

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

Athulya Ratnayake

e19328@eng.pdn.ac.lk

Multidisciplinary AI Research Center

Lelumi Edirisinghe

Multidisciplinary AI Research Center

Thisumi Wijesinghe

Multidisciplinary AI Research Center

Roshan Godaliyadda

University of Peradeniya

Parakrama Ekanayake

University of Peradeniya

Vijitha Herath

University of Peradeniya

Samath Dharmaratne

University of Peradeniya

Article

Keywords: Demographic change, Unsupervised learning, Spectral clustering, Demographic trajectories

Posted Date: October 23rd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11083501/v1>

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

[Read Full License](#)

Additional Declarations: No competing interests reported.

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

Athulya Ratnayake^{1,2*}, Lelumi Edirisinghe^{1,3},
Thisumi Wijesinghe^{1,4}, Roshan Godaliyadda^{5,1},
Parakrama Ekanayake^{5,1}, Vijitha Herath^{5,1}, Samath Dharmaratne⁶

^{1*}Multidisciplinary AI Research Center, University of Peradeniya,
Peradeniya, 20400, Sri Lanka.

²Farbe Technologies (Pvt.) Ltd., Peradeniya, 20400, Sri Lanka.

³Department of Mathematics, Faculty of Science, University of Ruhuna,
Matara, 81000, Sri Lanka.

⁴Department of Computer Science with Artificial Intelligence, NIBM
Kandy Innovation Center, Peradeniya, 20400, Sri Lanka.

⁵Department of Electrical and Electronic Engineering, Faculty of
Engineering, University of Peradeniya, Peradeniya, 20400, Sri Lanka.

⁶Department of Community Medicine, Faculty of Medicine, University
of Peradeniya, Peradeniya, 20400, Sri Lanka.

*Corresponding author(s). E-mail(s): e19328@eng.pdn.ac.lk;

Contributing authors: edirisinghe11783@usci.ruh.ac.lk;

thisumi.nimsani@gmail.com; roshang@eng.pdn.ac.lk;

mpbe@eng.pdn.ac.lk; vijitha@eng.pdn.ac.lk;

samath.dharmaratne@med.pdn.ac.lk;

Abstract

Population pyramids record how fertility, mortality and migration shape age structure, but international comparisons often reduce them to summary indicators or assign countries to predetermined transition stages. We used unsupervised machine learning to compare complete age profiles, distinguishing configurations that coexist at a given date from the longitudinal paths traced by countries. We analysed 200 countries and territories at decadal intervals from 1950 to 2020. Four principal components jointly represented 96.48% of age-profile

variation. Independently estimated decade-specific configurations aligned into three broad, interpretable families, while complete multivariate country histories formed three trajectory groups. The classifications were substantially associated but not interchangeable (pooled Cramér’s $V = 0.549$), showing that a contemporaneous configuration did not uniquely specify a national history. The dominant component described a youthful-to-mature gradient and preserved most trajectory-distance ordering (Spearman $\rho = 0.950$), yet reproduced the multivariate grouping only partly (adjusted Rand index = 0.545). Held-out migrant-stock prediction improved when all four components were added to demographic controls, whereas migration-flow evidence was weaker and transformation-sensitive. The framework therefore complements demographic-transition theory by allowing complete age structures and national pathways to be identified empirically before they are interpreted through established demographic concepts.

Keywords: Demographic change, Unsupervised learning, Spectral clustering, Demographic trajectories

1 Introduction

Population-age composition is a fundamental dimension of social and economic change, shaping patterns of dependency, labour supply, health-care demand and the provision of social support (Bloom et al. 2015). Population pyramids provide a compact demographic record of this composition, reflecting the cumulative influence of fertility, mortality and migration across successive cohorts (Wilson 2016). Since the mid-twentieth century, declining fertility and increasing longevity have substantially altered these profiles, producing a broad global shift towards older age structures (Lutz et al. 2008). Yet cross-national differences in the timing, pace and form of this reorganisation (Ma et al. 2025) raise a complementary question: what configurations and pathways can be identified from complete age profiles when transition stages are not assigned beforehand?

The demographic transition provides an enduring framework for interpreting long-term shifts from high to low mortality and fertility (Kirk 1996; Bongaarts 2009). Its broad stages offer valuable comparative reference points, but they are not expected to prescribe identical national sequences (Kirk 1996; Reher 2004). Transitions have varied in their onset, pace and duration, and may include periods of stagnation, acceleration or reversal as fertility, mortality and migration interact (Casterline 2001; Bongaarts 2006; Wu and Pesando 2024). Countries occupying comparable positions within the broad transition may therefore retain markedly different age profiles. Such heterogeneity does not invalidate the framework; it instead supports a complementary analysis in which groups are identified from complete age distributions and demographic-transition theory informs their interpretation.

Broad age shares, dependency ratios and median age provide interpretable summaries of population structure (Sanderson and Scherbov 2007). Their clarity, however, rests on compressing cohort-specific variation. Using the complete age distribution

as the analytical unit preserves differences in the concentration of children, young adults, working-age groups and older people among pyramids occupying similar positions along the transition gradient. Representing complete age profiles within a common multivariate space permits the dominant gradient to be examined alongside secondary differences in demographic form (Cecchini et al. 2019). A complementary cross-sectional approach has used Kullback–Leibler divergence between complete population pyramids to examine associations between age-structure morphology and cause-specific mortality (Fonseka et al. 2025). The importance of those differences for empirical classification must then be assessed rather than presumed. Yet this representation alone cannot distinguish contemporaneous global configurations from complete national histories.

Cross-sectional and longitudinal analyses address distinct dimensions of demographic change. The former is a snapshot of age structures coexisting at one date; the latter traces each country’s ordered history (Billari 2001). Countries may therefore resemble one another at one date despite following different paths, or begin similarly and later diverge. Here, “regime family” denotes an aligned, descriptive grouping of age-structure configurations, not a causal, institutional or predetermined transition state. Because the decade-specific configurations are estimated independently, they must be aligned through time before being reconnected to national trajectories. This comparison asks how much structure at either scale the dominant demographic gradient captures.

Because demographic transition reorganises age composition systematically, population-pyramid variation may be dominated by a leading youthful-to-mature gradient (Bongaarts 2009). Yet preserving countries’ broad ordering along this gradient does not ensure identical multivariate similarities or resulting group memberships. Secondary dimensions may retain cohort concentrations and other differences in shape that are orthogonal to the principal pattern (Jolliffe and Cadima 2016). Their value should not be judged solely by explained variance or cluster separation, nor need all retained components contribute equally. The question is whether the dominant gradient alone preserves similarities among national histories or whether lower-variance dimensions alter them in stable, interpretable and externally relevant ways.

International migration provides an external test of whether variation beyond the dominant gradient carries demographically relevant information. Because migration is strongly age-selective, inflows or outflows can modify age composition, particularly at young-adult and working ages, although fertility and mortality remain central (Dörflinger and Loichinger 2024; Raymer et al. 2025). Such variation need not align with a single component. We therefore examine migrant stock, migration flows and migrant-age profiles as complementary descriptive and predictive evidence. Migration data are used neither to construct the representation nor to define the clusters, and these comparisons do not establish migration as the cause of any configuration.

Here, we apply this framework to 200 countries and territories observed at decadal intervals from 1950 to 2020. Complete population pyramids are represented in a common principal-component space with directly interpretable age-specific loadings. Contemporaneous configurations are identified independently within each decade,

while national histories are analysed separately as multivariate trajectories. We reconnect the two scales through each country’s sequence of aligned age-structure families, test how fully the dominant youthful-to-mature gradient represents longitudinal similarity, and use migrant stock, migration flows and migrant-age profiles for external validation. This integration allows demographic configurations and national pathways to be identified before they are interpreted through demographic-transition theory, while keeping contemporaneous positions analytically distinct from longitudinal histories.

2 Results

Population pyramids revealed a dominant youthful-to-mature gradient and additional age-structure variation

Principal component analysis (PCA) was fitted once to all 1,600 country-year population profiles. Four principal components (PCs) were required to exceed the prespecified 95% variance threshold, explaining 81.4628%, 7.8779%, 4.9290% and 2.2101% individually, and 96.4798% cumulatively (Fig. 1a). This marked dominance of PC1 established the principal variance hierarchy, but explained variance alone did not determine substantive demographic importance. The age-specific loading patterns therefore provided the basis for interpreting the retained components.

PC1 contrasted child-heavy with mature and older profiles, with positive loadings at ages 0–4 (0.634), 5–9 (0.451) and 10–14 (0.313), and negative loadings from approximately age 25. PC1 can therefore be interpreted as an empirical youthful-to-mature gradient. PCs 2–4 resolved features of profile shape that position on this gradient did not capture. PC2 indicated working-age concentration, peaking at ages 25–29 (0.475), 20–24 (0.403) and 30–34 (0.397). PC3 contrasted positive loadings across ages 10–24, strongest at 15–19 (0.535), with a negative loading at ages 0–4 (-0.511), while PC4 showed finer oscillatory differences, including at ages 20–24 (0.485) (Fig. 1b–e). PCA orientation is arbitrary, so interpretation concerns relative age-specific contrasts; signs were retained for computational traceability.

Global demographic configurations were redistributed but remained heterogeneous

The analysis treated each decade as a separate cross-section. Cluster number and membership were therefore allowed to reflect the age-profile structure present in that decade, rather than being constrained to a common sequence of historical stages. Among 200 entities per decade, the selected fine partitions contained 9, 8, 2, 6, 8, 5, 7 and 8 clusters from 1950 to 2020, respectively, with decade-specific silhouette coefficients ranging from 0.214 to 0.634 (Supplementary Fig. S1). These fine clusters summarised within-decade structure, and their sensitivity to neighbourhood and kernel settings cautioned against treating their counts as stable historical categories. Cross-decade interpretation instead centred on alignment of their complete age-profile centroids.

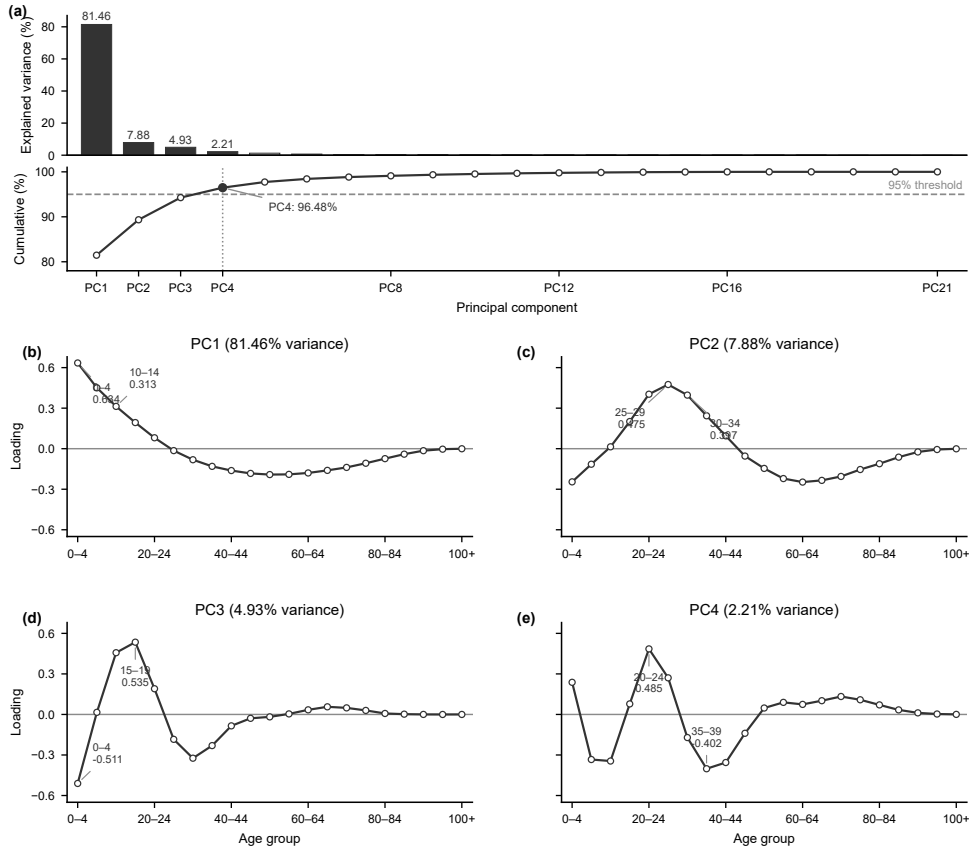


Fig. 1 Principal component analysis (PCA) revealed that population-age variation was strongly dominated by one dimension but was not exclusively one-dimensional. (a) Individual and cumulative explained variance for the PCA fitted to 1,600 country-year population profiles; the dashed line denotes the prespecified 95% cumulative-variance threshold. (b-e) Age-specific loading profiles for PC1-PC4, respectively. PC1 captures the dominant youthful-to-mature contrast, whereas the remaining components describe lower-variance differences in cohort concentration. Loading signs are arbitrary and interpretation concerns relative age-specific contrasts.

This alignment yielded three broad, interpretable families (alignment silhouette = 0.555565; Fig. 2a,b; Supplementary Fig. S2). The youthful centroid contained 40.98% aged 0–14 and 3.73% aged 65+, whereas the working-age-concentrated centroid contained 76.81% aged 15–64, including 57.83% aged 25–54. The mature-ageing centroid contained 22.54% aged 0–14 and 11.36% aged 65+. Their prevalence shifted from 154 youthful, 0 working-age-concentrated, and 46 mature-ageing entities in 1950 to 78, 6, and 116, respectively, in 2020. These labels describe centroid shapes rather than universal stages; their geographic distribution and representative cases further clarified this redistribution.

Redistribution was geographically uneven, with all three families represented in 2020 (Fig. 2c,d). Youthful membership in sub-Saharan Africa changed from 48/50 entities in 1950 to 45/50 in 2020. Mature-ageing membership expanded from 35/40 to 40/40 in Europe and from 1/7 to 7/7 in Eastern Asia. Qatar and the United Arab Emirates (UAE) entered the working-age-concentrated family in 1990; its 2020 members were Bahrain, Maldives, Oman, Qatar, Saudi Arabia and the UAE. The concentration of Gulf states among its members makes this configuration especially clear: a pronounced working-age concentration distinguishes the profile from both youthful and mature-ageing shapes, rather than placing it at a simple intermediate position along the dominant gradient. This category emerged from the complete age profiles and was not specified in advance. Nevertheless, 14/53 aligned nodes had lower-quartile margins, and one was nearer an alternative family centroid (Supplementary Fig. S2). The labels therefore provide defensible shorthand for heterogeneous contemporaneous states rather than taxonomic certainties or complete national histories.

Countries followed three broad multivariate demographic trajectories

Complete eight-decade trajectories in the four-component PCA space formed three broad trajectory groups ($k = 3$; eigengap = 0.110238; silhouette = 0.388135; Fig. 3a,c; Supplementary Fig. S3), containing 93, 44 and 63 entities. All 20 initialisation seeds produced identical assignments, giving an adjusted Rand index (ARI) of 1.000. The groups described three broad forms of long-term change: substantial maturation, further ageing from a relatively mature baseline, and persistent youthfulness. They summarise empirical patterns rather than rigid paths, and their demographic interpretation comes from age-profile evolution rather than numerical labels.

In the ageing-from-a-mature-baseline group ($n = 44$), the mean share aged 0–14 declined from 27.19% in 1950 to 16.32% in 2020, while the share aged 65+ rose from 8.09% to 18.79%; 36 members were European. In the persistently youthful group ($n = 63$), the corresponding shares changed only modestly, from 41.39% to 40.41% and from 3.74% to 3.34%; 45 members were African. These equal-country means are neither population-weighted global estimates nor evidence of complete demographic stasis.

The broad-maturation group ($n = 93$) showed greater variation in timing and final configuration (Fig. 3a,b). Its mean share aged 0–14 declined from 39.54% to 23.35%, while the shares aged 15–64 and 65+ rose from 56.19% to 67.65% and from 4.27% to 9.00%, respectively. Its membership spanned every region, with the largest representations in Asia (39 entities) and the Americas (34), and included Bahrain, Kuwait, Oman, Qatar, Saudi Arabia and the UAE. The primary three-group solution occurred in 10/15 sensitivity settings, with the remaining settings selecting four, five or eight groups (Supplementary Fig. S3). This specification sensitivity motivated examination of how much trajectory structure the dominant component alone preserved.

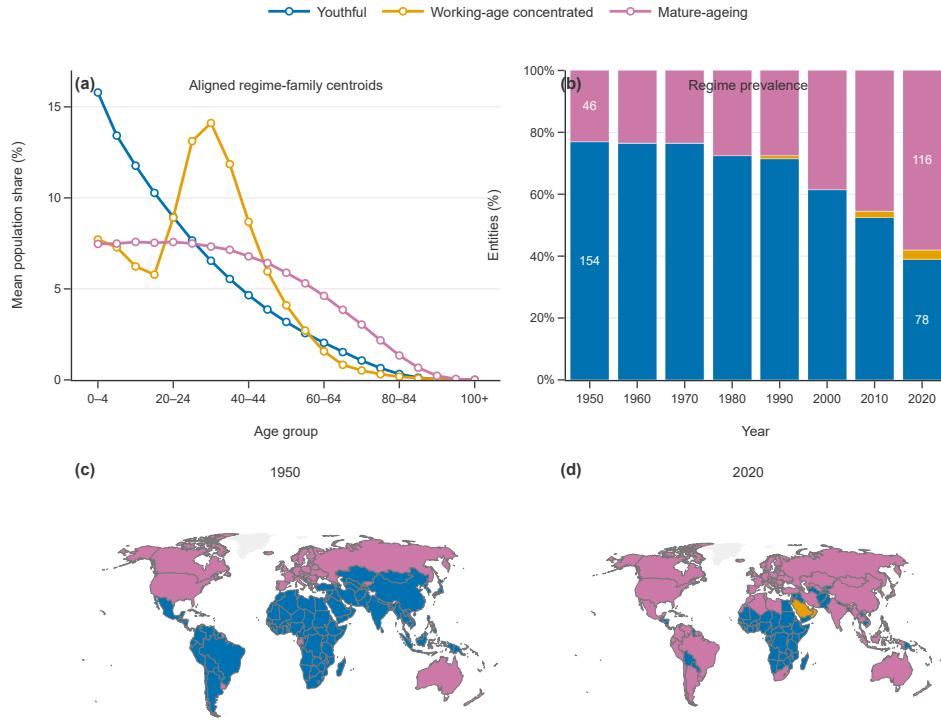


Fig. 2 Population-age configurations aligned into three broad families whose prevalence and geographic distribution changed markedly from 1950 to 2020. (a) Mean 21-age-group centroid profiles for the youthful, working-age-concentrated and mature-ageing regime families. (b) Proportion of the 200 entities assigned to each family at each decade. (c,d) Geographic distribution of regime-family membership in 1950 and 2020, respectively. Colours denote the same regime families throughout. Counts represent equally weighted entities rather than population-weighted global shares; Channel Islands was retained analytically but lacks a single corresponding polygon in the geographic boundary data used for mapping.

The dominant age-structure gradient preserved broad similarity but not the full grouping of trajectories

PC1 retained most pairwise trajectory ordering but did not reproduce the groups obtained from all four components. Its dynamic-time-warping (DTW) distance matrix correlated strongly with the four-PC distance matrix (Spearman $\rho = 0.950456$), whereas the resulting memberships agreed only partly (ARI = 0.544934; Fig. 3c; Supplementary Fig. S4). These statistics address different properties. Spearman's ρ measures whether pairs of countries retain their relative ordering from more to less similar, whereas the ARI measures whether the discrete group assignments coincide.

PC1 therefore preserved the broad ordering of trajectory similarity, while lower-variance dimensions changed enough local relationships to alter a meaningful share of memberships.

PC1 alone selected five groups with silhouette 0.389135, whereas four PCs selected three groups with silhouette 0.388135. The four-PC representation was therefore not uniformly superior by internal separation. Adding PC2 still produced five groups. When PC3 was included, the solution reorganised into three groups and closely matched the four-PC partition ($\text{ARI} = 0.966504$). PC3 consequently accounted for most of the reorganisation, while PC4 changed few final assignments. PC4 was retained because four components were required by the prespecified 95% variance threshold, not because it independently improved clustering. This distinction motivated the subsequent comparison of complete national trajectories with contemporaneous global configurations.

Global regimes and country trajectories were related but not interchangeable

Across 1,600 exact country-decade joins, with no unmatched entities, aligned regime-family membership was substantially associated with trajectory group (pooled Cramér’s $V = 0.549103$; Fig. 4a,b). The decade-specific association peaked in 1970 ($V = 0.959$) and was generally weaker thereafter as later configurations diversified. Pooled membership nevertheless remained distinctive: group 2 was almost entirely mature-ageing and group 3 almost entirely youthful, whereas group 1 contained a mixture dominated by youthful and mature-ageing observations. Because the pooled statistic includes repeated observations from the same countries, it is descriptive. These results describe association across two scales, not equivalence: a configuration observed at one date could occur within more than one long-term trajectory, and countries beginning with the same configuration could subsequently diverge.

The greatest sequence heterogeneity occurred in group 1 (Fig. 4c). Its modal sequence represented only 23/93 entities, 71/93 underwent one transition, and its 2020 assignments spanned all three families: 15 youthful, six working-age-concentrated and 72 mature-ageing. By contrast, 40/44 members of group 2 remained mature-ageing throughout and all 44 ended there, while 62/63 members of group 3 remained youthful throughout and all 63 ended youthful (Fig. 4c; Supplementary Fig. S5). Afghanistan, Qatar and Albania shared youthful assignments in 1950 but ended in youthful, working-age-concentrated and mature-ageing configurations, respectively; Australia and Albania also reached mature-ageing through different histories. These observations demonstrate multiple pathways without establishing deterministic path dependence or treating temporary switches as substantive reversals.

The zero-free 80+ and 85+ composition-aware analyses likewise selected three groups (Supplementary Table S1). Their pairwise distances remained strongly correlated with those of the primary analysis ($\rho = 0.899\text{--}0.938$), although trajectory membership was moderately dependent on representation ($\text{ARI} = 0.466\text{--}0.638$). Aligned regime families were more stable ($\text{ARI} = 0.857\text{--}0.873$), and pooled Cramér’s V remained 0.544–0.548. Constraining DTW to one or two decades retained three groups and closely reproduced the unrestricted solution ($\text{ARI} = 0.983$ and 1.000, respectively;

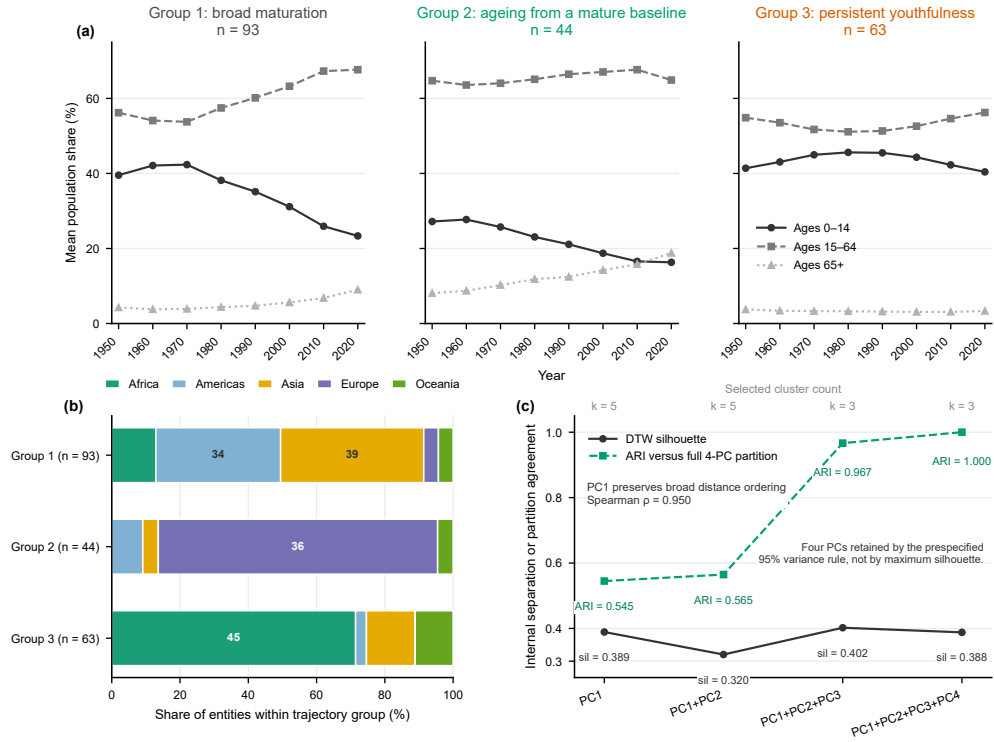


Fig. 3 Complete national histories formed three broad trajectory groups, and PC1 alone did not reproduce their full multivariate membership. (a) Mean shares aged 0-14, 15-64, and 65+ from 1950 to 2020 within the three trajectory groups containing 93, 44, and 63 entities. (b) Regional composition of the trajectory groups using United Nations M49 geographic classifications. (c) Sensitivity of trajectory clustering to progressive inclusion of PC1-PC4, showing selected cluster number, silhouette, and agreement with the full four-PC partition. PC1 preserved much of the broad trajectory geometry, whereas later components altered group membership; the four-PC representation was retained under the prespecified 95% variance criterion. ARI denotes adjusted Rand index; DTW, dynamic time warping; PC, principal component.

Supplementary Table S2). Consensus clustering across 15 graph specifications also selected three groups and closely matched the primary partition (ARI = 0.934). These robustness results support the broad cross-scale interpretation and motivate external validation of information beyond PC1.

Migrant stock provided the clearest validation of information beyond the dominant age-structure gradient

PC1 was associated with migrant stock, whereas the combined representation provided stronger held-out prediction (Fig. 5a). Across 796 observations from 199 entities in 1990, 2000, 2010 and 2020, the migrant-stock coefficient in the PC1 model had partial

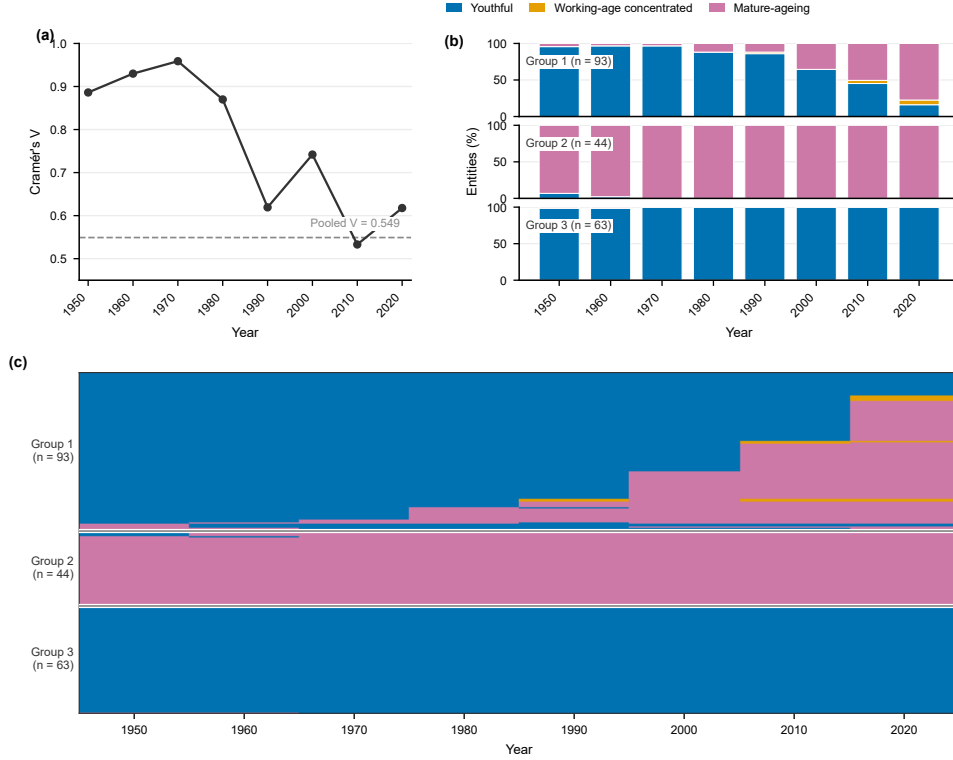


Fig. 4 Contemporaneous configurations and complete country trajectories were substantially related, but one did not uniquely determine the other. (a) Decade-specific Cramér's V between aligned regime-family membership and trajectory group; the horizontal reference denotes the pooled value ($V = 0.549103$). (b) Regime-family composition within each trajectory group across the eight decades. (c) Regime sequences for all 200 entities from 1950 to 2020, grouped by trajectory group. Colours denote global regime families throughout. The two classifications are substantially related but not interchangeable.

$R^2 = 0.019585$ (two-sided Wald $P = 0.0134$). Country-grouped cross-validation produced $R^2 = 0.170$ for demographic controls alone and 0.422 after adding all four PCs, reducing the sum of squared errors (SSE) by 30.32% across repeated folds. These associations neither identify PC1 as migration-specific nor support a causal interpretation of component coefficients.

The multivariate migrant-stock result generalised under nested validation: prediction error fell by 23.49%, reductions were positive in 88% of 25 outer folds, and all four PCs were selected in 100% of those folds (Fig. 5b). Its coefficient direction was also stable under country bootstrap resampling (Supplementary Fig. S6). The combined model also improved the prespecified high-stock classification beyond demographic controls, whereas the PC-only model did not (Supplementary Fig. S7). Structural back-projection nevertheless showed closer agreement between PC1 and the migrant-age profile than between that profile and the combined stock-associated direction, with

cosine similarities of 0.896757 across all 21 age groups, 0.903243 across ages 0–74 and 0.746025 for the combined direction (Fig. 5c). This comparison comprised 794 observations; Curaçao, Montenegro and South Sudan lacked profiles for 1990 and 2000. Stronger prediction therefore did not imply closer age-profile resemblance.

Migration-flow evidence was weaker, less generalisable and sensitive to target transformation (Fig. 5b,d). Across 1,393 observations spanning seven decadal intervals for 199 entities, PC2 had a partial R^2 of 0.016856 for raw flow, falling to 0.000672 after signed-log transformation. Under nested validation, adding PC2–PC4 reduced SSE by only 0.6272%, with positive reductions in 60% of outer folds. Detailed flow diagnostics are reported in the Supplementary Information. These associations provide no causal evidence, and no single PC captured a complete or stable migration-associated pattern. The clearest external validation concerned migrant-stock information distributed across the multivariate representation.

3 Discussion

Comparing complete population pyramids without predefined categories revealed contemporaneous configurations that redistributed over time and three broader national trajectory groups. The dominant shift from youthful towards mature structures accords with demographic-transition theory, which anticipates variation in onset, pace and sequence (Reher 2004; Casterline 2001; Frejka 2017), but the recovered organisation was neither one-dimensional nor a prescribed stage sequence. Unsupervised learning therefore served as an interpretable discovery layer, while demographic theory informed, rather than determined, the labels; the families remain descriptive rather than causal regimes.

Aligned-family membership showed substantial reorganisation without complete convergence (Fig. 2b–d): youthful configurations declined, mature-ageing configurations expanded and all three families persisted. This pattern accords with persistent youthfulness in much of sub-Saharan Africa and rapid ageing in Europe and Eastern Asia (Bongaarts 2017; Lutz et al. 2008; Balachandran et al. 2020). Because entities were equally weighted, counts represent configurations rather than population shares. Longitudinally, the three groups captured broad maturation, continued ageing from mature baselines and persistent youthfulness (Fig. 3a,b), while their internal heterogeneity reflected cross-national variation in transition timing and duration (Reher 2004; Casterline 2001; Wu and Pesando 2024). These empirical groups are therefore neither universal stages nor fixed pathways.

Contemporaneous configurations and complete trajectories were substantially related but not interchangeable (Fig. 4). The pooled association was appreciable, and the decade-specific association was strongest during 1960–1980, yet the largest trajectory group contained multiple regime sequences and 2020 destinations, and initially youthful entities ended in all three families. The working-age-concentrated family, especially among Gulf states, illustrates shape variation beyond the main youthful-to-mature gradient and is consistent with the demographic influence of predominantly working-age migrant labour (Shah 2012; Dörflinger and Loichinger 2024). More generally, similar endpoints need not imply identical histories (Wilson 2011; Dorius 2008;

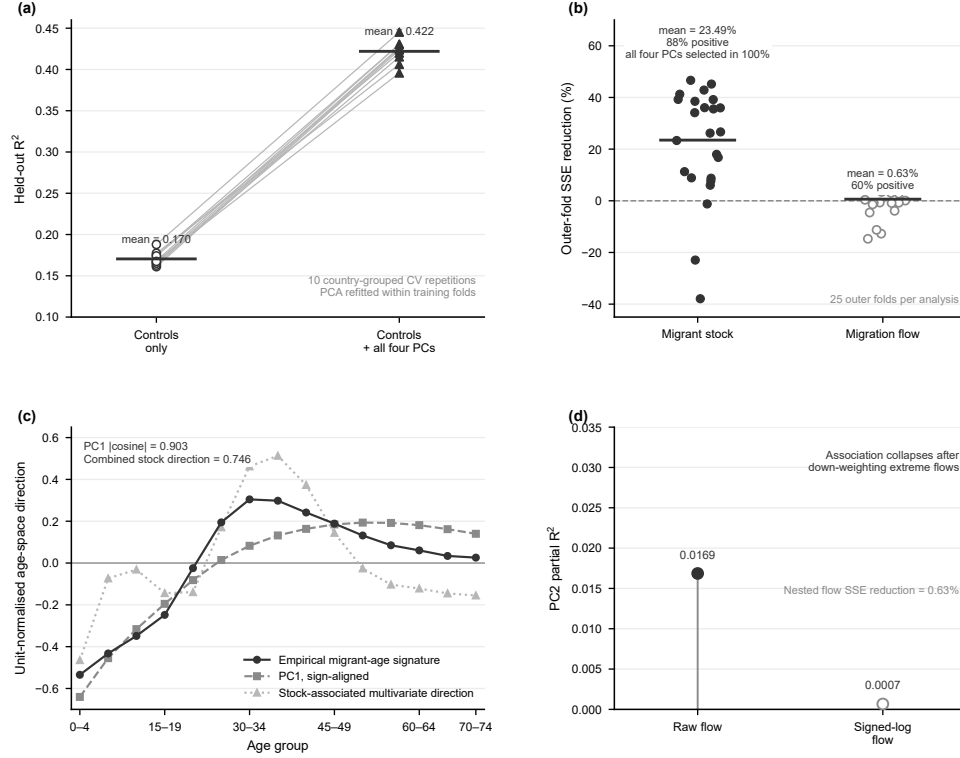


Fig. 5 Migrant-stock evidence supported the external relevance of the multivariate population-pyramid representation, whereas migration-flow evidence was weaker and sensitive to transformation. (a) Held-out-country migrant-stock prediction with PCA refitted within training-country folds, comparing demographic controls alone with controls plus all four PCs. (b) Nested-validation reduction in prediction error for migrant stock and migration flow. (c) Empirical migrant-age signature compared with PC1 and the back-projected multivariate stock-associated direction across ages 0-74. (d) Sensitivity of the PC2 migration-flow association to signed-log transformation. Migration was used only for external validation and was not an input to PCA or clustering; the results do not imply causal migration effects. PCA denotes principal component analysis; PC, principal component; SSE, sum of squared errors; CV, cross-validation.

Wu and Pesando 2024). These descriptive sequences demonstrate multiple pathways without establishing causal path dependence.

PC1 preserved the broad ordering of trajectory distances but not the complete four-PC partition (Figs. 1 and 3c): Spearman's ρ was 0.950, whereas ARI was 0.545. Lower-variance dimensions can preserve global distance ordering while altering local neighbourhoods near cluster boundaries (Jolliffe 1982; Ding and He 2004), and the slightly higher PC1-only silhouette confirms that additional dimensions did not uniformly improve separation. Adding PC3 changed the selected solution to three groups; the first-three-PC and four-PC partitions were nearly identical (ARI = 0.967), while PC4 was retained solely by the prespecified 95% variance criterion. Because loading

signs are arbitrary (Jolliffe and Cadima 2016), interpretation concerns relative age contrasts. Explained variance, separation, membership change and external relevance are distinct criteria (Handl et al. 2005), requiring the latter to be assessed independently.

Migration validation separated structural resemblance from predictive value (Fig. 5a–c). PC1 most closely resembled the empirical migrant-age structure, consistent with migration’s concentration at young-adult and working ages (Shen et al. 2024; United Nations Department of Economic and Social Affairs, Population Division 2020b), whereas all four PCs improved held-out migrant-stock prediction beyond controls. This gain does not identify a migration mechanism, and PCs alone did not outperform controls in the classification analysis. Flow evidence was weaker and transformation-sensitive: signed-log transformation greatly reduced the PC2 association, while nested validation reduced error by only 0.63% and improved 60% of outer folds. This contrast accords with migrant stock accumulating prior migration and survival, whereas net flow is volatile and may combine offsetting movements (United Nations Department of Economic and Social Affairs, Statistics Division 2022; Raymer et al. 2025). Fertility, mortality, shocks, cohort effects and data structure may also contribute; no causal effect is inferred.

The primary trajectory partition was identical across 20 random seeds but less stable across graph specifications (Supplementary Fig. S3): three groups were selected in 10/15 settings, with four, five or eight elsewhere. Fine within-decade counts also varied, and some aligned nodes had low margins (Supplementary Figs. S1 and S2). Initialisation stability therefore does not imply specification invariance (Ben-Hur et al. 2002; Maier et al. 2008). Parameter sensitivity qualifies the exact boundaries without erasing the broad demographic distinctions.

The analysis resolves decadal organisation rather than short-term dynamics or causal mechanisms. Decadal observations omit brief shocks, and unrestricted DTW may understate timing differences (Sakoe and Chiba 1978). Conventional PCA does not explicitly respect compositional geometry (Aitchison 1982, 1983), although zero-free log-ratio analyses recovered the three broad trajectory types and aligned families with strong distance and regime agreement; exact memberships and the small working-age-concentrated regime remained representation-sensitive. Alignment of independently estimated partitions introduces further uncertainty, and harmonised population estimates retain measurement and modelling uncertainty (United Nations Department of Economic and Social Affairs, Population Division 2024b). Migration analyses used narrower, non-identical samples. These constraints limit temporal precision, metric interpretation and generalisation, but the framework still coherently relates global configurations to national histories.

This study places interpretable unsupervised learning between empirical description and demographic theory. It identifies organisation in complete age profiles without predefined outcomes and makes that organisation inspectable through age-specific loadings, centroids, trajectory sequences and external validation (Figs. 1–5), but it neither replaces model-based analysis nor identifies mechanisms. Subject to choices of temporal resolution, weighting, representation and similarity, the framework may

extend to other structured longitudinal demographic data. Its contribution is to connect empirically recovered organisation with established theory while keeping the analytical path and inferential limits explicit.

4 Methods

Data sources and analytical population

Population estimates by five-year age group were obtained directly from United Nations World Population Prospects (WPP) 2024, File POP/02-1, using the both-sexes-combined estimates for 1950–2023 ([United Nations Department of Economic and Social Affairs, Population Division 2024c](#)). World Population Prospects 2024, File GEN/01/Rev.1, supplied total population and net migration (thousands), migration rates (per 1,000), fertility and life expectancy for 1950–2023 ([United Nations Department of Economic and Social Affairs, Population Division 2024c](#)). External validation used International Migrant Stock 2024, destination-and-origin Table 1 (persons, 1990–2024) ([United Nations Department of Economic and Social Affairs, Population Division 2024a](#)), and International Migrant Stock 2020: Age, sex and destination, Table 1 (persons, 1990–2020) ([United Nations Department of Economic and Social Affairs, Population Division 2020a](#)); United Nations M49 supplied geographic classifications ([United Nations Statistics Division 2026](#)). Named sheets and required columns were verified before harmonisation.

To compare population-age-profile shape rather than size, both-sex population estimates were normalised within each country-year into 21 shares ordered from 0–4 to 100+. The prespecified analytical entity list retained 200 countries or territories; 199 were matched to WPP 2024 country/area records, while Channel Islands was reconstructed by summing the corresponding Guernsey and Jersey age-group populations before normalisation. The year manifest retained eight decadal years from 1950 to 2020, yielding 1,600 profiles with identical temporal support. Analyses weighted entities equally. Profiles with duplicates, incomplete age coverage, negative or non-finite counts, or normalisation error exceeding 10^{-10} were rejected; only validated profiles entered the population-structure analyses.

Common representation of population age structures

Because adjacent age shares covary, principal component analysis represented complete profiles in a common lower-dimensional coordinate system ([Jolliffe and Cadima 2016](#)). In practical terms, PCA summarised the 21 correlated age shares using a smaller set of axes, with age-specific loadings showing how the age groups combine to define each axis. PCA was fitted once to the $1,600 \times 21$ matrix in chronological age order. Shares were centred but not variance-standardised, yielding a covariance decomposition with `svd.solver="full"`. Before clustering, we retained the minimum number of components explaining at least 95% of total variance. Conventional PCA of the centred age shares was retained as the primary representation because it provided directly interpretable age-specific contrasts. Compositional sensitivity was assessed using zero-free centred log-ratio representations after amalgamating the sparse upper-age tail

to 85+ or 80+; both 95%-variance and complete Aitchison-rank representations were evaluated. Full specifications and zero-free validation are reported in Supplementary Methods. Retained conventional PCA scores formed the primary cross-sectional and longitudinal representation.

Global demographic configurations

Global demographic evolution was examined as the changing configuration of entities around one another, rather than as individual national trajectories. Each decade was treated as an independent snapshot, with entities grouped by the structural similarity of their complete population pyramids as represented by the common four-PC scores. Comparing these partitions allowed examination of how contemporaneous configurations appeared, separated, expanded, contracted or redistributed. Because yearly clusters were estimated independently, continuity was not assumed; historical comparability was introduced subsequently by aligning complete 21-age-group centroids. This scale described global cross-sectional reorganisation, not the pathways followed by individual entities.

Spectral clustering was used because structurally related entities could form locally connected, non-spherical configurations in PC space (von Luxburg 2007). Within each decade, Euclidean distances were converted into a union-symmetrised 12-nearest-neighbour graph: an edge was retained when either entity selected the other. Retained edges received Gaussian affinities

$$w_{ij} = \exp\left(-\frac{d_{ij}^2}{2\sigma^2}\right),$$

where σ was the median positive retained-edge distance. After confirming graph connectedness, we constructed

$$L_{\text{sym}} = I - D^{-1/2} W D^{-1/2}.$$

For its ordered eigenvalues, the cluster count was selected by

$$\hat{k} = \arg \max_{k \in \{2, \dots, 10\}} (\lambda_{k+1} - \lambda_k).$$

Clustering used the precomputed affinity (`assign_labels="kmeans"`, `random_state=42`, `n_init=50`); silhouette was calculated from the original Euclidean distances (Rousseeuw 1987).

Each decade was clustered independently in four-PC space, allowing configurations to change rather than imposing fixed historical classes. Separate Euclidean-distance graphs and eigengap partitions were fitted. Robustness to graph construction was assessed across the 15 combinations of nearest-neighbour count and Gaussian-bandwidth multiplier (Maier et al. 2008) by recording the selected cluster count, Euclidean silhouette and adjusted Rand index relative to the corresponding primary yearly partition (Hubert and Arabie 1985) (Supplementary Fig. S1). Cluster

centroids were renormalised, equal-entity means of all 21 age shares. The 53 centroids were compared by base-2 Jensen–Shannon distance (Endres and Schindelin 2003) and grouped by average-linkage agglomerative clustering; the family count maximised Jensen–Shannon silhouette over $k = 2, \dots, 10$. Membership outputs included own-versus-nearest-family distance margins as alignment-uncertainty measures. Yearly cluster counts and aligned families were therefore estimated, not imposed. These configurations remained distinct from the complete entity trajectories analysed next.

Country-specific demographic trajectories

Whereas the global analysis represented the world through independently estimated cross-sectional configurations, the country-specific analysis treated each entity’s ordered demographic history as the object of comparison. A trajectory was the path traced by its complete population-pyramid structure through the common four-PC space at eight observations from 1950 to 2020. This distinction recognised that similarity at one date, or at a common endpoint, need not imply the same historical sequence. Trajectory comparison therefore preserved temporal order and multivariate structural change instead of reducing demographic history to isolated states. The two scales were complementary: one described successive global arrangements, while the other grouped long-term entity pathways. These trajectories represented observed decadal histories and did not reconstruct continuous change between observations or identify causal mechanisms.

Numerically ordered PC1-PC4 scores formed one complete 8×4 trajectory for each of 200 entities. Conceptually, dynamic time warping aligns comparable portions of two ordered histories before measuring their separation, allowing similar age-structural changes to occur in different observed decades. The trajectories were therefore compared by unrestricted multivariate dynamic time warping with Euclidean local geometry (Sakoe and Chiba 1978). Squared local distances were accumulated along the optimal path and the square root of the total returned; neither path-length normalisation nor a warping window was used. The 200×200 distance matrix was validated as symmetric, finite and zero-diagonal before applying the shared 12-nearest-neighbour spectral framework and eigengap rule. DTW silhouette and adjusted Rand agreement across 20 seeds assessed stochastic stability. Specification robustness was evaluated across 15 combinations of $k_{NN} \in \{8, 10, 12, 15, 20\}$ and Gaussian-bandwidth multipliers $\{0.5, 1, 2\}$, with pairwise co-assignment used to assess membership stability when the selected k changed. Timing sensitivity compared unrestricted DTW with one- and two-decade Sakoe-Chiba constraints and with unwarped time-aligned distance. The unrestricted four-PC specification remained primary; complete sensitivity procedures are given in Supplementary Methods.

Representation sensitivity and cross-scale association

The structural contribution of later components was assessed by repeating trajectory clustering with PC1 alone, the first two PCs, the first three PCs and all four PCs, using the same entity-year grid, DTW definition and graph conventions. Spearman correlation between upper-triangular distance matrices measured geometric agreement,

adjusted Rand index compared assignments, and silhouette evaluated each DTW partition. This sensitivity analysis did not determine dimensionality: four PCs had been retained independently by the 95% variance rule. Composition-aware sensitivity repeated the trajectory and global analyses under the zero-free 85+ and 80+ log-ratio representations described above; distance geometry, partitions, aligned regime families and cross-scale association were compared with the primary analysis. We then related the primary trajectories to contemporaneous regime families.

Aligned regime-family memberships were joined to trajectory groups by entity, preserving year and reconnecting contemporaneous positions with longitudinal pathways. Validation required 1,600 unique entity-year matches and rejected missing or duplicate memberships. Classical, uncorrected Cramér’s V quantified pooled and decade-specific association. Eight ordered memberships formed each entity’s regime sequence; dominant sequences, adjacent transitions, contiguous regime durations, starting regimes and 2020 destinations were summarised by trajectory group. Convergent and divergent histories were identified descriptively from sequences and endpoints. These measures described association, not causal path dependence. Migration was then examined as external evidence independent of both clustering solutions.

Migration-based external validation

Migration validation combined three forms of evidence: migrant stock represented accumulated presence, net migration represented decadal movement, and migrant-age profiles enabled direct comparison in age space. The stock panel comprised 199 entities in 1990, 2000, 2010 and 2020 ($n = 796$), with stock expressed as a population percentage. The flow panel paired cumulative net migration per 1,000 with interval PC changes across seven decadal intervals ($n = 1,393$; 199 entities). Both used fertility and life expectancy controls. Channel Islands was excluded from the stock/control and flow panels because WPP represents Guernsey and Jersey separately, but remained combined in the migrant-age data. That comparison contained 794 of 800 profiles; Curaçao, Montenegro and South Sudan were unavailable in 1990 and 2000.

Association models tested covariation rather than causal migration effects. For migrant stock, each PC score was modelled as the outcome and migrant-stock percentage as the predictor, adjusting for fertility, life expectancy, country fixed effects and year fixed effects. Flow models related each interval PC change to cumulative net migration per 1,000, with interval effects and interval-mean fertility and life expectancy. Models used ordinary least squares with country-clustered standard errors. Coefficient P values were obtained from two-sided Wald tests based on country-clustered standard errors and were evaluated at $\alpha = 0.05$. Partial R^2 was $(SSE_{\text{reduced}} - SSE_{\text{full}})/SSE_{\text{reduced}}$, and uncertainty was additionally examined using 500 country-level bootstrap resamples. Sensitivity analyses used $\log(1+x)$ -transformed stock and signed-log or annualised flow formulations. Benjamini–Hochberg correction (Benjamini and Hochberg 1995) covered the four primary flow tests; stock p -values remained nominal. Predictive validation subsequently reversed the analytical direction and treated migration measures as outcomes.

Out-of-country predictive value was assessed by comparing controls-only models with models adding each of the 15 non-empty PC subsets. For migrant-stock prediction, country-grouped five-fold cross-validation was repeated ten times, with PCA refitted on training-country population profiles within each split before projection of held-out countries. Nested validation used five repetitions of five outer folds and four inner folds, with PCA likewise refitted within the corresponding training folds before PC-subset selection. High migrant stock was defined within year using the training countries' top decile. Control-only, PC-only and combined logistic-regression classifiers were evaluated over five country-grouped folds repeated ten times using receiver operating characteristic and precision-recall areas under the curve (ROC-AUC and PR-AUC); numerical predictors were standardised using training-fold information, and the high-migration thresholds were derived exclusively from the training countries.

Age-structural validation formed $g = q - p$, where q and p were the migrant and host-population age distributions, using the common 0–74 age groups. Standardised stock-associated PC directions were returned to raw scale, projected through the PCA loadings and compared with g by absolute cosine similarity because PCA signs are arbitrary. Full 21-age-group redistribution, ridge models, bootstrap directions and high-stock threshold variation were sensitivity analyses. Migration supplied external validation, not causal evidence.

Computational implementation and reproducibility

Analyses ran in Python 3.12.13 with NumPy, pandas, SciPy, scikit-learn and statsmodels under a locked package environment. The single entry point, `run.all.py`, reproduced the workflow through `--main`, `--global` and `--migration`; stochastic procedures used deterministic seeds. Version-controlled entity and year manifests fixed the analytical grid, while the run manifest recorded input and output hashes. All 21 validation checks passed on a clean execution. Software versions, execution settings and a summary of the automated validation checks are provided in the Supplementary Methods.

Use of generative artificial intelligence

During the preparation and revision of this manuscript, the authors used OpenAI ChatGPT with the Codex agent (GPT-5; OpenAI; accessed September 2026). Prompts asked the tool to propose alternative wording and organisation; summarise the Abstract and selected passages; improve plain-language explanations and figure legends; standardise terminology; and reduce repetition. The tool was not used to generate or modify study data, execute the reported analyses, produce figures, select references or determine the scientific conclusions. All outputs were treated as provisional. The authors checked suggested text against the underlying analyses and source documents, independently verified all scientific claims, numerical values, citations and cross-references, revised or rejected suggestions as appropriate, and take full responsibility for the final content.

Code availability

The complete custom code, configuration files, manifests and locked computational environment underlying the reported analyses are available to editors and reviewers for confidential assessment upon request. The version associated with the accepted manuscript will be released publicly through the [project GitHub repository](#) and permanently archived in Zenodo with a version-specific DOI no later than publication.

Data Availability

The population-age profiles and demographic indicators analysed in this study were obtained from the United Nations World Population Prospects 2024, Files POP/02-1 and GEN/01/Rev.1 (<https://population.un.org/wpp/>). Migrant-stock data were obtained from International Migrant Stock 2024, destination-and-origin Table 1, and migrant-age profiles from International Migrant Stock 2020: Age, sex and destination, Table 1 (<https://www.un.org/development/desa/pd/content/international-migrant-stock>). United Nations M49 geographic classifications are available at <https://unstats.un.org/unsd/methodology/m49/>. The processed analytical datasets and derived outputs supporting the findings are available to editors and reviewers from the corresponding author upon reasonable request. Upon acceptance, the version supporting the accepted manuscript will be released publicly at <https://github.com/athulya24453/Global-demographic-configurations-and-national-trajectories-from-populationpyramids> and permanently archived in Zenodo with a version-specific DOI no later than publication.

References

- Aitchison J (1982) The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)* 44(2):139–160. <https://doi.org/https://doi.org/10.1111/j.2517-6161.1982.tb01195.x>
- Aitchison J (1983) Principal component analysis of compositional data. *Biometrika* 70(1):57–65. <https://doi.org/https://doi.org/10.1093/biomet/70.1.57>
- Balachandran A, de Beer J, James KS, et al (2020) Comparison of population aging in europe and asia using a time-consistent and comparative aging measure. *Journal of Aging and Health* 32(5–6):340–351. <https://doi.org/https://doi.org/10.1177/0898264318824180>
- Ben-Hur A, Elisseeff A, Guyon I (2002) A stability based method for discovering structure in clustered data. In: *Pacific Symposium on Biocomputing 2002*. World Scientific, Singapore, pp 6–17, https://doi.org/10.1142/9789812799623_0002
- Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)* 57(1):289–300. <https://doi.org/https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>

- Billari FC (2001) Sequence analysis in demographic research. *Canadian Studies in Population* 28(2):439–458. <https://doi.org/https://doi.org/10.25336/P6G30C>
- Bloom DE, Chatterji S, Kowal P, et al (2015) Macroeconomic implications of population ageing and selected policy responses. *The Lancet* 385(9968):649–657. [https://doi.org/https://doi.org/10.1016/S0140-6736\(14\)61464-1](https://doi.org/https://doi.org/10.1016/S0140-6736(14)61464-1)
- Bongaarts J (2006) The causes of stalling fertility transitions. *Studies in Family Planning* 37(1):1–16. <https://doi.org/https://doi.org/10.1111/j.1728-4465.2006.00079.x>
- Bongaarts J (2009) Human population growth and the demographic transition. *Philosophical Transactions of the Royal Society B: Biological Sciences* 364(1532):2985–2990. <https://doi.org/https://doi.org/10.1098/rstb.2009.0137>
- Bongaarts J (2017) Africa’s unique fertility transition. *Population and Development Review* 43(S1):39–58. <https://doi.org/https://doi.org/10.1111/j.1728-4457.2016.00164.x>
- Casterline JB (2001) The pace of fertility transition: National patterns in the second half of the twentieth century. *Population and Development Review* 27:17–52. Supplement: Global Fertility Transition
- Cecchini M, Cividino S, Turco R, et al (2019) Population age structure, complex socio-demographic systems and resilience potential: A spatio-temporal, evenness-based approach. *Sustainability* 11(7):2050. <https://doi.org/https://doi.org/10.3390/su11072050>
- Ding C, He X (2004) Cluster structure of k-means clustering via principal component analysis. In: *Advances in Knowledge Discovery and Data Mining, Lecture Notes in Computer Science*, vol 3056. Springer, Berlin, Heidelberg, pp 414–418, https://doi.org/10.1007/978-3-540-24775-3_50
- Dörflinger M, Loichinger E (2024) Fertility decline, changes in age structure, and the potential for demographic dividends: A global analysis. *Demographic Research* 50:221–290. <https://doi.org/https://doi.org/10.4054/DemRes.2024.50.9>
- Dorius SF (2008) Global demographic convergence? a reconsideration of changing intercountry inequality in fertility. *Population and Development Review* 34(3):519–537. <https://doi.org/https://doi.org/10.1111/j.1728-4457.2008.00235.x>
- Endres DM, Schindelin JE (2003) A new metric for probability distributions. *IEEE Transactions on Information Theory* 49(7):1858–1860. <https://doi.org/https://doi.org/10.1109/TIT.2003.813506>
- Fonseka T, Wijenayake B, Ratnayake A, et al (2025) Devising PoPStat: A metric bridging population pyramids with global disease mortality. *arXiv preprint*

- arXiv:250111514 <https://doi.org/https://doi.org/10.48550/arXiv.2501.11514>, URL <https://arxiv.org/abs/2501.11514>
- Frejka T (2017) The fertility transition revisited: A cohort perspective. *Comparative Population Studies* 42:89–116. <https://doi.org/https://doi.org/10.12765/CPoS-2017-09en>
- Handl J, Knowles J, Kell DB (2005) Computational cluster validation in post-genomic data analysis. *Bioinformatics* 21(15):3201–3212. <https://doi.org/https://doi.org/10.1093/bioinformatics/bti517>
- Hubert L, Arabie P (1985) Comparing partitions. *Journal of Classification* 2(1):193–218. <https://doi.org/https://doi.org/10.1007/BF01908075>
- Jolliffe IT (1982) A note on the use of principal components in regression. *Applied Statistics* 31(3):300–303. <https://doi.org/https://doi.org/10.2307/2348005>
- Jolliffe IT, Cadima J (2016) Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374(2065):20150202. <https://doi.org/https://doi.org/10.1098/rsta.2015.0202>
- Kirk D (1996) Demographic transition theory. *Population Studies* 50(3):361–387. <https://doi.org/https://doi.org/10.1080/0032472031000149536>
- Lutz W, Sanderson W, Scherbov S (2008) The coming acceleration of global population ageing. *Nature* 451(7179):716–719. <https://doi.org/https://doi.org/10.1038/nature06516>
- von Luxburg U (2007) A tutorial on spectral clustering. *Statistics and Computing* 17(4):395–416. <https://doi.org/https://doi.org/10.1007/s11222-007-9033-z>
- Ma J, Chen Q, Chen X, et al (2025) An inevitably ageing world: analysis on the evolutionary pattern of age structure in 200 countries. *Royal Society Open Science* 12(4):241988. <https://doi.org/https://doi.org/10.1098/rsos.241988>
- Maier M, von Luxburg U, Hein M (2008) Influence of graph construction on graph-based clustering measures. In: *Advances in Neural Information Processing Systems* 21, pp 1025–1032
- Raymer J, Guan Q, Shen T, et al (2025) Modelling the age and sex profiles of net international migration. *Demographic Research* 53:569–610. <https://doi.org/https://doi.org/10.4054/DemRes.2025.53.19>
- Reher DS (2004) The demographic transition revisited as a global process. *Population, Space and Place* 10(1):19–41. <https://doi.org/https://doi.org/10.1002/psp.313>

- Rousseeuw PJ (1987) Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* 20:53–65. [https://doi.org/https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/https://doi.org/10.1016/0377-0427(87)90125-7)
- Sakoe H, Chiba S (1978) Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 26(1):43–49. <https://doi.org/https://doi.org/10.1109/TASSP.1978.1163055>
- Sanderson WC, Scherbov S (2007) A new perspective on population aging. *Demographic Research* 16(2):27–58. <https://doi.org/https://doi.org/10.4054/DemRes.2007.16.2>
- Shah NM (2012) Socio-demographic transitions among nationals of gcc countries: implications for migration and labour force trends. *Migration and Development* 1(1):138–148. <https://doi.org/https://doi.org/10.1080/21632324.2012.709808>
- Shen T, Raymer J, Guan Q, et al (2024) The estimation of age and sex profiles for international migration amongst countries in the asia-pacific region. *Population, Space and Place* 30(1):e2716. <https://doi.org/https://doi.org/10.1002/psp.2716>
- United Nations Department of Economic and Social Affairs, Population Division (2020a) International migrant stock 2020: Age, sex and destination, table 1: International migrant stock at mid-year by age and sex and by region, country or area of destination, 1990–2020. <https://www.un.org/development/desa/pd/content/international-migrant-stock>, pOP/DB/MIG/Stock/Rev.2020; accessed 4 September 2026
- United Nations Department of Economic and Social Affairs, Population Division (2020b) International migration 2020 highlights. United Nations, sT/ESA/SER.A/452, New York
- United Nations Department of Economic and Social Affairs, Population Division (2024a) International migrant stock 2024: Destination and origin, table 1: International migrant stock at mid-year by sex and by region, country or area of destination and origin, 1990–2024. <https://www.un.org/development/desa/pd/content/international-migrant-stock>, pOP/DB/MIG/Stock/Rev.2024; accessed 4 September 2026
- United Nations Department of Economic and Social Affairs, Population Division (2024b) World population prospects 2024: Methodology of the united nations population estimates and projections. Tech. Rep. UN DESA/POP/2024/DC/NO. 10, United Nations, New York
- United Nations Department of Economic and Social Affairs, Population Division (2024c) World population prospects 2024, online edition. URL <https://population.un.org/wpp/>, files POP/02-1 and GEN/01/Rev.1; accessed 4 September 2026

United Nations Department of Economic and Social Affairs, Statistics Division (2022) Handbook on Measuring International Migration through Population Censuses. No. 115 in Studies in Methods, Series F, United Nations, New York

United Nations Statistics Division (2026) Standard country or area codes for statistical use (M49). <https://unstats.un.org/unsd/methodology/m49/>, online statistical classification; accessed 4 September 2026

Wilson C (2011) Understanding global demographic convergence since 1950. *Population and Development Review* 37(2):375–388. <https://doi.org/https://doi.org/10.1111/j.1728-4457.2011.00415.x>

Wilson T (2016) Visualising the demographic factors which shape population age structure. *Demographic Research* 35(29):867–890. <https://doi.org/https://doi.org/10.4054/DemRes.2016.35.29>

Wu H, Pesando LM (2024) Educational boundaries explain strength and variation in global fertility convergence. *Scientific Reports* 14:27323. <https://doi.org/https://doi.org/10.1038/s41598-024-78735-2>

Supplementary information Supplementary Information accompanies this article and provides extended methodological details, additional figures and tables, and sensitivity analyses concerning graph specifications, principal component retention, compositional representation and temporal alignment. It also presents diagnostics of regime alignment and trajectory stability, further results on regime sequences and migration-based external validation, and details of computational settings and reproducibility checks.

Author Contributions

All authors contributed to the conceptualisation of the study. A.R., L.E., T.W., R.G. and P.E. developed the methodology. A.R., L.E. and T.W. curated the data, developed the software, conducted the formal analyses and investigation, and wrote the original draft. A.R., L.E., T.W. and R.G. prepared the visualisations. A.R., L.E. and R.G. interpreted the findings. A.R., R.G. and P.E. validated the methods and results. R.G., P.E., V.H. and S.D. supervised the research and reviewed and edited the manuscript. R.G. administered the project. R.G., P.E. and V.H. provided resources. All authors reviewed and approved the final manuscript.

5 Declarations

5.1 Competing Interests

The authors declare no competing interests.

5.2 Ethics declaration

This study used only publicly available, aggregate secondary data and involved no human participants, individual-level personal data or animals. Institutional ethics approval and informed consent were therefore not required.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SIUnsupervisedlearningrevealsglobaldemographicconfigurationsandnationaltrajectoriesfrompopulationpyramids.pdf](#)