

The Shape of Taking

An event register, a measured trait space, and the limits of automatic taxonomy in the world history of land-taking and removal, c. 6500 BCE to the present

Craig Talbert

CONTENTS

- Abstract
- 1. The comparative question
- 2. What counts as an event?
 - 2.1 Worked comparisons
- 3. Register, sources and measurement
 - 3.1 The research universe and its limits
 - 3.2 Sixty dimensions and two kinds of missingness
 - 3.3 AI-assisted coding and agreement
 - 3.4 Analysis and sensitivity design
- 4. Results
 - 4.1 Current release: equal factor counts, changed relationships
 - 4.2 Chronology addendum and unavailable alternatives
 - 4.3 What the earlier diagnostics establish
- 5. From measurement to the public atlas
 - 5.1 Families are reviewed filing categories
 - 5.2 Eight public traits are descriptive summaries
 - 5.3 The atlas is an editorial sample
- 6. Limits and implications
- AI contribution and disclosure
- Data, source inventory and process history
- References
- Background reading

Working paper v2.19 — 15 September 2026. Not peer reviewed. This edition retains the argument organized around the completed statistical release `second_pass_v7_676` and updates the companion atlas and bibliography counts. It does not introduce a new statistical fit. The earlier v1.6 manuscript and process history remain available, including the superseded claims, corrections and dated diagnostics.

Research snapshot. The release contains 676 active modeled events, 672 reviewed family assignments and four family evidence holds. The companion atlas snapshot contains 438 authored cases and 448 playback stops. The remaining 238 events comprise 176 ordinary authoring candidates, 58 separate authoring holds and four family holds. A modeled event is not automatically ready for public authoring. The publication banner identifies this edition's build and whether it is an unpublished preparation.

AI disclosure. AI research agents carried out substantial source research, coding, analysis, verification, writing and implementation under the author's editorial direction. The agreement tests below measure consistency among coders, not accuracy against an expert gold standard. Independent verification in this project generally means another AI agent with a declared assignment; it does not mean external human peer review. The full disclosure appears below.

Original text and graphics © 2026 Craig Talbert, licensed under CC BY-SA 4.0 International. Third-party quotations and reproduced material retain their own rights. Structured research data are separately CC BY 4.0 and original software MIT. See the reuse guidance and exceptions.

Abstract

Land-taking and forced removal recur across widely separated historical settings, but the categories used to compare them often combine motives, means, settlement patterns and outcomes. This working paper describes a source-bound register of 676 events measured on sixty dimensions drawn from five disciplinary traditions. It asks what patterns are visible across those dimensions, how much they depend on event boundaries and evidence, and which can responsibly be presented in a public atlas. The current canonical analysis retains nine exploratory factors; declared sensitivity scenarios retain seven to nine, and agreement in factor count does not imply unchanged relationships. No tested automatic partition is robust enough to serve as the atlas's public taxonomy. The resulting architecture separates bounded historical events, eight descriptive trait profiles and six source-reviewed editorial families. Worked comparisons show why an event is neither a fixed period of time nor every separately named episode in a rich archive. AI-assisted coding made the register feasible but leaves material uncertainty: cross-model agreement was 76% exact on a thirty-event overlap and only 50% on the distinction between unknown and structurally inapplicable. The register supports comparison of documented positive cases. It has no representative denominator of all human encounters and cannot establish universal propensity, inevitability, or the rarity of peaceful alternatives.

1. The comparative question

Some historical processes move people onto land; others move people off it; many do both. Settler-colonial studies emphasizes the permanence of incoming populations (Wolfe 2006; Veracini 2010). Genocide and forced-migration studies centers destruction and flight (Lemkin 1944; Naimark 2001; Mann 2005). Political economy examines land and labor regimes (Nieboer 1900; Domar 1970; Acemoglu, Johnson, and Robinson 2001), historical demography and archaeogenetics measures population turnover (Haak et al. 2015; Reich 2018), and

accounts of state formation examine sovereignty, legibility and borders (Scott 1998; Tilly 1990; Maier 2016). Each tradition identifies consequential features that another may leave implicit.

The question is whether measuring these features together yields a useful comparative description without letting a familiar label decide the answer in advance. A deportation followed by settlement, a conquest that leaves most residents in place, and clearance for a dam can all involve forced removal; their incoming populations, institutions and territorial outcomes differ. Conversely, histories described as colonization can differ in their use of violence, labor and property law. Categories should not silently double as rankings of moral severity.

The historical claim here is recurrence with meaningful variation. It is not a claim that every society takes land whenever given the opportunity. The register selects documented qualifying cases rather than sampling taking and coexistence from a shared universe of opportunities. Adding cases improves coverage within that scope; it cannot create the missing comparative denominator.

The companion atlas, *Taking their land*, presents one chronological playthrough. Its analytical design separates three products that are easy to confuse: an event register whose boundaries require historical evidence; a measurement instrument whose patterns are exploratory; and a public filing system whose labels are editorial judgments. The purpose of the paper is to make those distinctions usable and their limitations testable.

2. What counts as an event?

An event is a linked territorial process with sufficiently continuous actors or program, affected population and land, mechanism, geography and chronology to support one analytical profile. Direct land taking or forced territorial removal is sufficient for inclusion when the evidence supports the boundary. Substantial demographic replacement or absorption may also qualify without proof of violent invasion. Settlement on demonstrably uninhabited land does not by itself establish a prior-population land transition.

The rule distinguishes scope from certainty. A source can establish migration without establishing displacement, or establish a qualifying removal without showing that several episodes form one event. A proposed successor needs its own bounded action, affected people or tract, evidence for an independent profile and an explicit relationship to the parent. A parent and its children cannot both carry weight for the same history. Unresolved cases remain holds; their evidence and exclusion reasons are retained.

Calendar span is a diagnostic, not the definition. A long interval can belong to one implementing institution; a short war can contain several independent programs. Four distinct quantities must remain separate: the event's calendar envelope, the precision of its dates, the tempo of the processes inside it, and the observation horizon appropriate to each coded outcome. Subtracting start from end measures an envelope, not the speed or duration of migration.

2.1 Worked comparisons

One program with phases: Veii and Three Gorges. The Veii row covers 396–387 BCE because conquest, captive sale, household land assignment, selective citizenship and tribal reorganization form one territorial program. Its cohorts and phases still matter; keeping one row does not erase them. The much later Three Gorges resettlement also remains one event despite staged implementation across jurisdictions: one state project, approved plan, inundation zone, funding system and irreversible land outcome connect the movements. Shared

duration is not what makes these cases comparable. The relevant comparison is the continuity of the territorial program.

A reversal creates a new event: Melos. The Athenian destruction and settlement of 416 BCE and the restoration of 405 BCE remain separate. Authority, affected cohort, movement direction, coercive act and land outcome reverse. Their broad parent stays inactive. In the earlier boundary diagnostic, substituting that parent for the two children moved the forced-eight factor space by 10.81 degrees and the corrected nine-factor space by 2.92 degrees. The historically supported split is therefore not described as weight-neutral. Its statistical consequence is reported rather than used to choose a more convenient history. The exact event decisions and comparisons are retained in the earlier fairness account.

A long institution and a short outbreak. The Habsburg Military Frontier remains one 1522–1881 event because a named land-for-service institution links its changing districts and populations. The four-day Osh outbreak also remains one event because the spread into Jalal-Abad belongs to a contiguous regional mobilization and displacement episode. Neither long duration nor short duration settles the boundary. By contrast, the Balkan, Croatian and Bosnian war wrappers contain independently organized programs, target populations and territorial purposes that require disaggregation.

Migration evidence does not establish every territorial claim. Malhi et al. (2008) support a small proto-Apachean movement into the Southwest and extensive admixture. The reviewed passages do not establish one continuous c. 1300–1700 displacement event. The broad Athabaskan row therefore remains an inactive evidence hold. Separately researched Sobaipuri-O'odham and Salinas relocations passed their own admission gates. Conversely, the absence of proof of coercive invasion does not disqualify directly supported prehistoric population replacement when the declared demographic-expansion rule is met. Mechanism uncertainty and event scope are separate judgments.

These comparisons illustrate two fairness obligations. Boundary fairness asks whether comparable reasoning defines the rows and whether any history is counted twice. Measurement fairness asks whether each dimension uses evidence and anchors appropriate to its era, including unknown values when the record cannot answer. Neither obligation is satisfied by making every row cover the same number of years.

3. Register, sources and measurement

3.1 The research universe and its limits

The project began with a country-and-territory audit of the atlas. That audit closed 254 geographic reviews, including six source-backed negative findings, and exposed both omissions and overbroad existing cases. Modern countries are geographic containers in that process; historical actors are named states, institutions, armies or population streams. The register does not attribute collective responsibility to present-day populations.

Subsequent searches sought gaps across regions, eras and kinds of removal. Examples included Soviet dekulakization beyond an earlier emphasis on ethnic deportation (Viola 2007; Polian 2004), Yaqui deportations (Hu-DeHart 1974), and the Argentine Chaco conquest (Gordillo 2004). Ancient expansion relied on source-ledger work, including Neo-Assyrian royal inscriptions (Oded 1979), Roman narrative and antiquarian evidence interpreted alongside Salmon (1969), and the Hellenistic settlement catalogues, including Cohen (2013). Classical passages and editions are recorded with the relevant case evidence; those author names are not interchangeable citations for every ancient event.

The active v7 release has 676 events. Its wider metadata register retains 95 inactive rows, for 771 records in total. The final correction cycle replaced 13 active parents with 14 bounded successors. Sixty-dimension coding of the new rows and bounded corrections to retained rows preceded fitting. Earlier release inputs remain immutable. This is a selected, source-uneven corpus; documented searches and negative findings improve accountability without proving worldwide saturation.

The bibliography follows the same distinction between evidence and counting. The August inventory's 854 URL records and 851 reviewed work groups remain a dated snapshot. The September inventory declares the current release source arrays, published-case source joins, held/inactive evidence and source-note inputs separately. It resolves the source references for all 438 authored cases. The current bibliography records 1,452 distinct access URLs across 1,434 conservative work or access records and 6,052 source occurrences, not 1,434 verified distinct publications. Original inventory files remain separately dated snapshots; the current bibliography and its update ledger record later uses and accepted metadata updates. The original duplicate decisions and work IDs are preserved; passages, editions and access URLs are not collapsed merely because their titles resemble one another.

Typed metadata is tiered: the previously reviewed thirty-record pilot is preserved, identifier-deposited metadata carries its provenance and limits, and missing author, edition or locus fields remain explicitly unresolved. The manuscript's cited works and uncited background reading are now separate. The bibliography methods and status explain the inputs, corrections and remaining research.

3.2 Sixty dimensions and two kinds of missingness

Five independently prepared disciplinary batteries contribute twelve dimensions each:

Battery	Principal concerns
Settler-colonial studies	Incoming populations, implantation and persistence
Political economy	Land, property, labor and incorporation
Historical demography	Population composition, movement and transformation
Genocide and forced migration	Removal, coercion, mortality and return
State formation and ideology	Authority, institutions and legitimation

The batteries were adopted side by side without removing overlapping constructs. That choice permits comparison across traditions but also gives repeated concepts additional influence in a covariance model; it is a design choice, not freedom from theory. The predictions preceded surviving coding outputs in a dated local file. They were not an externally preregistered, read-only specification. The later exploratory work and editorial corrections are identified as such.

Dimensions use explicit 0–4 anchors and era-sensitive evidence rules. An archaeogenetic case can use ancestry evidence where a modern case can use censuses or passenger records (Haak et al. 2015; Olalde et al. 2019). Unknown (–1) means the dimension has a referent but the evidence does not settle its value. Structurally

inapplicable (-2) means the referent does not exist—for example, an incoming-settler dimension in a pure expulsion. Neither code means a measured zero. The distinction is preserved in the typed missingness data.

Raw historical values are preserved as well: the final correction did not silently round the 100 inherited fractional values in otherwise unchanged cells. The codebook, score matrix and typed missingness matrix provide the exact definitions and values used.

3.3 AI-assisted coding and agreement

Research agents worked from the codebook and per-event evidence briefs, returned constrained score records, and recorded source references and qualifications. Initial production batches were balanced across era and mechanism. Later successor coding used independent coders followed by disagreement-only adjudication without averaging. The historical first-pass duplicate-reconciliation procedure was different and remains documented in the archived manuscript; it is not retrospectively described as the later protocol. Related work on LLM annotation provides context, not validation of this instrument (Gilardi, Alizadeh, and Kubli 2023; Ziems et al. 2024).

Check	Scope	Reported result	Interpretation
Same-model pilot	Six contrasting events; 345 scored comparisons	93% exact; 99% within one anchor step	Small instrument-debugging exercise; predates the final missingness rule
Candidate coder gate	Thirty-event reference set	67% exact; 22% unknown/inapplicable concordance	Candidate model rejected
Cross-model overlap	Thirty events; 1,704 numeric and 96 nonnumeric-involved comparisons	76% exact; 95% within one step; 50% unknown/inapplicable concordance	Material uncertainty remains, especially for missingness-based analysis

The raw inputs for the two model comparisons were recovered and both stored results reproduce from preserved copies. That resolves the earlier recomputation gap. It does not turn coder agreement into historical validity, or establish an expert gold standard. Correlated model errors can distort the measured structure; they are not assumed to be harmless random noise.

Consequential scope, inclusion, retirement, family-adoption and publication decisions were reserved for or confirmed by the human author. Many individual coding disagreements were resolved by model adjudicators under written rules. Human responsibility for the project is not a claim that its author personally checked every cell. The process record preserves failed returns as well as accepted work, including a disqualified mechanically complete source review that had opened no underlying sources.

3.4 Analysis and sensitivity design

The canonical numerical pipeline completes missing numeric entries with dimension medians and fits minimum-residual exploratory factor models with oblimin rotation, allowing correlated factors. Parallel analysis compares observed eigenvalues with the 95th percentiles from 500 matched-size normal null simulations (Horn 1965). Alternative treatments of missingness, era composition, linked histories and indicator overlap test dependence on those choices. Treating structural inapplicability as zero appears only as an extreme diagnostic alternative, not a preferred historical interpretation.

The comparisons distinguish named-axis congruence from whole-space movement. The first asks whether particular labeled directions remain similar; the second permits the fitted spaces to align and reports their principal angles. They answer different questions. Fixed eight- and nine-factor fits prevent a change in retained count from being mistaken for a direct comparison of named traits. K-means partitions and their stability checks are diagnostic tools, not procedures for assigning the public families.

The complete specifications, inputs and numerical outputs are available in the versioned release. The remainder of this paper reports the completed results; it does not rerun a model or select a boundary to improve its apparent stability.

4. Results

4.1 Current release: equal factor counts, changed relationships

Both the recomputed v6 baseline and the v7 candidate retain nine factors under canonical preprocessing. The fitted relationships change materially:

Comparison or diagnostic	Completed result
V6 to v7, largest nine-dimensional subspace angle	11.78°
V6 to v7, largest fixed-eight subspace angle	18.69°
Minimum absolute named-axis Tucker congruence	0.078
Shared events changing the aligned six-cluster diagnostic	17.98% of 662
Adjusted Rand index for that clustering comparison	0.632
Labeled sensitivity scenarios retaining nine / eight / seven factors	524 / 29 / 12

The six-cluster diagnostic is not the six-family public classification. No editorial family was changed merely to track the clustering output. The 565 labeled scenarios reuse 536 distinct preprocessed matrices; their counts are not independent replications or probabilities that a given factor count is correct. Separate indicator-battery omission tests retain six to eight factors. Agreement in the canonical count therefore cannot support a claim that nine unchanged historical traits have been discovered.

Linked-program weighting demonstrates the effect of granularity. Giving each declared program a total weight of one retains all 676 real rows but reduces their total weight to 583, with an effective sample size of 615.78. The largest fixed-eight subspace angle relative to the unweighted candidate is 44.67° . There are no 583 synthetic events behind that number. This is an alternative weighting of the same rows, not a new rule for defining the past.

The correction cycle covered 17 coding packages and 2,716 cells: 840 cells in the fourteen new events, plus 1,876 reviewed cells in retained events. Of the latter, 927 changed and 949 were confirmed. The other 37,844 cells in shared rows retain their stored values and missingness states. The numerical changes are tied to a finite, source-adjudicated correction scope.

4.2 Chronology addendum and unavailable alternatives

Two accepted dating corrections were applied in a separate addendum after the original diagnostic run. La Gomera uses the 1488–1489 revolt and punitive sequence rather than a 1449 parent-midpoint proxy. The West Bank and East Jerusalem case uses its 1967 onset rather than the 2024 reporting cutoff; its era moves from 1990–present to 1945–1989. Neither correction changes a raw code, family or event boundary, nor supplies a measured movement duration.

The addendum preserves the original 565 results and reuses the identical canonical fit. Only two new matrices require fitting: equal-era weighting and era-specific median imputation. Fourteen scenario aliases share those fits, and 511 affected date diagnostics were recalculated after reproducing their original values. Candidate date eligibility rises from 627 to 629; the baseline remains 628. The seven/eight/nine retention counts are unchanged.

Relative to their own pre-amendment versions, the largest eight-/nine-space angles are $0.39^\circ/0.30^\circ$ for equal-era weighting and $3.48^\circ/2.42^\circ$ for era-median imputation. These incremental changes do not cancel the larger sensitivities above. No universal angle threshold was predeclared to certify fairness. The addendum results and acceptance with limits are separate from the original verification.

No independently bound continuous movement-duration series exists for this release. Subtracting dates would answer a different question. The Egypt-through-642-only and internal Kitakami phase alternatives also remain unestimable because they lack separately coded profiles. No copied or invented scores fill those gaps.

4.3 What the earlier diagnostics establish

Earlier results retain their own release identities:

Historical checkpoint	Canonical retained count	Selected boundary evidence
Repaired 407 and expanded 490	8	At 407, eight factors account for 53% of variance
V3, 583 events	9	Ninth-factor margin +0.019
V4, different 583-event roster	8	Ninth-factor margin −0.00789
V5, 676 events	9	Nine in twenty repeated calibrations and every single-event deletion; several grouping/date alternatives return eight
V6, 675 events	9	Nine-space distance from v5: 2.765°; fixed-eight distance: 6.685°

At v5, all 676 single-event omissions move the nine-dimensional space by at most 4.808°, whereas forced eight reaches 86.489° and exceeds 10° in 36 runs. That exhaustive deletion result belongs to v5; it is not a fresh v7 test. Program substitutions, weighting and imputation remain material. The earlier durability direction was also era-confounded ($\rho = -0.51$ with start year in the v3 battery): older events have had longer to look permanent.

The clustering results are similarly bounded. At 490 events, 500 restarts produced 367 distinct near-tied partitions. The bootstrap tests distinguished stable neighborhoods from a unique defensible assignment (Hennig 2007, 2015). Replacing forty-four separately valid early Neo-Assyrian events with three ruler-level medians raised partition agreement with the earlier 446-event baseline from adjusted Rand index 0.661 to 0.928; nine coding-pattern medians gave 0.968, while minimum factor congruence remained at least 0.999. These medians were deliberately nonhistorical diagnostic aggregates, not admitted replacement events. In that experiment, a dense archive moved cluster boundaries much more than factor axes.

The result supports a narrow conclusion: no automatic partition tested here is robust enough to serve as the public taxonomy. It does not prove that history has no natural kinds, or that every possible clustering method must fail. The historical phrase “no natural species” is a metaphor, not an ontological finding.

5. From measurement to the public atlas

5.1 Families are reviewed filing categories

The current family table records source-reviewed editorial judgments and explicitly carried prior assignments:

Family	V7 events
Expulsion or flight without replacement	174
Conquest and incorporation	173
State-directed demographic remaking	149
Settler replacement	108
Low-state durable expansion	44
Land clearance for state or project use	24
Family evidence hold	4
Total	676

The four family holds concern Croton's foundation, the partial Paeonian return, Setia's foundation and Sullan Faesulae. They are distinct from other authoring holds. The family decision ledger preserves the judgments and qualifications.

A related normalized mechanism vocabulary was tested separately and failed its single permitted blinded reliability retest. Among 108 jointly reviewable events, micro kappa was 0.732, positive-label Dice 0.790 and tag Jaccard 0.653, against a required 0.80 on all three. It remains internal. This failure does not show that mechanism and family mean the same thing; it shows that the tested mechanism instrument did not justify another public analytical label. The failed and repaired attempts remain in the process history; they do not trigger recursive tuning in this release.

5.2 Eight public traits are descriptive summaries

The public profile summarizes government direction, legal and property machinery, settler implantation, elimination versus incorporation, lethal coercion, durability, separation from homeland, and imported unfree labor. These evidence-aware summaries of anchored judgments are neither the nine exploratory factors nor eight natural categories. Coverage and unknown ranges remain visible, and structurally inapplicable traits are suppressed.

Two corrections explain why source review takes precedence over a convenient factor label. The original seventh statistical direction could give a high “displacement” value to a case whose residents stayed. The public seventh trait therefore uses direct, evidence-gated separation from homeland, combining relocation distance and return foreclosure only when displacement exists. The statistical factor remains preserved for diagnosis.

Similarly, the earlier “Coerced outside labor” composite mixed unfree incomers with household, distance and subsistence correlates. The public eighth trait now asks whether a distinct population was brought from elsewhere as enslaved, indentured or otherwise unfree labor. Its applicability gate precedes its intensity calculation. Unknown applicability remains unknown; a known zero displays zero; structural N/A suppresses the

bar. Source-qualified exceptions are recorded separately from the general formula. Six older raw-coding follow-ups remain open despite those public display corrections.

The trait specification, current atlas profiles, the frozen release profile snapshot and open raw-coding follow-ups make those distinctions inspectable.

5.3 The atlas is an editorial sample

The 438 authored cases are a subset of the modeled register. Historical selection diagnostics found that the earlier 222-case atlas emphasized settler implantation, legal machinery and durability while underrepresenting conflict displacement and state demographic engineering. Those comparisons belong to that older sample; they are not newly calculated estimates for the current atlas. The remaining authoring queue uses balanced production while preserving chronological playback and source gates.

Retired cases remain as audit records, with successor relationships and stable-link explanations where appropriate. Map markers, route styles and date labels must communicate the precision the evidence supports. A timeline stop is a presentation decision, not another statistical row. Neither line density nor the number of stops measures population volume or event incidence.

The atlas provides a correction channel and adopts what the author calls a Cunningham's-Law posture: publish a strong provisional account that readers can challenge, preserve the evidence behind corrections, and revise visibly. The existence of that channel is not evidence that every report has received review. Public criticism supplements rather than replaces source research.

6. Limits and implications

The central limitations are substantive. Source survival and research attention shape the corpus. Ancient and modern evidence resolve events at different scales. Actor continuity, territorial scope and observed outcomes require judgments whose alternatives can change the model. The covariance structure depends on weighting, missingness treatment and overlapping indicators. Agreement among AI coders does not establish truth, and weak unknown/inapplicable agreement deserves particular caution.

The completed verification reconstructed scenario inputs and reported comparisons and reproduced selected fits and Horn simulations. It did not refit every result or repeat all historical research. The verifying AI agent's upstream contributions were disclosed. The later chronology addendum received a separate main implementation check rather than being retroactively included in that earlier independent assignment. The verification report and main interpretation state those boundaries.

The project therefore supports procedural auditability and bounded robustness, not a certificate of unbiased event units or complete global coverage. A source-supported boundary should not be changed merely because an alternative makes a factor solution move less. Conversely, historical plausibility is no reason to hide the numerical consequences of counting a connected history at one resolution rather than another.

What survives is a useful division of labor. The register states which histories are being compared and why. The instrument measures declared features with visible uncertainty. Sensitivity analysis shows what changes when defensible choices change. Reviewed families and descriptive traits make the results accessible without treating

an algorithm's partition as an inherited fact about the past. That architecture is an offer for criticism and reuse, rather than a universal theory of why people take land.

AI contribution and disclosure

The earlier record identifies research and coding agents from the Anthropic Claude family and OpenAI Codex. Sonnet 5 was the candidate coder excluded by the model-selection gate; Fable 5 supplied the banked reference set; Opus 5 produced the production coding matrix, as documented in the preserved process record. Codex contributed source research, integration, implementation, analysis, and audit work. This v2 editorial restructuring and bibliography maintenance were also prepared with Codex assistance. These model-role statements retain the project's recorded provenance rather than substituting current product names for the historical instruments.

Craig Talbert is the sole author. AI systems are disclosed as instruments, not co-authors. Consequential editorial decisions were reserved for or confirmed by the author; model adjudicators resolved many cell-level disagreements under written rules. The paper remains an AI-assisted working manuscript without external human peer review. Its responsibility and its limitations cannot be transferred to a model's claim of confidence.

Data, source inventory and process history

The data page supplies the immutable active release, original numerical inputs and outputs, typed missingness, editorial fields, source inventory and checksum manifest. The scenario table, diagnostic summary and compressed numerical results support the current results above. The chronology reproduction bundle preserves the original export and finite addendum path. Packaging checks do not claim a complete consumer replay of every numerical result.

The bibliography inventory preserves the August grouping decisions, current input coverage and metadata exceptions. The earlier manuscript retains detailed research and correction chronology, older result tables and edition cautions. Long codebook and sensitivity appendices are supplied through the versioned data rather than repeated as unfinished appendices here.

Source PDFs, browser captures and raw coder/model journals are excluded from the public download. An archival deposit remains a requirement before formal circulation; this working edition does not claim a published DOI or an external review that has not occurred.

References

Works cited in the manuscript. Bibliographic wording and edition cautions retain the prior citation-resolution review.

- Acemoglu, D., S. Johnson, and J. A. Robinson. 2001. "The Colonial Origins of Comparative Development: An Empirical Investigation." *American Economic Review* 91 (5): 1369–1401.
- Cohen, G. M. 2013. *The Hellenistic Settlements in the East from Armenia and Mesopotamia to Bactria and India*. University of California Press.
- Domar, E. D. 1970. "The Causes of Slavery or Serfdom: A Hypothesis." *Journal of Economic History* 30 (1): 18–32.

- Gilardi, F., M. Alizadeh, and M. Kubli. 2023. "ChatGPT Outperforms Crowd Workers for Text-Annotation Tasks." *PNAS* 120 (30): e2305016120.
- Gordillo, G. R. 2004. *Landscapes of Devils: Tensions of Place and Memory in the Argentinean Chaco*. Duke University Press.
- Haak, W., et al. 2015. "Massive Migration from the Steppe Was a Source for Indo-European Languages in Europe." *Nature* 522: 207–211.
- Hennig, C. 2007. "Cluster-Wise Assessment of Cluster Stability." *Computational Statistics & Data Analysis* 52 (1): 258–271.
- Hennig, C. 2015. "What Are the True Clusters?" *Pattern Recognition Letters* 64: 53–62.
- Horn, J. L. 1965. "A Rationale and Test for the Number of Factors in Factor Analysis." *Psychometrika* 30 (2): 179–185.
- Hu-DeHart, E. 1974. "Development and Rural Rebellion: Pacification of the Yaquis in the Late Porfiriato." *Hispanic American Historical Review* 54 (1): 72–93.
- Lemkin, R. 1944. *Axis Rule in Occupied Europe: Laws of Occupation, Analysis of Government, Proposals for Redress*. Washington, DC: Carnegie Endowment for International Peace.
- Maier, C. S. 2016. *Once Within Borders: Territories of Power, Wealth, and Belonging since 1500*. Belknap Press of Harvard University Press.
- Malhi, R. S., et al. 2008. "Distribution of Y Chromosomes among Native North Americans: A Study of Athapaskan Population History." *American Journal of Physical Anthropology* 137 (4): 412–424. <https://doi.org/10.1002/ajpa.20883>.
- Mann, M. 2005. *The Dark Side of Democracy: Explaining Ethnic Cleansing*. Cambridge University Press.
- Naimark, N. M. 2001. *Fires of Hatred: Ethnic Cleansing in Twentieth-Century Europe*. Harvard University Press.
- Nieboer, H. J. 1900. *Slavery as an Industrial System: Ethnological Researches*. Martinus Nijhoff.
- Oded, B. 1979. *Mass Deportations and Deportees in the Neo-Assyrian Empire*. Reichert.
- Olalde, I., et al. 2019. "The Genomic History of the Iberian Peninsula over the Past 8000 Years." *Science* 363: 1230–1234.
- Polian, P. 2004. *Against Their Will: The History and Geography of Forced Migrations in the USSR*. CEU Press.
- Reich, D. 2018. *Who We Are and How We Got Here: Ancient DNA and the New Science of the Human Past*. Pantheon.
- Salmon, E. T. 1969. *Roman Colonization under the Republic*. Thames & Hudson.
- Scott, J. C. 1998. *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed*. Yale University Press.
- Tilly, C. 1990. *Coercion, Capital, and European States, AD 990–1990*. Blackwell.
- Veracini, L. 2010. *Settler Colonialism: A Theoretical Overview*. Palgrave Macmillan.
- Viola, L. 2007. *The Unknown Gulag: The Lost World of Stalin's Special Settlements*. Oxford University Press.
- Wolfe, P. 2006. "Settler Colonialism and the Elimination of the Native." *Journal of Genocide Research* 8 (4): 387–409.
- Ziems, C., et al. 2024. "Can Large Language Models Transform Computational Social Science?" *Computational Linguistics* 50 (1): 237–291.

Background reading

These works were retained in the earlier anchor list but are not cited in the manuscript body. They are separated here without inventing an influence on the project or adding citations merely to retain them.

- Allentoft, M. E., et al. 2024. "100 Ancient Genomes Show Repeated Population Turnovers in Neolithic Denmark." *Nature* 625: 329–337.
- Belich, J. 2009. *Replenishing the Earth: The Settler Revolution and the Rise of the Anglo-World, 1783–1939*. Oxford University Press.
- Brace, S., et al. 2019. "Ancient Genomes Indicate Population Replacement in Early Neolithic Britain." *Nature Ecology & Evolution* 3: 765–771.
- Bulutgil, H. Z. 2016. *The Roots of Ethnic Cleansing in Europe*. Cambridge University Press.
- Cavanagh, E., and L. Veracini, eds. 2017. *The Routledge Handbook of the History of Settler Colonialism*. Routledge.
- Fieldhouse, D. K. 1966. *The Colonial Empires: A Comparative Survey from the Eighteenth Century*. Weidenfeld & Nicolson.
- Hubert, L., and P. Arabie. 1985. "Comparing Partitions." *Journal of Classification* 2: 193–218.
- Lipson, M., et al. 2018. "Ancient Genomes Document Multiple Waves of Migration in Southeast Asian Prehistory." *Science* 361: 92–95.
- Olalde, I., et al. 2018. "The Beaker Phenomenon and the Genomic Transformation of Northwest Europe." *Nature* 555: 190–196.
- Osterhammel, J. 2005. *Colonialism: A Theoretical Overview*. 2nd ed. Markus Wiener.
- Posth, C., et al. 2018. "Language Continuity despite Population Replacement in Remote Oceania." *Nature Ecology & Evolution* 2: 731–740.
- Rousseeuw, P. J. 1987. "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis." *Journal of Computational and Applied Mathematics* 20: 53–65.
- Seymour, D. J., ed. 2012. *From the Land of Ever Winter to the American Southwest: Athapaskan Migrations, Mobility, and Ethnogenesis*. University of Utah Press.
- Sokoloff, K. L., and S. L. Engerman. 2000. "Institutions, Factor Endowments, and Paths of Development in the New World." *Journal of Economic Perspectives* 14 (3): 217–232.