

Online Resource 1

Supplementary methods, denominator reconciliation and audit diagnostics

Companion to: Contributor and city concentration shape apparent urban differences in the call–song balance of a global sound archive

Journal: Ornithology Research | Authors: Lingjun Dai and Rongzhen Dai

Affiliations: Xi'an Jiaotong-Liverpool University, Suzhou, China; Independent researcher, Yangzhou, China

Corresponding author: Rongzhen Dai, dairongzhen3@gmail.com

S1. Data lineage and eligibility

The research object is an opportunistically assembled collection of Xeno-canto recordings. All estimates below concern the proportion of retained recordings carrying contributor-supplied call/song labels, not vocal time budgets or the incidence of true acoustic behaviours. Two retrieval routes were combined with recording-ID deduplication, retaining the targeted eleven-species retrieval when a record occurred in both original lists.

Population	Recordings	Species	Eligibility or meaning
Derived master archive	86,931	543	Unique recording IDs; 20 input columns
Binary call/song study set	66,741	227	At least 20 call and 20 song labels per species
Rural–large-city descriptive set	14,965	73	Both endpoints; no valid-month requirement
Rural–large-city adjusted-model set	14,944	73	Both endpoints and valid recording month
Species meta-gradient subset	66,119	221	Stable estimable species gradients
Contributor-mean decomposition	63,649	192	Same species for two-stage and pooled decomposition

The final 66,741 records comprise 42,764 from the targeted eleven-species source and 23,977 from the country-targeted source. The original retrieval lists contained 23,375 repeated recording IDs between them; the derived master archive contains no duplicate recording IDs. Among 86,931 master records, the stored city category matched its stored nearest-city distance and population threshold rule in all 86,931 checks. The original historical city inventory has subsequently been recovered and a population-specific alternative has been recomputed (S10); this archive sample remains non-probabilistic.

Development subset (pipeline pilot). The eleven species are *Columba palumbus*, *Turdus philomelos*, *Erithacus rubecula*, *Pycnonotus sinensis*, *Turdus merula*, *Pycnonotus jocosus*, *Parus major*, *Sturnus vulgaris*, *Cyanistes caeruleus*, *Passer domesticus* and *Corvus corone*. The archived pilot fixed-effect gradient was OR 0.984 ($p = 0.15$; $I^2 = 90\%$; five positive and six negative slopes). All eleven species occur in the full 227-species dataset, so this subset is not independent replication.

S2. Rural-large-city sample and contributor-deletion accounting

Table S1. Distinct samples and deletion scopes. The same top three contributors are selected using the adjusted model's 983 large-city records. Their record counts differ because the descriptive sample admits recordings with an invalid or missing month. Only model-based results are expressed as adjusted odds ratios.

Statistic	Descriptive composition	Adjusted model
Rural observations	13,976	13,961
Large-city observations	989	983
Total observations	14,965	14,944
Species represented	73	73
Records from top three: large-city	532	526
Records from top three: rural	182	182
Full three-contributor deletion	714	708
Only their large-city records deleted	532	526
Large-city labelled-call share	0.713	0.711
Rural labelled-call share	0.526	0.526

The adjusted model is a logistic regression of the call label on large-city exposure, species and month fixed effects, and absolute latitude, with a species-clustered sandwich covariance. The fitted OR is 1.589778 (nominal 95% CI 1.056729–2.391717), versus 1.054161 (0.718167–1.547350) after removing all 708 records from the three contributors or 1.043709 (0.708486–1.537543) after removing only their 526 large-city records. The latter is the sample-size-matched reference for random deletion of 526 high-city records, not the full 708-record removal.

The descriptive pooled rural-to-large-city call-share gap is approximately 0.187. Removing the same contributors from the 14,965-record descriptive set changes this gap to approximately 0.066. The resulting composition terms are 0.0041077 at rural rates and –0.0044556 at large-city rates. These are descriptions of different selected samples, not estimates of a counterfactual biological effect.

S3. Conditional comparison and loss of overlap

Table S2. Exposure overlap in the 14,944-record regression sample, based on recorded contributor, species, and month. A two-endpoint overlap layer does not imply balance in year, city, recording effort, habitat, or outcome. Calendar month matching does not require the same year.

Matched unit	Layers with both endpoints	Large-city recordings with rural match	Share of 983
Same contributor, species unrestricted	77	803	81.7%
Same contributor × species	78	495	50.4%
Same contributor × species × month	56	174	17.7%

Of the 78 contributor×species strata, 54 were outcome-informative and contained 413 large-city recordings; the Mantel–Haenszel (MH) pooled OR was 0.849 (nominal 95% CI 0.584–1.233). Under same-contributor×species×month matching, 31 of 56 strata were outcome-informative, involving 126 large-city recordings, and the MH OR was 1.125 (0.614–2.063). Of all 174 matched high-city records, 132 belonged to the top-three contributors. After removing those contributors, 42 large-city records retained month-matched rural counterparts, 28 of which belonged to 15 outcome-informative strata; their MH OR was 0.805 (0.320–2.030). None of these estimates calibrates dependence among strata sharing a contributor, and their target population is not the same as that of the full regression.

S4. Original eleven-species annotation inclusion audit

Table S3. Raw retrieval records classified using the archived token rule: a song-only or call-only label enters the binary analysis; mixed and other labels are excluded. Counts below are for the original eleven-species retrieval, not all 227 study species.

Urban category	Raw recordings	Mixed / other	Excluded	Call share among binary
Rural	12,144	833	6.86%	51.88%
Suburban	17,316	1,324	7.65%	52.22%
Urban	14,251	1,613	11.32%	53.05%
City core	2,156	165	7.65%	42.09%
Large-city core	935	103	11.02%	72.84%

Across the 46,802 raw targeted records: call 22,367; song 20,397; mixed 2,930; other 1,108. The unadjusted retained-binary call-share difference between the highest category and rural is +20.96 percentage points. If every mixed/other label is assigned an extreme call or song outcome as a pure arithmetic bound, that difference ranges from +9.63 to +27.51 percentage points. This is not a sensitivity bound on adjusted regression odds ratios, nor evidence of label correctness. Mixed labels may contain both calls and songs.

S5. Date-quality check on the same adjusted model

Table S4. The calendar date and month filter target different questions. Both models use the same 73 species; the stricter filter removes six rural records with malformed complete dates.

Filter	N records	Large-city N	Adjusted OR	Nominal 95% CI
Valid month field	14,944	983	1.589778	1.056729–2.391717
Valid full calendar date	14,938	983	1.588123	1.054897–2.390882

The OR difference is -0.001656 , roughly -0.10% relative to the first estimate. Across all 66,741 eligible records, 114 complete dates are unparseable; 44 lack a valid month, and these subsets overlap. No date was imputed. This only tests full-date invalidity within the fixed rural–large-city model, not broader temporality or archive selection.

S6. Two-way clustered pooled decomposition: unit accounting

The ordinary species-gradient subset contains 221 stable species slopes from 66,119 recordings; it contains 2,515 identified contributors and seven unmatched singleton groups (2,522 contributor clustering units). The decomposition eligibility rule leaves 192 species. The pooled model with species fixed effects is built on the same 192 species and their 63,649 recordings. The source recordist_map.csv identifies 2,457 distinct contributor IDs within this subset. The seven remaining recordings have missing contributor IDs and were assigned one synthetic singleton clustering label per record, so there are 2,464 contributor clustering units rather than 2,464 known people. Treating these 7 IDs as singleton groups is an analytic convention that requires acknowledgement in any count or covariance interpretation.

Pooled estimand	OR	Nominal 95% CI	Nominal p
Within-contributor deviation	0.984	0.922–1.050	0.630
Contributor mean	1.030	0.919–1.155	0.611

The between-minus-within contrast was 0.04564 log-odds units (SE 0.06338; nominal $p = 0.471$).

Two-way covariance clustering spans species and contributor units, but this pooled recording-weighted model is not the same estimand as an inverse-variance-weighted summary of 192 separate species-specific slopes. Neither can be substituted for the other as a fully adjusted causal estimate.

S7. City-classification scope and remaining provenance

The original city assignment selected one nearest city with population $\geq 100,000$ and assigned the five categories from that city's distance and population. The 48,059-row historical SimpleMaps city workbook has now been recovered and independently read back against its recorded SHA256; the archived BallTree/Haversine algorithm reproduced 86,931/86,931 stored classes. The separate nearest-city population-threshold sensitivity analysis is presented in S10. This verifies the implemented exposure algorithm, not urban built cover or historical environmental change; taxonomy-workbook version documentation remains incomplete. The original city inventory is licence-restricted and is not included in this PDF.

These alternative definitions do not isolate recording preferences from geography, season, microphone conditions or differences in archive inclusion. The results remain conditional on the selected metadata and are not an estimate of an individual bird's sound-use allocation.

S8. Reproduction status and scope

The following frozen derived files were preserved privately, and their SHA256 values are given below for checking the analysis provenance. The recovered SimpleMaps workbook, licence terms and city reclassification code were separately verified on 26 October 2026. Full source files, raw audio, contributor pseudonym maps and the licensed city inventory are not reproduced in this publicly viewable supplementary PDF. Researchers must obtain any restricted inputs lawfully.

master_recordings.csv: a1f1ac34d0138b4d7d6ae49b20c86bc621588c79d9628c3df441cab70ee75846

xc_all_species_urban.csv: f50c86cb333a9341ccb67a92ec82c0d985a0dc1e8f31d0edde655cf26445a646

recordist_map.csv: e1114f7dc759d1322d3d916d93c334dcd239e5ffaf60ab5c45eb15bdfac612df

Historical baseline overlap and date-quality audits were reproduced against the frozen master; additional 2026-10-26 scripts independently reran nearest-city assignment with separate population thresholds, fitted logistic sensitivity models and checked a contributor-city bipartite support graph. Outputs are summarized in S10–S11. These checks do not reconstruct the exact historical taxonomy workbook or a built-cover exposure series, and they do not certify reproducibility of every exploratory model from a clean installation.

S9. Targeted audio audibility review

A blinded-to-metadata, purposively stratified subset of 120 Xeno-canto recordings was prepared for the first author to check sound audibility, rather than species or call/song correctness. Forty sample entries came from concentrated contributor sampling blocks and 80 from other strata. The rater received anonymised L08-001 through L08-120 audio files and recorded the clarity rating and whether bird sound was heard. These nonrandom strata do not give a population-representative estimate of sound quality or of metadata label validity.

Table S5. First-author audibility ratings of the 120-file targeted sample. Values are descriptive counts and are not estimates of Xeno-canto label classification accuracy.

Rating	Recordings	Percent of targeted sample
Clear	115	95.8%
Moderate	5	4.2%
Unclear	0	0.0%
Unplayable	0	0.0%

Bird sounds were marked audible in all 120 recordings. The rater added optional text comments for 22 recordings, while 98 comment fields were empty. The five moderate ratings were L08-025, L08-026, L08-041, L08-053 and L08-054. This nonrandom audibility check does not independently validate the call/song metadata labels. The original review spreadsheet and source audio are not part of this supplementary PDF.

S10. Population-specific nearest-city sensitivity analysis

Using the recovered historical SimpleMaps city inventory (48,059 city records), we independently formed nearest-city indexes with population $\geq 100,000$ (6,135 cities), $\geq 500,000$ (1,437), and $\geq 1,000,000$ (746). The original single-nearest-city rule reproduced all 86,931 stored urban categories. Under independent population-threshold searches, 950 master records moved upward in category (870 urban $\rightarrow$ city core, 76 urban $\rightarrow$ large-city core, four city core $\rightarrow$ large-city core); none moved down. Eligible large-city recordings rose from 983 to 1,061 within the same 73-species, 13,961-rural-record model. The original three focal contributors continued to supply 526 large-city recordings (49.6% of the expanded group).

Table S6. City-proxy sensitivity of the adjusted rural–large-city call-label odds ratio. All intervals below are nominal Wald 95% intervals. The species $\times$ contributor covariance model changes the uncertainty estimate, not the fitted coefficient. Removing all focal contributors eliminates 708 observations in the two-endpoint model; removal of large-city records only excludes 526.

Model sample	Original city proxy: species-clustered OR (95% CI)	Population-specific proxy: species-clustered OR (95% CI)	Population-specific proxy: two-way-clustered OR (95% CI)
All observations	1.590 (1.057–2.392)	1.489 (1.007–2.200)	1.489 (0.952–2.328)
Exclude three contributors at both endpoints	1.054 (0.718–1.547)	0.996 (0.710–1.396)	0.996 (0.669–1.483)
Exclude only 526 focal large-city records	1.044 (0.708–1.538)	0.983 (0.699–1.382)	0.983 (0.657–1.470)

This reclassification uses the same historical city inventory but changes the definition of the upper urban tiers. The change in odds ratios includes 78 newly eligible recordings, not an independent population survey; the uncertainty intervals do not correct selection bias, and both definitions can misclassify ecological urban exposure.

S11. Contributor–city overlap in the original large-city stratum

Within the original 983 large-city recordings, 168 contributors and 99 nearest-city strata formed 205 occupied contributor–city links. Their bipartite graph had 63 connected components and cycle-space dimension one (cross-checked with two graph implementations). A total of 143/168 contributors appeared in only one qualifying large-city stratum and supplied 857/983 records. Three highly influential contributors supplied 526/564 records in Sevilla, Paris and Berlin; only 38 came from other contributors in those same three cities. Requiring the same city, species and calendar month to contain both focal and other contributors left only 11 such strata, with 55 and 18 recordings respectively. The graph uses only the selected large-city observations, not their rural recordings, and its cycle count is not a regression degree of freedom. These summaries measure sample connectivity and cannot separate the contributors' preferences from the effects of location.