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.

A provenance-linked dataset of Cup of Excellence award-winning coffee lots from 2016 to 2025

Zhengyang Zhang

zzy5222008@gmail.com

Independent Researcher

Data Note

Keywords: coffee, Cup of Excellence, data provenance, sensory scores, open data, reproducibility

Posted Date: November 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11669953/v1

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

Additional Declarations: The authors declare no competing interests.

Abstract

Cup of Excellence award records describe selected high-quality coffees, but heterogeneous webpages complicate reproducible comparison. This resource covers 2,711 awarded lots from 102 observed events and 127 award groups in 16 countries and territories during 2016–2025. It retains selected factual labels and values, documented normalization, field-level provenance, missingness and conflicts, while withholding organizer webpage copies and narrative text. A separate common-factor layer distinguishes reported lot attributes from environmental grid context, external experiments, derived quantities and assumptions. Versioned corrections preserve source evidence and identify changes to derived values. Structural checks, source replay within documented access limits, and regeneration of stored computational products assess consistency and reproducibility. No human source-audit observations have been recorded. Reuse remains subject to award-selection bias, score-system differences, incomplete coverage and uncertain identity links. The resource supports inspection and reuse without manufacturing missing measurements or treating external study samples as awarded lots.

BACKGROUND & SUMMARY

Cup of Excellence (COE) results and farm-profile pages document awarded coffee lots, ranks, scores and reported attributes [1]. Their heterogeneous formats and changing pages complicate comparisons across competitions. A reusable compilation must identify the awarded entity, distinguish split auction offerings from independent lots, retain conflicting evidence and separate source reports from measured attributes.

Previous COE analyses studied auction prices in 2004–2010 [2], 2,123 observations from 11 countries in 2004–2015 [3], and Mexican records inspected for 2012–2019 but restricted to 2017–2019 after eligibility checks [4]. This resource extends the documented record to 2016–2025 and supplies inspectable identities, field provenance, normalization, missingness and versioned computational records.

The cohort contains 2,711 source-backed awarded lots from 102 observed events, 127 award groups and 16 countries and territories. It includes every identified eligible award record rather than a top-rank sample, but the discovery process does not certify exhaustive historical coverage. For example, the Brazil pulped-naturals 2017 announcement reports 25 winners, whereas the recovered official table has 24 unique lots and no rank 14 [1, 5]. No missing award row was fabricated. Figure 1 summarizes observed coverage. Reported agronomic attributes, contextual environmental data and external study records remain separately identified. Stored reuse analyses are retained for reproducibility; this Data Descriptor presents no hypothesis test, location-to-score predictor or planting-suitability recommendation.

METHODS

Population and provenance

Official event results establish award status, rank and published score. Event and category identifiers distinguish award groups; auction sub-lots are linked to their award entity. National-winner, other-auction and non-awarded records remain separately classified. The event registry contains 104 discovery records, of which two candidate pages lacked a parsable scored award table. Their exclusion does not establish that no contest occurred. Dates denote contest year unless labelled as harvest, publication, retrieval, auction or measurement time.

Deterministic parsing produced relational records from inventoried source URLs. Retained provenance includes source identities, requested/resolved URLs, retrieval states, hashes and field locators where available. Repeated pages, revised pages and conflicting event/profile values remain distinguishable; a recorded precedence rule selects an analytical value. Acquisition failure is not source absence. Historical identifiers link through a versioned crosswalk, and ambiguous farm names or unmatched records are quarantined rather than joined on name similarity. The Alliance for Coffee Excellence (ACE) identifies farm and lot information as farmer-submitted [6]; these attributes are therefore farmer-reported information published by ACE/COE unless a separate measurement or verification is documented.

Frozen replay inputs include reduced document-object-model and sitemap representations and profile bodies re-fetched on 25 October 2026. They are not claimed to be untouched original HTTP responses. Four removed narrative values, comprising three history values and one sensory_overall value, cannot be regenerated from these retained snapshots. One exact profile response for the Taiwan 2025 rank-1 lot is also unavailable. Hashes identify retained objects and documented gaps; they do not recover missing content. Snapshot-dependent reconstruction and fresh network acquisition therefore have different limits.

Field selection and interpretation

Selected factual values are retained alongside normalized labels. Variety mappings preserve mixtures and do not establish genetic identity. Processing families and modification mentions do not imply measured fermentation time, temperature or inoculum. Evidence categories distinguish lot/sample measurements, reported lot/farm attributes, regional or gridded context, derived calculations, and assumed or literature-based parameters. Entity, period, units and missingness remain attached to each value; farm context is not silently converted to a lot measurement. The historical common-factor ledger contains 90 conceptual variables rather than 90 measured variables for every lot (Fig. 2).

Two source declarations cannot support reconciled numerical altitude bounds: Guatemala 2019 rank 17 reports “4500 – 600 feet”, while Guatemala 2023 rank 20 labels 1500 m as maximum and 1890 m as minimum. The corrected scientific database v9-rc1 preserves the original strings and provenance but sets the derived numerical bounds to null. This exclusion reduces the explicit-unit altitude cohort from

628 to 626 lots; affected stored model products were recomputed. A separate qualification applies to 30 explicit-metre declarations containing 33 tokens with one period followed by exactly three digits. Within this bounded set, a token such as 1.500 is interpreted as grouped thousands. The convention is recorded as an assumption and changes neither the existing numerical values nor eligibility; it does not establish source accuracy or apply to unit-absent declarations.

The distribution selection retains factual records, selected whole pre-existing ordinary sensory-label values and project-derived outputs. A retained sensory value must satisfy the documented 401-term vocabulary, at most six existing list items and 120 characters, with only case and whitespace normalization. An existing parenthesized integer from 1 to 99 is retained without assigning it an intensity, panel-size or measurement meaning. These are explicit representation rules, not a legal safe-harbour threshold. Original webpage copies, images, narrative prose and unnecessary contacts are withheld. Nulls carry release-scope provenance that distinguishes withholding from source absence and from a false or zero observation. The scope decision does not constitute blanket ACE permission or resolve all source-specific reuse rights [6].

Environmental context

NASA Prediction Of Worldwide Energy Resources (POWER) responses provide regional context for four registry coordinates from 1 January 2016 to 31 December 2025, using an explicit Coordinated Universal Time (UTC) daily boundary and Daily Point application programming interface (API) v2.10.1 [7]. The integrated nine-parameter module contains 14,612 point-day and 480 point-month records. Monthly means use available days; a complete-month mean or cumulative precipitation/solar-energy value requires all days for that variable. This batch has no missing parameter-day values. A separate ten-parameter NASA-only module adds native dewpoint responses and is not a replacement for the mixed COE database. SoilGrids 2.0 predictions are 250 m grid context with property, depth, uncertainty and conversion metadata [8]. Neither product establishes plot-level measurements or repairs uncertain farm coordinates.

Historical arrays from Anacafé's meteorological station network (Red de Estaciones Meteorológicas de Anacafé; REMA) are excluded. The Clima Café notice describes research use without granting public redistribution of those observations [9]. The retained manifest concerns four stations and 20 station-years in 2021–2025. Two rainfall summaries represented by four stored copies and all original station/API arrays are withheld; 140 series references retain metadata only. The current help page describes monthly mean wind and maximum gust in km h^{-1}, wind direction as azimuth and maximum solar radiation in W m^{-2} [10]. It does not resolve 40 historical station-metadata gaps covering channel meanings, sensor heights, calibration, relocation, sampling windows and timezone. The vientoR and vientoV meanings remain unresolved. NASA quantities cannot recover gusts, subdaily radiation maxima or original station observations. Nineteen separately assessed literature facts remain contextual citations.

The separate ten-parameter NASA continuation uses the same daily period and adds four native dewpoint responses, with 71 CSV field definitions. Dewpoint was not inferred from temperature and humidity. Identical grid-response series at some registry points are recorded as dependencies, not independent measurements. Historical relocation, equivalence to farm subplots and native grid-cell-centre identity are not established. Neither NASA module automatically links to lots or refits score models.

External study resources

External resources retain their original sample identities, experimental units, citations and licences: a water-sorption, attenuated total reflectance Fourier transform infrared (ATR-FTIR) spectroscopy and roasting workbook [11]; Fourier transform near-infrared (FT-NIR) spectra with sensory labels [12]; Chiapas soil/bean summaries [13]; published soil/chemistry/sensory summaries [14]; de-identified consumer preferences [15]; European Nucleotide Archive (ENA) run/sample metadata [16]; and numerical roast-condition or benchmark transcriptions [17, 18]. File hashes and source locators accompany parsing. Raw observations, repeated representations, summaries and derived quantities remain distinct. ENA metadata are not sequence reads or a processed microbiome matrix; consumer preferences are not COE jury scores; external measurements are not imputed into awarded lots.

The volatile-fingerprint study by Taiti et al. [19] is methodological background only. Its supplementary workbook, sample values, imported numerical records and derivatives are excluded. No sample-by-compound matrix of volatile organic compounds (VOCs) is claimed as acquired or supplied, and no Taiti measurements enter the COE score models.

Derived records and optional reuse analyses

Frozen lexical rules record descriptor occurrence, not sensory intensity, chemical concentration or consumer preference. Categories may overlap, and an absent word does not establish an absent sensation. Summaries use the historical eligible-text denominator and retain missingness and note-length qualifications. The retained analytical inputs support checks of flags and matched terms. Complete historical text-to-feature re-extraction for 2,434 of the 2,476 eligible text records requires withheld original text. The frozen lexicon and the executable label-selection masks identify this boundary.

The database also preserves previously generated selected-award regression outputs and environment, postharvest and roast calculation products. Their full specifications, tuning procedures, sensitivity analyses and numerical results are retained as optional reuse material in support/Extended_Reuse_Analysis.md and the corresponding code and machine-readable outputs. They are not used here to test new hypotheses, select a superior production method or infer a causal coffee-quality effect. The technical checks below assess regeneration of these stored objects. Physical

calculation kernels remain subject to their recorded assumptions and synthetic-test limits; empirical calibration to cup quality has not been established.

Computational validation design and AI assistance

Frozen frames support a 300-lot simple random sample from 2,711 awarded lots, a 240-record stratified sample from a historical 1,205,559-record extension frame, and a separate 593-case purposive risk census. The extension frame uses ten records without replacement in each of 24 evidence-layer/record-grain strata, using Algorithm R streams from master seed 202610252. Nine core checks and six extension dimensions specify representation, units, entity/spatial scale, time/depth, linkage and missingness where applicable. The frames, identifiers, hashes and blank entry forms are frozen for potential future audit. All reported source comparisons or semantic dispositions are computational or AI-assisted. No returned human audit form has been recorded, including for the proposed single-author blinded protocol. The checked records also informed checker revisions, so these are not an unseen evaluation set. No human-reference agreement rate, population error estimate or error confidence interval is reported.

ChatGPT and OpenAI agent tools assisted source discovery, extraction, data organization, software implementation, verification, documentation and writing. The model reported for this revision is GPT-6 Astra Pro; earlier identifiers are given only where preserved in execution records. An additional pre-submission assessment was produced with Claude; the supplied report does not identify its exact model version. Separate programs and AI agents can share assumptions and correlated errors. Their checks are computational assessments, not independent human reviewers or external laboratories. Completed human source-audit judgments and external human replications remain zero. AI tools are not authors; Zhengyang Zhang is responsible for the work, claims, attribution and publication [20, 21].

No new human-participant or animal experiment, recruitment or intervention is reported. Public award-level scores and published sensory summaries are reused as secondary source records. The separate consumer-preference resource is the public Dryad dataset of Ristenpart and colleagues [15], reused under CC0. Its record reports that personal identifiers were removed, that the University of California, Davis Institutional Review Board approved the tasting study (IRB no. 1082568), and that participants gave informed consent. These statements concern the original study. This secondary curation asserts no project-specific institutional ethics approval or exemption. Withholding source-page copies and unnecessary contact values is a distribution decision, not an ethics approval. Producer names retained from public records may identify individuals. No new participant consent, independent legal review or ethics determination is inferred from public availability.

DATA RECORDS

This article revision and review packet are v9-rc2. They describe the unchanged corrected scientific database v9-rc1, published at https://doi.org/10.5281/zenodo.24054588 [22] and derived from the v8

deposit at https://doi.org/10.5281/zenodo.24036944. COE_Review_Data_v9_RC2.zip supplies COE.sqlite, coe_data.py, README.md and Supporting_materials.zip. It retains 39 original table interfaces, 285 columns and six scientific views, with four additional reviewer entry views. Table 1 identifies the principal layers. Physical compaction preserves the corrected logical records; it does not make them identical to uncorrected v8.

Table 1
Principal data layers in the corrected scientific database v9-rc1

Layer and unit	Extent	Principal interface
Awarded lots	2,711 lots; 102 observed events; 127 award groups; 16 countries/territories	award_lots; events; award_groups
Field evidence and selection	57,972 field observations; 51,509 selection records	field_observations; field_selections
Imported evidence and provenance	1,202,242 records; 3,236 catalogue entries; 5,707 source documents; 469,392 award links	ext_records; ext_dataset_catalog; ext_source_documents; ext_record_awards
Optional reuse products	2,711 input rows; 10 stored parameter rows; 69,246 out-of-fold rows; 62 validation rows	current_model_inputs; current_model_parameters; current_oof_predictions; current_validation_metrics
NASA grid context	4 coordinates; 14,612 point-days; 480 point-months; 9 parameters	nasa_v2_points; nasa_v2_daily; nasa_v2_monthly

Identifiers preserve record grain: award_id joins current lot records; event_id and group_id identify editions and categories; historical identifiers require ext_lot_crosswalk. Repeated sources, predictions and external samples do not add awarded lots. The structural catalogue lists all current table interfaces, columns and views. Semantic documentation has narrower coverage: the source-core dictionary describes 16 tables and 111 columns, the selected-variable dictionary describes 19 fields, and factor_definitions contains 90 conceptual definitions. The historical ledger has 243,990 lot-variable positions, including 207,486 missing positions and 39 wholly missing variables in that snapshot. These are not measured samples or a certified current completeness census. Unknown meanings, units, timing and identity remain qualified; logical JSON/code relationships are not invented as SQL foreign keys.

Table 2
Distinct sensory-text populations and distribution states

State	All awarded lots	Historical eligible analytical-text subset
Population size	2,711	2,476
Absent before selection	233	Not a separate category in this subset
Fully retained	42	42
Partly retained	1,109	1,109
All original text withheld	1,327	1,325

These lot-level and analytical-text denominators are distinct. At the original profile-field level, 7,954 sensory rows comprise 1,706 retained whole values, 6,245 withheld values and three already missing values. A further 283 contact values and their dependent copies are withheld. These counts describe the enumerated projection; they do not certify coverage of every historical page or nested copy. The masks retain source-payload hashes and changed paths, with withheld_by_release_scope_not_source_missing states. The sensory projection changes 2,436 lot-variable records and no other variable in the historical 90-factor ledger.

Supporting_materials.zip separates current corrected files under COE_Review_v9/ from retained historical files under COE_Zenodo_v8/. Current code, catalogue, correction and validation paths refer to COE_Review_v9/; original source policies and audit records refer to COE_Zenodo_v8/. The existing catalog/database_*_v9_rc1 files describe the unchanged scientific database, and corrections/CHANGELOG_v9_rc1.json records the altitude changes. Historical LICENSE_SCOPE.txt and SOURCE_SCOPE_AND_ATTRIBUTION.txt retain source-specific conditions. The outer support/ directory supplies Validation_Status.json, Audit_Evidence_Index.json and Rights_Position.md. COE_Generation_Code_v9_RC2.zip separately provides the generation/reconstruction code and its reproduction manifest. Historical acceptance labels are retained as provenance of internal project decisions; they are not evidence of human validation, journal acceptance or permission from a source provider.

TECHNICAL VALIDATION

Technical validation is limited to structural consistency, traceable transformations and computational reproducibility. The current correction log identifies two altitude exclusions, a bounded punctuation convention and regeneration of dependent stored products. Raw altitude strings, field locators, award identities and scores are retained. The external-study scope, NASA context and ACE/Taiti/REMA exclusions remain unchanged. Checksums establish the identity of supplied files; row-level comparisons establish logical equivalence within explicit tolerances and runtime-field exclusions. Neither type of check establishes that a farmer report or source value is factually correct.

On 8 November 2026, regeneration of the optional stored computational products passed all 29 comparison groups. Across 1,135,992 numerical comparisons, the maximum absolute difference from the corresponding references was zero; 2,275,986 other scalar comparisons were exact. A separately written singular-value-decomposition implementation reproduced all 69,246 current out-of-fold predictions with maximum absolute difference 1.279×10^{-13} score points. The execution receipts, model reference outputs and calculation details identify the tested inputs, environments and tolerances. These comparisons support implementation agreement and preservation of stored outputs, not empirical accuracy, independent human replication or a scientific model-selection claim.

Historical reconstruction depends on restricted frozen snapshots. A 14-stage internal rebuild and a repeat on the same host reproduced the 36-table historical baseline within their recorded environment. The additional Claude assessment reported logical v8-to-v9 replay after path/hash adaptations and SQLite-dependent byte identity in the core rebuild; these remain attributed AI-review results. For this revision, tools/project_v2_to_v8.py was reconstructed from the differences between preserved v2 and published v8. Its explicit projection removes 3,317 extension records belonging to 48 datasets and records the remaining row updates and added NASA context. Applied to frozen v2, it reproduces all 39 published v8 table interfaces and six scientific views as exact typed row multisets, with no runtime fields ignored. validation/v2_to_v8_replay.json in the generation-code archive records that comparison. The physical file hashes differ. This reconstructed projection documents the release differences; it does not recover the missing original v6/v7/v8 producer or establish source accuracy. The four reviewer views were separately recreated; comparison of all 49 exposed table/view interfaces confirmed unchanged typed rows despite a different physical file hash.

The frozen core, extension and risk frames retain method-specific computational dispositions (Table 3). The core source checks compared 5,252 cells and rehashed 342 objects; risk checks compared 23,423 cells and rehashed 795 objects. Extension checks distinguish 257,668 source-value leaves from 124 wrapper/locator leaves. A disposition such as source-unknown, inapplicable or quarantined records a limitation; it does not convert an unresolved scientific meaning into a verified measurement. Human entry forms remain blank. A prior internal completion decision based on automated work does not satisfy the unperformed single-author blinded audit protocol.

Table 3
Frozen computational check frames and remaining limits

Frame	Recorded computational or AI-assisted outcome	Limit
Core sample 300 lots	2,700 representation checks; 25 initial representation issues addressed; 243 outcomes retain unknown or withheld facts	Used in checker revision; no human-reference error rate
Extension sample 240 records	480 representation/linkage checks and 960 semantic dispositions: 679 existing evidence, 188 source-unknown, 93 inapplicable	AI dispositions do not supply new source observations
Risk census 593 cases	147 quarantined; 445 retained with limitations; one source-supported display-label correction	Purposive census; does not estimate population error

Historical integration records retain 647 discrepancy dispositions and two unresolved reference occurrences, 1,944 source-specific semantic qualifications, 263 SoilGrids timing-method qualifications, and 39 nodes in three comparison contexts with unestablished generating formula or unit. These have different counting units and are not additive. For a 319-component historical binding increment, 318 components had unique bindings; one drying-duration component remains explicitly excluded from interpretation as boundary-layer thickness. A completed documentary field, automated pass or populated disposition is not a resolution of these scientific uncertainties.

External-data checks expose dependence and source-grain limits. Each FT-NIR workbook has 64 author-labelled sample groups with three labelled replicates on a 2,001-point grid from 12,000 to 4,000 cm^{-1} at 4 cm^{-1} intervals [12, 23]. All 16 representations passed recorded header/axis checks. The 192 sensory rows repeat one score within each group. In both raw green and roasted workbooks, sample 4/replicate 2 and sample 5/replicate 3 are exact duplicates at all ordinates. The cause is unestablished, and preprocessing variants do not add independent outcomes. The Chiapas workbook provides 41 variables summarized by five statistics, not the 139 individual soil–coffee pairs described in the underlying study [13, 24]. Those individual observations were not reconstructed. Other external records retain their original units, experimental grain, licence and uncertainty.

USAGE NOTES

Comparisons must preserve competition, category and year. Bridge samples and complete judge-level records are not included in the dataset, so group centring, ranks or standardized scores cannot be described as absolute cross-panel calibration. Analyses are conditional on award selection and do not identify a farm's probability of winning, causal cultivar/process effects or the expected score of an unobserved site. Uneven coverage, missingness, label quality and repeated farms limit transfer. Prediction or generalization claims require an appropriate out-of-sample design and target population.

Source coordinates and environmental grid values retain their original precision, spatial support and period; simulations remain uncalibrated to cup quality.

For read-only access and checks, use the coe_data.py wrapper and the package README. For regeneration of optional stored analysis outputs, run python coe_data.py reproduce --out NEWDIR, where NEWDIR is a new directory. The recorded execution used Python 3.12.14, NumPy 2.3.5, pandas 2.2.3 and SciPy 1.17.0 with numerical-library threads fixed to one. Generation and historical rebuild dependencies are separately recorded in COE_Generation_Code_v9_RC2.zip. Exact database-file hashes may depend on SQLite version and physical layout even when typed rows agree; the accompanying logical comparator records the schema, values and runtime metadata it compares.

Original project software uses MIT and project-controlled compilation/documentation uses CC BY 4.0 within the declared scope; upstream materials retain their own terms. Missingness, conflicts, quarantined links and withholding states should remain visible in reuse. Zhengyang Zhang receives corrections at zzy5222008@gmail.com. Corrections will be documented in a new version without silently replacing cited scientific files. Data, code and supporting provenance will be retained for at least five years after publication of the associated article, subject to the existing restrictions on excluded source materials.

Declarations

DATA AVAILABILITY

The corrected scientific database v9-rc1 and the v9-rc2 generation-code archive are publicly available at https://doi.org/10.5281/zenodo.24054588 [22]. The deposited files are COE_Review_Data_v9_RC2.zip and COE_Generation_Code_v9_RC2.zip. The published v8 predecessor remains available at https://doi.org/10.5281/zenodo.24036944. Access-pending, unpublished-status and publication-pause statements in the sealed supporting inventories describe the packaging-time state; the published record supersedes those operational statements. Original webpage prose, Taiti numerical payloads, historical REMA observations and restricted replay snapshots remain excluded. The documented scientific and rights limitations remain applicable; public deposition does not make restricted source replay possible. This deposit supplies data and code access for editorial assessment under the journal’s repository and data-sharing policies [25, 26].

COMPETING INTERESTS

The author declares no competing interests.

FUNDING

This research received no external funding.

AUTHOR CONTRIBUTIONS

Zhengyang Zhang conceived the project, defined its research objectives and publication scope, directed project priorities and supervised the research workflow. He directed manuscript preparation and revisions and is responsible for the data, code and article. Software implementation, computational checking and drafting used the AI assistance disclosed in Methods. No personal source-audit observations are attributed to the author.

CODE AVAILABILITY

COE_Generation_Code_v9_RC2.zip is publicly available with the data at https://doi.org/10.5281/zenodo.24054588 [22]. It contains historical/core/run_rebuild.py, historical/v2/candidate/rebuild_all.py and retained acquisition/parsing code; current altitude-correction, model-refresh and finalization scripts; compaction code; and model_reuse/. The tools/ directory provides project_v2_to_v8.py, add_reviewer_views.py and logical_sqlite.py. The manifest identifies original versus reconstructed code and the tested environments. Preserved historical scripts retain documented path and restricted-input requirements; their inclusion does not promise a turnkey public raw-source rebuild. The README gives executable entry points and states which checks were run. The data package additionally supplies coe_data.py and the unchanged original model/reuse engines with their corrected reference products. The documented current-reuse and projection commands write fresh outputs separately from their input database. Original project software is supplied under MIT; dependencies and upstream materials retain their own terms.

References

1. Alliance for Coffee Excellence and Cup of Excellence Official competition results and farm-profile sources. https://allianceforcoffeeexcellence.org/. Individual source URLs, recorded retrieval context and field locators are retained in the database source and field-observation records and related provenance documentation
2. Wilson AP, Wilson NL (2014) W. The economics of quality in the specialty coffee industry: insights from the Cup of Excellence auction programs. Agric Econ 45(S1):91–105. https://doi.org/10.1111/agec.12132
3. Traore TM, Wilson NLW, Fields D III (2018) What explains specialty coffee quality scores and prices: a case study from the Cup of Excellence program. J Agric Appl Econ 50(3):349–368. https://doi.org/10.1017/aae.2018.5
4. Gumecindo-Alejo AL et al (2021) Factors related to coffee quality, based on the Cup of Excellence contest in Mexico. Coffee Sci 16:e161887. https://doi.org/10.25186/.v16i.1887

5. Associação Brasileira de Cafés Especiais. Produtores de Patrocínio (MG) e Domingos Martins (ES) vencem o Cup of Excellence – Brazil 2017. 23 (2017) October https://www.bsca.com.br/produtores-de-patrocinio-mg-e-domingos-martins-es-vencem-o-cup-of-excellence-brazil-2017/. The retained discrepancy record is accessible in the v8 database table ext_records at dataset_id = 'data/Source_Discrepancies_v8.json' and source_locator = '/1'; this locator does not imply a standalone file with that name. A readable export is supplied as data/source_scope_discrepancies.json
6. Alliance for Coffee Excellence (2026) Acronyms: farm-information notes. https://allianceforcoffeeexcellence.org/acronyms/. Accessed 5 November
7. NASA POWER Referencing guidance and provider-managed data listing. https://power.larc.nasa.gov/docs/referencing/; https://registry.opendata.aws/nasa-power/
8. Poggio L et al (2021) SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty. SOIL 7:217–240. https://doi.org/10.5194/soil-7-217-2021. Product documentation: https://docs.isric.org/globaldata/soilgrids/
9. Asociación Nacional del Café Política de privacidad Clima Café, effective 17 September 2025. https://www.anacafe.org/pol%C3%ADtica-de-privacidad-clima-caf%C3%A9/ (accessed 6 November 2026).
10. Asociación Nacional del Café (2026) Ayuda de la Red de Estaciones Meteorológicas de Anacafé, version 2.1.2. https://meteorologia.anacafe.org/ayuda.html(accessed 6 November
11. Collazos-Escobar GA, Bahamón-Monje AF (2025) Gutierrez Guzman, N. Statistical dataset on water sorption isotherms in roasted specialty coffee (Coffee arabica L.). Mendeley Data, version 2. https://doi.org/10.17632/6dnw4dymkp.2
12. Collazos-Escobar GA, Morales-Angulo EM, Bahamón Monje AF (2025) & Gutierrez Guzman, N. Fourier Transform Near Infrared (FT-NIR) spectra and sensory scores in green and roasted specialty coffee for machine learning-based quality monitoring. Mendeley Data, version 3. https://doi.org/10.17632/nz2fr76trm.3
13. Morales-Ramos V (2020) Characteristics of soils and coffee beans from the State of Chiapas, México. Mendeley Data, version 2, https://doi.org/10.17632/s5twjrhg7n.2
14. Viana GS et al (2025) Association between higher coffee quality, bioactive chemical profile and sustainable practices. Front Agron 7:1645329. https://doi.org/10.3389/fagro.2025.1645329
15. Ristenpart W, Cotter AR, Guinard J-X (2023) Consumer preference data for black coffee. https://doi.org/10.25338/B8993H. Dryad
16. Kutos S, Bennett RE, Santos D, Botero-Delgadillo E, Muletz-Wolz CR (2025) Soil and cherry bacterial communities predict flavor on coffee farms. Sci Rep 15:19387. https://doi.org/10.1038/s41598-025. https://www.ebi.ac.uk/ena/browser/view/PRJNA1200072 -03665-6; associated public ENA records
17. Alstrup J, Petersen MA, Larsen FH, Münchow M (2020) The Effect of Roast Development Time Modulations on the Sensory Profile and Chemical Composition of the Coffee Brew as Measured by NMR and DHS-GC-MS. Beverages 6(4):70. https://doi.org/10.3390/beverages6040070

18. Al-Shemmeri M, Fryer PJ, Farr R, Lopez-Quiroga E (2025) Batch-Scale Simulation of Heat and Mass Transfer of Coffee Roasting in Spouted Bed Roasters. Beverages 11(6):162. https://doi.org/10.3390/beverages11060162
19. Taiti C, Vivaldo G, Mancuso S, Comparini D, Pandolfi C (2025) Volatile organic compounds (VOCs) fingerprinting combined with complex network analysis as a forecasting tool for tracing the origin and genetic lineage of Arabica specialty coffees. Sci Rep 15:13709. https://doi.org/10.1038/s41598-025-97162-5
20. Nature Portfolio Artificial intelligence policies. https://www.nature.com/nature-portfolio/editorial-policies/ai. Accessed 25 October 2026
21. Scientific Data (2026) Editorial & Publishing Policies. https://www.nature.com/sdata/policies/editorial-and-publishing-policies. Accessed 7 November
22. Zhang Z (2026) A provenance-linked dataset of Cup of Excellence award-winning coffee lots from 2016 to 2025 (v9-rc1). Zenodo https://doi.org/10.5281/zenodo.24054588
23. Bahamón-Monje AF, Morales-Angulo EM, Collazos-Escobar GA (2025) Gutiérrez-Guzmán, N. Large dataset on Fourier transform near infrared (FT-NIR) spectroscopy of green and roasted specialty coffee: Preprocessed infrared spectra and sensory scores for machine learning-based quality monitoring. Data Brief 60:111609. https://doi.org/10.1016/j.dib.2025.111609
24. Morales-Ramos V et al (2020) On the soil–bean–cup relationships in Coffea arabica L. J Sci Food Agric 100:5434–5441. https://doi.org/10.1002/jsfa.10594
25. Scientific Data Submission guidelines. https://www.nature.com/sdata/submission-guidelines. Accessed 8 November 2026
26. Scientific Data (2026) Data repository guidance and data policies. https://www.nature.com/sdata/policies/repositories; https://www.nature.com/sdata/policies/data-policies. Accessed 5 November

Figures

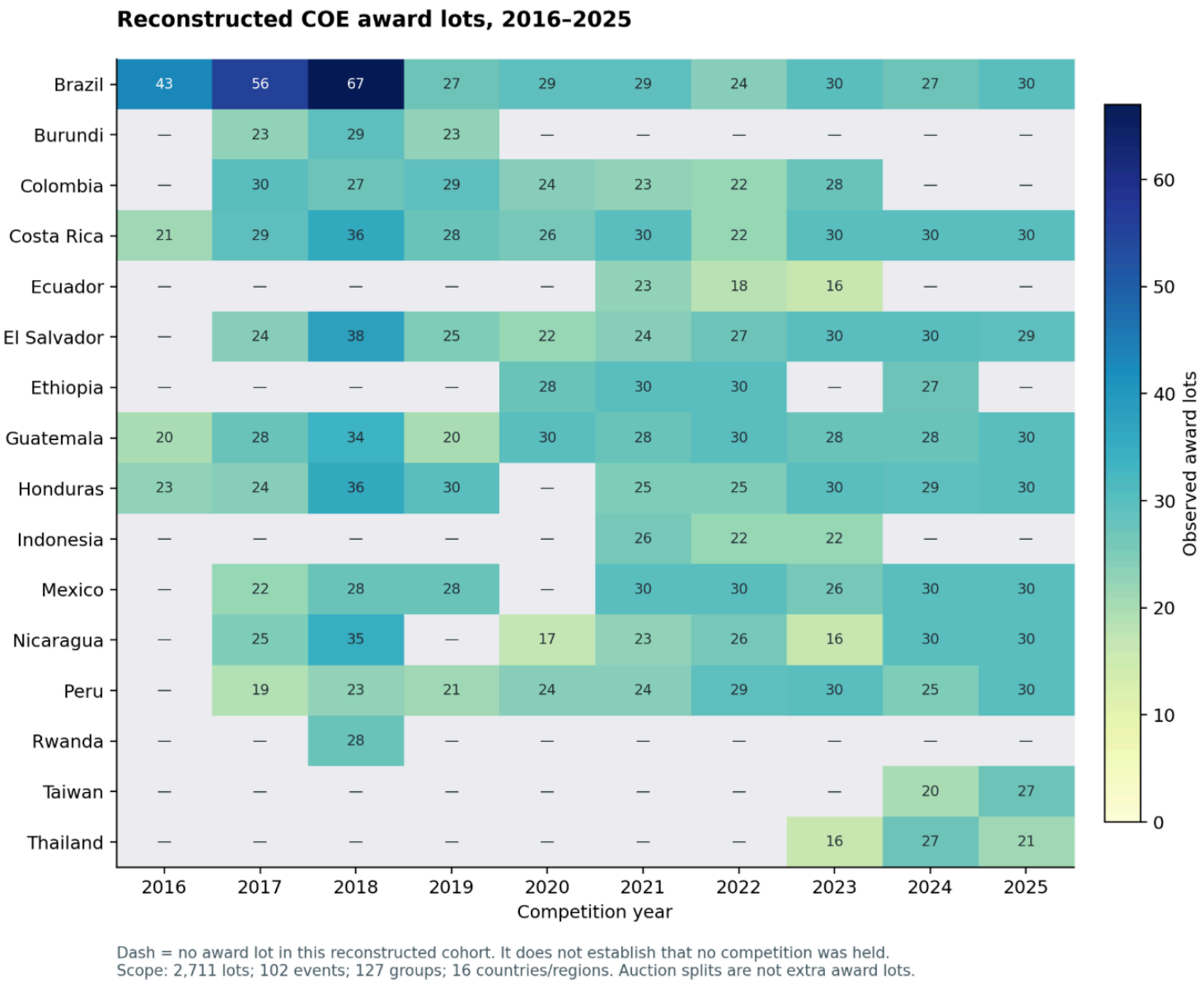


Figure 1

Observed awarded-lot coverage by country and year. An unobserved cell does not establish that a contest did not occur; unknown coverage and confirmed non-occurrence remain distinct. The 2,711-lot coverage is unchanged by the altitude corrections. The original figure is COE_Zenodo_v8/manuscript/figure_assets/main_figure_1.png in the supporting archive.

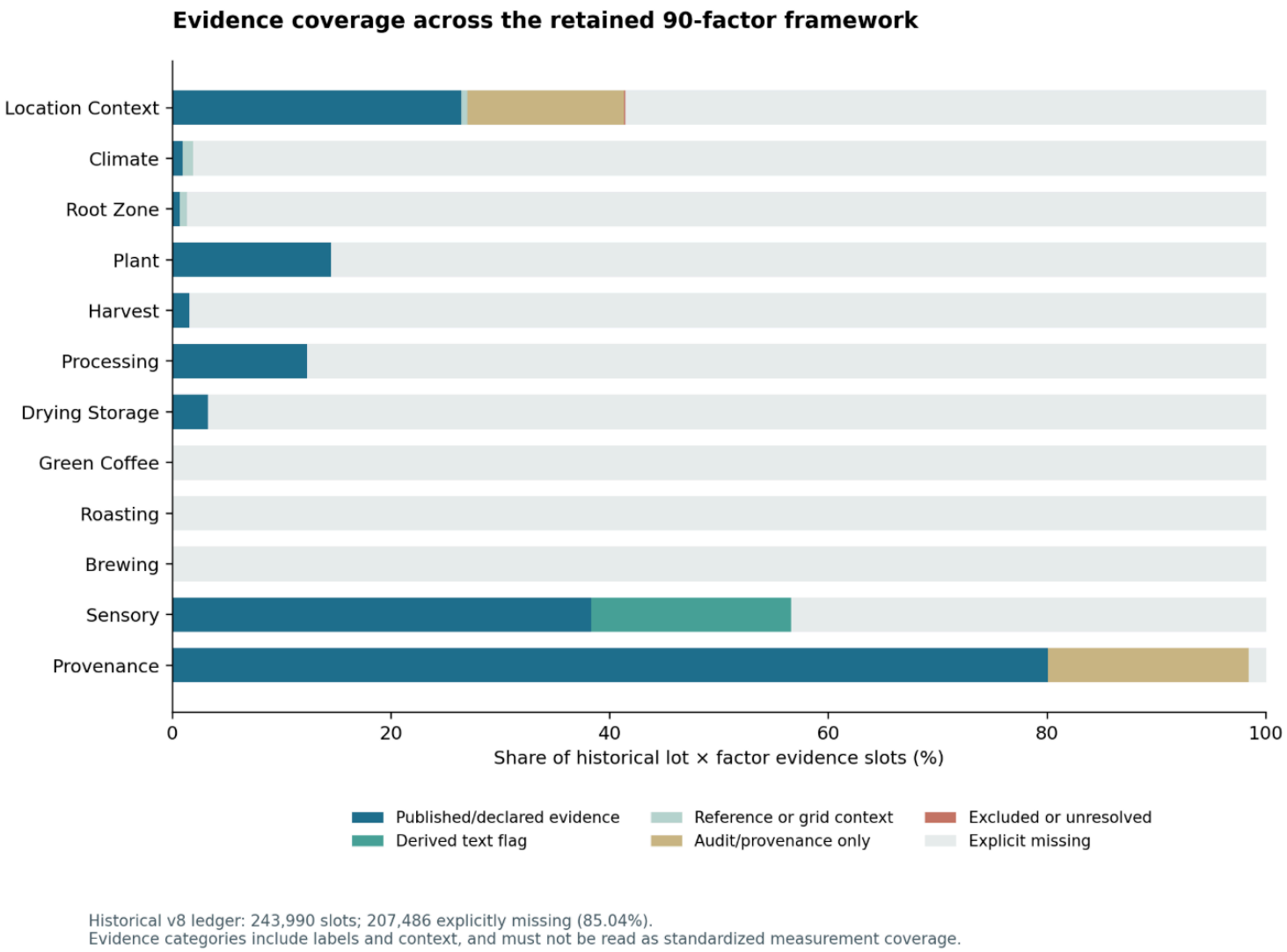


Figure 2

Historical common-factor evidence states. The ledger distinguishes reported attributes, contextual values, derived quantities, assumptions and missing information. It does not represent standardized measurement completeness; environmental context is not a new direct farm measurement. This historical ledger is unchanged. The original figure is COE_Zenodo_v8/manuscript/figure_assets/main_figure_2.png in the supporting archive.