

Supplementary Information

Unsupervised learning reveals global demographic configurations and national trajectories from population pyramids

Athulya Ratnayake, Lelumi Edirisinghe, Thisumi Wijesinghe, Roshan Godaliyadda, Parakrama Ekanayake, Vijitha Herath, and Samath Dharmaratne

Supplementary Methods

Composition-aware sensitivity analyses

Population-age profiles were represented by 21 age-group shares constrained to sum to one and therefore constituted compositional data¹. Because the centred log-ratio (CLR) transformation requires every component to be strictly positive, sparse categories in the upper-age tail were amalgamated under two alternative schemes: ages 85–89 through 100+ were combined into an 85+ category, or ages 80–84 through 100+ were combined into an 80+ category. . Following each amalgamation, every country-year profile was renormalised by its total so that the retained shares again summed to one. Only profiles with strictly positive shares after this procedure were used in the centred log-ratio sensitivity analyses.

For a zero-free composition $\mathbf{x} = (x_1, \dots, x_D)$, the centred log-ratio transformation was defined as

$$\text{clr}(x_j) = \log\left(\frac{x_j}{g(\mathbf{x})}\right) = \log(x_j) - \frac{1}{D} \sum_{\ell=1}^D \log(x_\ell),$$

where $g(\mathbf{x})$ is the geometric mean of its D components. Because the transformed coordinates sum to zero, their complete Aitchison rank is $D - 1$. For each upper-age amalgamation, two representations were evaluated: one retaining the minimum number of principal components (PCs) required to explain at least 95% of the variance, and another retaining all $D - 1$ non-zero components, thereby preserving the complete Aitchison geometry². The 85+ and 80+ schemes contained 18 and 17 age categories, respectively, giving complete ranks of 17 and 16. Principal component analysis (PCA), which summarises the main ways in which complete age profiles differ, was fitted independently to each sensitivity dataset after the CLR transformation; coordinates were mean-centred across the 1,600 profiles but not variance-standardised. Consequently, no PCA loadings or scaling parameters were shared between the primary analysis and any alternative sensitivity representation.

The global-configuration, trajectory and cross-scale analyses were repeated independently under each of the four zero-free representations. For the trajectory analysis, dynamic time warping (DTW) compared complete sequences while allowing similar changes to occur in different decades; Spearman correlation then measured whether entity pairs retained their relative ordering from more to less similar. The adjusted Rand index (ARI) compared trajectory-group assignments and aligned regime-family memberships with their corresponding primary classifications. Selected k and silhouette values were recorded for the independently estimated within-decade and trajectory partitions. After confirming all 1,600 entity-decade joins, pooled Cramér's V was recalculated between the sensitivity-specific trajectory groups and aligned regime families. Together, these comparisons tested whether conclusions about pairwise similarity, clustering and cross-scale association depended materially on applying conventional PCA directly to the original age shares.

Graph and temporal-specification sensitivity

Because graph construction can influence graph-based clustering results³, sensitivity was evaluated across all 15 combinations of nearest-neighbour count:

$$k_{\text{NN}} \in \{8, 10, 12, 15, 20\}$$

and Gaussian-bandwidth multiplier:

$$m \in \{0.5, 1, 2\}.$$

For each combination, the union-symmetrised nearest-neighbour graph was reconstructed, retaining an edge when either entity selected the other as a neighbour. Gaussian affinities were calculated using a bandwidth equal to m times the median

positive retained-edge distance, while the remaining spectral-clustering and eigengap-selection procedures were held fixed. The 12-neighbour graph with bandwidth multiplier $m = 1$ was designated as the primary specification.

For global-configuration robustness, each of the eight decades was reclustered under all 15 graph specifications. For every decade and specification, the eigengap-selected cluster count, Euclidean silhouette and ARI relative to the corresponding primary decade-specific partition were recorded. The primary within-decade clusters were then represented by renormalised, equal-entity mean profiles across all available age groups. These centroids were compared across decades using base-2 Jensen–Shannon distance⁴ and realigned by average-linkage agglomerative clustering, with the number of broad regime families selected by maximum silhouette over $k = 2, \dots, 10$ ⁵. For each aligned centroid, uncertainty was quantified using its distance to the assigned-family centroid and to the nearest alternative-family centroid; the assignment margin was defined as the latter minus the former, so smaller or negative margins indicated less-certain alignment. These diagnostics are reported in Supplementary Figs. S1 and S2.

The same 15 graph specifications were applied to the unrestricted four-PC DTW distance matrix. For each specification, the eigengap-selected k , DTW silhouette and ARI relative to the primary trajectory partition were recorded. To construct a consensus partition, a country-pair co-assignment matrix was calculated across the 15 trajectory graph specifications, with each entry giving the proportion of specifications in which the corresponding countries were assigned to the same group. A consensus distance matrix was defined as one minus the co-assignment probability and clustered using average-linkage agglomerative clustering. Candidate solutions with $k = 2, \dots, 10$ were evaluated using silhouette width, and agreement between the selected consensus partition and the primary 12-neighbour, bandwidth-multiplier-1 trajectory partition was quantified using the adjusted Rand index⁶. Stochastic stability, a recognised means of assessing clustering structure⁷, was assessed by repeating the primary specification with 20 different k -means initialisation seeds while holding the distance matrix, graph construction and selected k fixed; each resulting partition was compared with the primary assignment using ARI. Temporal sensitivity was evaluated by recomputing DTW distances under Sakoe–Chiba constraints⁸ of one and two observations, restricting admissible alignments to $|i - j| \leq 1$ and $|i - j| \leq 2$, corresponding to one- and two-decade displacement. An unwarped comparison permitted only contemporaneous matches ($i = j$), thereby measuring time-aligned trajectory distance without temporal realignment. The resulting partitions were compared with the unrestricted solution. Separately, PC1-only DTW distances and assignments were compared with the four-PC results using Spearman correlation and ARI, respectively. Graph-specification diagnostics are reported in Supplementary Fig. S3, the PC1-only comparison is reported in Supplementary Fig. S4, and the temporal-specification results are reported in Supplementary Table S2.

Supplementary migration-validation procedures

Stability of the migrant-stock-associated direction was assessed using 1,000 country-level bootstrap resamples. Countries were sampled with replacement, and all available observations for each selected country were retained together to preserve within-country dependence. In every bootstrap sample, the same migrant-stock model was refitted with the demographic controls and all four standardised PC predictors, and the resulting PC coefficient vector $\beta^{(b)}$ was extracted. Its normalised direction was

$$\mathbf{w}^{(b)} = \frac{\beta^{(b)}}{\|\beta^{(b)}\|_2},$$

with the primary full-sample direction $\mathbf{w}^{(0)}$ defined analogously. Directional agreement was measured by the cosine similarity

$$c^{(b)} = \mathbf{w}^{(b)\top} \mathbf{w}^{(0)}.$$

The normalised component weights were the four elements of $\mathbf{w}^{(b)}$, while sign consistency for each PC was the proportion of bootstrap estimates having the same coefficient sign as its primary estimate. All 1,000 bootstrap fits were successful; their cosine distribution, component weights and sign-consistency values are reported in Supplementary Fig. S6.

High migrant stock was defined separately within each year using the top 5%, 10% or 20% of the migrant-stock distribution, with the top-decile definition prespecified as primary. For each country-grouped five-fold split, repeated ten times, the relevant percentile thresholds were estimated using training countries only and then applied to the held-out countries. PCA fitting, numerical-predictor standardisation and logistic-regression estimation were likewise confined to the training fold. Three classifiers were compared: demographic controls only, the four PCs only, and demographic controls plus all four PCs. Area under the receiver operating characteristic curve (ROC-AUC) measured the probability that a randomly selected high-stock observation received a higher predicted score than a non-high-stock observation, whereas area under the precision–recall curve (PR-AUC) summarised the precision–recall trade-off within each prevalence definition⁹. Incremental gain was calculated within each fold as the combined-model AUC minus the controls-only AUC, and fold-wise consistency was the proportion of held-out folds in which this difference was positive. Threshold-specific mean performance, incremental gains and improvement frequencies are reported in Supplementary Fig. S7.

Migration-flow sensitivity was evaluated using three formulations of cumulative net migration per 1,000 over each decadal interval. The raw outcome was used on its original scale; the signed-log outcome was defined as

$$y_{\text{slog}} = \text{sign}(y) \log(1 + |y|),$$

thereby retaining the distinction between net immigration and emigration while reducing the influence of extreme values; and the annualised outcome divided the cumulative value by the corresponding interval duration in years. Models included interval-mean fertility and life expectancy as demographic controls together with interval effects. Generalisation was assessed using nested country-grouped validation, with countries kept intact across folds, outer folds reserved for evaluation and inner folds used for PC-subset selection. PCA fitting and predictor standardisation were repeated using training data only. Held-out improvement in the sum of squared errors (SSE) was quantified as

$$\Delta_{\text{SSE}} = 100 \times \frac{\text{SSE}_{\text{controls}} - \text{SSE}_{\text{controls+PCs}}}{\text{SSE}_{\text{controls}}},$$

so that positive values indicated lower prediction error after adding the selected PCs. Numerical results for these comparisons are reported separately in the detailed migration-flow diagnostics.

Computational implementation and reproducibility

Analyses were executed using Python 3.12.13 with NumPy 2.4.4, pandas 3.0.3, SciPy 1.17.1, scikit-learn 1.9.0 and statsmodels 0.14.6 under the locked computational environment supplied with the project. The reported workflow was reproduced from the project root using

```
python run_all.py --main --global --migration
```

whereas the optional `-sex-difference` branch was not used for the reported analyses. Primary spectral-clustering assignments used `random_state=42` with `n_init=50` as recorded in the project configuration. Repeated country-grouped cross-validation and high-stock classification used a base seed of 42, incremented deterministically by the repeat number. Nested country-grouped validation used the same base seed with deterministic offsets for the outer and inner splits. The 500-replicate migrant-stock and migration-flow association bootstraps both used seed 42. The 1,000-replicate directional-stability analysis used seed 42 for migrant stock and seed 10,042 for migration flow. The trajectory initialisation-stability analysis deliberately used seeds 0–19 while holding the distance matrix, graph specification and selected number of clusters fixed. Deterministic input ordering, fixed entity and year manifests, and the locked package environment were used to reproduce the reported outputs.

Fixed entity and year manifests defined the analytical grid of 200 entities and eight decadal observations from 1950 to 2020. Automated validation checks covered required-file availability, analytical dimensions, complete and unique entity-year coverage, age-share normalisation, finite values, PCA outputs, DTW-matrix symmetry and diagonal values, graph construction, decade-specific assignments, cross-scale joins, geographic mappings, migration-panel dimensions and required analytical outputs. All 21 checks passed during the clean execution used to generate the reported results.

Composition-aware sensitivity results

All four zero-free log-ratio representations selected three trajectory groups and three aligned regime families (Supplementary Table S1). Relative to the primary conventional four-PC analysis, trajectory-distance correlations ranged from 0.899 to 0.938, trajectory ARI from 0.466 to 0.638, pooled aligned-regime ARI from 0.857 to 0.873, and pooled Cramér’s *V* from 0.544 to 0.548.

Temporal-specification sensitivity

The one- and two-decade-constrained DTW specifications retained three trajectory groups and closely reproduced the unrestricted primary partition (Supplementary Table S2). The unwarped time-aligned comparison preserved most pairwise distance ordering but selected six groups and showed lower membership agreement with the primary partition.

Trajectory graph-consensus robustness

Consensus clustering across the 15 trajectory graph specifications selected $k = 3$, with a consensus-distance silhouette of 0.828. The resulting consensus partition closely agreed with the primary trajectory solution (ARI = 0.934).

Table S1. Composition-aware sensitivity of the trajectory, regime-alignment and cross-scale results. Zero-free profiles were obtained by amalgamating the upper-age tail into either an 85+ or 80+ category. CLR-PCA representations retained either the minimum number of components explaining at least 95% of variance or the complete Aitchison rank.

Log-ratio representation			Trajectory analysis				Global/cross-scale	
Tail	Retention	d	k_{traj}	Silhouette	ρ_{dist}	ARI_{traj}	ARI_{reg}	V_{pooled}
85+	95% variance	3	3	0.344	0.899	0.466	0.860	0.544
85+	Full rank	17	3	0.326	0.910	0.591	0.857	0.548
80+	95% variance	4	3	0.360	0.932	0.627	0.873	0.548
80+	Full rank	16	3	0.332	0.938	0.638	0.871	0.548

Notes: d is the number of retained CLR-PCA coordinates. ρ_{dist} is the Spearman correlation between the upper-triangular elements of the alternative and primary DTW distance matrices. The trajectory-group and regime-family ARIs compare alternative trajectory-group assignments and pooled aligned regime-family memberships, respectively, with the primary conventional four-PC analysis. All four representations produced three aligned regime families. The primary pooled Cramér's V was 0.549103. Values in the table are rounded to three decimal places.

Table S2. Sensitivity of the trajectory solution to temporal alignment constraints. The unrestricted multivariate DTW analysis was the primary specification. Constrained analyses limited temporal displacement to one or two decadal observations, whereas the unwarped comparison permitted only contemporaneous matches.

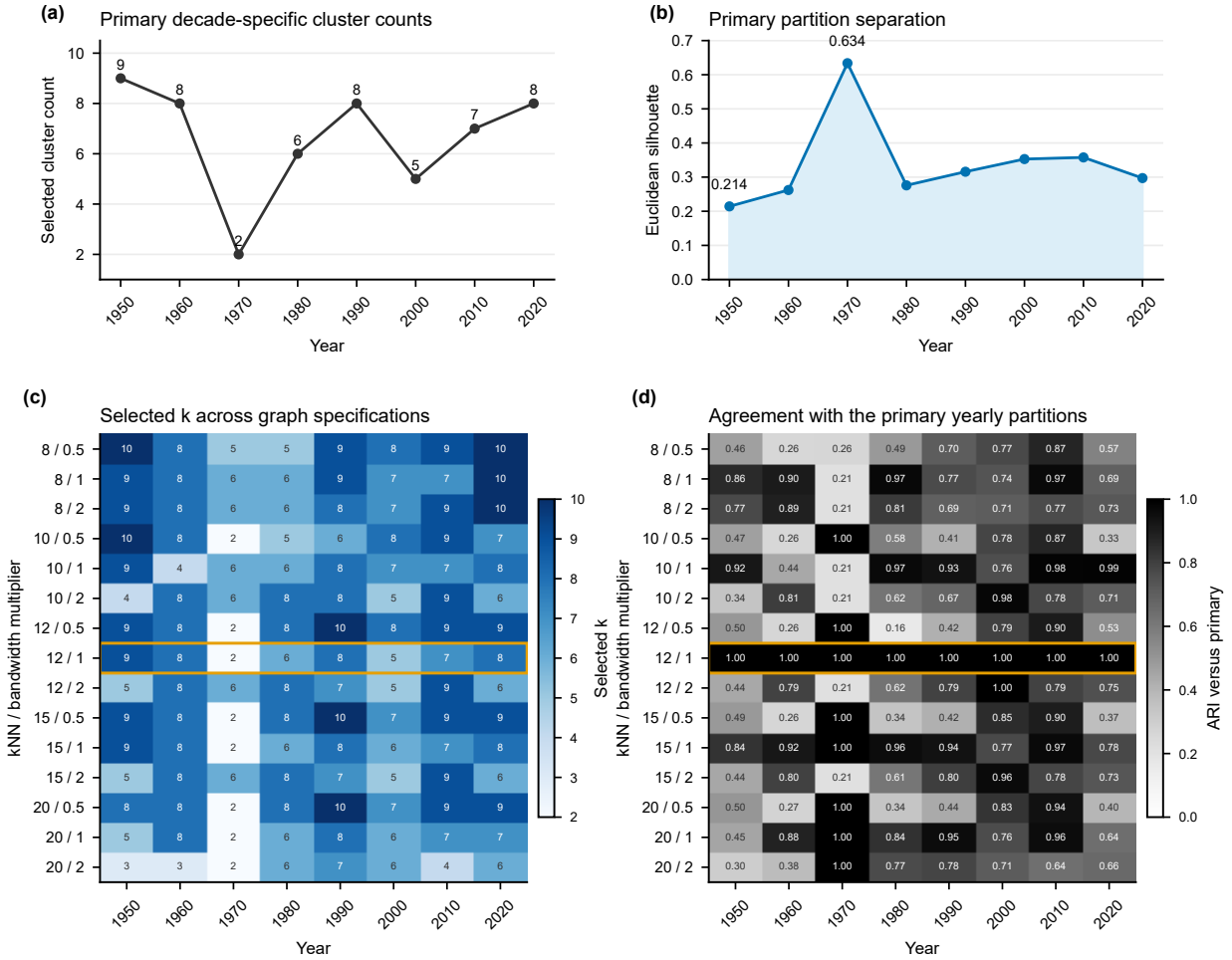
Specification	Allowed displacement	Selected k	Eigengap	Silhouette	ρ_{dist}	ARI_{traj}
Unrestricted DTW	Unrestricted	3	0.110	0.388	1.000	1.000
Two-decade window	$ i - j \leq 2$	3	0.109	0.391	0.999	1.000
One-decade window	$ i - j \leq 1$	3	0.099	0.385	0.996	0.983
Unwarped, time-aligned	$i = j$	6	0.080	0.207	0.986	0.490

Notes: ρ_{dist} is the Spearman correlation between the upper-triangular elements of each distance matrix and the primary unrestricted-DTW distance matrix. ARI_{traj} is the adjusted Rand index comparing the resulting trajectory partition with the primary unrestricted-DTW partition. Each observation represents one decade.

Detailed migration-flow diagnostics

The migration-flow analysis included 1,393 country-interval observations from 199 entities across seven decadal intervals. After adjustment for demographic controls, PC2 explained little additional variation in raw net migration flow (partial $R^2 = 0.016856$), and this association nearly disappeared following signed-log transformation (partial $R^2 = 0.000672$). In nested country-grouped validation, adding PC2-PC4 reduced held-out SSE by only 0.6272%, with positive reductions in 60% of outer folds. Migration-flow evidence was therefore weak, less generalisable and sensitive to target transformation, and should be interpreted descriptively rather than causally.

Supplementary Figures



Orange outline denotes the primary 12-NN, bandwidth-multiplier 1 specification.

Figure S1. Fine within-decade partitions were sensitive to graph construction. (a) Eigengap-selected cluster counts for the primary 12-nearest-neighbour, bandwidth-multiplier-1 specification in each decade, based on 200 entities. (b) Euclidean silhouette coefficients for the corresponding primary decade-specific partitions. (c) Selected cluster count across 15 graph specifications formed from $k_{NN} \in \{8, 10, 12, 15, 20\}$ and Gaussian-bandwidth multipliers $\{0.5, 1, 2\}$. (d) Adjusted Rand index (ARI) between each sensitivity partition and the primary partition for the same decade. The orange outlines identify the primary specification.

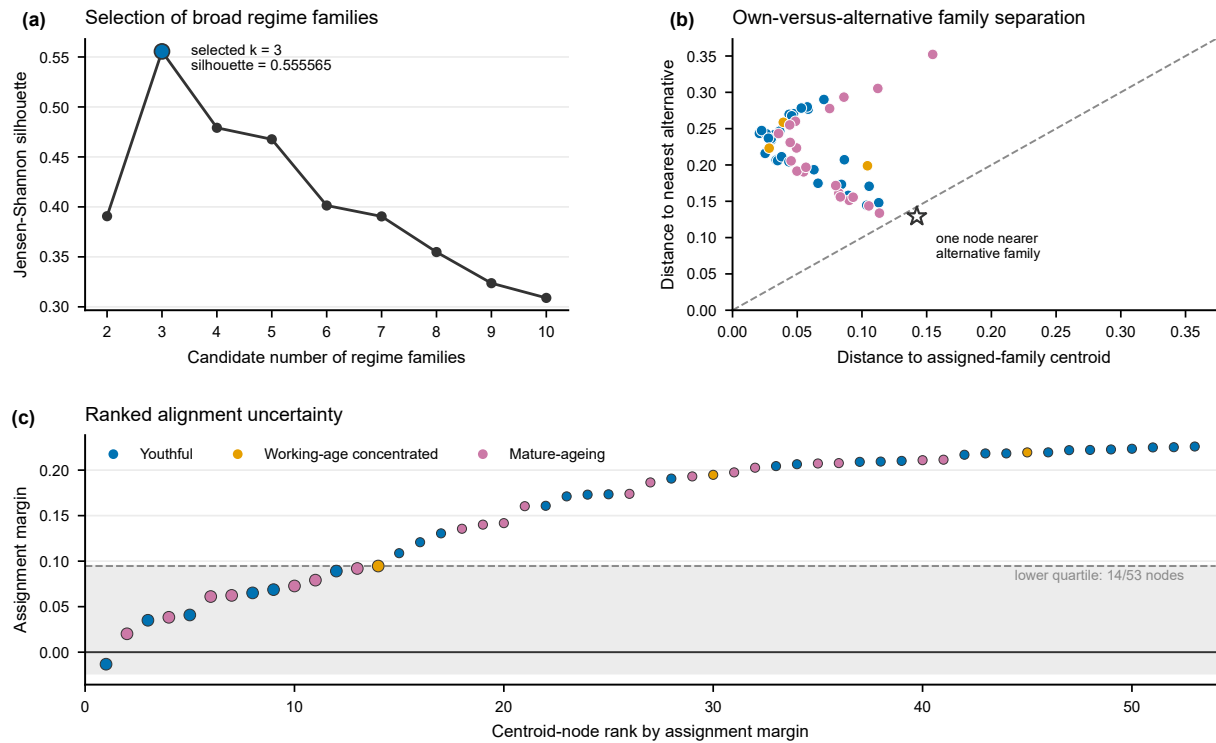
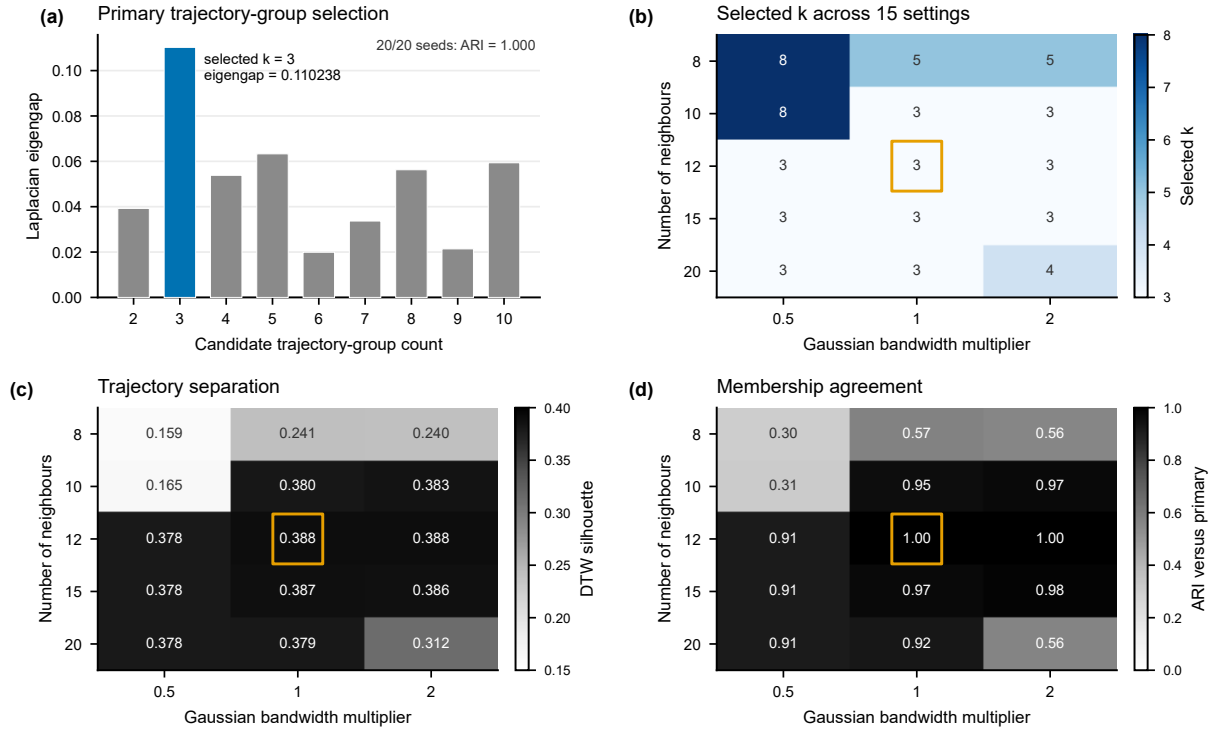


Figure S2. The centroid alignment recovered three broad regime families, with uncertainty concentrated in a minority of assignments. (a) Jensen–Shannon silhouette across candidate numbers of broad regime families obtained by average-linkage clustering of the 53 complete age-profile centroids. Three families were selected, with silhouette 0.555565. (b) Jensen–Shannon distance from each centroid to its assigned-family centroid versus its nearest alternative-family centroid. Points above the diagonal are nearer their assigned family; the starred centroid below the diagonal is nearer an alternative-family centroid despite its hierarchical assignment. (c) Centroids ranked by assignment margin, defined as the distance to the nearest alternative family minus the distance to the assigned family. The shaded region and dashed line identify the lower quartile, containing 14 of 53 centroids. Colours denote youthful, working-age-concentrated and mature-ageing families.



Orange outline denotes the primary 12-NN, bandwidth-multiplier 1 specification; $k=3$ was selected in 10/15 settings.

Figure S3. The three-group trajectory solution was stable across initialisations but not invariant across graph specifications. (a) Laplacian eigengaps for candidate trajectory-group counts under the primary four-PC dynamic-time-warping (DTW) specification. The largest eigengap selected $k = 3$; all 20 initialisation seeds reproduced the primary assignments exactly (ARI = 1.000). (b) Selected k across the 15 combinations of nearest-neighbour count and Gaussian-bandwidth multiplier. (c) DTW silhouette for each graph specification. (d) ARI between each sensitivity partition and the primary trajectory partition. The orange outlines identify the primary 12-nearest-neighbour, bandwidth-multiplier-1 specification. Three groups were selected in 10 of the 15 settings.

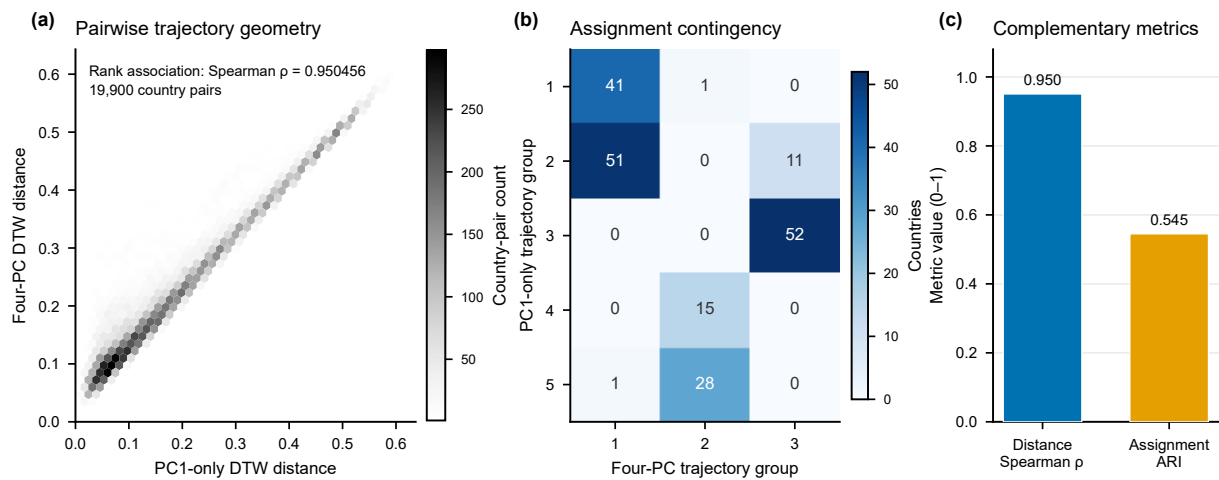


Figure S4. PC1 preserved pairwise trajectory ordering but reproduced four-PC group membership only partially. (a) Hexagonal-bin comparison of PC1-only and four-PC DTW distances for all 19,900 unordered pairs of 200 entities. Colour intensity represents the number of entity pairs; Spearman's $\rho = 0.950456$ indicates strong agreement in pairwise distance ordering. (b) Contingency table comparing the independently selected five-group PC1-only partition with the three-group four-PC partition; cell values are entity counts, and cluster numbers are arbitrary identifiers within each solution. (c) Complementary agreement measures for distance geometry and cluster membership. Although PC1 preserved most relative trajectory geometry, assignment concordance was only partial (ARI = 0.544934). Spearman's ρ and ARI quantify different properties and should not be interpreted as directly comparable effect sizes.

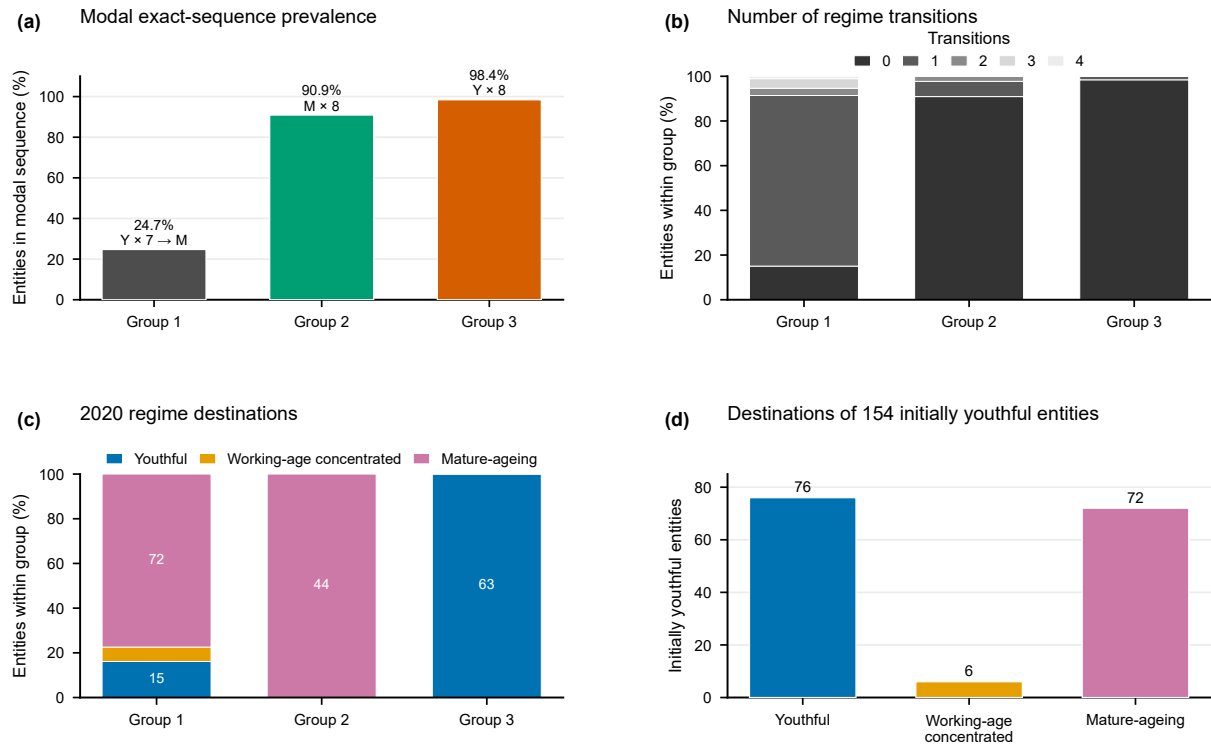


Figure S5. Contemporaneous regime families and complete trajectory groups were related but not interchangeable. (a) Percentage of entities following the modal exact eight-decade regime sequence in each trajectory group. Here, Y, W and M denote youthful, working-age-concentrated and mature-ageing regime families, respectively; $\times 8$ denotes membership in the same family at all eight observations. (b) Within-group distribution of the number of adjacent regime transitions from 1950 to 2020. (c) Composition of each trajectory group by its 2020 regime-family destination; numbers within sufficiently large segments are entity counts. (d) 2020 destinations of the 154 entities assigned to the youthful regime family in 1950: 76 remained youthful, 6 became working-age concentrated and 72 became mature-ageing. Family colours are consistent across panels.

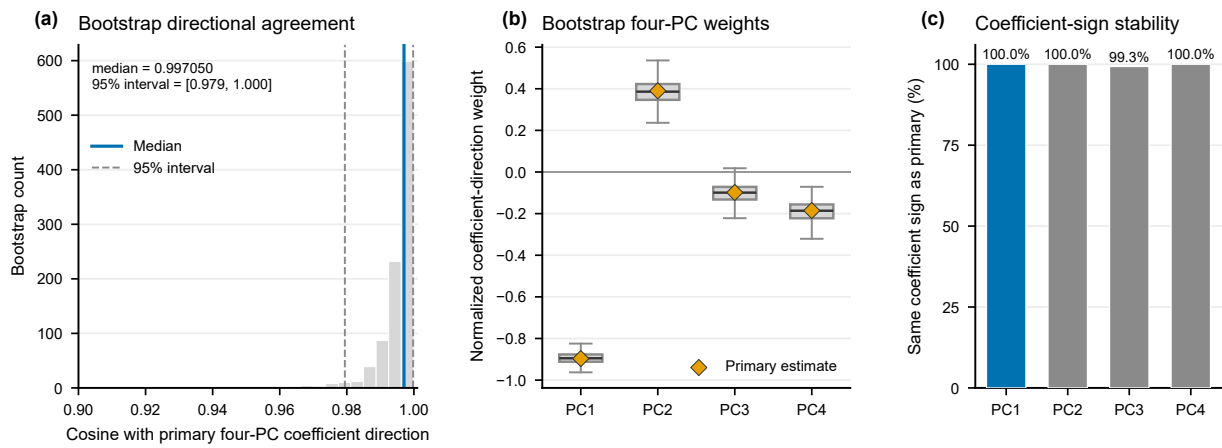


Figure S6. The migrant-stock-associated four-PC direction was highly stable under country resampling. The analysed direction is the unit-normalised vector of the four standardised PC coefficients from the migrant-stock model. **(a)** Distribution of cosine similarity between the primary direction and directions re-estimated in 1,000 successful country-level bootstrap samples. The solid blue line marks the median (0.997050), and dashed lines mark the empirical 95% interval (0.979–1.000). **(b)** Bootstrap distributions of the normalised coefficient-direction weights. Boxes show the interquartile range, central lines show medians, whiskers extend to 1.5 times the interquartile range, and orange diamonds denote the primary estimates. **(c)** Percentage of bootstrap estimates retaining the sign of the corresponding primary coefficient.

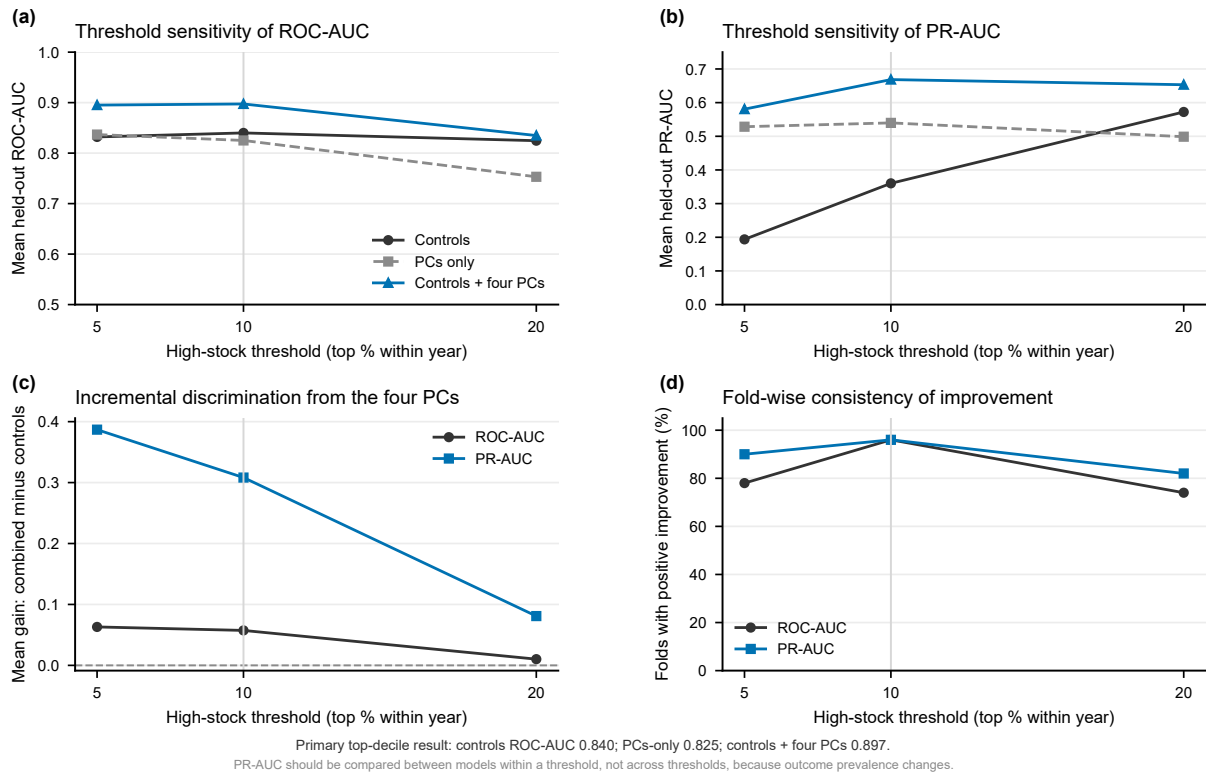


Figure S7. Adding four PCs to demographic controls improved high-stock classification, although the gain weakened as the threshold broadened. High migrant stock was defined within year using the top 5%, 10% or 20% of the training-country distribution, with the top-decile definition prespecified as primary. Country-grouped five-fold validation was repeated ten times; PCA fitting, predictor standardisation and outcome-threshold estimation were confined to the training folds. Models used demographic controls only, all four PCs only, or controls plus all four PCs. **(a)** Mean held-out area under the receiver-operating-characteristic curve (ROC-AUC). **(b)** Mean held-out area under the precision–recall curve (PR-AUC). **(c)** Mean incremental AUC of the combined model relative to controls alone. **(d)** Percentage of held-out folds in which the combined model improved upon controls. At the primary top-decile threshold, ROC-AUC was 0.840 for controls, 0.825 for PCs only and 0.897 for controls plus four PCs. Because positive-class prevalence changes with the threshold, PR-AUC should be compared between models within each threshold rather than directly across thresholds.

References

1. Aitchison, J. The statistical analysis of compositional data. *J. Royal Stat. Soc. Ser. B (Methodological)* **44**, 139–160, DOI: [10.1111/j.2517-6161.1982.tb01195.x](https://doi.org/10.1111/j.2517-6161.1982.tb01195.x) (1982).
2. Aitchison, J. Principal component analysis of compositional data. *Biometrika* **70**, 57–65, DOI: [10.1093/biomet/70.1.57](https://doi.org/10.1093/biomet/70.1.57) (1983).
3. Maier, M., von Luxburg, U. & Hein, M. Influence of graph construction on graph-based clustering measures. In *Advances in Neural Information Processing Systems 21*, 1025–1032 (2008).
4. Endres, D. M. & Schindelin, J. E. A new metric for probability distributions. *IEEE Transactions on Inf. Theory* **49**, 1858–1860, DOI: [10.1109/TIT.2003.813506](https://doi.org/10.1109/TIT.2003.813506) (2003).
5. Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65, DOI: [10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7) (1987).
6. Hubert, L. & Arabie, P. Comparing partitions. *J. Classif.* **2**, 193–218, DOI: [10.1007/BF01908075](https://doi.org/10.1007/BF01908075) (1985).
7. Ben-Hur, A., Elisseeff, A. & Guyon, I. A stability based method for discovering structure in clustered data. In *Pacific Symposium on Biocomputing 2002*, 6–17, DOI: [10.1142/9789812799623_0002](https://doi.org/10.1142/9789812799623_0002) (World Scientific, 2002).
8. Sakoe, H. & Chiba, S. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoust. Speech, Signal Process.* **26**, 43–49, DOI: [10.1109/TASSP.1978.1163055](https://doi.org/10.1109/TASSP.1978.1163055) (1978).
9. Saito, T. & Rehmsmeier, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE* **10**, e0118432, DOI: [10.1371/journal.pone.0118432](https://doi.org/10.1371/journal.pone.0118432) (2015).