Shen, Tianyu; Wang, Xueqing; Baselini, Ugofilippo; Alexander, Monica; Chen, Irena

Working Paper

Leveraging population change to improve estimation of mortality in data-scarce context

Vienna Institute of Demography Working Papers, No. 02/2025

Provided in Cooperation with:
Vienna Institute of Demography (VID), Austrian Academy of Sciences

Suggested Citation: Shen, Tianyu; Wang, Xueqing; Baselini, Ugofilippo; Alexander, Monica; Chen, Irena (2025) : Leveraging population change to improve estimation of mortality in data-scarce context, Vienna Institute of Demography Working Papers, No. 02/2025, Austrian Academy of Sciences (ÖAW), Vienna Institute of Demography (VID), Vienna, https://doi.org/10.1553/0x00408a51

This Version is available at:
https://hdl.handle.net/10419/333470

ÖAW AUSTRIAN ACADEMY OF SCIENCES

VIENNA INSTITUTE OF DEMOGRAPHY

WORKING PAPERS

02/2025

LEVERAGING POPULATION CHANGE TO IMPROVE ESTIMATION OF MORTALITY IN DATA-SCARCE CONTEXT

TIANYU SHEN, XUEQING WANG, UGOFILIPPO BASELINI, MONICA ALEXANDER, IRENA CHEN

Vienna Institute of Demography
Austrian Academy of Sciences
Dominikanerbastei 16 | 1010 Vienna, Austria
vid@oeaw.ac.at | www.oeaw.ac.at/vid

VIENNA INSTITUTE OF DEMOGRAPHY

https://doi.org/10.1553/0x00408a51

Abstract

In countries with incomplete vital registration, contemporary mortality estimation models such as the log-quadratic model the and Singular Value Decomposition (SVD)-Comp model often produce sizable errors, particularly at prime adult ages (20-45) and older ages (70+), where mortality varies substantially due to country-specific epidemiological profiles. This paper introduces a novel model-agnostic sorting approach that incorporates population change information, derived from decennial censuses and typically available even in data-scarce settings, to improve mortality estimation accuracy. Our method uses cosine similarity of z-score standardized population change patterns to identify the most demographically similar country-years (top 10%) in the training dataset. Using data from the Human Mortality Database and from HIV-affected countries from UN World Population Prospects, we test this approach with both log-quadratic and SVD-Comp models. Results demonstrate consistent improvements in prediction accuracy, with median RMSE for probability of dying (q_x) decreasing by 17% (log-quadratic) and 13% (SVD-Comp), respectively. Gains are most apparent at young adult and older ages, particularly in countries with distinct mortality patterns, such as those affected by HIV. This approach offers a practical, easily implementable solution that can be integrated into existing mortality estimation workflows without additional data requirements or model restructuring.

Keywords

Model life tables, Mortality, Population change, Data-scarce settings, Demographic similarity

Authors

Tianyu Shen*, Vienna Institute of Demography/Austrian Academy of Sciences
E-Mail: *Tianyu.Shen@oeaw.ac.at*

Xueqing Wang*, Office of Population Research, Princeton University, USA
E-Mail: *xueqingw@princeton.edu*

Ugofilippo Basellini, Max Planck Institute for Demographic Research, Rostock, Germany
E-Mail: *basellini@demogr.mpg.de*

Monica Alexander, Department of Sociology, University of Toronto, Canada
E-Mail: *monica.alexander@utoronto.ca*

Irena Chen, Max Planck Institute for Demographic Research, Rostock, Germany
E-Mail: *chen@demogr.mpg.de*

* Shen and Wang contributed equally.

Acknowledgements

We would like to thank Emilio Zagheni, Marília R. Nepomuceno, Julio Romero-Prieto, Isabella Buber-Ennser, and participants of the summer incubator program 2024 at the Max Planck Institute for Demographic Research for their helpful comments and suggestions. Thanks to Charles Cui and Nathaniel Darling for their input to the early stages of this work. An earlier version of this work was presented at the Population Association of America Annual Meeting at Washington D.C. in 2025.

LEVERAGING POPULATION CHANGE TO IMPROVE ESTIMATION OF MORTALITY IN DATA-SCARCE CONTEXT

TIANYU SHEN, XUEQING WANG, UGOFILIPPO BASELINI, MONICA ALEXANDER, IRENA CHEN

1 INTRODUCTION

Mortality estimation is central to demographic analysis, public health planning, and policy development. Complete age-specific mortality schedules underpin key indicators like life expectancy and inform decisions around resource allocation and health system design. While countries with complete vital registration systems can construct these schedules directly from administrative records, such systems cover only about 30% of the world's population (Mikkelsen et al., 2015). The remaining 70%, concentrated in regions like sub-Saharan Africa, lack reliable adult mortality data at all ages, relying instead on sparse surveys or incomplete records (Li, 2015; Wilmoth et al., 2012; Clark, 2019). This widespread data scarcity necessitates the use of model-based approaches to estimate mortality patterns where direct data are unavailable.

To fill this gap, demographers developed model life tables, which leverage observed relationships between child and adult mortality levels to generate complete mortality schedules when only mortality data for certain age groups are available. Early approaches, such as those by Coale and Demeny (1966, 1983) and the United Nations (1982), organized mortality patterns into discrete regional "families" based on geographic similarities. Later innovations introduced more flexible statistical methods, including Brass's (1975) relational logit model and Murray et al.'s (2003) modified logit system, both of which use parameters to relate observed child and adult mortality to a standard life table. While these models moved beyond rigid regional classifications, they still assume that mortality patterns across populations differ primarily in level and shape, and can be derived from linear transformations of a standard schedule.

The log-quadratic model (Wilmoth et al., 2012) represents a significant advance by using a quadratic function on the log scale to relate child mortality to mortality at other ages. While the model offers flexibility and ease of use, it treats the relationship

between child mortality and each age group independently, thus lacking structural coherence across ages. More importantly, it applies this relationship uniformly across all target populations, regardless of whether the empirical link between child and adult mortality holds in each setting. Clark's (2019) SVD-Comp model adopts Singular Value Decomposition (SVD) to represent mortality patterns through weighted components, introducing stronger cross-age coherence. Clark demonstrates better performance compared to the log-quadratic model with lower total absolute errors, reduced systematic bias at older ages, and more plausible predictions for distinctive populations like HIV-affected South Africa. However, this approach also faces a key limitation: while later SVD components can theoretically capture distinctive mortality patterns like adult HIV/AIDS peaks, the first component dominates all predictions and reflects typical Western mortality patterns. Since the model learns how to weight these components from global data, where such distinctive features are relatively rare, it under-weights precisely the components that could identify these atypical patterns. This results in systematic underestimation of mortality at prime adult ages (20-40) and overestimation at older ages (70+), with deviations reaching 30% for certain age groups, as starkly apparent when applied to South Africa during the HIV/AIDS epidemic.

The core challenge lies in how current models use the training data. Both SVD-Comp and log-quadratic models learn child-adult mortality relationships from the entire Human Mortality Database (HMD), then apply these globally averaged patterns uniformly to every new population. This approach assumes that a single prediction rule can capture mortality across diverse epidemiological contexts, even when populations differ markedly in their health profiles. This reflects a fundamental statistical principle: while global models trained on large, heterogeneous datasets produce stable predictions overall, they often perform poorly for populations with distinctive characteristics. Localized approaches that prioritize similar training data typically reduce these prediction errors for atypical cases (Hastie et al., 2009). Sharrow et al. (2014) demonstrated this by manually restricting their training data to HIV-affected countries, substantially improving mortality predictions for similar contexts. However, such manual curation, while effective, is ad hoc and difficult to generalize. It depends on prior knowledge and categorical grouping with no systematic way to identify relevant training data when epidemiological similarities are less obvious.

This paper introduces a new economical solution: a model-agnostic preprocessing step that automatically selects and weights training data based on demographic similarity, without requiring changes to the model structure or additional input data. The approach builds on the insight that population change serves as a powerful proxy for demographic and epidemiological characteristics, reflecting cumulative fertility, mortality, and migration processes that distinguish between contexts dominated by communicable disease, chronic illness, or demographic aging. This is supported by empirical evidence: population age structure explains a large share of variation in global disease burden (Chang et al., 2019), and epidemiological transitions create context-specific mortality patterns that distinguish between populations at different stages of demographic development (Omran, 1971). Our sorting mechanism operationalizes this principle by leveraging population change derived from decennial censuses, typically available even in countries lacking complete vital registration, to identify and weight the most demographically similar country-years in the training dataset.

Our method improves the predictive accuracy of both log-quadratic and SVD-Comp models, especially in atypical settings where global relationships perform poorly, while offering a scalable and transparent improvement that can be easily integrated into existing mortality estimation workflows, extending the reach and relevance of demographic models in data-scarce contexts.

2 DATA AND METHOD

2.1 DATA

We use the Human Mortality Database (HMD) as our primary data source, following established models like the log-quadratic model (Wilmoth et al., 2012) and SVD-comp (Clark, 2019). However, to capture more irregular mortality patterns, such as those driven by the AIDS epidemic, we supplement HMD with data from the World Population Prospects 2024 (UN WPP). Specifically, we include 40 countries with high AIDS-related mortality (over 5 deaths per 10,000 in 2000, according to UNAIDS 2024), following a similar approach to Sharrow et al. (2014). These countries span sub-Saharan Africa, Southeast Asia, and the Caribbean. While WPP estimates are modelled, they represent the best available mortality data for many lower- and middle-income countries (LMICs) and have been used to improve mortality models. We focus on the years 1970–2020, as AIDS was first identified in 1981 (CDC, 1981).

Our final dataset combines population estimates and life tables from two sources: the Human Mortality Database (HMD) and the United Nations World Population Prospects (WPP). The HMD covers 41 populations, each with multiple years of annual data, often spanning more than a century for many countries, compiled from historical mortality and population records. The WPP contributes data for 40 HIV-affected populations between 1970 and 2020. There is no overlap between the HMD and WPP populations. After calculating 10-year population change, the dataset contains 5,841 country–year observations, of which 4,201 come from the HMD and 1,640 from the WPP. On average, the HMD observations are from 1959, whereas the WPP observations are from 2000.

2.2 METHOD

Change in mortality, fertility and net migration determines the change in size of any population. This relationship is captured in the demographic balancing equation, a fundamental concept in demography:

$$Current\ population = Starting\ population + Fertility - Death + Net\ Migration$$

where net migration can be positive or negative (e.g. Preston et al. 2001; Raymer et al., 2015). This equation can be developed further for age-specific change, where previous studies have been used to explore mortality or migration (e.g., Canudas Romo et al. 2022; Raymer et al. 2022; Shen et al. 2023). Population data is often used in demographic adjustment techniques, such as the Death Distribution Method (DDM), to validate and adjust mortality estimates (Brass 1975; Hill et al. 2009). These approaches exploit the fact that age-specific population data inherently encode a population's mortality history. Building on this idea, we use the decennial cohort change, the difference between the number of people aged x in year t and those aged x–10 in year t–10, as a summary of age-specific population dynamics. This cohort change reflects cumulative effects of mortality and migration over the decade (i.e., $population(t, x) - population(t - 10, x - 10)$, where t is time and x is age). This decennial population cohort change, calculated using this approach, would contain both mortality and net migration over the decade.

Traditionally, population change has been used to estimate mortality rates mathematically, particularly within indirect estimation techniques developed for data-scarce contexts. Classic methods such as the Preston-Coale approach (Preston and Coale, 1982) and the Bennett-Horiuchi variable-r methods (Bennett and Horiuchi, 1981, 1984) directly convert age-specific population growth rates from consecutive censuses into mortality estimates. These methods apply the demographic balancing equation to decompose population change into its constituent components—births, deaths, and migration—using population change as the primary input for mortality calculation (Preston et al., 2001, Chapters 8 and 11.5) but depending on assumptions of population stability (Preston-Coale) or generalized stable population relations (Bennett-Horiuchi).

In contrast, our approach departs from these methods in both purpose and design. Instead of calculating mortality directly from population change, we use population change patterns as signals of demographic and epidemiological similarity, and match the target population to past country-years with similar patterns. This is grounded in the idea that population change reflects a country's stage in the demographic transition. Cohort growth rates capture the consequences of mortality, fertility, and migration over time, and thus serve as meaningful signatures of health and epidemiological profiles. For example, Chang et al. (2019) found that age structure change alone explains over 27% of global variation in disease burden. Similarly, younger populations tend to exhibit high burdens of communicable disease (e.g., malaria, tuberculosis, diarrheal diseases), while aging populations face greater non-communicable disease risk (e.g., cardiovascular disease, cancer, dementia), patterns directly encoded in population age dynamics. Our own analysis confirms this: populations with similar cohort change profiles tend to exhibit more similar age-specific mortality schedules than those matched by region or time alone.

A potential concern is that net migration may distort mortality signals in cohort change. However, this proves largely immaterial in practice. Migration is typically small relative to population size in most countries (median <0.1% annually, WPP 2024), and in high-migration settings like Ukraine or South Sudan, large population shifts usually co-occur with major mortality disruptions (e.g., war, displacement). Rather than confounding, migration in these cases reinforces demographic distinctiveness. This represents a fundamental advantage over traditional indirect methods like Preston-Bennett, which require minimal migration assumptions and treat migration as bias to be avoided[1]. Moreover, our goal is not to isolate mortality, but to match demographic contexts holistically, migration included. Unlike traditional indirect methods, which assume minimal migration and treat it as noise, we view migration as part of the underlying pattern we aim to recognize.

2.3 SIMILARITY METRIC

Each country-year's cohort change profile can be represented as an n-dimensional vector, where n is the number of age groups. To compare these vectors and identify similar demographic patterns, we require a metric of similarity. We evaluate both Euclidean distance and cosine similarity to compare population change profiles. Even after standardization, Euclidean distance focused too much on small, localized differences across age groups, often failing to capture meaningful demographic patterns. Cosine similarity, by contrast, emphasizes the overall shape of population change rather than pointwise differences, making it better suited for identifying structurally similar populations. Prior work supports its use in comparing standardized high-dimensional data (Lee et al., 2015; Nahm, 2004), and our empirical tests showed that it produced more interpretable and epidemiologically relevant clusters. We therefore use cosine similarity as the basis for matching country-years. Eq (1) shows the calculation of the cosine similarity (S_C):

$$S_C(P_a, P_b) = \frac{P_a \cdot P_b}{\|P_a\|\|P_b\|} = \frac{\sum_{i=1}^{n} p_{a,i} p_{b,i}}{\sqrt{\sum_{i=1}^{n} {p_{a,i}}^2} \cdot \sqrt{\sum_{i=1}^{n} {p_{b,i}}^2}}, \quad (1)$$

where P_a and P_b are vectors of population change of n age groups for two country-years a and b.

The population change needs to be standardized so that population size is not affecting the cosine similarity. One potential solution is to compute the rate using population change divided by the population at time t. However, this method produces a measure that overemphasizes the change in older ages as the population at older ages (denominator) is usually much smaller. In contrast, the rate of change in younger ages would be marginal. To give equal emphasis to all ages within a specific country-year, we instead compute the z-score of population change across age for each country-year. Thus, P_a and P_b in Formula (1) are

[1] Consider two populations both showing declining young adult populations: one experiencing HIV-related mortality and another experiencing labour out-migration. Classic methods would apply the same mathematical transformation to both cases, treating identical population change patterns as equivalent mortality signals when the underlying demographic processes are fundamentally different. The Preston-Bennett formula cannot distinguish between deaths and migration when converting population decline into mortality rates, leading to systematic overestimation of mortality in populations with high out-migration.

vectors of z-score standardized population change of n age groups. Cosine similarity ranges from -1 to 1, with 1 meaning a perfect match. We proportionally rescale[2] this range to 0 and 1 where 0 indicates no similarity.
All pairwise similarity values between the observed data and the target population are calculated. We rank all the similarity index and apply the top 10% of the similar country-year (about 500) as our target population. As noted by Lee et al. (2015), the homogeneity of the training data can influence prediction accuracy; a larger sample may reduce homogeneity and lead to poorer performance. However, a very small training sample can lead to unstable results. Consequently, there is no universal rule of thumb for this percentage; the optimal choice depends on the size of the full dataset and its internal heterogeneity (see Appendix 1 for results using the top 20% of similar country–years). We use these country-years equally as the training data tailored to the target population. This sub-sample is then used as input for the log-quadratic model by Wilmoth et al. (2012) and the SVD-Comp model developed by Clark (2019) to predict the mortality schedule for the target country.

2.4 EVALUATION

To evaluate the effectiveness of the new approach estimated by similar countries, we draw 80% of the full dataset as our training set and use it to predict the remaining 20% for out-of-sample testing. We train Log-quad and SVD-Comp models with both the full training sample and the customized sample to predict the probability of dying (q_x) of the 20% of country-years. We construct the child mortality ($0q_5$) and adult mortality ($15q_{45}$) from the observed life table as the prediction input for both Log-quad and SVD-Comp. The model adequacy is examined by the close fit between predicted and observed probability of dying and life expectancy at birth (e_0) using root mean squared error (RMSE). We repeat this process for 50 iterations with different 80% of the full dataset.

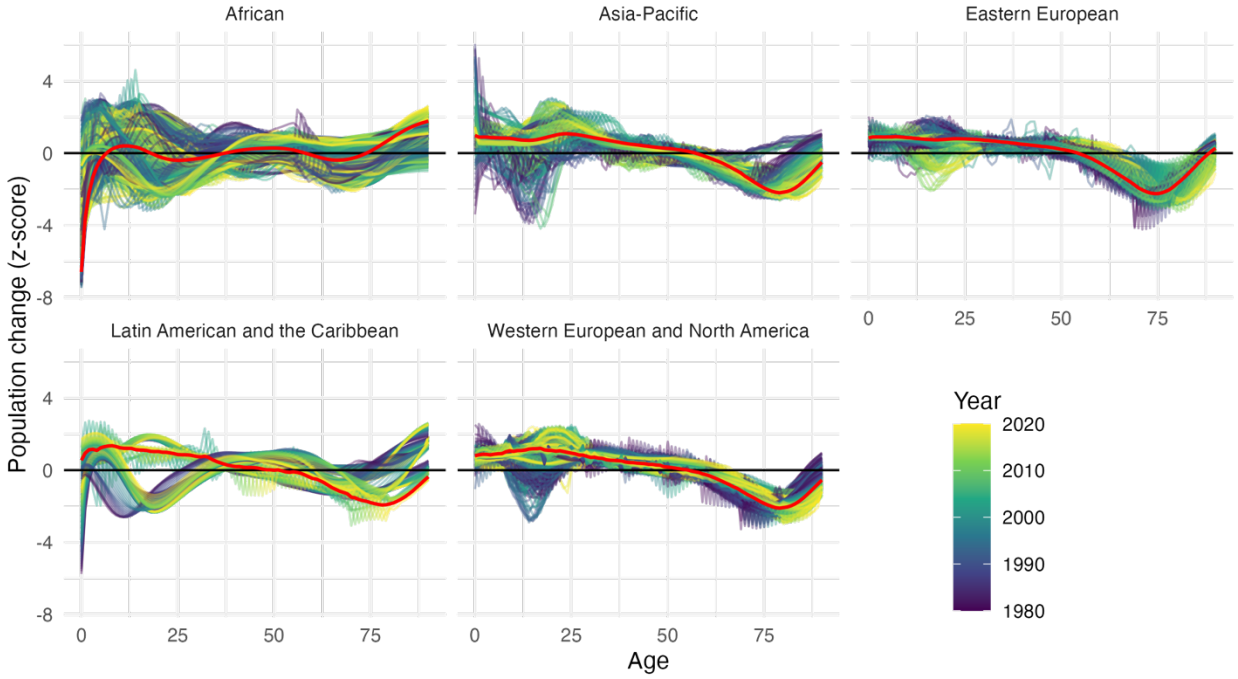
[2] Rescaled $S_c = \frac{S_C+1}{2}$

3 RESULTS

Figure 1 displays z-score standardized population change by age group across five global regions, based on 10-year intervals from 1980 to 2020. Each line represents a country-year, with colours indicating the year. Red lines show the regional mean across all country-years. Standardization allows meaningful comparison by removing differences in population scale and highlighting relative age-specific change patterns. Western European and North American populations exhibit consistently negative z-scores in older ages (60–80), indicating steady population declines driven by mortality at those ages, a signature of aging, low-mortality societies.

By contrast, African and Latin American countries show less negative or near-zero z-scores in older age groups, suggesting that population change is not primarily driven by mortality at older ages. Instead, larger variation and steeper declines at younger ages point to greater influence from early-life mortality or net migration. Eastern Europe shows a shift of negative peak to older ages especially in the post-Soviet era, indicative of delayed adult mortality. These patterns align with Riley's (2005) observations that mortality transitions occurred at different times and paces across regions. Despite regional grouping, substantial within-region variation remains, particularly in younger (<30) and older (>60) age groups, highlighting both heterogeneity in demographic transitions and the value of population change profiles as demographic signatures.

FIGURE 1: Z-SCORE OF POPULATION COHORT CHANGE BY AGE AND REGION FROM 1980 TO 2020, FEMALES



Note:
Red lines represent the region average for these years. Data are from the Human Mortality Database (HMD) and the United Nations World Population Prospects (UN WPP) for 88 countries or regions. Z-scores represent the standardized population change across different age groups. A z-score of 0 indicates the average population change, while positive and negative values show deviations from this average.

Figure 2 provides a more detailed view of how these patterns evolve over time within specific countries as examples. In 1870, Sweden's population change (panel a, red line) was characterized by significant reductions in younger age groups, particularly ages 0-10, reflecting high infant mortality typical of early-stage demographic transitions. By 2010, Sweden's largest population

decreases had shifted to much older ages (75+) (panel a, blue line), demonstrating the "rectangularization" of the mortality curve (Rossi et al. 2012; Kannisto 2000) that is characteristic of later-stage transitions (Riley 2005). South Africa (panel b) presents a contrasting case: while its 1980 pattern (red line) reflected an earlier transition stage, by 2010 (blue line), the impact of the HIV pandemic created distinctive population decreases among young adults (20-35 age range), markedly different from patterns observed in countries at similar levels of development without high HIV prevalence. The two examples illustrate how age-specific patterns of population change over time reflect different stages of demographic transition and epidemiological profiles, which are key factors linked to a population's mortality patterns. Although in a completely different period, Sweden in 1870 is very similar to South Africa in 1980 with a high population decrease in the young childhood followed by the decrease in the older age hump.

FIGURE 2: Z-SCORE OF POPULATION COHORT CHANGE BY AGE FOR SELECTED COUNTRIES AND YEARS, FEMALES

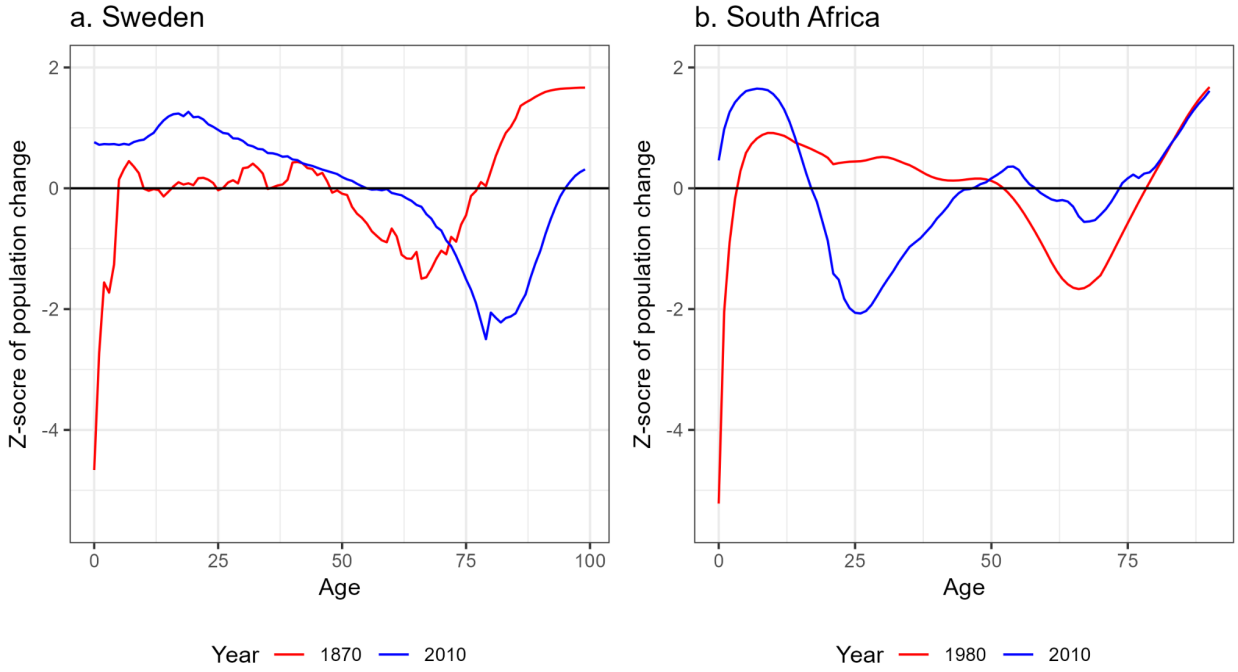
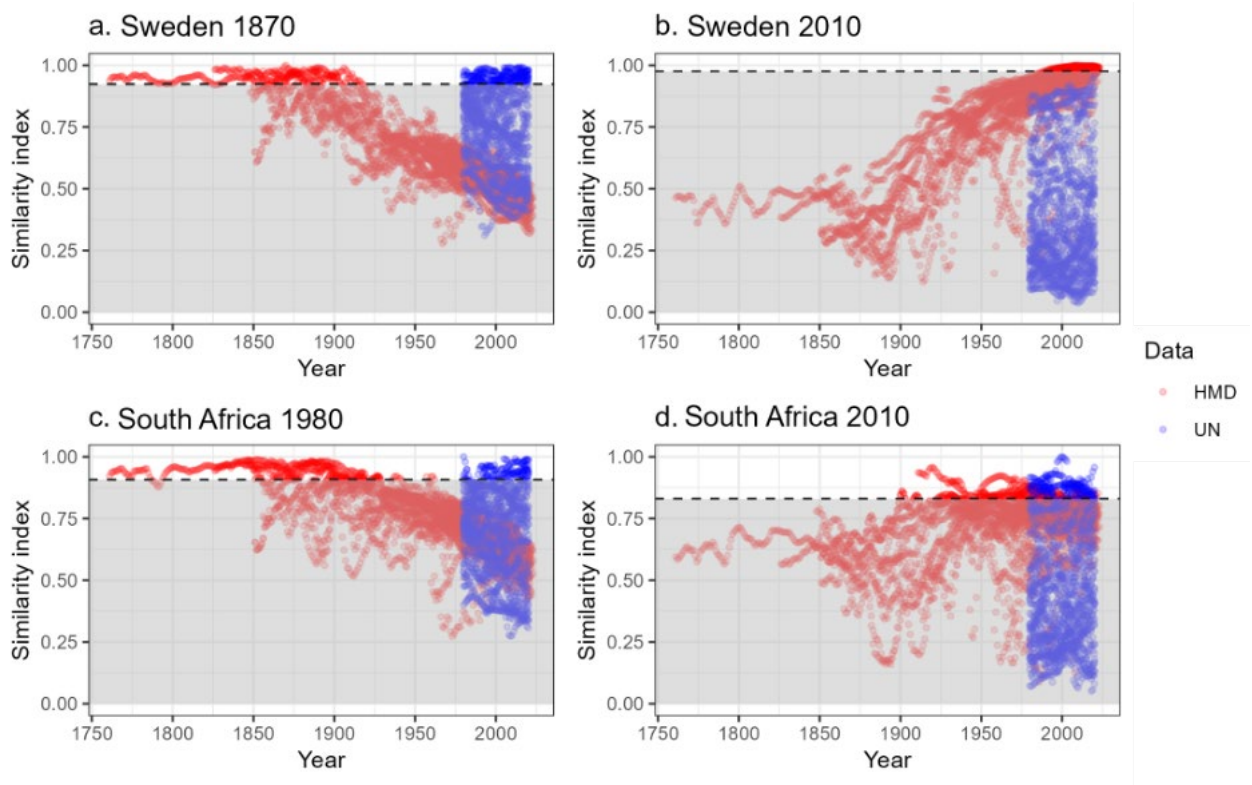


Figure 3 presents the similarity indices of population change across different country-year pairs, focusing on the examples of Sweden in 1870 and 2010, and South Africa in 1980 and 2010. The visualization quantifies similarities through our cosine similarity index, revealing how effectively population change can identify countries with comparable mortality profiles regardless of geographical proximity. Each dot represents the similarity between the target country-year and other country-years, with red dots for HMD data and blue for UN WPP data. The dashed line is the cutoff at 90% of the similarity index. Only the top 10% of the country-year is used in later models. Sweden in 1870 showed high similarity to other European countries in the 19th century but progressively lower similarity to more contemporary populations. South Africa in 2010 demonstrates high similarity specifically to other HIV-affected countries (clustered blue dots) while showing low similarity to both historical patterns and non-HIV-affected contemporary populations.

The descriptive evidence suggests that countries at similar stages of the demographic transition show comparable mortality profiles despite geographical separation. Thus, using population change as external information has the potential to improve mortality prediction by identifying country-years with similar profiles, reducing errors that arise when predictions are based on a heterogeneous mix of similar and dissimilar mortality patterns.

FIGURE 3: SIMILARITY INDEX FOR THE SELECTED COUNTRIES AND YEARS SHOWN IN FIGURE 2



Note:
Each dot represents the similarity of the target country-year (specified in the panel title) compared to other country-years, with red dots corresponding to data from the HMD and blue dots from the UN. The dashed line is the cutoff at 90% of the similarity index. Only the top 10% of the country-year is used in later models. HMD = Human Mortality Database; UN WPP = United Nations World Population Prospects.

TABLE 1 COMPARISON OF MEDIAN RMSE BETWEEN ORIGINAL AND SIMILARITY-BASED ESTIMATION MODELS, FEMALES

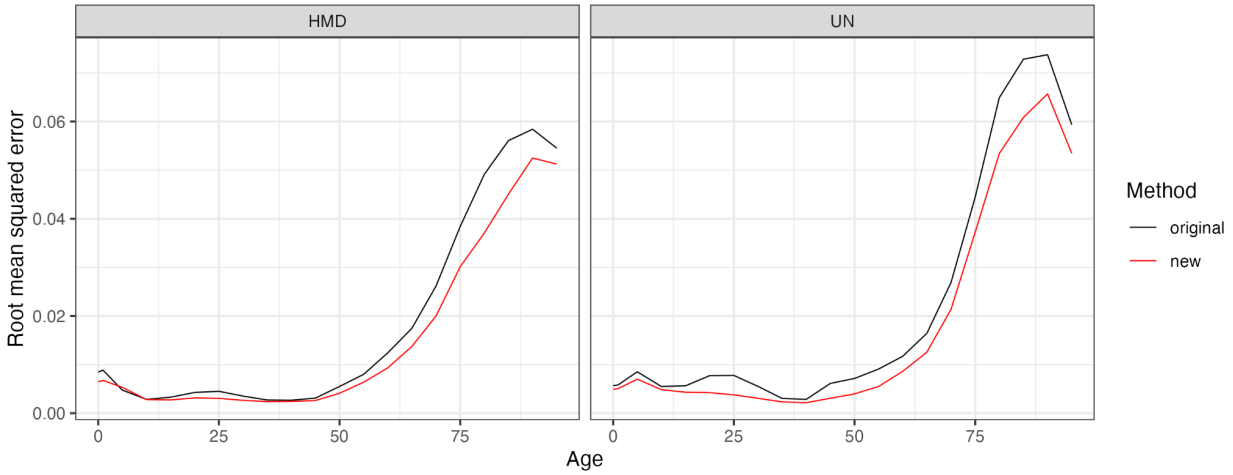
MODEL	LOG-QUAD			SVD-COMP		
DATA GROUP	ALL	HMD	UN WPP	ALL	HMD	UN WPP
	PROBABILITY OF DYING (q_X)					
ORIGINAL	0.029	0.027	0.033	0.016	0.014	0.020
NEW	0.024	0.020	0.028	0.014	0.013	0.016
	LIFE EXPECTANCY AT BIRTH (e_0)					
ORIGINAL	0.88	0.76	1.11	1.03	0.97	1.13
NEW	0.68	0.61	0.81	0.71	0.68	0.80

Note:
"Original" refers to models trained on the full dataset; "New" refers to models trained on the top 10% most similar country-years based on population change patterns. Results based on 50 iterations of 80/20 train-test splits. Life expectancy RMSE values

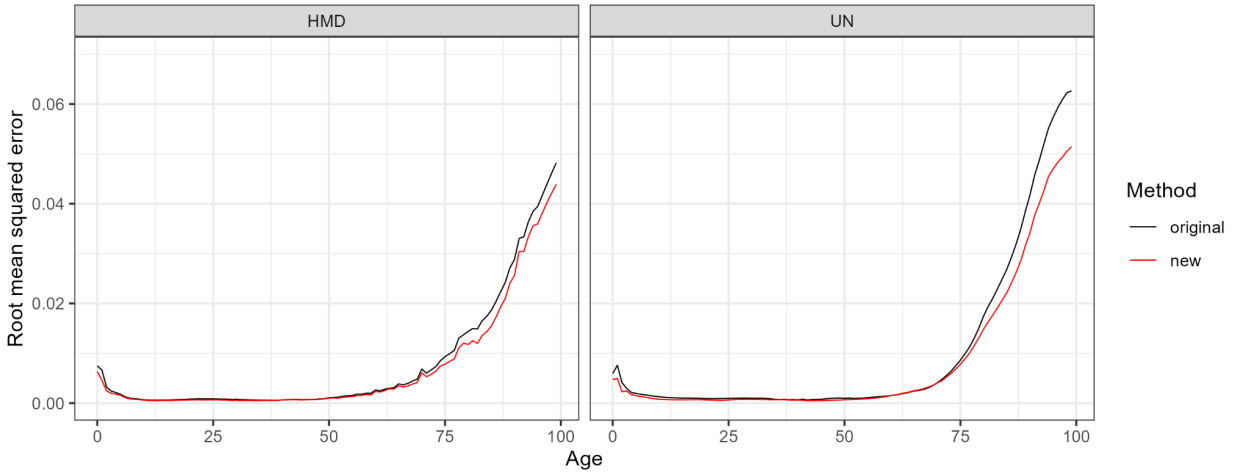
represent the median prediction error in years. For example, an RMSE of 0.88 years indicates that the typical prediction differs from the observed life expectancy by approximately 0.88 years. HMD = Human Mortality Database; UN WPP = United Nations World Population Prospects.

FIGURE 4: ROOT MEAN SQUARED ERROR (RMSE) BY AGE AND MODEL, FEMALES

a. LOG-QUAD



b. SVD-COMP



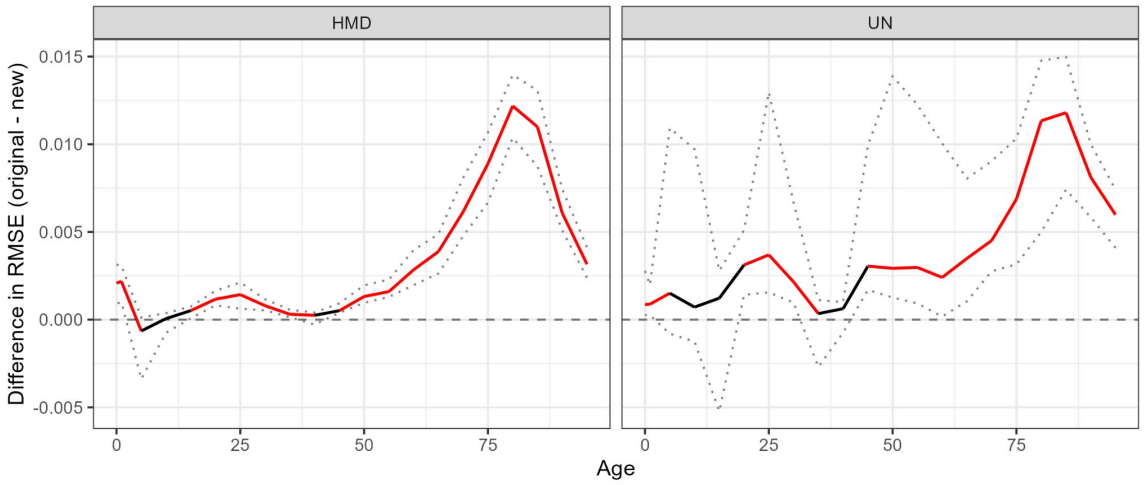
Note:
"Original" refers to models trained on the full dataset; "New" refers to models trained on the top 10% most similar country-years based on population change patterns. Because RMSE is very sensitive to extreme values, when the results from two methods differ by a factor of 7 or more, they are excluded from the analysis (affecting approximately 0.1% of all results). See Figure 5 for significance and Appendix 1 for the difference in RMSE by age. HMD = Human Mortality Database; UN WPP = United Nations World Population Prospects.

Table 1 presents the root mean squared error (RMSE) for probability of dying (q_x) and life expectancy at birth (e_0) comparing our approach against the original methods using the full training dataset. We observe a consistent and significant improvement across models. For females, median RMSE for probability of dying decreases from 0.029 to 0.024 in log-quadratic models (17% improvement) and from 0.016 to 0.014 in SVD-Comp models (13% improvement). The benefits are particularly pronounced for

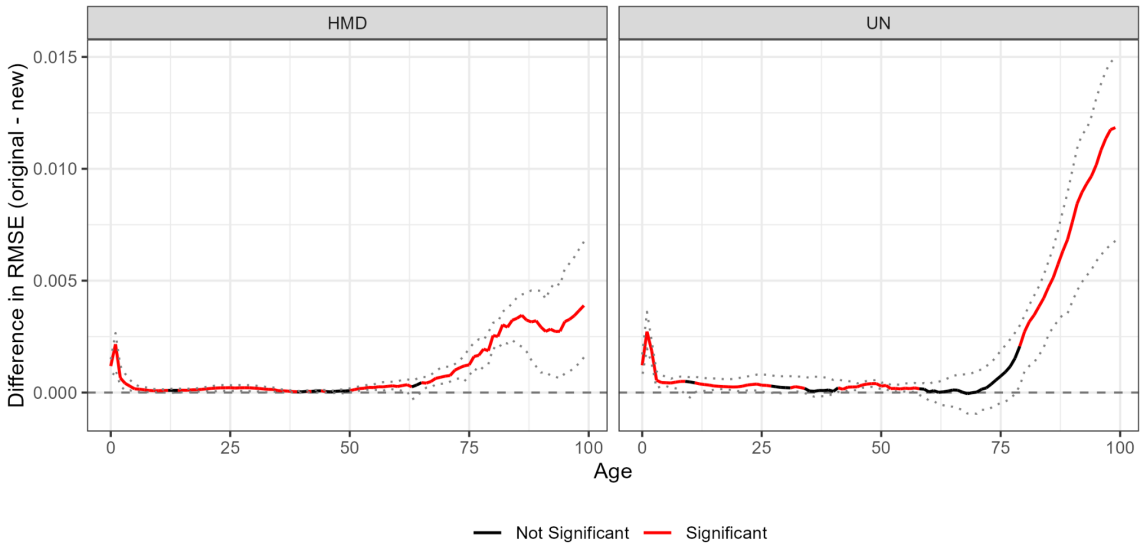
UN WPP countries compared to HMD countries. For log-quadratic models, UN WPP countries show RMSE reductions from 0.033 to 0.028, while HMD countries improve from 0.027 to 0.022. Similar patterns emerge for life expectancy predictions, where UN WPP countries achieve substantial improvements: from 1.11 to 0.81 years (log-quadratic) and from 1.13 to 0.80 years (SVD-Comp). These findings indicate that our approach provides the greatest benefit for populations with distinctive mortality profiles that diverge from patterns typical in high-quality vital registration systems.

FIGURE 5: DIFFERENCE IN RMSE BY AGE AND MODEL, WITH SIGNIFICANCE AT 95%, FEMALES

a. LOG-QUAD



b. SVD-COMP



Note:
Because RMSE is very sensitive to extreme values, when the results from two methods differ by a factor of 7, they are excluded from the analysis, affecting approximately 0.1% of all results. (We only show significance if the new method performs better. The new method never performs significantly worse than the original method in any age)

Figure 4 displays age-specific RMSE by data group for log-quadratic (panel A) and SVD-Comp (panel B) models. Both models show substantial improvements at older ages, with the log-quadratic model also demonstrating clearer gains at younger ages, though these younger-age improvements are less apparent in panel B due to scaling differences. Figure 5 presents the statistical significance of these improvements, revealing that our approach significantly enhances prediction accuracy at three key age ranges: older ages (65+), young adult ages (20-35), and infancy (0-1). These age groups represent periods where mortality varies substantially across populations due to different epidemiological profiles, from early-life infectious diseases and maternal health conditions to adult HIV/AIDS mortality and age-related chronic diseases. The improvements are particularly pronounced for non-HMD countries, especially for log-quadratic models at early ages, highlighting the method's effectiveness for populations whose mortality patterns diverge from those typical of Western historical experience.

4 DISCUSSION

This paper introduces a novel sorting approach that leverages population change patterns to improve mortality estimation accuracy in data-scarce settings. Our method addresses longstanding challenges in the mortality modelling literature while leveraging the established demographic principle that population change reflects underlying mortality, fertility, and migration processes.

Our first key contribution is providing a model-agnostic solution that integrates into existing workflows established by Wilmoth et al. (2012) and Clark (2019). Rather than developing entirely new estimation frameworks, our preprocessing step preserves the mathematical foundations of established models while enhancing their performance. This design choice draws on lessons from the statistical learning literature, where localized approaches have shown effectiveness for populations with distinctive characteristics (Hastie et al., 2009). By requiring only population data from decennial censuses, typically available even where vital registration systems are incomplete (Mikkelsen et al., 2015), our method avoids the additional data requirements that often limit model applicability in low-resource settings.

Our results demonstrate sizable improvements over existing methods. The 17% reduction in median RMSE for log-quadratic models and 13% for SVD-Comp models represent meaningful gains, particularly given the inherent challenges of mortality estimation in data-scarce contexts (Li, 2015). To put this in perspective, these improvements translate to reducing life expectancy prediction errors from 1.11 to 0.81 years for UN WPP countries using log-quadratic models, a practically significant improvement for public health planning and resource allocation. While these improvements are more modest than the substantial 30-61% error reductions achieved by major methodological breakthroughs like the modified logit system (Murray et al., 2003) or SVD-Comp introduction (Clark, 2019) when compared to their traditional predecessors, they represent meaningful advances within the typical range reported for methodological refinements in the demographic literature. These gains are particularly notable given that they build upon already state-of-the-art approaches, demonstrating the flexibility of our model-agnostic preprocessing step to enhance existing methods without structural changes.

The largest gains occur at older ages (70+) and young adult ages (20-45), which correspond to age ranges where Clark (2019) identified significant limitations in current approaches. These age groups exhibit the greatest variability in mortality patterns across populations due to differences in epidemiological contexts, from HIV/AIDS affecting young adults to varying chronic disease burdens at older ages. By identifying demographically similar training populations, our method better captures these context-specific mortality patterns than approaches that average across all available data. In addition, the particularly strong improvements for UN WPP countries, with life expectancy RMSE decreasing from 1.11 to 0.81 years for log-quadratic models, underscore the method's value for populations most dependent on model-based estimation. These populations often have distinctive mortality profiles that diverge from the Western historical patterns dominating the Human Mortality Database, explaining why similarity-based matching proves especially beneficial.

For practitioners working in data-scarce settings, our approach offers a transparent, step-by-step process: (1) calculating age-specific cohort change using consecutive censuses, (2) standardizing changes using z-scores across age groups, (3) computing cosine similarity with historical training data, and (4) selecting the top 10% most similar country-years for model training. This methodology is most beneficial for populations suspected of deviating from typical Western mortality patterns, particularly those affected by HIV/AIDS, experiencing rapid demographic transitions, or exhibiting distinctive epidemiological profiles that challenge existing model assumptions.

Several limitations warrant acknowledgment. First, population change captures not only mortality but also net migration, which could distort similarity metrics in high-migration contexts. While our analysis suggests this concern is largely immaterial in practice, given that migration typically represents less than 0.1% of population size annually (UN WPP, 2024), future refinements might explore methods to isolate mortality signals more precisely or develop migration-adjusted similarity metrics. Second, like all statistical models relying on historical patterns, our approach assumes that relevant training data exist for target populations. Populations experiencing truly unprecedented mortality crises may still present challenges, echoing broader limitations in the mortality modelling literature noted by Clark (2019). We partially address this through our inclusion

of diverse WPP data, but the fundamental challenge of extrapolating beyond observed experience remains. Third, the differences between observed census data and modelled population estimates, including enumeration errors and timing inconsistencies, could affect performance in real-world applications. Future research should test our framework using raw census data and explore robustness to data quality issues commonly encountered in demographic analysis (Hill et al., 2009).

Despite the limitations, this framework opens several promising research directions. Extension to subnational mortality estimation could leverage population change patterns to identify comparable regions despite limited data availability, addressing a key gap identified by Li (2015) in current estimation capabilities. The same logic of contextual similarity could be adapted to estimate other demographic indicators, building on the broader literature on similarity-based prediction methods in demographics and related fields. More broadly, our results support the value of incorporating demographic transition stages into mortality modelling, moving beyond the geographical classifications that dominated early model life table development (United Nations, 1982) toward more nuanced approaches that recognize the complex interplay between epidemiological context and mortality patterns. By demonstrating that decennial population change can effectively identify demographically similar training cases, our approach offers a practical tool for improving model performance where direct mortality data remain sparse or unreliable, addressing a persistent challenge in global demographic analysis.

REFERENCES

Bennett, N. G., & Horiuchi, S. (1981). Estimating the completeness of death registration in a closed population. Population Index, 47(2), 207–221. https://doi.org/10.2307/2736447

Bennett, N. G., & Horiuchi, S. (1984). Mortality estimation from registered deaths in less developed countries. Demography, 21(2), 217–233. https://doi.org/10.2307/2061041

Brass, W. (1975). Methods for estimating fertility and mortality from limited and defective data. Laboratories for Population Statistics Occasional Publication.

Canudas-Romo, V., Shen, T., & Payne, C. F. (2022). The components of change in population growth rates. Demography, 59(2), 417–431. https://doi.org/10.1215/00703370-9765067

Centers for Disease Control and Prevention. (1981). Pneumocystis pneumonia — Los Angeles. Morbidity and Mortality Weekly Report, 30, 250–252.

Chang, A. Y., Skirbekk, V. F., Tyrovolas, S., Kassebaum, N. J., & Dieleman, J. L. (2019). Measuring population ageing: An analysis of the Global Burden of Disease Study 2017. The Lancet Public Health, 4(3), e159–e167. https://doi.org/10.1016/S2468-2667(19)30019-2

Clark, S. J. (2019). A general age-specific mortality model with an example indexed by child mortality or both child and adult mortality. Demography, 56(3), 1131–1159. https://doi.org/10.1007/s13524-019-00785-3

Coale, A. J., & Demeny, P. G. (1966). Regional model life tables and stable populations. Princeton University Press.

Coale, A. J., Demeny, P. G., & Vaughan, B. (1983). Regional model life tables and stable populations (2nd ed.). Academic Press.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7

Hill, K., You, D., & Choi, Y. (2009). Death distribution methods for estimating adult mortality: Sensitivity analysis with simulated data errors. Demographic Research, 21, 235–254. https://doi.org/10.4054/DemRes.2009.21.9

Kannisto, V. (2000). Measuring the compression of mortality. Demographic Research, 3, 6. https://doi.org/10.4054/demres.2000.3.6

Lee, J., Maslove, D. M., & Dubin, J. A. (2015). Personalized mortality prediction driven by electronic medical data and a patient similarity metric. PLoS One, 10(5), e0127428. https://doi.org/10.1371/journal.pone.0127428

Li, N. (2015). Estimating life tables for developing countries (Technical Paper No. 2014/4). United Nations, Department of Economic and Social Affairs, Population Division. http://www.un.org/en/development/desa/population/publications/pdf/technical/TP2014-4.pdf

Mikkelsen, L., Phillips, D. E., AbouZahr, C., Setel, P. W., de Savigny, D., Lozano, R., & Lopez, A. D. (2015). A global assessment of civil registration and vital statistics systems: Monitoring data quality and progress. The Lancet, 386(10001), 1395–1406. https://doi.org/10.1016/S0140-6736(15)60171-4

Murray, C. J. L., Ferguson, B. D., Lopez, A. D., Guillot, M., Salomon, J. A., & Ahmad, O. (2003). Modified logit life table system: Principles, empirical validation, and application. Population Studies, 57(2), 165–182. https://doi.org/10.1080/0032472032000097083

Nahm, U. Y. (2004). Text mining with information extraction [Doctoral dissertation, The University of Texas at Austin]. ProQuest Dissertations & Theses.

Omran, A. R. (1971). The epidemiologic transition: A theory of the epidemiology of population change. The Milbank Memorial Fund Quarterly, 49(4), 509–538. https://doi.org/10.2307/3349375

Preston, S. H., & Coale, A. J. (1982). Age structure, growth, attrition, and accession: A new synthesis. Population Index, 48(2), 217–259. https://doi.org/10.2307/2736093

Preston, S. H., Heuveline, P., & Guillot, M. (2001). Demography: Measuring and modeling population processes. Blackwell Publishers.

Raymer, J., Rees, P., & Blake, A. (2015). Frameworks for guiding the development and improvement of population statistics in the United Kingdom. Journal of Official Statistics, 31(4), 699–722. https://doi.org/10.1515/jos-2015-0041

Raymer, J., Wiśniowski, A., Forster, J. J., Smith, P. W., & Bijak, J. (2013). Integrated modeling of European migration. Journal of the American Statistical Association, 108(503), 801–819.

Riley, J. C. (2005). The timing and pace of health transitions around the world. Population and Development Review, 31, 741–764. https://doi.org/10.1111/j.1728-4457.2005.00096.x

Rossi, I. A., Rousson, V., & Paccaud, F. (2012). The contribution of rectangularization to the secular increase of life expectancy: An empirical study. International Journal of Epidemiology, 42(1), 250–258. https://doi.org/10.1093/ije/dys219

Sharrow, D. J., Clark, S. J., & Raftery, A. E. (2014). Modeling age-specific mortality for countries with generalized HIV epidemics. PLoS One, 9(5), e96447. https://doi.org/10.1371/journal.pone.0096447

Shen, T., Raymer, J., Guan, Q., & Wiśniowski, A. (2024). The estimation of age and sex profiles for international migration amongst countries in the Asia-Pacific region. Population, Space and Place, 30, e2716. https://doi.org/10.1002/psp.2716

UNAIDS. (2024, July 31). AIDS mortality per 1000 population. https://aidsinfo.unaids.org/?did=5f011137629b296603d55ab7&r=world&t=null&tb=d&bt=dnli&ts=0,0&tr=world&aid=5f01116b629b296603d55ab8&sav=Population

United Nations. (1982). Model life tables for developing countries. Population Division, Department of Economic and Social Affairs, United Nations.

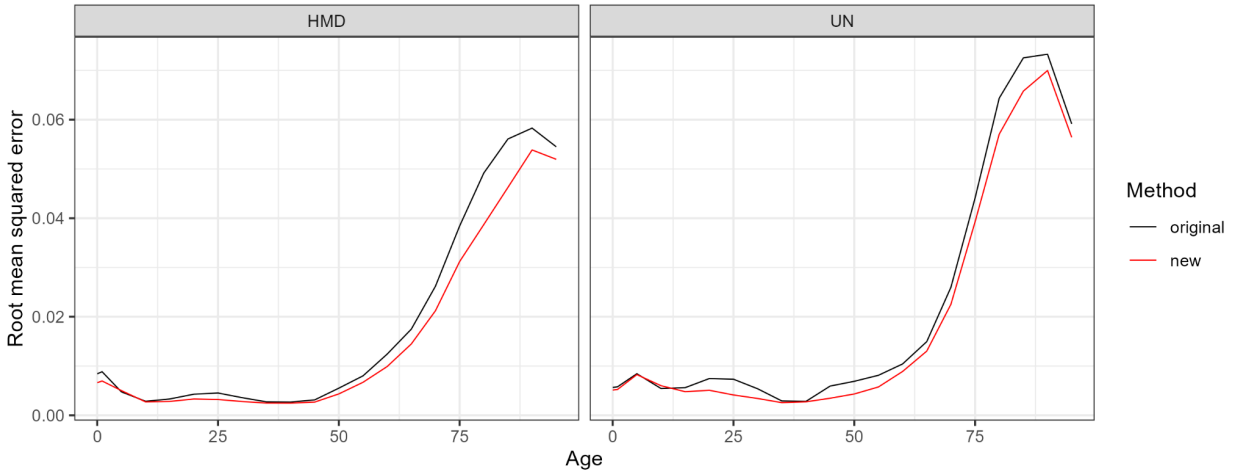
United Nations. (2024). World Population Prospects 2024. Population Division, Department of Economic and Social Affairs, United Nations.

Wilmoth, J., Zureick, S., Canudas-Romo, V., Inoue, M., & Sawyer, C. (2012). A flexible two-dimensional mortality model for use in indirect estimation. Population Studies, 66(1), 1–28. https://doi.org/10.1080/00324728.2011.611411

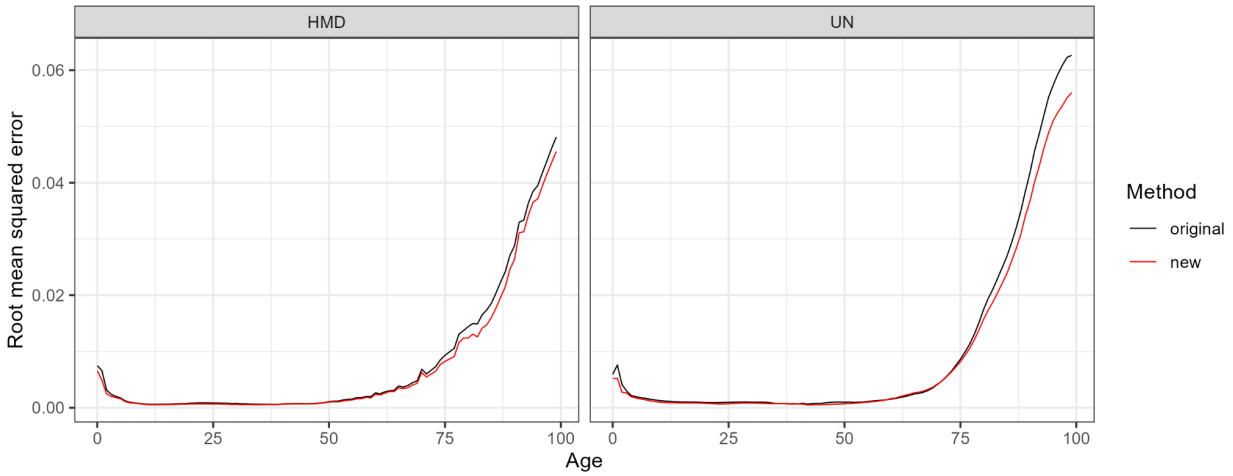
APPENDIX

APPENDIX 1: ROOT MEAN SQUARED ERROR (RMSE) WITH TOP 20% SIMILAR COUNTRIES BY AGE AND MODEL , FEMALES

a. LOG-QUAD



b. SVD-COMP



Note:
"Original" refers to models trained on the full dataset; "New" refers to models trained on the top 10% most similar country-years based on population change patterns. Because RMSE is very sensitive to extreme values, when the results from two methods differ by a factor of 7 or more, they are excluded from the analysis.


VIENNA INSTITUTE
OF DEMOGRAPHY