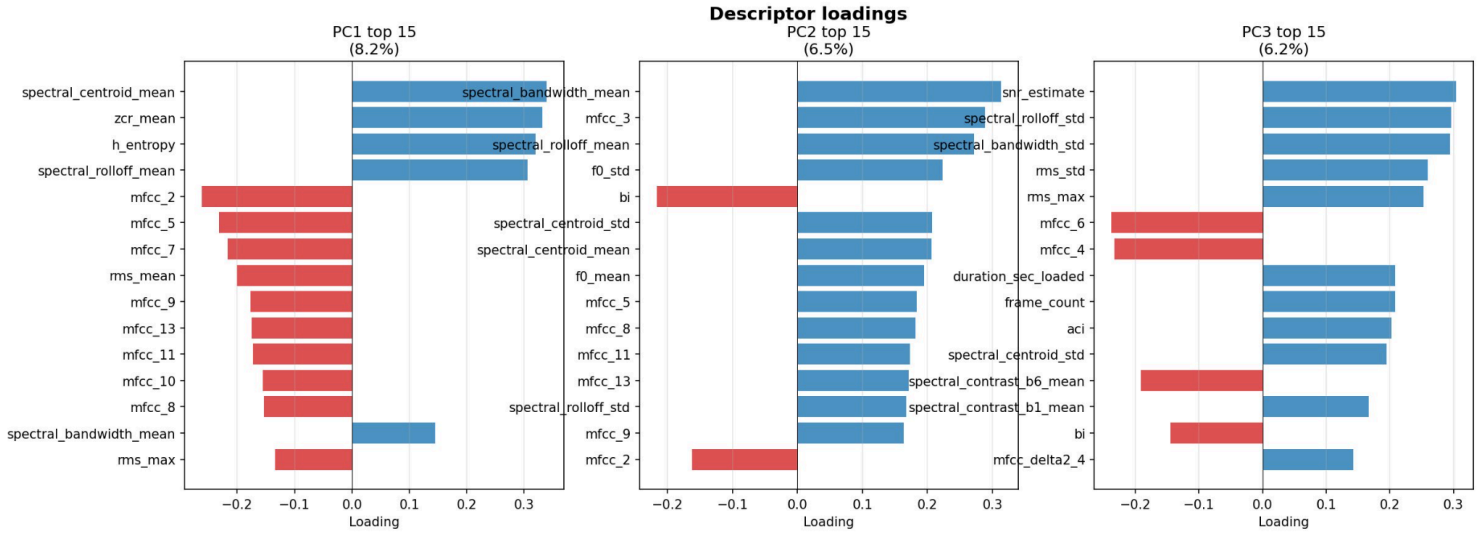


Figure 6

Largest positive and negative descriptor loadings for PC1, PC2 and PC3. The plot provides an empirical basis for interpreting the component axes.

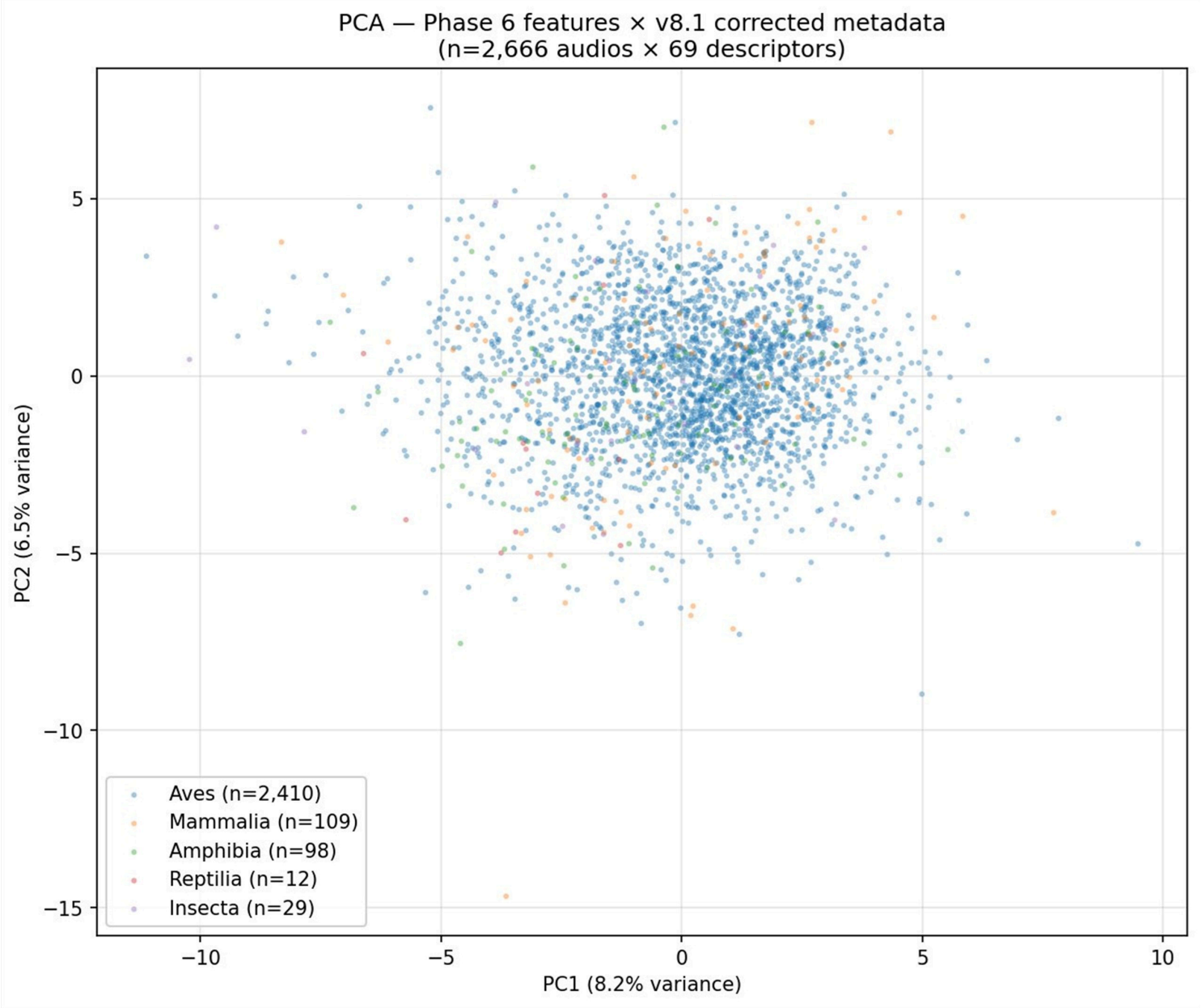


Figure 7

Phase 8b projection of 2,666 recordings in the PC1-PC2 space, classified by taxonomic group.

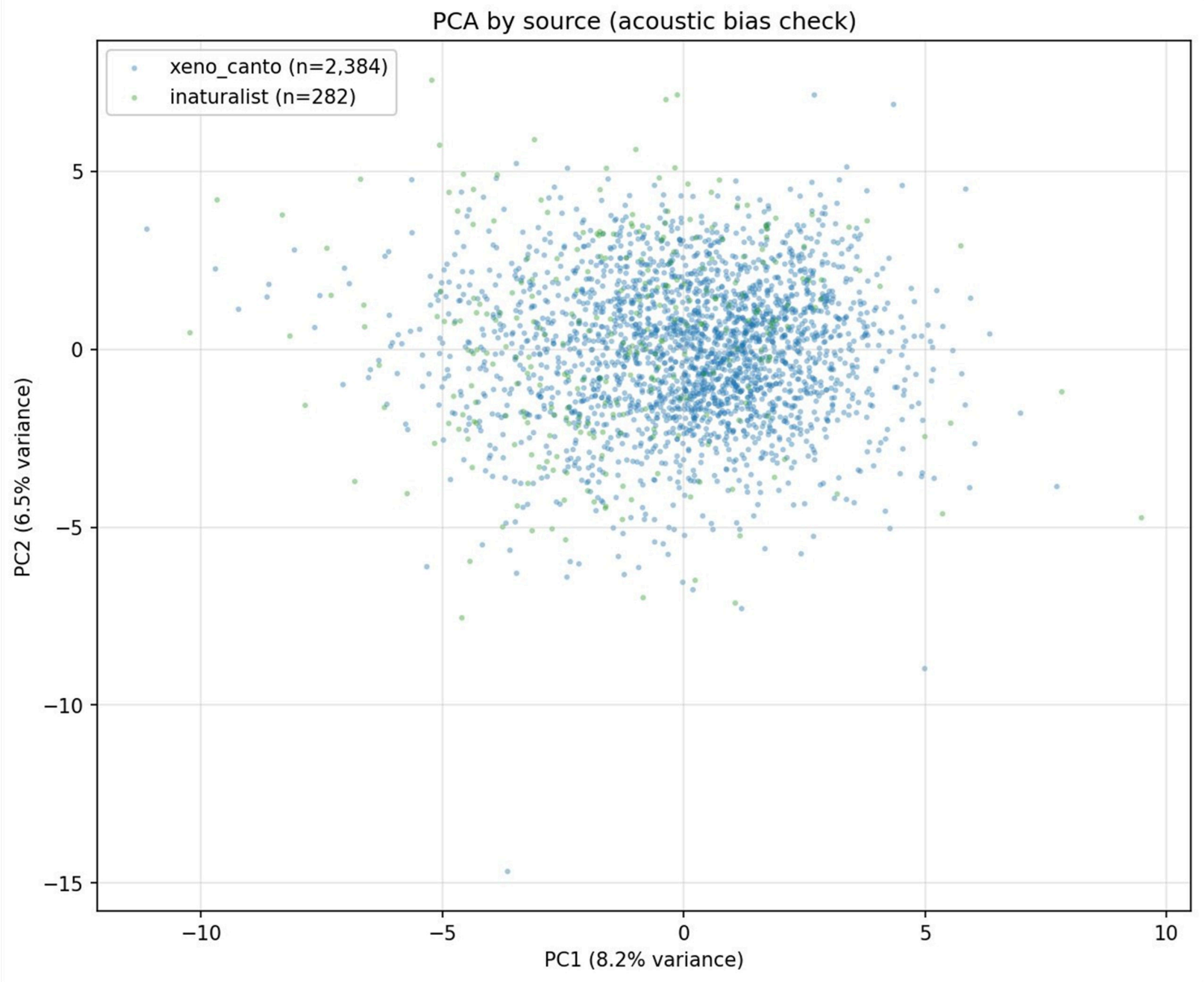


Figure 8

Phase 8b PCA classified by source. Xeno-Canto and iNaturalist overlap broadly, although source imbalance remains substantial.

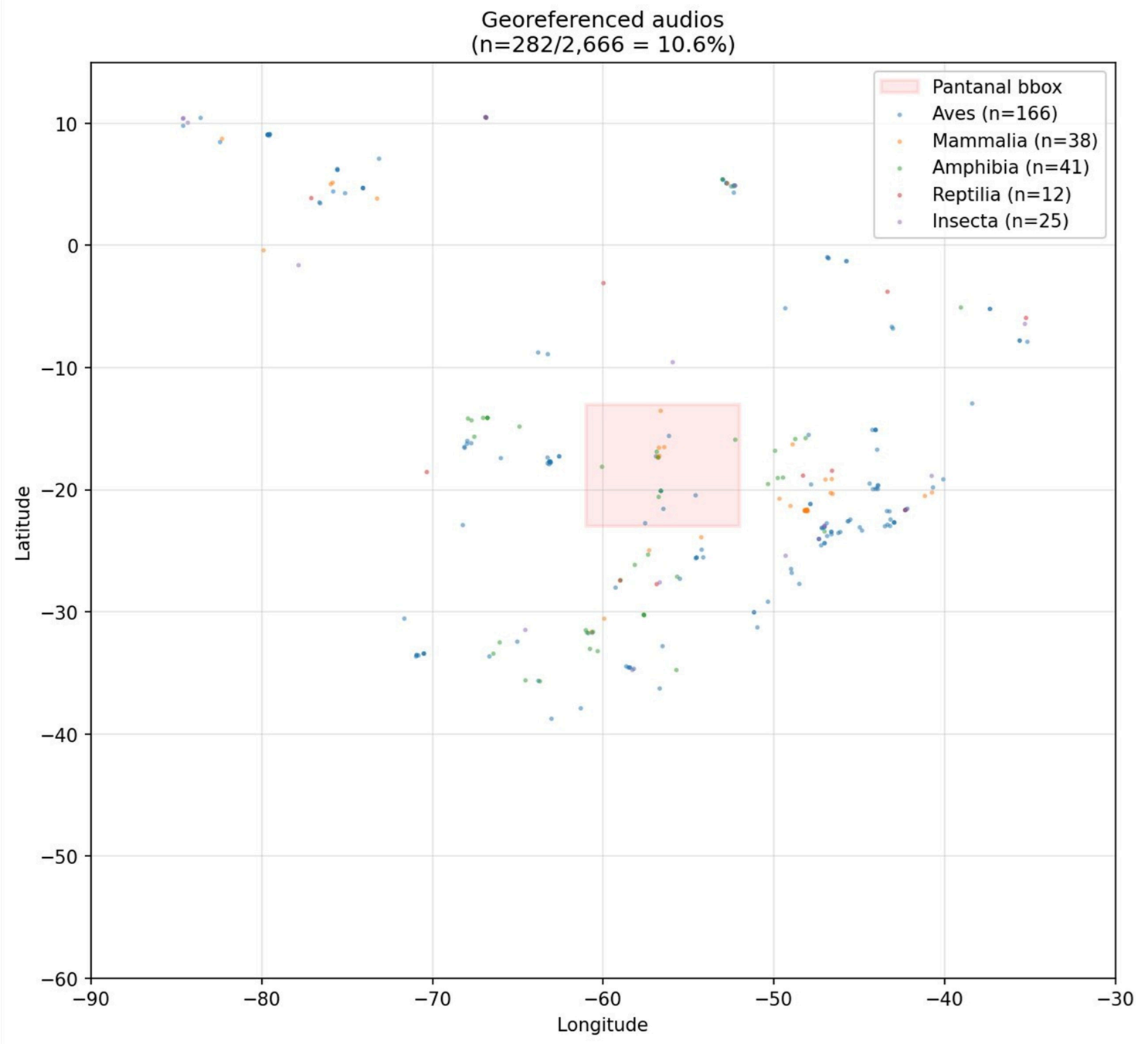


Figure 9

Geographic distribution of the 282 georeferenced Phase 8b recordings. The shaded rectangle shows the strict Pantanal bounding box; 19 plotted records fall within it.

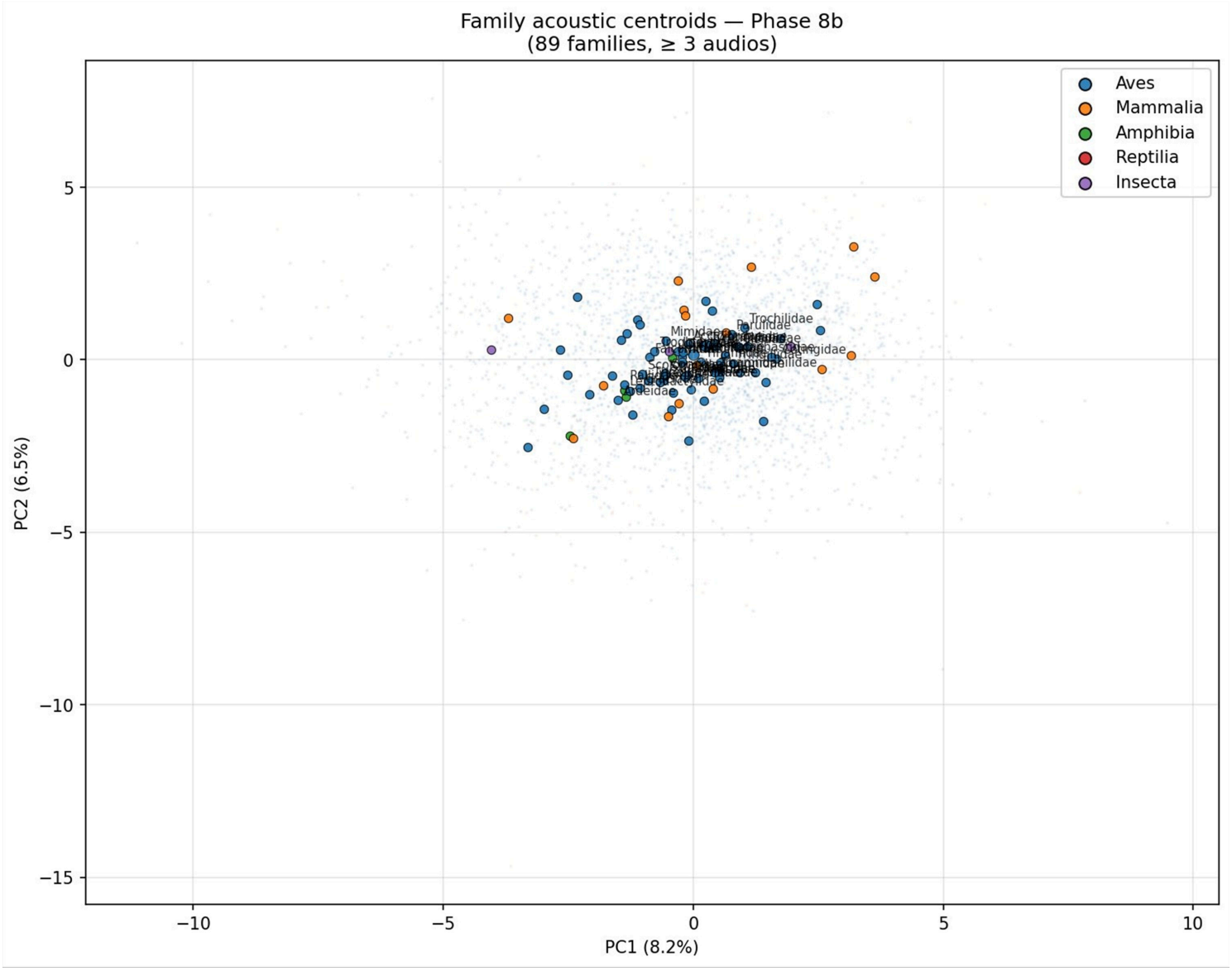


Figure 10

Acoustic family centroids in the PC1-PC2 space for 89 families with at least three recordings. The pale background points are individual recordings; labelled centroids correspond to better represented families.

Supplementary Files

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

- fig01taxonomiccomposition.png
- fig03phenologypantanal.png
- fig05iucngap.png
- fig06ndsidistribution.png
- fig07vocalnicheqvalues.png

- fig02familycoverage.png
- fig04migrationsensitivity.png