Automated Survey Harmonization via Source-Consistent Entity Resolution

Dawson Verley*[1], Sara Constantino[2], Brandon de la Cuesta[3], Amanda Kennard[4], and Forrest Owens[5]

[1]Energy and Resources Group, UC Berkeley

[2]Doerr School of Sustainability, Stanford University

[3]Center on Food Security and the Environment, Stanford University

[4]Department of Political Science, Stanford University

[5]Department of Political Science, University of California, Davis

Thursday 23rd July, 2026

Abstract

Survey harmonization is a common challenge in applied social science research, particularly for multi-wave surveys in which question wording varies across rounds. Existing approaches rely on manual coding, making them difficult to scale and reproduce. We develop a fully automated method by framing survey harmonization as a problem of multi-source entity resolution. The approach combines sentence-embedding similarity with a source-constrained agglomerative clustering algorithm that enforces a one-per-source matching rule. Evaluated on Afrobarometer data, the method achieves a precision of 0.941 and a recall of 0.969 (F1 = 0.955), closely matching human-coded results while operating at substantially lower cost. Applied to Latinobarómetro, the method performs comparably, achieving a precision of 0.945, a recall of 0.999, and an F1 of 0.971. The procedure scales efficiently to thousands of items and enables the construction of harmonized indices across survey rounds. More broadly, the framework applies to data integration problems involving multiple sources without reliable identifiers.

*Authors are listed in alphabetical order after the first author.

1 Introduction

In large-scale survey projects such as Afrobarometer and Latinobarómetro, many questions are repeated across survey rounds. These repeated items offer valuable opportunities to study change over time and to construct comparable measures across waves. However, combining data from repeated questions is difficult in practice. Survey questions are rarely assigned stable identifiers that persist across rounds, and even minor differences in punctuation, capitalization, or phrasing prevent exact string matching. As a result, researchers cannot reliably use the question text itself as a unique identifier.

Existing approaches to survey harmonization are often prohibitively time-consuming. Although some projects provide harmonized datasets as a public good, their coverage is limited and updates are infrequent.[1] For most applications, researchers must manually identify corresponding questions across datasets and construct their own crosswalks. While hand-matching may be feasible for a small number of items, it becomes intractable at scale. Matching thousands of questions across many surveys quickly exceeds the capacity of manual annotation. In addition, hand-labeling raises concerns about inter-coder reliability (Granda et al., 2010) and produces outputs that are difficult to validate, document, and reproduce.

In this paper, we develop a fully automated pipeline for multi-wave survey harmonization. We conceptualize this task as a special case of multi-source entity resolution, a problem with a well-developed literature spanning probabilistic record linkage (Fellegi and Sunter, 1969; Enamorado et al., 2019), Bayesian multi-file models (Sadinle and Fienberg, 2013), and graph-based clustering approaches. Our contribution is threefold. First, we formalize multi-wave survey harmonization as a problem of source-consistent multi-source entity resolution, in which transitive matches must satisfy the constraint that each cluster contains at most one question from each survey wave. Second, we introduce an end-to-end matching procedure that combines sentence-transformer embeddings (Reimers and Gurevych, 2019) with a source-constrained agglomerative clustering algorithm, producing high-quality matches without requiring hand-labeled training data. Third, we provide a systematic empirical evaluation using labeled data from the first eight waves of the Afrobarometer, benchmarking our approach against representative methods from three families of entity resolution algorithms, and we demonstrate the substantive value of the resulting crosswalk by constructing harmonized indices of African public opinion across survey rounds.

[1] The Afrobarometer, for example, publishes Merged Round files that combine country-level data within a single round but do not harmonize question wording across rounds.

Section 4 describes the matching procedure at the core of the pipeline. The algorithm takes as input the full set of questions pooled across survey waves and computes pairwise cosine similarities between their sentence-transformer embeddings. It then processes candidate matches in descending order of similarity, merging each pair into a cluster only if doing so does not violate the one-per-source constraint. Enforcing source consistency during the merging process, rather than through post hoc correction, ensures that the output respects the structure of the data by construction. This design also admits an efficient union-find implementation that scales near-linearly in the number of candidate edges above a similarity threshold. In practice, the full Afrobarometer crosswalk can be generated in seconds on standard hardware.

Section 5 benchmarks the pipeline against four alternative methods drawn from the three families of multi-source entity resolution approaches reviewed earlier. Using labeled data from eight rounds of the Afrobarometer, our method achieves a precision of 0.941 and a recall of 0.969 ($F_1 = 0.955$), outperforming each alternative and closely matching the output of a team of human annotators working from the same materials. To assess generalizability, we also apply the pipeline to Latinobarómetro. The method achieves a precision of 0.945, a recall of 0.999, and an F1 of 0.971, indicating that it transfers well across survey contexts.

Section 6 illustrates the substantive value of the resulting crosswalk. Using the harmonized data, we construct twenty-two thematic indices of African public opinion, covering domains such as trust in institutions, political participation, perceptions of corruption, and economic evaluations. Each index spans all survey rounds in which its component items appear. The indices exhibit strong psychometric properties: Cronbach's α exceeds the conventional 0.7 threshold in most cases, and partial least squares factor analysis shows that items within each index load consistently and in the same direction on a common latent factor. The resulting country-by-round series reveal systematic patterns across countries and over time that are not observable within any single wave. Constructing such series manually would be prohibitively costly.

Finally, Section 7 considers extensions of the method to other domains. Many areas of research, including official statistics, conflict studies, and public health, require linking records across multiple sources in the absence of reliable identifiers. The framework developed here can be adapted to these settings with minimal modification. To facilitate adoption, we provide an open-source implementation of the pipeline that allows researchers to apply the method to their own data.

2 The Challenge of Automated Survey Harmonization

Consider the following example from the Afrobarometer project. Table 1 shows the text of eight questions, one from each survey round. Each question measures a similar underlying concept: whether individuals must be careful when discussing politics, a commonly used proxy for democratic freedom. However, the questions have distinct variable names and are phrased in several slightly different ways.[2] These differences make it difficult to identify corresponding items across waves. In practice, researchers must rely on manual inspection and subjective judgment to determine which questions are sufficiently similar to be treated as the same measure. This type of harmonization task arises frequently in the ex post analysis of large-scale survey data and becomes increasingly burdensome as the number of items grows.

Round	Question No.	Question Text
1	dmpsay	In this country, you must be very careful of what you say and do with regards to politics.
2	q41A	In this country, how often: Do people have to be careful of what they say about politics?
3	q53a	In this country, how often: Do people have to be careful of what they say about politics?
4	q46	In this country, how often: do people have to be careful of what they say about politics?
5	q56a	In your opinion, how often, in this country: do people have to be careful of what they say about politics?
6	q51a	In your opinion, how often, in this country: do people have to be careful of what they say about politics?
7	q42a	In your opinion, how often, in this country: Do people have to be careful of what they say about politics?
8	q40d	In your opinion, how often, in this country: Do people have to be careful of what they say about politics?

Table 1: Text of Eight Afrobarometer Questions

Survey harmonization is part of a broader class of record linkage problems in which items must be matched across datasets in the absence of unique identifiers. Existing record linkage methods are typically designed for merging two datasets at a time (Fellegi and Sunter, 1969; Enamorado et al., 2019; Binette and Steorts, 2022a), whereas survey harmonization requires linking many waves simultaneously. Standard pipelines also rely on *blocking* to reduce the computational cost of pairwise comparisons, restricting attention to records that share values on structured covariates (Christen, 2012; Campbell et al., 2021). This approach presumes the availability of auxiliary fields that are informative for matching. In survey harmonization, however, the primary content of each record is free-form question text, and such structured fields are

[2]The short variable labels for these eight questions, which we include in Appendix Table 4, are no more informative.

often unavailable or uninformative.

A natural alternative is to apply pairwise record linkage sequentially across survey waves. For example, one might match wave one to wave two, wave one to wave three, and so on, and then combine the resulting matches. However, this approach does not scale well and fails to enforce consistency across waves. As the number of survey rounds increases, the number of pairwise comparisons grows rapidly, and inconsistencies across matches become more likely. In the next section, we show that this problem is best understood as a multi-source matching task with additional structural constraints.

We formalize multi-wave survey harmonization as a problem of multi-source entity resolution. Let $X = \{x_1, \ldots, x_n\}$ denote the set of survey questions pooled across m waves, and let $c_i \in \{1, \ldots, m\}$ indicate the wave in which question x_i was asked. Each question belongs to an unobserved semantic cluster indexed by $z_i \in \{1, \ldots, k\}$, where two questions match if and only if they share the same cluster assignment. Because each wave includes at most one instance of a given concept, the latent partition must satisfy a one-per-source constraint. Formally, if $z_i = z_j$, then $c_i \neq c_j$. This imposes the requirement that each cluster contains at most one question from each wave. The goal is to recover the partition $\{X_{z=1}, \ldots, X_{z=k}\}$ that satisfies this constraint without *ex ante* knowledge the assignments z or the number of clusters k.

In practice, cluster membership must be inferred from pairwise comparisons among question texts. Table 2 illustrates the challenge using a simple example with three survey waves. Let D_i denote wave i and $d_{i,j}$ the jth question in that wave. The table reports hypothetical similarity scores for all pairwise comparisons across waves. In the first row, the matches are consistent and transitive: $d_{1,1}$ is most similar to $d_{2,1}$ and $d_{3,1}$, and these items form a coherent cluster. In the second row, however, pairwise matches are inconsistent. While $d_{1,2}$ is most similar to $d_{2,2}$ and $d_{3,2}$, the item $d_{2,2}$ instead matches more closely to $d_{3,3}$. As a result, pairwise matching does not guarantee transitivity and can produce clusters that violate the one-per-source constraint. These inconsistencies motivate a joint matching procedure that enforces structural constraints across all waves.

3 Computational Approaches to Multi-Source Entity Resolution

Multi-source entity resolution is a challenging computational problem. While efficient algorithms exist for pairwise matching between two sources, finding optimal matches across three or more sources is

	d_1				d_2				d_3			
d_1	.	.	.	.	1	0	0	0	1	0	0	0
	.	.	.	.	0	1	0	0	0	1	0	0
	.	.	.	.	0	0	1	0	0	0	1	0
	.	.	.	.	0	0	0	1	0	0	0	1
d_2	.	.	.	.	.	.	.	.	1	0	0	0
	.	.	.	.	.	.	.	.	0	0	1	0
	.	.	.	.	.	.	.	.	0	1	0	0
	.	.	.	.	.	.	.	.	0	0	0	1
d_3	.	.	.	.	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.	.	.	.	.

Table 2: Hypothetical Similarity Across Survey Items

$\mathcal{NP}$-complete (Karp, 1972; Spieksma, 2000). As a result, exact optimization quickly becomes infeasible as the number of sources grows. In survey harmonization, tasks involve many waves. Some form of approximation is therefore unavoidable. The existing literature has developed three broad families of approaches, each with distinct trade-offs in terms of scalability, statistical interpretability, and suitability for text-based inputs.

The first and most widely used family is *probabilistic record linkage*, descended from the Fellegi–Sunter framework (Fellegi and Sunter, 1969; Winkler, 1988). These methods treat record pairs as draws from a mixture of matches and non-matches and estimate match probabilities based on comparison vectors constructed from multiple covariate fields. Extensions have been developed in political methodology (Enamorado et al., 2019) and reviewed in recent surveys (Binette and Steorts, 2022a). However, these approaches are not well suited to our setting. They are designed for pairwise merges, requiring repeated application across source pairs, and they depend on structured covariates to evaluate similarity. In survey harmonization, by contrast, we must link many sources simultaneously and typically have only a single free-form text field per record.

A second family uses fully Bayesian generative models for multi-file entity resolution (Sadinle and Fienberg, 2013; Aleshin-Guendel and Sadinle, 2023; Marchant et al., 2021); see Binette and Steorts (2022b) for a recent review. These models are well aligned with the structure of the problem, as they natively accommodate multiple sources and provide principled uncertainty quantification. However, they face substantial computational challenges at scale. Posterior inference via MCMC becomes impractical for datasets with thousands of records across many sources, particularly when compared to methods that operate in a single pass. Like probabilistic record linkage, these models also assume tabular inputs with

multiple covariate fields rather than relying primarily on free-form text.

A third family frames entity resolution as a graph-clustering problem. In this approach, a similarity graph is constructed over candidate record pairs, and clustering or community-detection algorithms are used to partition the graph. Examples include simple thresholding of similarity scores, transitive closure, and community-detection methods such as Louvain (Blondel et al., 2008). These methods are computationally efficient and scale well to large datasets. However, they do not impose structural constraints linking records to their sources. As a result, they often produce overly large clusters by merging records that are jointly similar, leading to high recall but reduced precision.

What is missing from existing approaches is a method that combines scalability, compatibility with text-based similarity measures, and enforcement of the structural constraints inherent to multi-source data. The procedure we develop in Section 4 addresses this gap. It builds on recent work on source-consistent multi-source entity resolution (Saeedi et al., 2017; Lerm et al., 2021), which imposes the constraint that each cluster may contain at most one record from each source—the *one-per-source* constraint. We adapt this framework to the survey harmonization setting by integrating sentence-embedding similarity with a source-constrained agglomerative clustering procedure that can be implemented in a single pass. This design exploits the sparsity structure of thresholded similarity graphs while ensuring that all matches satisfy the required structural constraints. More broadly, the approach can be interpreted as a form of constrained agglomerative clustering with cannot-link constraints, where records from the same source are prohibited from appearing in the same cluster (Wagstaff and Cardie, 2000; Basu et al., 2008). In Section 5, we evaluate this procedure against representative methods from each of the three families described above.

4 A Heuristic Approach for Multi-Source Entity Resolution

To identify transitive item sets across survey waves, we propose a simple heuristic that combines sentence embeddings with constrained agglomerative clustering. At a high level, the procedure is analogous to Kruskal's algorithm: we sort candidate matches by similarity and iteratively merge them, subject to a constraint that prevents combining items from the same source. This yields a fast, single-pass method that produces globally consistent clusters from local similarity scores.

We begin by constructing similarity scores for all pairs of survey questions, as illustrated in Table 2. Classical approaches in record linkage rely on comparison vectors derived from structured covariates

(Fellegi and Sunter, 1969; Winkler, 1988). In our setting, however, the primary information is free-form question text. We therefore use sentence embeddings to represent each question in a latent semantic space. Pre-trained models such as Sentence-BERT achieve strong performance on semantic similarity tasks (Reimers and Gurevych, 2019) and have been successfully applied to entity resolution in social science contexts (Arora and Dell, 2023; Silcock et al., 2023; Dell et al., 2023).

We embed each survey question into a d-dimensional vector space using Sentence-BERT and compute pairwise cosine similarities between the resulting vectors.[3] These embeddings provide a flexible representation of semantic similarity between questions. However, similarity scores alone are not sufficient to recover a harmonized crosswalk. As shown in Section 2, pairwise comparisons may be non-transitive and may violate the one-per-source constraint. We therefore require a procedure that aggregates pairwise similarities into clusters while enforcing these structural constraints.

To address this problem, we propose a greedy heuristic that interleaves similarity-based merging with constraint enforcement. Rather than constructing clusters first and correcting them afterward, our approach enforces the one-per-source constraint at the moment of each merge. This design follows the general logic of source-consistent clustering (Saeedi et al., 2017) but is adapted to the sparsity structure of sentence-embedding similarity graphs. Candidate edges are processed in descending order of similarity, and each proposed merge is accepted only if it does not introduce a conflict between sources. This ensures that high-confidence matches are incorporated first and that all resulting clusters satisfy the structural constraint by construction.

Formally, the procedure implements greedy agglomerative clustering with cannot-link constraints induced by source labels. Let $G = (V, E)$ denote a weighted graph in which each vertex $i \in V$ corresponds to a survey question x_i with source label c_i, and each edge $(i, j) \in E$ represents a pair of items with similarity s_{ij} exceeding a threshold λ. The goal is to recover a partition of V that satisfies the one-per-source constraint in (**??**). The algorithm proceeds as follows. We initialize each item as its own cluster and sort all edges in descending order of similarity. We then iterate through the sorted edge list and, for each edge (i, j), check whether the clusters containing i and j can be merged without introducing duplicate source labels. If so, we merge the clusters; otherwise, we skip the edge. After processing all edges, the resulting connected components define the final clustering. Algorithm 1 presents the procedure in pseudocode.

[3]While our baseline implementation does not incorporate Fellegi–Sunter-style comparison vectors, the procedure can be extended to include covariate information when available.

Algorithm 1 Greedy source-consistent entity resolution

Require: Items $X = \{x_1, \ldots, x_n\}$ with source labels $\{c_1, \ldots, c_n\}$; pairwise similarity $s(\cdot, \cdot)$; threshold λ
Ensure: Partition $\{X_{z=1}, \ldots, X_{z=k}\}$ of X satisfying $z_i = z_j \Rightarrow c_i \neq c_j$

Compute $s_{ij} \leftarrow s(x_i, x_j)$ for all $i < j$
$E \leftarrow \{(i, j) : s_{ij} \geq \lambda\}$
Sort E in descending order of s_{ij}
Initialize union-find structure with each x_i in its own singleton cluster
for each edge $(i, j) \in E$ in sorted order **do**
    $C_i \leftarrow \textsc{Find}(i)$; $C_j \leftarrow \textsc{Find}(j)$
    if $C_i \neq C_j$ **and** $\{c_k : k \in C_i\} \cap \{c_k : k \in C_j\} = \emptyset$ **then**
        $\textsc{Union}(C_i, C_j)$
    end if
end for
return connected components of the union-find structure

This procedure has several useful properties. First, because the one-per-source constraint is enforced during merging, the final clustering satisfies the constraint by construction and does not require post hoc correction. Second, processing edges in descending order ensures that high-similarity matches are prioritized, reducing the likelihood that weaker matches block stronger ones. Third, the algorithm admits an efficient union-find implementation, allowing cluster membership and constraint checks to be performed in near-constant amortized time. As a result, runtime scales near-linearly in the number of edges above the similarity threshold. We refer to this procedure as `greedy` and benchmark it in Section 5 against alternative approaches.

A final point concerns what `greedy` does not provide. The procedure yields a hard partition of survey items rather than probabilistic match scores, and it does not produce uncertainty estimates of the type associated with probabilistic record linkage (Fellegi and Sunter, 1969; Enamorado et al., 2019). This trade-off is deliberate. In the survey harmonization setting, the primary constraints are scale and the need to operate on free-form text. Methods that provide full uncertainty quantification, such as MCMC-based approaches, are not computationally competitive at this scale. Prioritizing scalability and text compatibility therefore provides a practical solution for the class of problems we consider. We return to this trade-off, and to possible extensions, in Section 7.

5 Tests

To evaluate our approach, we conduct two empirical tests on real-world survey harmonization tasks. First, we harmonize two decades of Afrobarometer data and compare our results to a hand-coded crosswalk constructed by a team of research assistants. Second, we apply the method to Latinobarómetro and

assess performance by manually reviewing a random sample of proposed matches. Together, these tests evaluate both accuracy and generalizability. Across both settings, the procedure achieves high precision and recall.

Our first test uses data from the Afrobarometer project. We apply our method to the full set of questions from the first eight survey rounds and compare the resulting clusters to those produced by four undergraduate research assistants working independently from the same materials. In our primary specification, we use the full text of each survey question, scraped from the codebooks. We also report results using only the short variable labels available in the survey data files. This comparison allows us to assess how sensitive performance is to the quality of textual inputs.

Standard evaluation metrics for multi-source entity resolution are not yet well established (Hand and Christen, 2018; Saeedi et al., 2017). We therefore evaluate performance using precision and recall, defined in the usual way as functions of true positives, false positives, and false negatives:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

We also report the F1 score, the harmonic mean of precision and recall. All metrics are computed over the set of matched question pairs implied by the clustering.

Our results are reported in Table 3. Using the full question text, we obtain a precision of 0.941 and a recall of 0.969, yielding an F1 score of 0.955. This F1 score outperforms all other algorithms tested. The next best approach, the `St_sbert` algorithm, produces an F1 score of .950. Using the short question labels from the survey data, we obtain more false negatives, giving us a precision of 0.944, a recall of 0.890, and an F1 score of 0.916. The merge involved the comparison of 890 questions and took approximately 63 seconds on a 2025 MacBook Pro with 32GB of RAM.

To assess the importance of the one-per-source constraint, we benchmark our method against four alternative algorithms drawn from the three families discussed in Section 3. We refer to our method as `greedy`. The comparison methods are: simple thresholding of the similarity matrix (`thresh`); thresholding with post hoc transitivity correction (`thresh_t`); Louvain community detection (Blondel et al., 2008) applied to the thresholded graph (`louvain`); and mutual nearest-neighbor matching with post hoc correction (`st`). Each method is evaluated using both Sentence-BERT and TF-IDF cosine similarity at a common

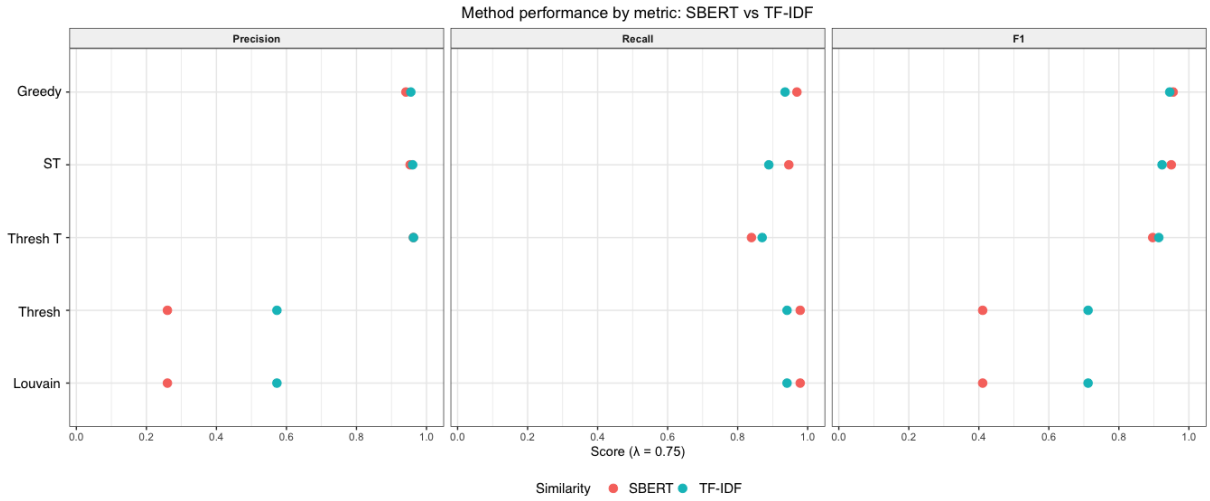

Figure 1: Precision, recall, and F1 for five matching algorithms on the Afrobarometer labeled data at similarity threshold $\lambda = 0.75$, comparing SBERT (red) and TF-IDF (teal) similarity measures. Methods that enforce a one-per-source constraint (`st`, `greedy`) substantially outperform methods that do not (`thresh`, `louvain`); transitive closure on top of thresholding (`thresh_t`) narrows but does not close the gap. Within each method, the choice of similarity measure matters far less than the choice of matching algorithm.

threshold of $\lambda = 0.75$.

Figure 1 highlights three key patterns. First, the choice of matching algorithm matters far more than the choice of similarity measure: variation in F1 across algorithms substantially exceeds the gap between SBERT and TF-IDF within any given method. Second, methods that do not enforce the one-per-source constraint (`thresh`, `louvain`) exhibit high recall but low precision, as they tend to merge all mutually similar items into large clusters. Post hoc correction (`thresh_t`) mitigates this problem but does not eliminate it. Third, the two methods that enforce the constraint (`st` and `greedy`) perform consistently well, with `greedy` achieving the highest F1 score. These results indicate that enforcing structural constraints is central to performance in multi-source settings.

To evaluate generalizability, we apply the method to Latinobarómetro data. This dataset includes 844 questions across the first 11 survey rounds. We compare the resulting clusters to the crosswalk created by the Latinobarómetro Corporation. Similar to Afrobarometer, we rely primarily on the full question text for surveys through 2006. This allows us to continue to calculate precision, recall, and F1 scores. The codebook formatting challenges limit the utilization of more contemporary Latinobarómetro survey waves.

The results are also reported in Table 3. On Latinobarómetro, the method achieves an F1 score of 0.971

percent, comparable to or slightly exceeding performance on Afrobarometer. Several other algorithms achieve similarly high performance. The best-performing algorithm for the Latinobarómetro data is `thresh_t_sbert` with an F1 score of 0.975. The `greedy` algorithm we develop performs comparably well while producing more interpretable group assignments. The full merge completes in 116 seconds on a 2025 MacBook Pro with 32 GB of RAM. These findings suggest that the approach generalizes well across survey contexts, even when input quality is lower.

6 Harmonized Indices for African and Latin American Public Opinion

The harmonization procedure developed above produces a crosswalk linking questions across all eight Afrobarometer rounds. To demonstrate its substantive value, we use the harmonized data to construct a set of thematic composite indices, validate them psychometrically, and examine the resulting cross-country and temporal patterns. The goal is illustrative rather than exhaustive: to show how automated harmonization enables analyses that would otherwise be difficult to implement. In particular, we emphasize both measurement validity and the interpretability of the resulting time series. Together, these exercises highlight the practical utility of the method.

6.1 From Matched Questions to Composite Indices

Using the harmonized output, we group matched questions into twenty-two thematic indices covering domains such as trust in institutions, political participation, evaluations of government performance, perceptions of corruption, and satisfaction with the economy. A full catalog of these indices—including the rounds in which each is observed and the number of constituent items per round—is provided in Appendix B (Figure 5). The indices are constructed to preserve comparability across waves while maximizing the use of available information. Each reflects a concept that recurs, though not always identically, across survey rounds.

Survey	Data Source	Precision	Recall	F1
Afrobarometer	Full Question Text	0.941	0.969	0.955
Afrobarometer	Short Variable Labels	0.944	0.890	0.916
Latinobarómetro	Full Question Text	0.945	0.999	0.971
Latinobarómetro	Short Variable Labels	0.934	0.997	0.965

Table 3: Results on Afrobarometer and Latinobarómetro

A key feature of the Afrobarometer is that not all survey items appear in every round. Some concepts are introduced only in later waves, while others evolve through minor rewordings. To balance coverage and comparability, we report two versions of each index where relevant. *Long-period* indices use only items that appear in every round in which the concept is measured, producing continuous series at the cost of excluding later refinements. *Short-period* indices incorporate a richer set of items but are restricted to the subset of rounds in which all items are available. For concepts observed only within a limited set of rounds, we report a single long-period version. This distinction is made explicit in all figures below.

6.2 Psychometric Validity

A natural concern with automated harmonization is whether the resulting indices are psychometrically coherent. That is, do the items grouped together by the algorithm behave as if they measure a common underlying construct? We evaluate this using two standard diagnostics. The first assesses internal consistency across items, and the second examines the structure of latent factors within each index. Together, these tests provide complementary evidence on measurement quality.

Figure 2 reports Cronbach's α for each index, computed separately for its long-period and short-period versions. Most indices exceed the conventional $\alpha = 0.7$ threshold for acceptable internal consistency, and many exceed 0.8. For indices with both versions, the short-period specification typically yields slightly higher values, consistent with the inclusion of additional conceptually related items in later rounds. These results suggest that the algorithm successfully groups items that behave similarly across respondents.

As a second diagnostic, we perform partial least squares (PLS) factor analysis on each index and examine the loadings of constituent items on the first latent factor. Figure 3 shows that within-index loadings are generally large and share the same sign, with only a small number of exceptions. This pattern is consistent with the presence of a common latent construct and inconsistent with the alternative that the algorithm has grouped heterogeneous items. Additional diagnostics—including item-level contributions to the first factor and coverage by round—are reported in Appendix C.

Taken together, the evidence from α and PLS analysis indicates that the harmonization procedure does more than identify surface-level textual similarity. The resulting indices exhibit the internal structure expected of coherent measurement instruments, supporting their use in downstream analysis.

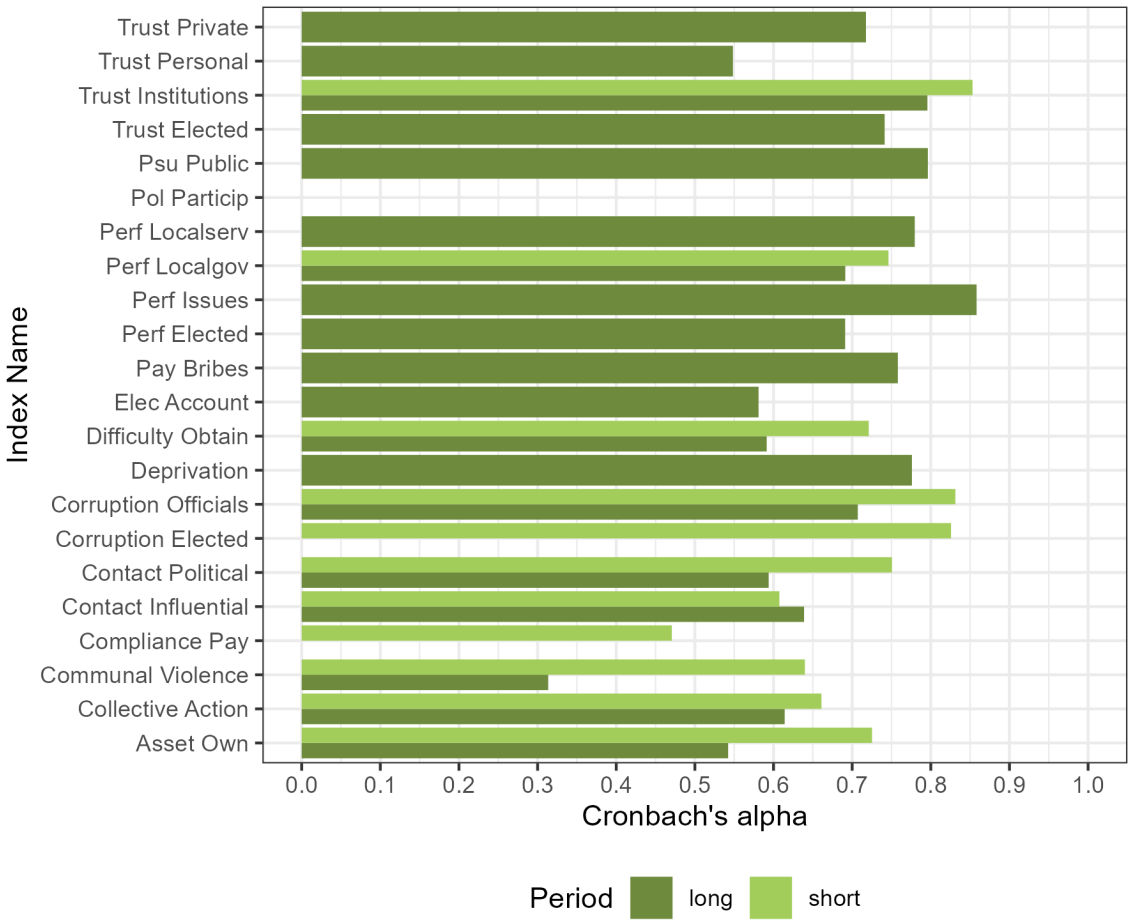

Figure 2: Cronbach's α for twenty-two thematic indices, computed for each index's long-period version (all available rounds) and short-period version (subset of rounds with a richer item set), where applicable. Indices above the conventional $\alpha = 0.7$ threshold are internally consistent by standard psychometric criteria.

6.3 Substantive Patterns Across Countries and Rounds

Once validated, the indices can be used to study variation across countries and over time at the full scale of the Afrobarometer. Figure 4 illustrates this with a measure of collective-action capacity, constructed from items on attending community meetings, joining others to raise issues, and participating in protest. The figure reports country-level values for each of the seven rounds in which the index is observed (long-period specification), along with an unweighted cross-country time series.

Two patterns emerge. First, there is substantial and persistent cross-country variation in collective-action capacity that is not reducible to region or period: neighboring countries often differ, and within-country rankings change across rounds. Second, the time series displays structured movement that is neither

purely noisy nor monotonic, suggesting that the index captures meaningful changes in political behavior over time. Similar patterns are observed for indices of economic satisfaction and institutional trust, reported in Appendix D.

None of this analysis would be feasible at comparable scale without automated harmonization. Constructing indices spanning multiple survey rounds requires a correspondence table linking matched items across waves, and building such a table manually for twenty-two indices would be prohibitively costly. The procedure developed here generates that crosswalk automatically, producing outputs that are reproducible, auditable, and readily extendable as new data become available.

7 Discussion

In this paper, we develop a method for automated survey harmonization that combines sentence embeddings with a source-constrained matching algorithm. Applied to two large survey projects, the approach achieves high levels of precision and recall while operating at a scale that would be infeasible with manual coding. Although the results are not perfect, they substantially reduce the cost of constructing crosswalks across survey waves. The method is computationally efficient, requires no hand-labeled training data, and can be implemented with standard hardware. We provide an open-source implementation to facilitate adoption and replication.

While our application focuses on survey harmonization, the underlying problem arises in many areas of social science research. In conflict studies and human rights research, analysts often need to link events across multiple databases that lack consistent identifiers. In demography and economic history, researchers face similar challenges when tracking individuals across census waves. In international political economy, linking products or tariff schedules across datasets presents analogous difficulties. These problems can all be framed as instances of multi-source entity resolution. The framework developed here can be adapted to these settings with minimal modification, particularly when the underlying data are text-based or weakly structured.

A central feature of the approach is its modularity with respect to the similarity measure. In our empirical application, we use Sentence-BERT because it provides strong performance on short-text semantic similarity and is computationally efficient. However, the matching procedure itself is agnostic to this choice. Any method that produces a reliable ranking of candidate pairs can be substituted without altering the clustering algorithm. As embedding models and language-based similarity measures continue

to improve, the overall performance of the pipeline can be expected to improve as well. This separation between representation and clustering is an important advantage of the approach.

At the same time, the method is a heuristic and does not provide formal guarantees or well-characterized statistical properties. In particular, it produces deterministic cluster assignments rather than probabilistic match scores and does not quantify uncertainty. Developing theoretical guarantees, characterizing error rates under different data-generating processes, and integrating uncertainty quantification are important directions for future research. More broadly, the results suggest that combining modern text representations with structure-aware matching algorithms is a promising direction for large-scale data integration in the social sciences.

References

Serge Aleshin-Guendel and Mauricio Sadinle. Multifile Partitioning for Record Linkage and Duplicate Detection. *Journal of the American Statistical Association*, 118(543):1786–1795, July 2023. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.2021.2013242. URL https://www.tandfonline.com/doi/full/10.1080/01621459.2021.2013242.

Abhishek Arora and Melissa Dell. LinkTransformer: A Unified Package for Record Linkage with Transformer Language Models, September 2023. URL http://arxiv.org/abs/2309.00789. arXiv:2309.00789 [cs].

Sugato Basu, Ian Davidson, and Kiri L. Wagstaff, editors. *Constrained Clustering: Advances in Algorithms, Theory, and Applications*. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series. Chapman & Hall/CRC, Boca Raton, FL, 2008. ISBN 978-1-58488-996-0.

Olivier Binette and Rebecca C. Steorts. (Almost) all of entity resolution. *Science Advances*, 8(12): eabi8021, March 2022a. doi: 10.1126/sciadv.abi8021. URL https://www.science.org/doi/10.1126/sciadv.abi8021. Publisher: American Association for the Advancement of Science.

Olivier Binette and Rebecca C. Steorts. Modern Bayesian Entity Resolution. In *Wiley StatsRef: Statistics Reference Online*, pages 1–9. John Wiley & Sons, Ltd, 2022b. ISBN 978-1-118-44511-2. doi: 10.1002/9781118445112.stat08367. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118445112.stat08367. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118445112.stat08367.

Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10): P10008, October 2008. ISSN 1742-5468. doi: 10.1088/1742-5468/2008/10/P10008.

Scott R. Campbell, Dean M. Resnick, Christine S. Cox, and Lisa B. Mirel. Using supervised machine learning to identify efficient blocking schemes for record linkage. *Statistical journal of the IAOS*, 37 (2):673, June 2021. doi: 10.3233/sji-200779. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8371678/. Publisher: NIH Public Access.

Peter Christen. *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and*

Duplicate Detection. Springer, Berlin, Heidelberg, 2012. ISBN 978-3-642-31163-5 978-3-642-31164-2. doi: 10.1007/978-3-642-31164-2. URL https://link.springer.com/10.1007/978-3-642-31164-2.

Melissa Dell, Jacob Carlson, Tom Bryan, Emily Silcock, Abhishek Arora, Zejiang Shen, Luca D'Amico-Wong, Quan Le, Pablo Querubin, and Leander Heldring. American Stories: A Large-Scale Structured Text Dataset of Historical U.S. Newspapers, August 2023. URL http://arxiv.org/abs/2308.12477. arXiv:2308.12477 [cs, econ, q-fin].

Ted Enamorado, Benjamin Fifield, and Kosuke Imai. Using a Probabilistic Model to Assist Merging of Large-Scale Administrative Records. *American Political Science Review*, 113(2):353–371, May 2019. ISSN 0003-0554, 1537-5943. doi: 10.1017/S0003055418000783. URL https://www.cambridge.org/core/product/identifier/S0003055418000783/type/journal_article.

Ivan P. Fellegi and Alan B. Sunter. A Theory for Record Linkage. *Journal of the American Statistical Association*, 64(328):1183–1210, December 1969. ISSN 0162-1459. doi: 10.1080/01621459.1969.10501049. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1969.10501049. Publisher: Taylor & Francis _eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1969.10501049.

Peter Granda, Christof Wolf, and Reto Hadorn. Harmonizing Survey Data. In Janet A. Harkness, Michael Braun, Brad Edwards, Timothy P. Johnson, Lars Lyberg, Peter Ph. Mohler, Beth-Ellen Pennell, and Tom W. Smith, editors, *Survey Methods in Multinational, Multiregional, and Multicultural Contexts*, pages 315–332. Wiley, Hoboken, NJ, 2010.

David Hand and Peter Christen. A note on using the F-measure for evaluating record linkage algorithms. *Statistics and Computing*, 28(3):539–547, May 2018. ISSN 1573-1375. doi: 10.1007/s11222-017-9746-6. URL https://doi.org/10.1007/s11222-017-9746-6.

Richard M. Karp. Reducibility among Combinatorial Problems. In Raymond E. Miller, James W. Thatcher, and Jean D. Bohlinger, editors, *Complexity of Computer Computations: Proceedings of a symposium on the Complexity of Computer Computations, held March 20–22, 1972, at the IBM Thomas J. Watson Research Center, Yorktown Heights, New York, and sponsored by the Office of Naval Research, Mathematics Program, IBM World Trade Corporation, and the IBM Research Mathematical Sciences Department*, pages 85–103. Springer US, Boston, MA, 1972. ISBN 978-1-4684-2001-2. doi: 10.1007/978-1-4684-2001-2_9. URL https://doi.org/10.1007/978-1-4684-2001-2_9.

Stefan Lerm, Alieh Saeedi, and Erhard Rahm. Extended affinity propagation clustering for multi-source entity resolution. In *Datenbanksysteme für Business, Technologie und Web (BTW 2021)*, Lecture Notes in Informatics (LNI), pages 217–236. Gesellschaft für Informatik, 2021.

Neil G. Marchant, Andee Kaplan, Daniel N. Elazar, Benjamin I. P. Rubinstein, and Rebecca C. Steorts. d-blink: Distributed End-to-End Bayesian Entity Resolution. *Journal of Computational and Graphical Statistics*, 30(2):406–421, February 2021. ISSN 1061-8600. doi: 10.1080/10618600.2020.1825451. URL https://doi.org/10.1080/10618600.2020.1825451. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10618600.2020.1825451.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410.

Mauricio Sadinle and Stephen E. Fienberg. A Generalized Fellegi–Sunter Framework for Multiple Record Linkage With Application to Homicide Record Systems. *Journal of the American Statistical Association*, 108(502):385–397, June 2013. ISSN 0162-1459. doi: 10.1080/01621459.2012.757231. URL https://doi.org/10.1080/01621459.2012.757231. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.2012.757231.

Alieh Saeedi, Eric Peukert, and Erhard Rahm. Comparative Evaluation of Distributed Clustering Schemes for Multi-source Entity Resolution. In Mārīte Kirikova, Kjetil Nørvåg, and George A. Papadopoulos, editors, *Advances in Databases and Information Systems*, pages 278–293, Cham, 2017. Springer International Publishing. ISBN 978-3-319-66917-5. doi: 10.1007/978-3-319-66917-5_19.

Emily Silcock, Luca D'Amico-Wong, Jinglin Yang, and Melissa Dell. Noise-robust de-duplication at scale. In *Proceedings of the Eleventh International Conference on Learning Representations (ICLR)*, 2023.

Frits C. R. Spieksma. Multi Index Assignment Problems: Complexity, Approximation, Applications. In Ding-Zhu Du, Panos M. Pardalos, Panos M. Pardalos, and Leonidas S. Pitsoulis, editors, *Nonlinear Assignment Problems*, volume 7, pages 1–12. Springer US, Boston, MA, 2000. ISBN 978-1-4419-4841-0 978-1-4757-3155-2. doi: 10.1007/978-1-4757-3155-2_1. URL http://link.springer.com/10.1007/978-1-4757-3155-2_1. Series Title: Combinatorial Optimization.

Kiri Wagstaff and Claire Cardie. Clustering with Instance-level Constraints. In *Proceedings of the Seventeenth International Conference on Machine Learning (ICML)*, pages 1103–1110, Stanford, CA, 2000. Morgan Kaufmann.

William E. Winkler. Using the EM algorithm for weight computation in the Fellegi–Sunter model of record linkage. In *Proceedings of the Section on Survey Research Methods*, pages 667–671. American Statistical Association, 1988.

A Short Variable Labels for Table 1 Questions

Round	Years	Short Question Label
r1	dmpsay	Careful what you say and do regarding politics
r2	q41A	Careful about what you say
r3	q53a	Careful about what you say
r4	q46	How often careful what you say
r5	q56a	How often careful what you say
r6	q51a	Q51a. How often careful what you say
r7	q42a	Q42a. How often careful what you say
r8	q40d	Q40d. How often careful what you say

Table 4: Short Question Labels from Eight Afrobarometer Questions

B Index Catalog

Figure 5 provides the full catalog of the twenty-two thematic indices constructed from the harmonized Afrobarometer data, as referenced in Section 6. For each index, the figure reports the rounds in which the constituent items appear and the number of items available per period (long versus short).

C Additional Psychometric Diagnostics

Figure 6 reports the contribution of each item to the first latent factor of its respective index, providing an item-level view of the aggregate loadings shown in Figure 3. Figure 7 shows the number of observations available per round for each index, which helps diagnose whether short-period versus long-period distinctions are driven by true coverage gaps versus sample-size fluctuations.

D Additional Substantive Indices

Figures 8 and 9 display two additional spatial-temporal indices constructed from the harmonized Afrobarometer data. Both exhibit cross-country and over-time variation consistent with the patterns discussed in Section 6.

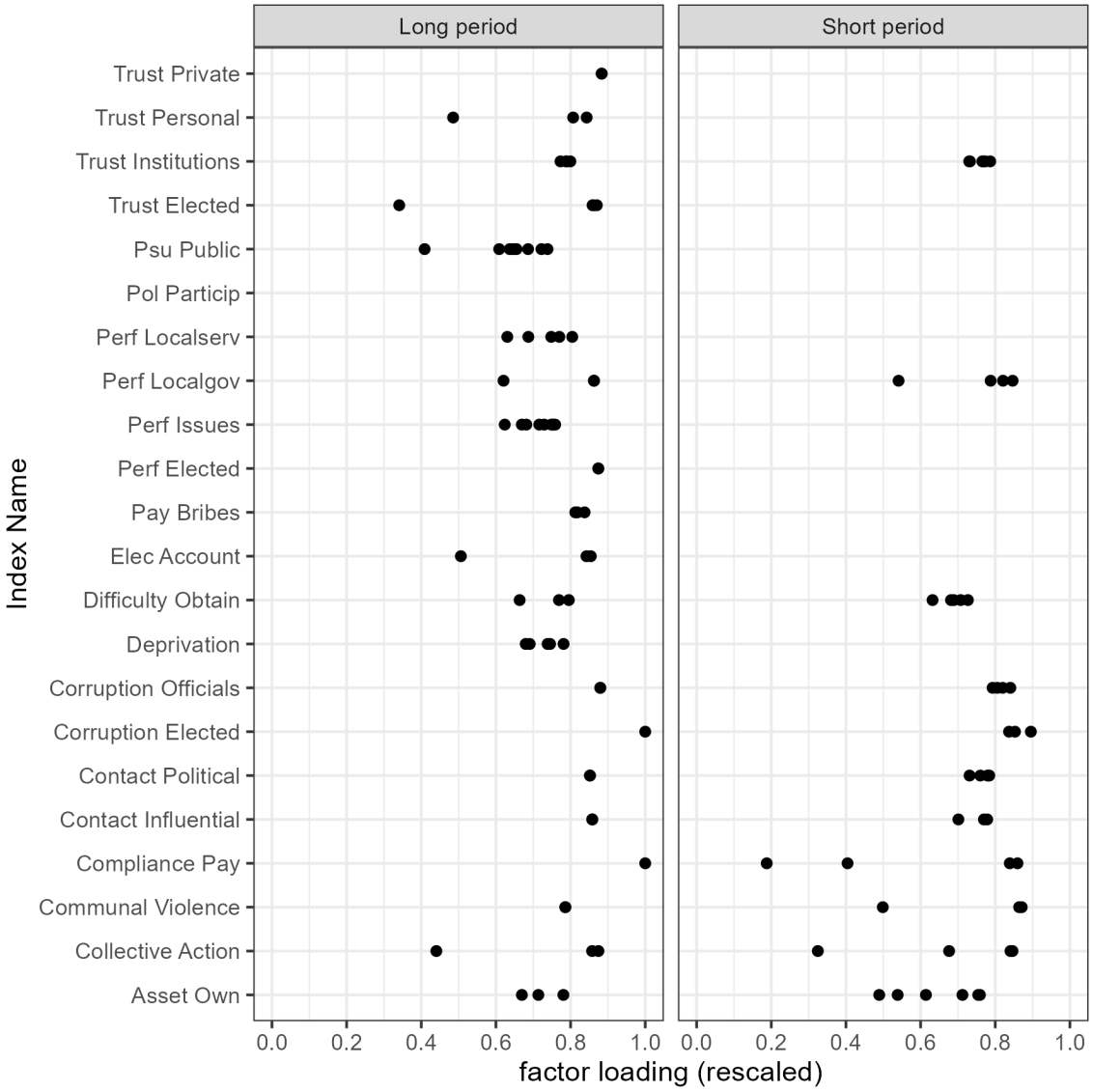

Figure 3: PLS factor loadings on the first latent factor, by item, for each index. Within-index loadings are consistently large and of the same sign, consistent with the items measuring a common underlying construct.

Figure 4: Collective-action index across seven rounds of Afrobarometer data (long-period specification). Maps show country-level means in each round; the lower-right panel shows the unweighted cross-country average over time. Spatial variation is substantial and persistent, and the time series shows structured within-country movement over two decades.

	Index Name	Rounds (long period)	N var (long period)	Rounds (short period)	N var (short period)
1	Asset Own	r4|r5|r6|r7|r8	3	r7|r8	6
2	Collective Action	r2|r3|r4|r5|r6|r7|r8	3	r4|r5|r6|r7	4
3	Communal Violence	r2|r3|r4|r5|r6|r7	2	r5|r6|r7	3
4	Compliance Pay	r2|r3|r4|r5|r6|r7|r8	1	r5|r6	4
5	Contact Influential	r2|r3|r4|r6|r7	2	r2|r3|r4	3
6	Contact Political	r2|r3|r4|r5|r6|r7|r8	2	r2|r3|r5|r6|r7	4
7	Corruption Elected	r2|r3|r4|r5|r6|r7|r8	1	r5|r6|r7	3
8	Corruption Officials	r2|r3|r4|r5|r6|r7	2	r4|r5|r6	4
9	Deprivation	r2|r3|r4|r5|r6|r7|r8	5	NA	NA
10	Difficulty Obtain	r3|r5|r6|r7|r8	3	r3|r5|r6	5
11	Elec Account	r3|r4|r6|r8	3	NA	NA
12	Pay Bribes	r2|r3|r4|r5|r6|r7	3	NA	NA
13	Perf Elected	r2|r3|r4|r5|r6|r7	2	NA	NA
14	Perf Issues	r2|r3|r4|r5|r6|r7|r8	8	NA	NA
15	Perf Localgov	r4|r5|r6	3	r4|r5	4
16	Perf Localserv	r3|r5	5	NA	NA
17	Pol Particip	r2|r3|r4|r5|r6|r7|r8	1	NA	NA
18	Psu Public	r2|r3|r4|r5|r6|r7|r8	8	NA	NA
19	Trust Elected	r2|r3|r5|r6|r7|r8	4	NA	NA
20	Trust Institutions	r2|r3|r4|r5|r6|r7|r8	4	r5|r6|r8	6
21	Trust Personal	r3|r5	3	NA	NA
22	Trust Private	r6|r7|r8	2	NA	NA

Figure 5: Catalog of the twenty-two thematic indices. Bars indicate the rounds in which each index is observed and the number of constituent items in its long-period versus short-period specifications.

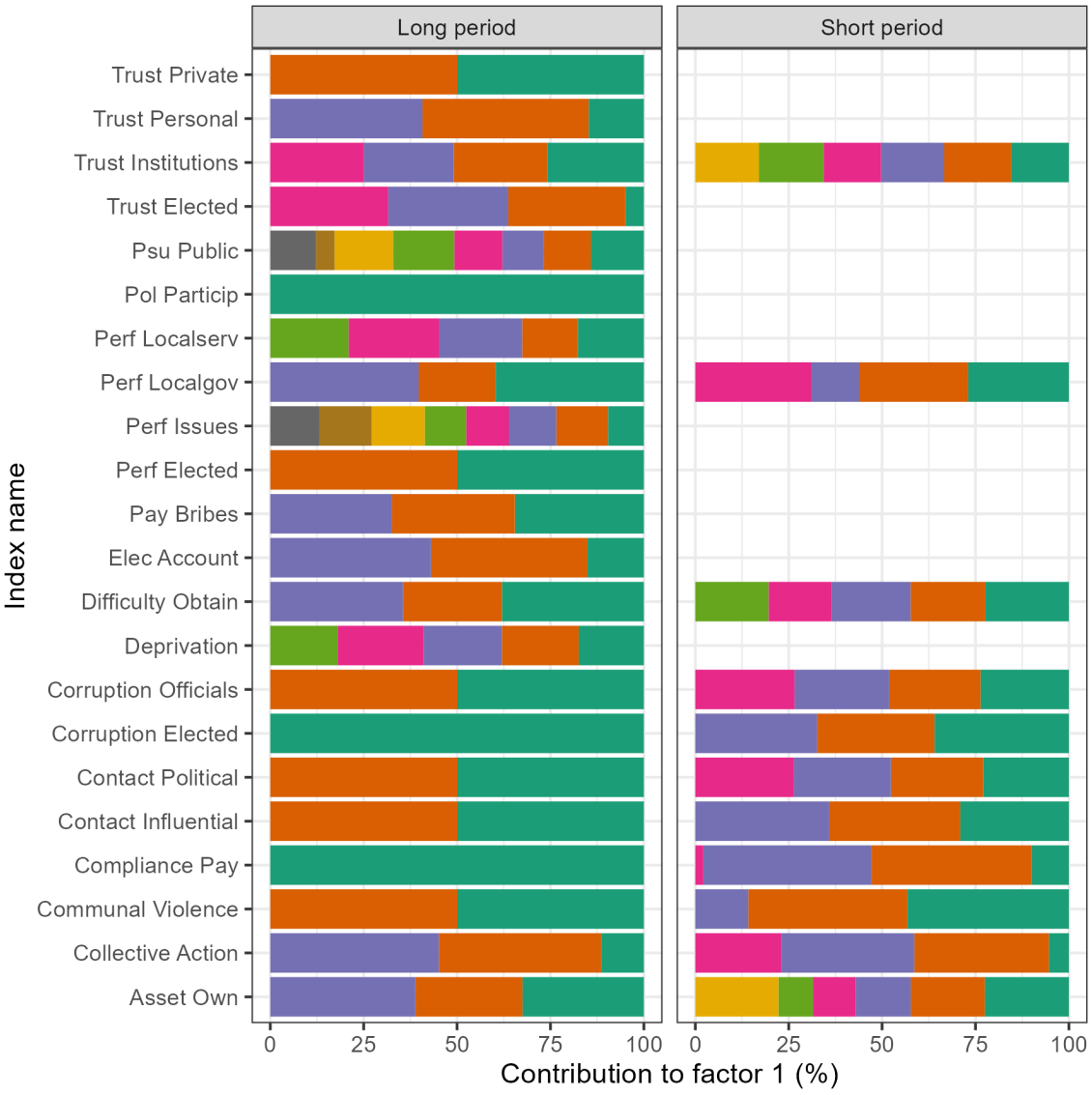

Figure 6: Item-wise contribution to the first latent factor, by index. Complements the loadings shown in Figure 3.

Figure 7: Observation counts by round for each thematic index. Gaps reflect rounds in which the constituent items are not asked, and motivate the long-period / short-period distinction in Section 6.

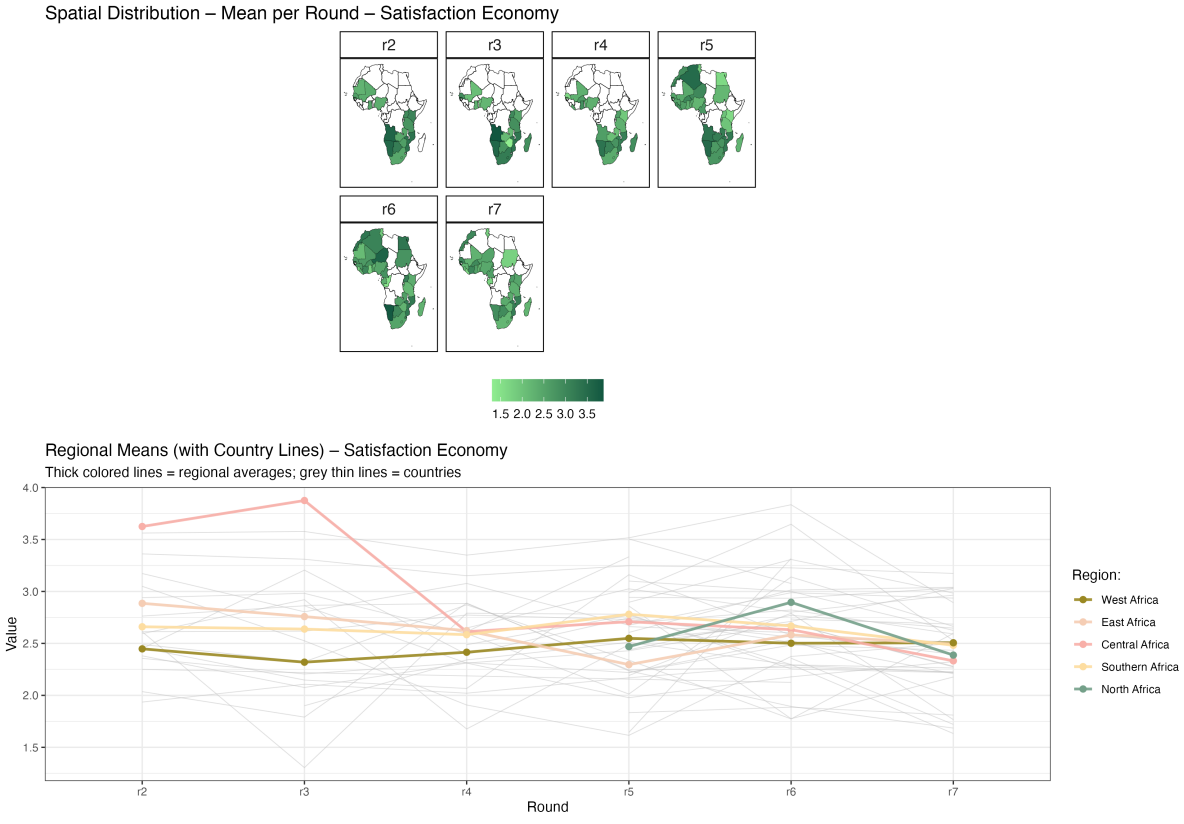

Figure 8: Economic satisfaction index across rounds of Afrobarometer data. Maps show country-level means per round; the final panel shows the unweighted cross-country time series.

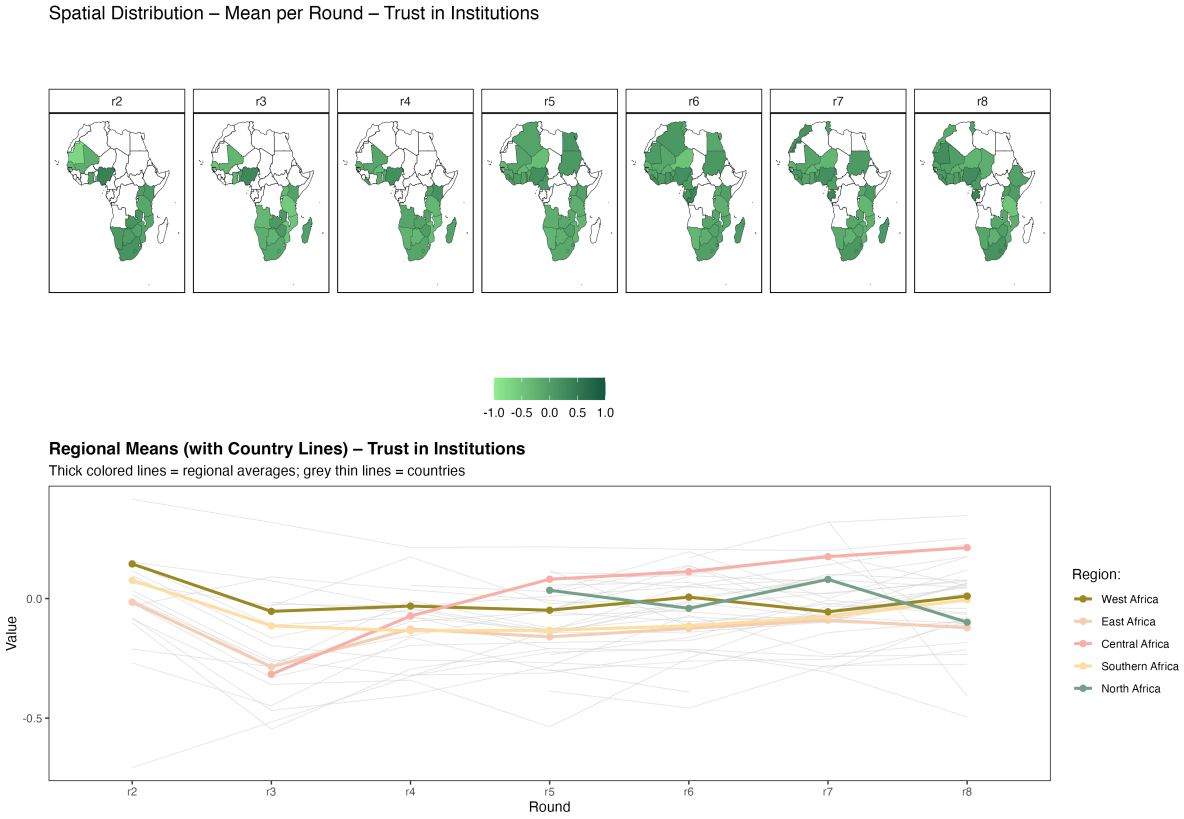

Figure 9: Trust-in-institutions index (long-period specification) across seven rounds of Afrobarometer data. Maps show country-level means per round; the final panel shows the unweighted cross-country time series.