DEMOGRAPHIC RESEARCH
A peer-reviewed, open-access journal of population sciences

DEMOGRAPHIC RESEARCH

VOLUME 54, ARTICLE 28, PAGES 877–934
PUBLISHED 23 APRIL 2026
http://www.demographic-research.org/Volumes/Vol54/28/
DOI: 10.4054/DemRes.2026.54.28

Research Article

Bayesian multidimensional mortality reconstruction

Andrea Tamburini

Arkadiusz Wiśniowski

Dilek Yildiz

Contents

Demographic Research: Volume 54, Article 28
Research Article

Bayesian multidimensional mortality reconstruction

Andrea Tamburini[1]

Arkadiusz Wiśniowski[2]

Dilek Yildiz[3]

Abstract

BACKGROUND
Even though mortality differentials by socioeconomic status and educational attainment level have been widely examined, the research is often limited to developed countries and recent years. This is primarily due to the absence of consistent and good-quality data. Systematic studies with a broad geographical and temporal spectrum that engage with the link between educational attainment and mortality are still lacking.

OBJECTIVE
We aim to develop a statistical model that uses multiple patchy data sources to reconstruct mortality rates by age, sex, and the level of education.

METHODS
The proposed approach is a hierarchical Bayesian model that combines the strengths of multiple sources in order to estimate mortality rates by time periods, age groups, sex, and educational attainment.

RESULTS
We apply the model in a case study that includes 13 countries across South-East Europe, Western Asia, and North Africa, and calculate education-specific mortality rates for females in five-year age groups starting at age 15 for the 1980–2015 period. We then validate our estimates via posterior predictive checks.

[1] International Institute for Applied Systems Analysis, Wittgenstein Centre for Demography and Global Human Capital (IIASA, OeAW, University of Vienna), Laxenburg, Austria. Email: tamburini@iiasa.ac.at.
[2] Social Statistics Department, The University of Manchester, Manchester, United Kingdom.
[3] International Institute for Applied Systems Analysis, Wittgenstein Centre for Demography and Global Human Capital (IIASA, OeAW, University of Vienna), Laxenburg, Austria.

CONTRIBUTION
There are two contributions of this work. First, we propose a novel and flexible probabilistic method to inform the research on the relationship between education and adult mortality and, second, we provide age, sex, and education-specific mortality estimates for 13 countries. Our study addresses the lack of education-specific mortality differentials by providing a flexible method for their estimation.

1. Introduction

1.1 Motivation

There is a growing body of literature showing that education has a direct impact on mortality (Baker et al. 2011). Although this relationship has been reported globally (Pradhan et al. 2017; Gakidou et al. 2010; Byhoff et al. 2017), nationally (Montez, Hummer, and Hayward 2012; Krueger et al. 2015), and subnationally (Bora, Lutz, and Raushan 2018; Sasson and Hayward 2019), the research to date has focused only on specific populations (e.g., subgroups of the adult population, infants), and has not addressed the systematic reconstruction of the age- and education-specific mortality rates for adults. Moreover, most of the previous studies have analysed the association between education and mortality, or have quantified the positive effects of education on a population's health and survival rates only at aggregate levels. The primary obstacle that has constrained the existing research in terms of both the spatial and the historical scope is the incompleteness of mortality data by educational attainment. To the best of our knowledge, there are no databases or collections of datasets that provide mortality rates or counts of deaths by educational attainment for a large group of countries especially in the Global South and for more than 15 years. Nonetheless, such data, analysed either in isolation or in combination with other indicators, are needed (1) to understand how the interaction between education and mortality has evolved for subpopulations in different countries; (2) to study socioeconomic disparities in mortality over a broader geography and longer time periods; and (3) to provide more accurate baseline estimates to project multidimensional populations.

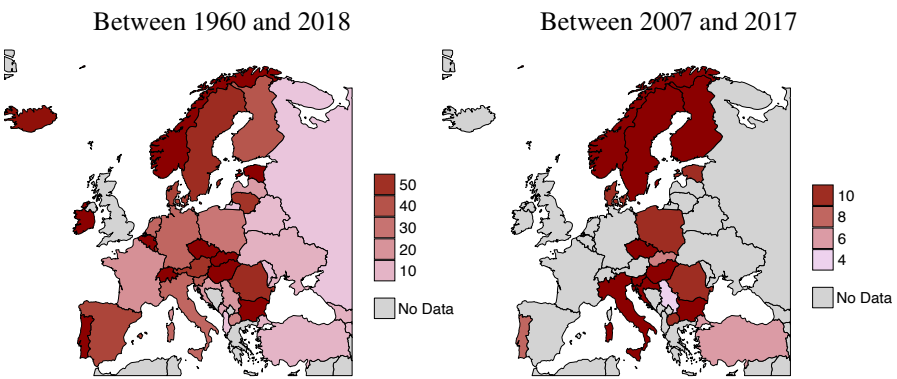
Globally, the main source of comparable mortality data is the United Nations World Population Prospects (UN WPP) (United Nations 2022). The UN WPP provides population counts, vital rate estimates, and projections from 1950 to 2100 for 235 countries or areas. However, these estimates are not broken down by levels of educational attainment. The most comprehensive and systematically verified estimates and projections of population counts and consistent demographic rates, such as total fertility rates and age-specific survival ratios disaggregated by education level, come from the Wittgenstein

Centre Data Explorer (WCDE, Wittgenstein Centre for Demography and Global Human Capital 2018). Although survival rate data based on educational attainment are available for the reference period (2015–2020) and for future projections under various shared socioeconomic pathways scenarios, these datasets are not available for historical periods.

In Europe, mortality data segmented by educational attainment are available only for the recent years and primarily through national statistical offices (see Figure 1, right panel). The Eurostat data collection (Eurostat 2025) represents a recent effort to address these gaps. Despite these initiatives, harmonized high-quality data remain scarce, even for broad age groups.

For countries in the Global South, the situation is more challenging due to the lack of reliable civil registration systems. Mortality data are derived primarily from nationally representative surveys such as the Demographic and Health Surveys (DHS, ICF 2022). However, these surveys rarely collect adult mortality information. Instead, mortality estimates often rely on indirect methods, such as life tables and the sisterhood method for maternal deaths (Graham, Brass, and Snow 1989; United Nations 1983).

Figure 1: Number of age- and sex-specific life tables available in the Eurostat database, without (left panel) and with (right panel) the educational attainment attribute



Source: Authors' own calculations based on Eurostat data (Eurostat 2025).

In this article, we propose a probabilistic hierarchical model to estimate past mortality rates by five-year age groups and by educational levels between the years 1980 and 2015. We apply the model to a case study that integrates data from Eurostat, the DHS, and the UN WPP. Our contribution is twofold. First, we propose a method that fills the gap in the literature on reconstructing multidimensional mortality rates with a systematic procedure for constructing inputs when data are missing. Second, we apply the method and

reconstruct mortality rates by educational levels for a set of countries, including measures of uncertainty that take into account the quality of the input data.

1.2 Modelling framework

Historically, the research on demographic data reconstruction has focused more on population sizes than on separate estimates of fertility, mortality, and migration rates (for an overview, see Wheldon et al. 2013). Two main methods are used to reconstruct past population sizes: (1) demographic back-projection attempting to revert the relationships between population size and composition and mortality, fertility, and migration rates based on the cohort components method for population projections (Wrigley and Schofield 1983; Lutz et al. 2007; Goujon et al. 2016; Lutz et al. 2018; Speringer et al. 2021), and (2) inverse projection (Lee 1974, 1985). While these advanced methods have been validated with historical data, they are both deterministic and do not fully account for uncertainties related to data quality and assumptions.

Another group of researchers used Bayesian inference to simultaneously reconstruct population sizes and demographic rates (mortality and fertility rates and net migration flows) by combining incomplete data sources (Wheldon et al. 2013, 2016, 2015). While they take into account the possible uncertainties in the modelling process, their methodology has so far been limited to analysing age and sex structures and has not been applied to multistate populations, such as to populations disaggregated by the level of education.

Bayesian hierarchical models have also been utilised to reconstruct fertility rates. Durowaa-Boateng, Yildiz, and Goujon (2023) and Yildiz et al. (2023) disaggregate the UN WPP age-specific fertility rates and total fertility rates (TFR) by four levels of educational attainment for countries in Latin America and sub-Saharan Africa. Alkema et al. (2012) account for deficiencies in data sources to estimate TFRs by combining information from multiple surveys, including DHS and World Fertility Surveys. They produce estimates for West Africa, a region characterised by data scarcity. In an alternative approach, Schmertmann and Hauer (2019) combine information regarding the age-sex population structure and the child-to-woman ratio to infer the TFR.

Missing and incomplete data are more pronounced in the study of migration (for a detailed review, see Willekens et al. 2016). To address this, Bayesian hierarchical models have been developed to integrate various types of migration data (Aparicio-Castro, Wiśniowski, and Rowe 2023; Raymer et al. 2013; Wiśniowski 2017; Yildiz et al. 2024; Wiśniowski 2021). These models correct for measurement errors and impute missing information, and can incorporate information derived from social media, such as the Facebook Advertising Platform, to complement traditional data sources.

Finally, Bayesian approaches have also been employed to estimate mortality rates. Alkema and New (2014) and Alexander and Alkema (2018) deal with limited data avail-

ability and data quality issues in developing countries to estimate under-five and neonatal mortality rates, respectively. In these works, the authors use Bayesian regression spline models. They take into account data quality issues and various sources of error, as well as the considerable differences in data availability across various countries. Their methodology permits the borrowing of information from multiple countries and over time, and is designed to prevent the over-representation of countries with better data, by adjusting the predictive intervals according to the amount and the quality of the available information. A similar approach of borrowing information from other data sources is proposed by Alexander, Zagheni, and Barbieri (2017). They address the small sample size problem in subpopulations (the population of the United States split up at the county level) by sharing information across different geographical levels. They used state-level mortality profiles to inform estimates of mortality rates in counties (small areas) via singular value decomposition (SVD), which allowed for the imputation of missing observations and the correction of measurement irregularities in small area data.

Building upon the above-mentioned literature, we propose a multidimensional hierarchical Bayesian model to reconstruct mortality by level of education. The model integrates available population and mortality data drawn from multiple sources, exploits their strengths, and compensates for their limitations by borrowing information over time and across countries through its hierarchical structure. The model also takes into account the uncertainty arising from the variability of the quality and the precision of the data, and the uncertainty about the model parameters. It generates age-country-specific mortality rates for multiple countries (five-year age groups starting at age 15) by two levels of education: (1) completed primary education or less and (2) more than completed primary education.

Our modelling framework is similar to that developed by Alexander, Zagheni, and Barbieri (2017), as it also uses the SVD to extract information on mortality age profiles. However, our objective is to reconstruct mortality rates by level of education. Hence, the SVD was performed on the estimates of age- and education-specific mortality rates in order to borrow information from various countries. The year- and education-specific mortality rates are then shaped by additional inputs.

The general reconstruction model specification is outlined in Section 2. The case study application is presented in Section 3. The model validation, performance analysis, and results are summarised in Section 4, and we conclude in Section 5.

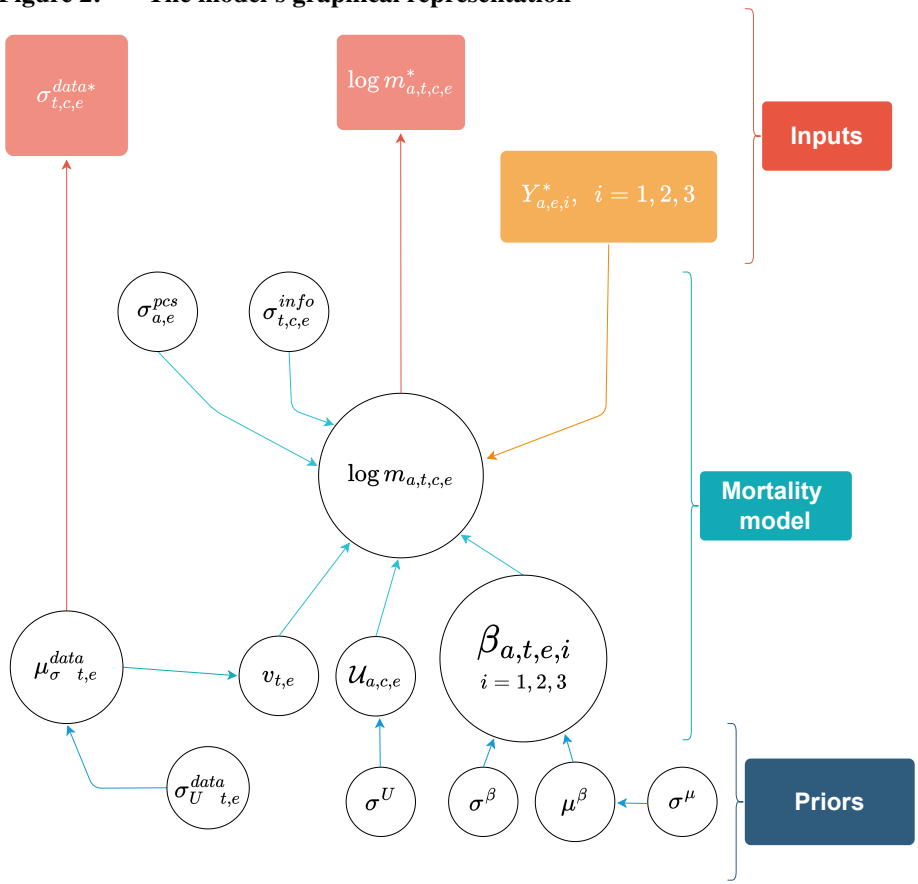
2. Reconstruction model setup

We propose a model that captures the unknown educational differentials in age-specific adult mortality rates. Our model formulation is not sex-specific and has been tested for the female population in our case study. However, the same modelling technique can be used to produce estimates for the male or the total population. Notably, the data sources

employed in this study do not indicate any potential quality degradation for the male population, and no additional or different step would be required to apply this modelling technique to the male population. Our case study is presented, together with a detailed description of the input data preparation, in Section 3.

The model is presented in Figure 2, which shows that we are reconstructing unknown (latent) log-mortality rates, $\log m_{a,t,c,e}$, that are specific to age group (a), year (t), country (c), and level of education (e), by using a variety of inputs and relying on informative prior distributions.

Figure 2: The model's graphical representation



Notes: The different components are visualised as follows: red square: quantities estimated outside of the model that are used as data; orange square: quantities estimated outside of the model that are used as fixed covariates; circle: random variables.

In the model, the age-, year-, country-, and education-specific input log-mortality rates $\log m^*_{a,t,c,e}$ rely on information from other countries and various sources (calculations for our case study are presented in Appendix D). We constructed them by applying the estimated mortality schedules to the 15–19 age group log-mortality rates as starting points. These starting points were derived from a log-linear model implemented outside of the principal one (for details, see Section 3.2.2, Appendix C, and Appendix D). Then, we assume $\log m^*_{a,t,c,e}$ are normally distributed (as in Equation 1):

$$\log m^*_{a,t,c,e} \sim \mathcal{N}\Big(\log m_{a,t,c,e},\ \sigma^{info}_{t,c,e}\Big), \tag{1}$$

with the expected value being the key quantity of interest, that is, unobserved (reconstructed) log-mortality rates $\log m_{a,t,c,e}$. Throughout this article, we use an asterisk $*$ in the superscript to denote an input rather than a model parameter. An illustration of how to create this input is presented in Section 3.2. The initial rates are broken down by age and education, and need to be estimated for each year and country. Consequently, uncertainty due to modelling or due to measurement errors in data sources need to be taken into account. In Equation 1, parameter $\sigma^{info}_{t,c,e}$ denotes the standard deviation reflecting the uncertainty around $\log m_{a,t,c,e}$ that can be derived from the aforementioned log-linear model (Appendix C).

The outcome of the log-linear model shapes the development of the entire schedule, as the initial log-mortality rate for the '15–19' age group serves as the starting point for constructing the mortality schedules used to estimate log-mortality rates. The associated uncertainty is estimated using a weakly informative prior for the corresponding precision:

$$\begin{aligned} \tau^{info}_{t,c,e} &\sim \Gamma(0.1, 0.1), \\ \sigma^{info}_{t,c,e} &= 1/\sqrt{\tau^{info}_{t,c,e}}. \end{aligned} \tag{2}$$

Next, the unobserved mortality rates are reconstructed by using information derived from various data sources (Equation 3). Here we assume that the reconstructed mortality rates are informed by three age- and education-specific principal components ($Y^*_{a,e,i}$), obtained from two separate sets of mortality curves differentiated by the educational attainment. These provide a time-independent basic structure of the mortality curves together with their time-dependent loads ($\beta_{a,t,e,i}$), and a set of random effects:

$$\log m_{a,t,c,e} \sim \mathcal{N}\Big(\sum_{i=1}^{3} \beta_{a,t,e,i} \cdot Y^*_{a,e,i} + u_{a,c,e} + \nu_{t,e},\ \sigma^{pcs}_{a,e}\Big). \tag{3}$$

The mean of the normal distribution in Equation 3 is derived from an expansion of the principal components structure as outlined by Alexander, Zagheni, and Barbieri (2017). The most notable differences are the education-specific formulation of the principal components and the structure of the random effects. The principal components are specified according to the educational attainment to ensure enough flexibility in describing the log-mortality rates for the subpopulations we focus on. Random effects $u_{a,c,e}$ capture deviations from the education-specific profiles described by the principal components for each country (Equation 4), and are informed by the data through weakly informative hierarchical priors that are uniform distributions for their standard deviation $\sigma^u_{a,e}$ (Equation 5):

$$u_{a,c,e} \sim \mathcal{N}\Big(0,\ \sigma^u_{a,e}\Big) \tag{4}$$

$$\sigma^u_{a,e} \sim \mathcal{U}[0, 40]. \tag{5}$$

Furthermore in Equation 3, we introduce random effects $\nu_{t,e} \sim \mathcal{N}\Big(0,\ {\mu^{data}_{\sigma}}_{t,e}\Big)$, which depend on the available data through the ${\mu^{data}_{\sigma}}_{t,e}$ parameter, denoting the standard deviation associated with the estimation of the log-mortality rates for the 15–19 age group and that is also a random variable (estimand). It has a prior informed by the uncertainty around the estimates used to fit the log-linear model denoted by $\sigma^{data*}_{t,c,e}$ (cf. Figure 2). This implies that not just the uncertainty stemming from the log-linear estimation (Equation 2) is considered but also the one directly linked to the estimates used for it. To ensure the standard deviation is positive, we assume it is normally distributed on the logarithmic scale, with $\log \sigma^{data*}_{t,c,e} = \log(\sigma^{data*}_{t,c,e})$:

$$\log \sigma^{data*}_{t,c,e} \sim \mathcal{N}\Big(\log {\mu^{data}_{\sigma}}_{t,e}, {\sigma^{data}_{u}}_{t,e}\Big), \tag{6}$$

with the priors defined as

$$\begin{aligned} {\sigma^{data}_{u}}_{t,e} &\sim \mathcal{U}[0, 40], \\ \log {\mu^{data}_{\sigma}}_{t,e} &\sim \mathcal{N}(0, 1). \end{aligned} \tag{7}$$

We selected this specification to ensure that variations between countries from the SVD structure are effectively captured (via random effect $u_{a,c,e}$), with differentiation between the levels of educational attainment and age groups. Additionally, we also account for uncertainty arising from external data sources (DHS infant mortality in our case) and

the availability of this information across time and education (as available in the DHS) via random effect $\nu_{t,e}$.

The factor loadings $\beta_{a,t,e,i}$ we assume have a hierarchical prior distribution, as in (Alexander, Zagheni, and Barbieri 2017: Equations 9–12):

$$\beta_{a,t,e,i} \sim \mathcal{N}\Big(\mu^{\beta}_{t,e,i},\ \sigma^{\beta}_{t,e,i}\Big), \tag{8}$$

$$\sigma^{\beta}_{t,e,i} \sim \mathcal{U}[0,40], \tag{9}$$

$$\mu^{\beta}_{t,e,i} \sim \begin{cases} \mathcal{N}(0, \sigma^{\mu}_{e,i}) & t \in \{1,2\}, \quad (10) \\ \mathcal{N}(2 \times \mu^{\beta}_{t-1,e,i} - \mu^{\beta}_{t-2,e,i}, \sigma^{\mu}_{e,i}) & t > 2, \quad (11) \end{cases}$$

$$\sigma^{\mu}_{e,i} \sim \mathcal{U}[0,40], \tag{12}$$

where $2 \times \mu^{\beta}_{t-1,e,i} - \mu^{\beta}_{t-2,e,i}$ denotes a random walk specification of the first differences of the time effects over time. The priors for the variance parameters are weakly informative leveraging on the data to shape the posterior distribution.

Standard deviation $\sigma^{pcs}_{a,e}$ accounts for the potential variation resulting from the selection of the curves employed in the SVD, which are used to derive principal components that are used in the model.[4] These have been derived for two education-specific mortality profiles that we investigate in our case study (Section 3) separately and independently. They are estimated separately by educational attainment to provide the model with maximum flexibility in defining mortality schedules according to the level of education, as if representing two distinct populations within the same country. This standard deviation is also estimated from the data:

$$\begin{aligned} \tau^{pcs}_{a,e} &\sim \Gamma(0.1, 0.1) \\ \sigma^{pcs}_{a,e} &= 1/\sqrt{\tau^{pcs}_{a,e}}. \end{aligned} \tag{13}$$

The full model specification is presented in Appendix A.

[4] Hereinafter, a mortality curve refers to an age-specific mortality profile.

3. Case study and data

3.1 Case study setting

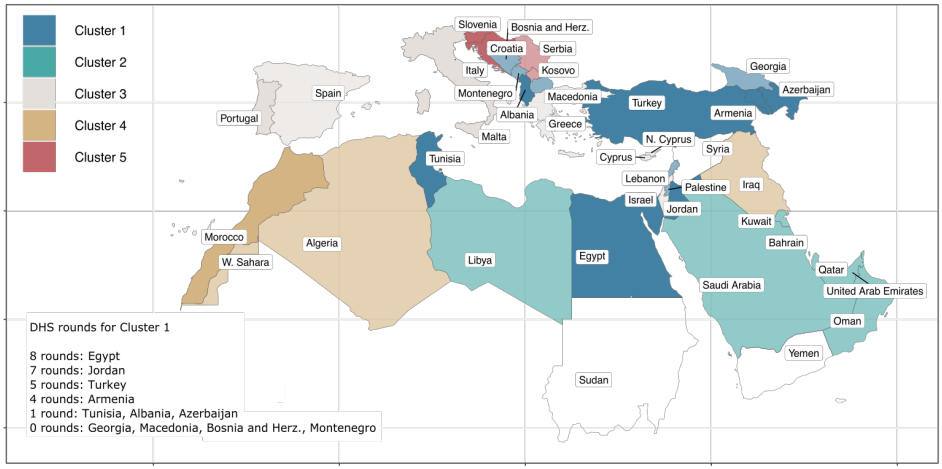
In our case study, we apply the model described in the previous section to a group of countries that have been selected to represent a wide range of geographical locations, socioeconomic levels of development, and data quality and availability. The countries have been chosen as described briefly below and in Appendix B, in a way that ensured efficient borrowing of information across countries and over time.

In order to demonstrate the borrowing of information between Eurostat and DHS, we have selected a macro region comprising Southern Europe, Western Asia, and Northern Africa. By employing a hierarchical clustering algorithm (Nielsen 2016), the countries within this macro region have been arranged into five clusters. Each cluster contains countries that have similarities, as measured through variables such as the socioeconomic status, mortality, and schooling trends. Details of the geographical setting and the clustering procedure are presented in Appendix B. Figure 3 shows the countries included in our case study and the number of DHS waves available for each of them. In the rest of this paper, we focus on the female population for the countries belonging to cluster 1, which are Albania (ALB), Armenia (ARM), Azerbaijan (AZE), Bosnia and Herzegovina (BIH), Egypt (EGY), Georgia (GEO), Jordan (JOR), Lebanon (LBN), Montenegro (MNE), North Macedonia (MKD), the State of Palestine (PSE), Tunisia (TUN), and Turkey (TUR). These countries have a noteworthy range of data availability in DHS, and are distributed across various geographical locations within the group of countries comprising our study region. The same analysis can be replicated for the other clusters.

For the case study, the indices (in the model in Section 1.2) represent the following:

- c: countries (ALB, ARM, AZE, BIH, EGY, GEO, JOR, LBN, MNE, MKD, PSE, TUN, TUR)
- a: five-years age groups from 15–19 until $85+$
- t: years 1980, 1985, 1990, 1995, 2000, 2005, 2010, and 2015
- e: educational attainment "No education or primary" and "More than primary"

Figure 3: Clustering results and information about DHS data availability for cluster 1



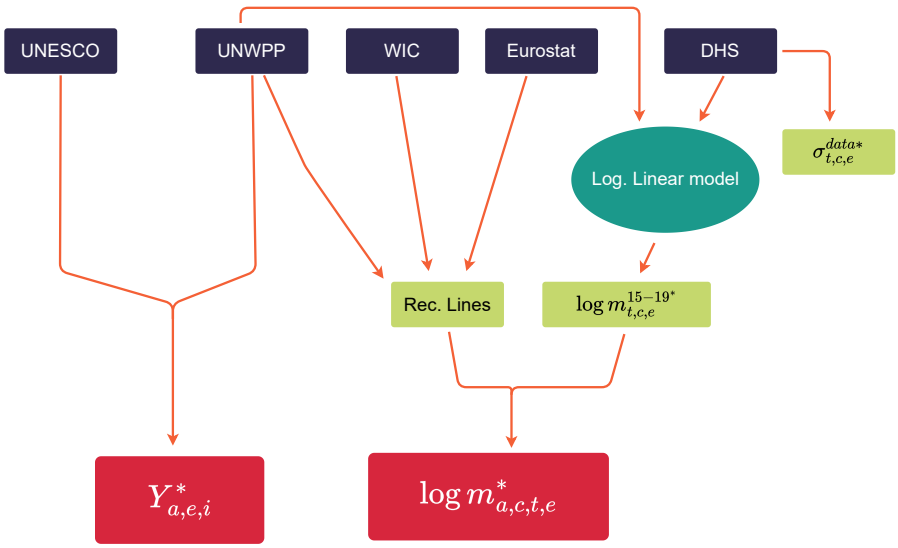
Source: Authors' own calculations and DHS.
Notes: Solid colours represent the availability of DHS rounds or Eurostat data. More information about the DHS data is available in Appendix G.

3.2 Data sources and model inputs

Given the scarcity of mortality data disaggregated by levels of education, such as death counts or mortality rates, we use information from various sources. Considering the systematic nature of our approach and our desire to ensure replicability in other countries, we have used the data sources that are available for different regions of the world. The Bayesian inferential framework facilitates the data integration in our study, while taking into account their quality and the uncertainty generated by their integration.

Two main data inputs for the model are the age- and education-specific principal components $Y^*_{a,e,j}$ and the age-, year-, country-, and education-specific reconstructed log-mortality rates $(\log m^*_{a,c,t,e})$. Figure 4 depicts the data sources used in our case study and a schematic approach to their integration to generate input to the model.

Figure 4: The input construction scheme



Notes: The different components are visualised as follows: purple squares: data sources; green squares: uncertainty estimations and intermediate estimation steps; red squares: estimated inputs; green oval: additional model for input estimation.

3.2.1 Data sources

The data used for the case study include the following:

1. Eurostat Database: life expectancy by age, sex, and level of education for 19 countries between 2007 and 2017 (Eurostat 2023).[5]
2. United Nations World Population Prospects (UN WPP): mortality rates by age, sex, period, and country. These are collected for the case study countries, and are available for five-year intervals from 1980 to 2015 (United Nations 2022).
3. Demographic and Health Surveys (DHS): infant mortality rate by mother's level of education (ICF 2022). A detailed description of the DHS data is presented in Appendix G.
4. Wittgenstein Centre for Demography and Global Human Capital (WIC): population counts by five-year age group, sex, country, five-year period and educational

[5] Bulgaria, Croatia, Czechia, Denmark, Estonia, Finland, Greece, Hungary, Italy, Malta, Norway, Poland, Portugal, Romania, Serbia, Slovakia, Slovenia, Sweden, Turkey.

attainment, and mean years of schooling for 15+ population, sex, country, and period (Wittgenstein Centre for Demography and Global Human Capital 2018).
5. United Nations Educational, Scientific and Cultural Organization (UNESCO): the duration of study cycles in different countries (e.g., for Georgia, six years for primary education and a further six years for secondary education) (UNESCO Institute for Statistics 2023).

3.2.2 Model inputs

Considering the temporal coverage of the DHS waves and the recall period of 10 years before each survey date, we focus on the 1980–2015 period.[6] The variables with an asterisk $*$ in Figures 2 and 4 are employed either as hyperparameters defining the distributions or as data informing them. A detailed explanation of the procedure for the estimation of other necessary quantities – that is, the reconstruction curves and the mortality rates for the 15–19 age group – can be found in Appendix C and Appendix D; generation of inputs can also be generalised and expanded to other periods and countries. In particular, the education-specific mortality rates for the 15–19 age group were estimated using a separate log-linear model. To inform this model, we extracted infant mortality rates by mother's level of education from the DHS STATcompiler, focusing on countries included in our case study. Assuming that these infant mortality differentials by education (not the rates values) approximate those for the 15–19 age group, we used them to derive education-specific (log) mortality rate estimates for the countries analysed.

As introduced in Section 2, superscript pcs is used for quantities that stem from the UN WPP and the UNESCO data (which are employed for the SVD), while $info$ denotes the outputs from the Bayesian log-linear model that is used to estimate the log-mortality rates for the 15–19 age group (the information which we need as a starting point for the reconstruction lines). Superscript $data$ marks the quantities obtained from the DHS database. The superscripts are not directly tied to specific data sources, but rather represent the roles they play within the model and the information they convey, independent of any particular data source.

The inputs are summarised as follows:

1. $\log m^{*}_{a,t,c,e}$: the log-mortality rates resulting from the application of the region-specific reconstruction curves to the estimated starting points $\log m^{15-19^{*}}_{t,c,e}$; for details, see Appendix C. The reconstruction curves are obtained following a procedure developed by Sauerberg (2021). More specifically, $\log m^{*}_{a,t,c,e}$ are based on data stemming from 18 European countries, which we have grouped into four re-

[6] The values provided by DHS are derived from inquiries that solicit information pertaining to both the current year and the preceding decade from the date of survey deployment.

gions.[7] In our case study, we used the reconstruction curves for Central-Southern Europe (see Figure A-5 and Appendix 6), given the geographical location of the countries.[8] These profiles are applied to the log-mortality rates for the 15–19 age group by level of education, which are obtained via Bayesian log-linear modelling (for details see Appendix C). Then, adjusted to match WIC population data and the UN WPP age-specific mortality rates, we obtained log-mortality rates for all 13 countries. These rates are the key inputs for disaggregating mortality schedules by the level of education, and rely on multiple sources: DHS, Eurostat, UN WPP and WIC Data Explorer (see Figure 4). We present details of the construction of $\log m^*_{a,t,c,e}$ in Appendix D.

In Figure 5, we present $\log m^*_{a,t,c,e}$ reported for 1980 for all case study countries. As shown in the figure, our reconstruction method defines sets of mortality profiles that are country- and year-specific. As expected, the mortality profiles referring to the most represented level of education in the population are those most similar to the total profile (for population composition, see Figure A-16 in Appendix H). For example, Tunisia's mortality curve is very close to the curve for the Tunisian population with no or primary education, who make up a large majority of the country's total population. Our model takes the uncertainty of this input construction into account.

2. $Y^*_{a,e,i}$: principal components extracted from the collection of mortality rates referring to the relevant time span and the region. We use the underlying mortality rates from female population life tables from the UN WPP database for the case study countries for the 1980–2015 period. We separate mortality curves into two groups based on information from WIC Data Explorer on mean years of schooling and primary education duration in years from UNESCO database. The first group includes year-country combinations in which the average years of schooling exceed the duration of primary school (country-specific), and the second group includes instances in which the average years of schooling fall below this threshold. Because the information is available for different time intervals, we performed one-year interval interpolations of the values prior to this step, which resulted in datasets that could be combined. The principal components were obtained via SVD of these two collections of education-specific (log-)mortality curves to effectively represent their key characteristics. Essentially, age-specific mortality rates over time can be decomposed into a linear combination of principal components. The approach is conceptually similar to the Lee–Carter approach (Lee and Carter 1992). In our case, the principal components depict how the log-mortality curves develop in a given set of countries over a specified time interval (1980–2015) according to the

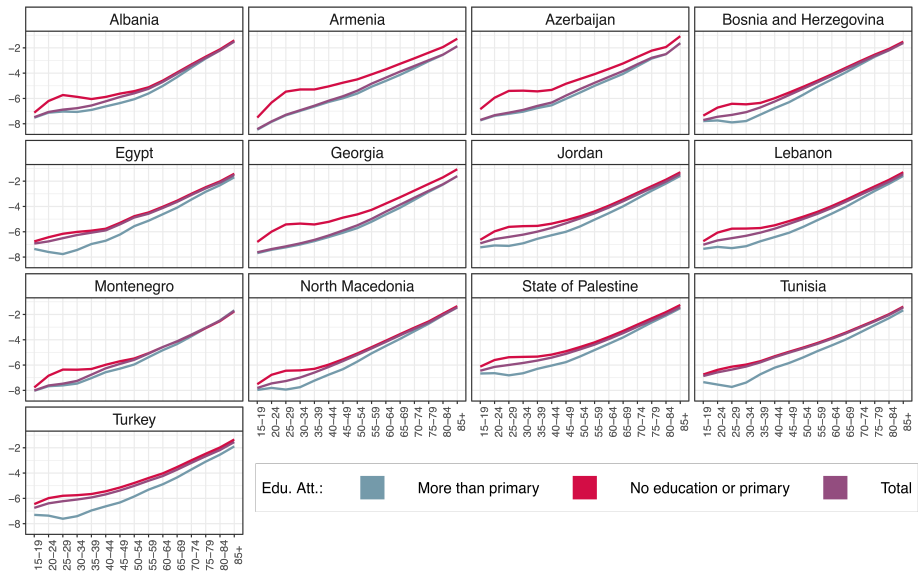
[7] **North**: DNK, EST, FIN, NOR, SWE. **South**: ITA, GRC, PRT, MLT. **Central East**: BGR, HUN, POL, ROU, SVN, SVK. **Central South**: SRB, HRV, TUR.

[8] ALB, ARM, AZE, BIH, EGY, GEO, JOR, LBN, MKD, MNE, PSE, TUN, TUR.

estimated average level of education. Further details of their derivation can be found in Appendix E.

3. $\sigma_{t,c,e}^{data*}$: standard deviations obtained from the confidence intervals published by the DHS concerning estimates of infant mortality by mother's level of education.

Figure 5: Constructed inputs for the education-specific log-mortality rates. Case study countries, 1980, female population



Source: Authors' own calculations based on the DHS, Eurostat, WIC, and the UN WPP data.

3.3 Implementation

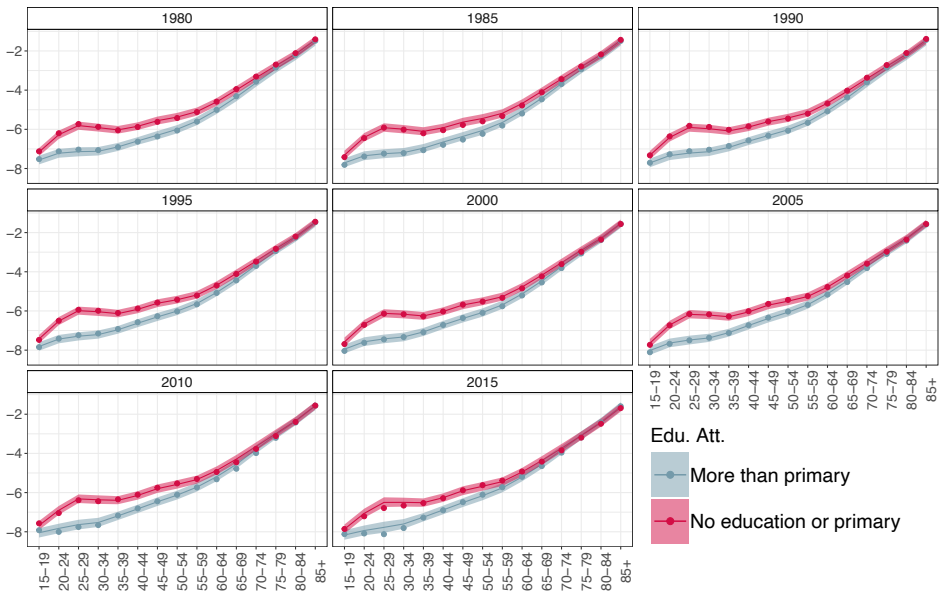
To produce characteristics of the model parameters and estimated mortality rates, we sampled from the posteriors by using Markov Chain Monte Carlo within JAGS software (Plummer 2003), implemented in package `rjags` in the `R` environment. For the convergence checks, we relied on indicators such as the Gelman and Rubin diagnostic (Gelman and Rubin 1992), $\hat{R}$ statistic, and the visual inspection of trace plots. The model was executed on a system equipped with an AMD EPYC 7H12 64-core 2.6 GHz processor, 2 TB of RAM, and running Windows Server 2019 Standard. Following a burn-in period of 10,000 iterations and a total of 75,000 iterations, convergence was achieved in 23.4 minutes with a maximum $\hat{R}$ of 1.086.

4. Results

4.1 Reconstructed mortality rates

Our results are a set of age-, year-, and education-specific reconstructed mortality rates for females in Albania, Armenia, Azerbaijan, Bosnia and Herzegovina, Egypt, Georgia, Jordan, Lebanon, Montenegro, North Macedonia, the State of Palestine, Tunisia, and Turkey, which are the countries in our case study for 1980–2015. In Figure 6, we present posterior medians of log-mortality rates for Albania. These estimates are shaped by the country-level mortality rates during the specified period, and are sensitive to shifts in the level of educational composition of the population (Figure A-17 in the Appendix H).

Figure 6: Log-mortality rates by level of education, age group, and time. Albania, female population



Notes: The solid lines present the estimated medians, while the 95% credible intervals are visualised via shaded areas. The dots represent the inputs to our model.

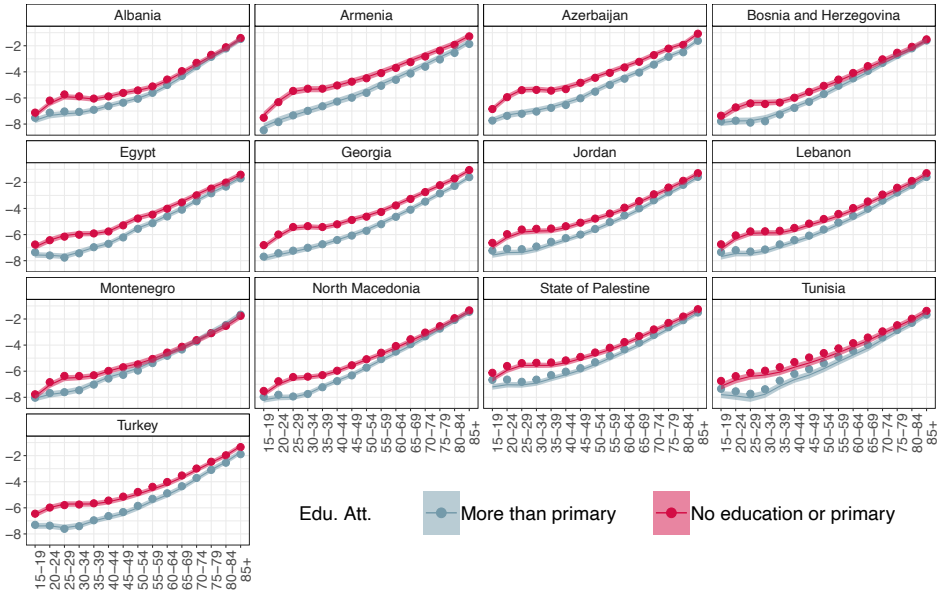
In Figure 7, we present the estimated mortality rates for all case study countries for the year 1990. While the common features derived from the principal components are maintained, the mortality rates are differentiated for each country under consideration. For example, a higher level of education is typically associated with a lower level of

mortality. Specifically, in former Yugoslavian countries such as Albania, Montenegro, North Macedonia, and Bosnia and Herzegovina, relatively small differences in mortality rates are observed across different educational levels. In contrast, Tunisia, Turkey, and Egypt stand out as having large differentials in mortality rates across educational strata.

While investigating the specific causes of these variations at the country level falls outside the scope of our study, we can offer some suggestions regarding potential contributing factors. The countries with narrower differentials may have more equitable access to healthcare and education, resulting in a relatively homogeneous distribution of health outcomes. Conversely, in countries with wider education differentials, disparities in socioeconomic status and access to healthcare may be more pronounced, leading to significant variations in mortality rates. Additionally, cultural and societal factors can play a role, influencing health behaviours and healthcare-seeking patterns across educational levels.

Some valuable insights emerge from the posteriors of the random effects $u_{a,c,e}$ and $\nu_{t,e}$. The first (Figure A-13) captures systematic country-specific variations with respect to mortality curves, deviating from the average principal component structure. The most significant deviations are observed at younger ages, particularly among individuals with lower educational attainment – for example, in Armenia and Azerbaijan, where elevated mortality levels are evident. At the same time, random effects also reflect downward adjustments, such as those observed in the younger age groups in Montenegro, or generally higher values, as in the case of the State of Palestine. The second random effect, $\nu_{t,e}$, captures the interaction between time and education (Figure A-14). These random effects primarily reflect a reduction in mortality over time for the lower-educated population. This pattern is consistent with general trends in mortality decline.

Figure 7: Log-mortality rates by level of education, age group, and time. Year 1990, female population



Notes: The solid lines present the estimated medians, while the 95% credible intervals are visualised via shaded areas. The dots represent the inputs to our model.

4.2 Model performance

Due to the aforementioned lack of data, we do not have a gold standard data against which we can evaluate our estimates. Furthermore, the measurement of goodness-of-fit is complicated by the fact that the inputs to the model are derived from a variety of data sources. Hence, we assess the performance of the model and the robustness of resulting estimates first by applying posterior predictive checks – that is, by generating new data from the model. Then, we test the sensitivity of the results when the inputs are partially removed, both at random and systematically. To externally validate our estimates, we also calculate the total mortality resulting from our estimates, and compare it with the data available in the UN WPP.

4.2.1 Posterior predictive checks for model inputs

First, we assess the performance of our model through posterior predictive distributions (PPD) for the model inputs. New (predicted) inputs are generated from PPDs (analogous to fitted values). For instance, in Figure 8 we present the PPDs for the $\mu_{\sigma\ t,e}^{data}$ as defined in Equation 6 for each year along the inputs $\left(\sigma_{t,c,e}^{data*}\right)$ (i.e., observed data), which are represented by triangles. In the plot, the circular markers represent the median values of the PPD, with the solid lines indicating credible intervals. The triangular markers depict the model inputs inferred from the DHS data. Notably, the credible intervals widen for lower levels of education, mirroring the increased dispersion observed in mortality values within this category. Conversely, higher education levels exhibit less susceptibility to this widening of values. We observe that only 8% of the total of available observations fall outside of the aggregated PPDs for any given year. This suggests that our model is well-calibrated and reproduces the inputs well.

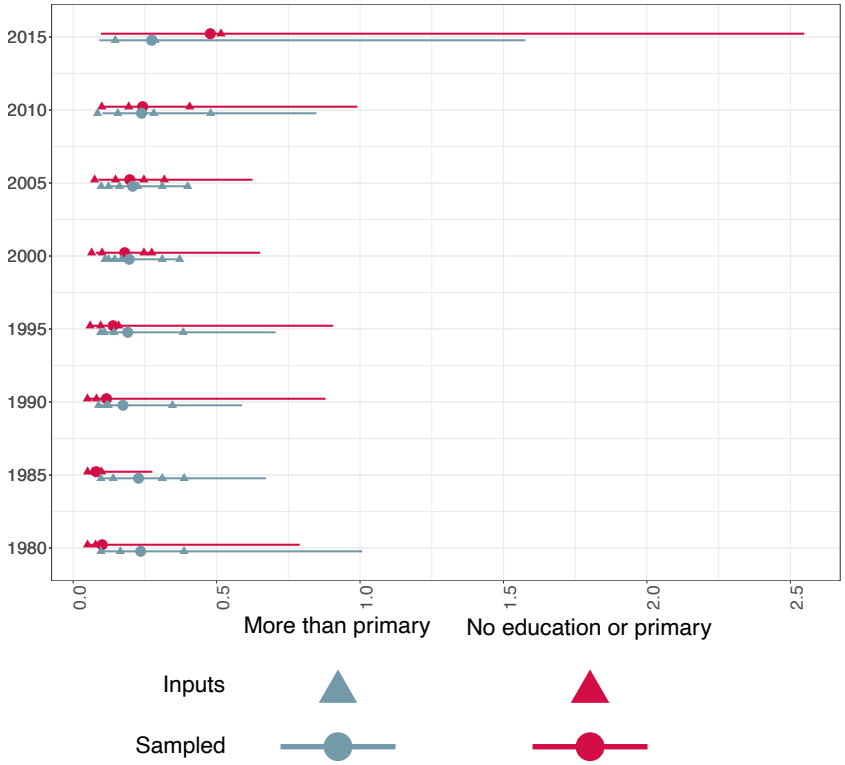
This is further confirmed by the results in Table 1, which presents the percentage of inputs $\log m_{a,t,c,e}^{*}$ falling within the credible interval (CI) for our quantities of interest, $\log m_{a,t,c,e}$, based on the interval's coverage. The results align closely with the expected CI percentages, indicating good calibration.

Table 1: Percentage of inputs falling in the % credible interval (CI)

% CI	% contained in the CI
50	55.8%
80	83.7%
95	93.6%

Note: The percentage is calculated as the percentage of data falling into CI.

Figure 8: Standard deviation associated with the DHS mortality estimates $\mu^{data}_{\sigma\ t,e}$



Notes: The solid lines present the 95% credible intervals. The circles represent the medians of the posterior predictive distributions. The triangles represent the inputs to our model.

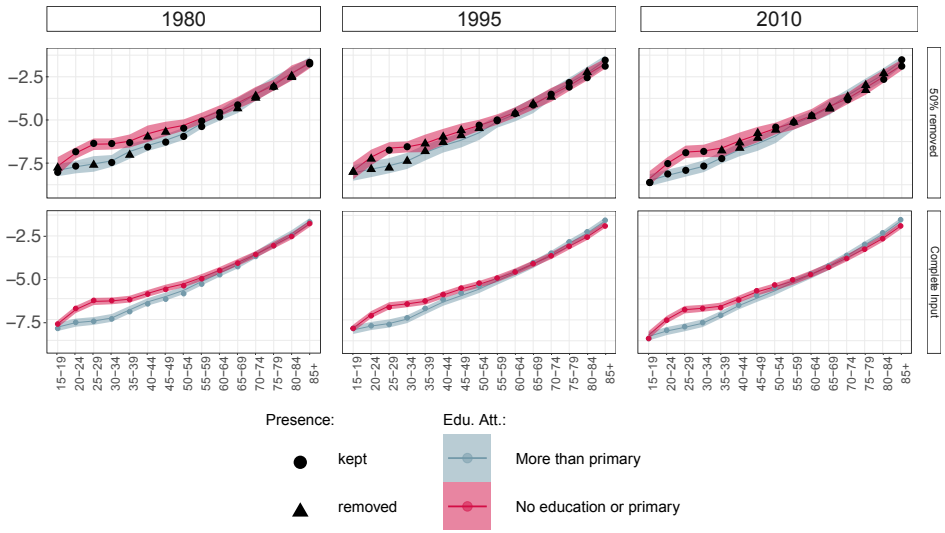
4.2.2 Randomised input reduction

Next, we randomly removed a portion of the data used to inform the age- and education-specific mortality rates. That is, we progressively removed 20%, 50%, and 75% of the inputs obtained from the reconstruction of $\log m^*_{a,t,c,e}$, and then assess how the reduction of inputs affects the model estimates and imputation of missing information.

Figure 9 shows the mortality rates obtained for Montenegro when all the input are used and when the estimates are based on only 50% of the $\log m^*_{a,t,c,e}$ inputs. Although the plot shows less regular predictive intervals than those obtained from the full input, the model still performs reasonably well. The uncertainty increases where information is

removed. However, the fundamental structure and the year- and country-specific profiles continue to be clearly discernible and distinct, organised in accordance with educational levels.

Figure 9: Log-mortality rates by level of education, age group, and time. Montenegro, female population



Notes: Estimates based on 50% of the $\log m^*_{a,t,c,e}$ inputs removed. The solid lines present the estimated medians, while the 95% credible intervals are visualised via shaded areas. The dots represent the inputs to our model, the triangles represent the inputs which were removed.

Generally, reducing inputs at random does not present major systemic problems for the model. The differences between lower and higher levels of education remain unchanged. In Table 2, we present the percentage of inputs contained in the 75% credible interval (i.e., coverage) according to the percentage of reduced inputs for all estimates. We observe that the coverage for the model with full inputs is around 72%, which suggests that our model fits the inputs correctly as already reported in Table 1. With a random reduction in inputs, the width of the CIs increases, as does their coverage. This occurs for both retained and removed inputs, suggesting that the estimated mortality profiles remain stable in terms of general shape and characterization of the reconstructed curves. However, uncertainty expands considerably when input data reductions are substantial, such as with a removal of three-quarters of observations.

Table 2: Assessment of model performance: percentage of inputs falling in the 75% credible interval (CI)

% reduced inputs	% contained in the CI (removed)	% contained in the CI (kept)
0% (full)	0	79.7
20%	73.6	84.2
50%	83.9	89.7
75%	84.7	93.7

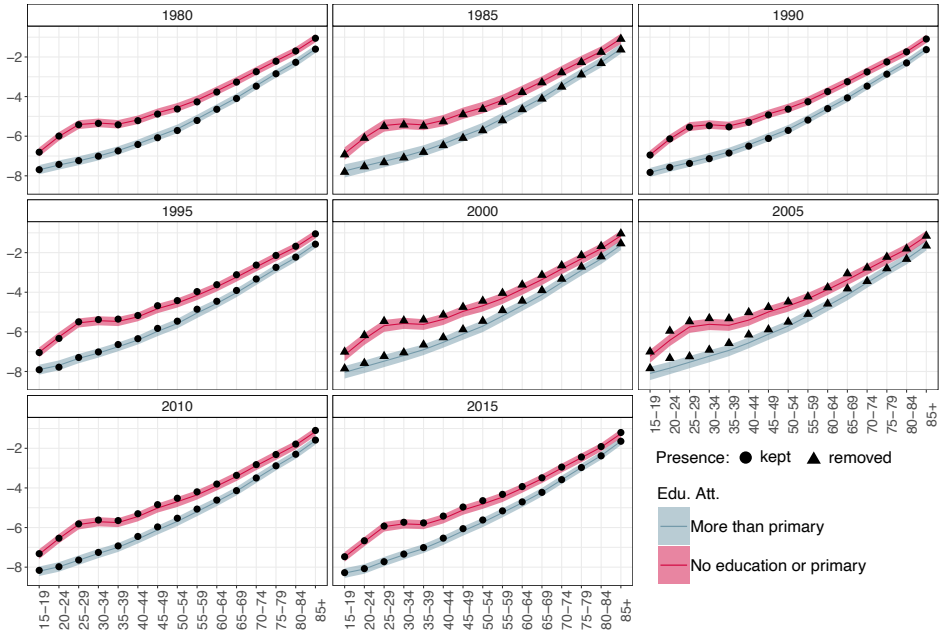
Note: The performance is calculated as the percentage of data falling into the 75% CI according to the amount of inputs employed.

4.2.3 Systematic input reduction

In addition to randomly reducing the inputs $\log m^*_{a,t,c,e}$, we also test the sensitivity of the model to the removal of inputs in a systematic way, for instance, for a given country for specific periods. This assessment evaluates the efficacy of our model in performing geographical pooling and temporal smoothing. Additionally, it allows us to assess the effectiveness of the model in reconstructing mortality even in the near absence of information about education-specific mortality for a given country or year.

In Figure 10, we present the results for Georgia from a model with the input data for the years 1985, 2000, and 2005 removed for Azerbaijan, Georgia, North Macedonia, and Tunisia. Even the total absence of information for a designated country does not lead to modelling failures (see also Appendix A-10). The new estimates are characterised by increased uncertainty for the years in which data are removed, and seem to rely to a greater extent on the mortality profiles derived from other countries. This observation indicates that the model tends to utilise information from different countries to a greater extent than it does from different time periods.

Figure 10: Log-mortality rates by level of education, age group, and time. Georgia, female population



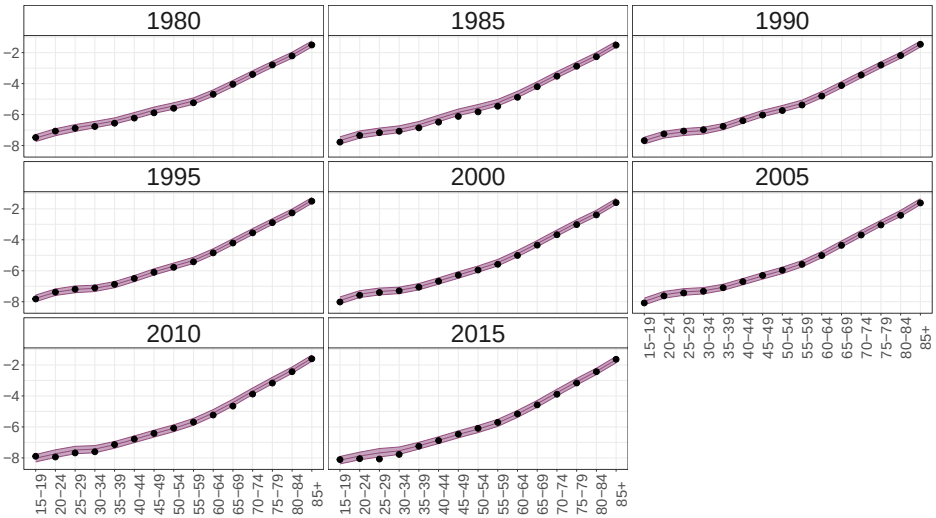
Notes: The solid lines present the estimated medians, while the 95% credible intervals are visualised via shaded areas. The dots represent the inputs to our model, the triangles represent the inputs which were removed. Inputs for years 1985, 2000, and 2005 were removed for Azerbaijan, Georgia, North Macedonia, and Tunisia.

4.2.4 Coherence with total mortality

The last check we carried out addresses the coherency of the results with the (only) available data: namely, the overall mortality rates (i.e., not education-specific). These comparisons assess the consistency of our estimates with total mortality rates after taking into account the uncertainty of our estimates. To do this, we calculate the total mortality rates by sampling from the posterior distributions of our education-specific estimates and weighing the components according to the available population composition by education. Then, we compared resulting total mortality with those published by the UN WPP. This is an indirect approach used to externally validate our results. An example of this approach is shown in Figure 11, in which the two education-specific mortality rates for Albania are summed up and compared with the total mortality rates. We observe that our estimates match the UN WPP estimates well, with slight overestimation for the age

groups older than 40–44, especially in the 1980s. Similar results are obtained for other countries and years.

Figure 11: Log-mortality rates by age group and time. Albania, female population. Comparison between the weighted sum of the estimates and the total log-mortality rates from UN WPP



Note: The solid lines present the estimated medians, while the 95% credible intervals are visualised via shaded areas. The dots represent the UN WPP data.

5. Conclusions

Our work makes a methodological contribution by proposing a modelling framework that includes a Bayesian hierarchical model and a mechanism for constructing its inputs. It fills an important research gap in the systematic study of mortality differentials by level of education, and of the role of education in determining demographic change in particular countries or regions. Our case study application can be easily adapted and extended to other countries and periods, and can be used to predict mortality differentials by other characteristics, such as socioeconomic status. The model exploits available information on adult and child mortality, and on their relationships with the level of education. Given the widespread availability of the DHS and the UN WPP data, it would be possible to generalise information on education differentials available from Eurostat and other sources to reconstruct estimates of mortality by education for all countries. An additional po-

tential extension of our work involves incorporating a broader range of educational categories. The main challenge in this endeavour would be identifying a set of more refined, education-specific mortality curves, along with ensuring the availability of a sufficient number of mortality curves that align with realistic years of education.

In our case study, the only information available for several countries (like Albania, Armenia, Azerbaijan, Egypt, Jordan, and Tunisia) was, to the best of our knowledge, on the link between the mother's level of education and infant mortality obtained from the DHS at irregular intervals. For other countries (like Bosnia and Herzegovina, Georgia, Lebanon, North Macedonia, Montenegro, and the State of Palestine), no information on the link between mortality and level of education was available, and it was imputed within the model by borrowing information from other countries. Our results thus also fill a data gap.

The results we presented reaffirm a well-documented relationship between education and mortality which is consistent with the established literature. Specifically, our analysis underscores that higher educational attainment is closely linked to reduced mortality rates. Notably, the differentials in education-specific mortality appears to be most pronounced within the 20–24 to 45–49 age groups, as the patterns in the graphical representations clearly show. Several different mechanisms might contribute to this phenomenon. First and foremost, as outlined in Karlsen et al. (2011), women with higher education tend to enjoy improved access to essential health information and healthcare services. This advantage facilitates the early detection and more effective management of health issues, particularly during the childbearing years. Moreover, higher educational attainment is associated with the adoption of healthier lifestyles, like lower alcohol consumption (Murakami and Hashimoto 2019) or smoking rates (Assari and Mistry 2018), and more informed health-related decision-making, which can influence overall well-being (Luy et al. 2019). Additionally, education is often correlated with enhanced socioeconomic conditions, which can give women the resources necessary to access better healthcare, improved nutrition, and safer living environments (Fard et al. 2021). Finally, higher education fosters greater awareness of health risks and preventive measures.

Remarkably, the level of education seems to have less impact on mortality among older age groups, for whom the influence of behavioural risk factors is less pronounced (Cutler et al. 2011; Herd 2006). The diminishing educational differentials in mortality observed among older age groups can be attributed to several factors. These include survival bias, as individuals who reach older ages may possess certain advantages in terms of health and healthcare access. Additionally, more equitable access to healthcare services among older adults, changing cohort effects, cumulative life exposures, and the increasing influence of age-related factors such as chronic diseases and genetics all contribute to the reduction of educational disparities in mortality at older ages.

The main limitation of the proposed framework is related to the validation of the estimates, as data on mortality by educational attainment are sparse and are usually limited

to developed countries. Therefore, we relied on internal model validation through posterior predictive checks and sensitivity analysis. We also analysed how well the model predicts total mortality rates (not broken down by education) that are available in the UN WPP. A second limitation is that the proposed model relies on having data available for certain countries that span most of the period of interest. However, it is important to note that the model generates significantly higher levels of uncertainty when data for specific years or periods are missing. Third, the model is compelled to utilise information derived from the European context due to the lack of relevant information in the corresponding geographical region. Although this information is adjusted to the total mortality rates for the selected countries, it originates from a socioeconomic context that differs from that of the studied population. Overcoming these gaps in the data would significantly enhance the potential for the widespread application of our technique. The fourth limitation of this study lies in the model's extensive parameterization, partly driven by the numerous and complex pre-processing and estimation steps required for the analysis. Furthermore, the model and some pre-processing steps rely on external quantities, such as population sizes by educational attainment from WIC, total mortality rates from the UN WPP, and standard deviations from the DHS – all derived from other modelling efforts. Although we account for the uncertainty introduced by these data sources in our model, a comprehensive assessment of model sensitivity to these inputs, as well as the inclusion of uncertainty from these external modelling processes, is not covered in this study.

6. Acknowledgements

This work was carried out partially within the BayesEdu project at the Vienna Institute of Demography. Authors gratefully acknowledge funding received from the 'Innovation Fund Research, Science and Society' established by the Austrian Academy of Sciences (ÖAW).

References

Alexander, M. and Alkema, L. (2018). Global estimation of neonatal mortality using a Bayesian hierarchical splines regression model. *Demographic Research* 38(15): 335–372. doi:10.4054/DemRes.2018.38.15.

Alexander, M., Zagheni, E., and Barbieri, M. (2017). A flexible Bayesian model for estimating subnational mortality. *Demography* 54(6): 2025–2041. doi:10.1007/s13524-017-0618-7.

Alkema, L. and New, J.R. (2014). Global estimation of child mortality using a Bayesian b-spline bias-reduction model. *The Annals of Applied Statistics* 8(4): 2122–2149. doi:10.1214/14-aoas768.

Alkema, L., Raftery, A.E., Gerland, P., Clark, S.J., and Pelletier, F. (2012). Estimating trends in the total fertility rate with uncertainty using imperfect data: Examples from West Africa. *Demographic Research* 26(15): 331–362. doi:10.4054/DemRes.2012.26.15.

Aparicio-Castro, A., Wiśniowski, A., and Rowe, F. (2023). A Bayesian approach to estimate annual bilateral migration flows for South America using census data. *Journal of the Royal Statistical Society Series A: Statistics in Society* 187(2): 410–435. doi:10.1093/jrsssa/qnad127.

Assari, S. and Mistry, R. (2018). Educational attainment and smoking status in a national sample of American adults: Evidence for the Blacks' Diminished Return. *International Journal of Environmental Research and Public Health* 15(4): 763. doi:10.3390/ijerph15040763.

Baker, D., León, J., Greenaway, E., Collins, J., and Movit, M. (2011). The education effect on population health: A reassessment. *Population and Development Review* 37(2): 307–332. doi:10.1111/j.1728-4457.2011.00412.x.

Bora, J.K., Lutz, W., and Raushan, R. (2018). Contribution of education to infant and under-five mortality disparities among caste groups in India. Vienna: Austrian Academy of Sciences (ÖAW), Vienna Institute of Demography (VID) (VID working paper 2021/09/14). doi:10.1553/0x003ccd42.

Byhoff, E., Hamati, M.C., Power, R., Burgard, S.A., and Chopra, V. (2017). Increasing educational attainment and mortality reduction: A systematic review and taxonomy. *BMC Public Health* 17(719). doi:10.1186/s12889-017-4754-1.

Caldwell, J. and McDonald, P. (1982). Influence of maternal education on infant and child mortality: Levels and causes. *Health Policy and Education* 2(3–4): 251–267. doi:10.1016/0165-2281(82)90012-1.

Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software* 76(1): 1–32. doi:10.18637/jss.v076.i01.

Cutler, D., Lange, F., Meara, E., Richards-Shubik, S., and Ruhm, C. (2011). Rising educational gradients in mortality: The role of behavioral risk factors. *Journal of Health Economics* 30(6): 1174–1187. doi:10.1016/j.jhealeco.2011.06.009.

Dubow, E., Boxer, P., and Huesmann, L. (2009). Long-term effects of parents' education on children's educational and occupational success: Mediation by family interactions, child aggression, and teenage aspirations. *Merrill-Palmer Quarterly* 55(3): 224–249. doi:10.1353/mpq.0.0030.

Durowaa-Boateng, A., Yildiz, D., and Goujon, A. (2023). A Bayesian model for the reconstruction of education- and age-specific fertility rates: An application to African and Latin American countries. *Demographic Research* 49(31): 809–848. doi:10.4054/DemRes.2023.49.31.

Eccles, J. (2005). Influences of parents' education on their children's educational attainments: The role of parent and child perceptions. *London Review of Education* 3(3): 191–204. doi:10.1080/14748460500372309.

Eurostat (2023). Life expectancy by age, sex and educational attainment level. Luxembourg: Eurostat. doi:10.2908/demo_mlexpecedu.

Eurostat (2025). Deaths by age, sex and educational attainment level. Luxembourg: Eurostat. doi:10.2908/demo_maeduc.

Fard, N.A., Morales, G.D.F., Mejova, Y., and Schifanella, R. (2021). On the interplay between educational attainment and nutrition: A spatially-aware perspective. *EPJ Data Science* 10(18). doi:10.1140/epjds/s13688-021-00273-y.

Gakidou, E., Cowling, K., Lozano, R., and Murray, C. (2010). Increased educational attainment and its effect on child mortality in 175 countries between 1970 and 2009: A systematic analysis. *The Lancet* 376(9745): 959–974. doi:10.1016/S0140-6736(10)61257-3.

Gelman, A. and Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science* 7(4): 457–472. doi:10.1214/ss/1177011136.

Goujon, A., K.C., S., Speringer, M., Barakat, B., Potancoková, M., Eder, J., Striessnig, E., Bauer, R., and Lutz, W. (2016). A harmonized dataset on global educational attainment between 1970 and 2060 – an analytical window into recent trends and future prospects in human capital development. *Journal of Demographic Economics* 82(3): 315–363. doi:10.1017/dem.2016.10.

Graham, W., Brass, W., and Snow, R.W. (1989). Estimating maternal mortality: The sisterhood method. *Studies in Family Planning* 20(3): 125–135. doi:10.2307/1966567.

Green, T. and Hamilton, T. (2019). Maternal educational attainment and infant mortality in the United States: Does the gradient vary by race/ethnicity and nativity? *Demographic Research* 41(25): 713–752. doi:10.4054/DemRes.2019.41.25.

Herd, P. (2006). Do functional health inequalities decrease in old age? Educational status and functional decline among the 1931–1941 birth cohort. *Research on Aging* 28(3): 375–392. doi:10.1177/0164027505285845.

ICF (2022). The DHS program STATcompiler. Funded by USAID. Rockville, MD: ICF International. https://www.statcompiler.com/en/.

Karlsen, S., Say, L., Souza, J., Hogue, C., Calles, D., Gülmezoglu, A., and Raine, R. (2011). The relationship between maternal education and mortality among women giving birth in health care institutions: Analysis of the cross sectional WHO global survey on maternal and perinatal health. *BMC Public Health* 11(606). doi:10.1186/1471-2458-11-606.

Kiross, G., Chojenta, C., Barker, D., Tiruye, T., and Loxton, D. (2019). The effect of maternal education on infant mortality in Ethiopia: A systematic review and meta-analysis. *PLOS One* 14(7): e0220076. doi:10.1371/journal.pone.0220076.

Krueger, P.M., Tran, M.K., Hummer, R.A., and Chang, V.W. (2015). Mortality attributable to low levels of education in the United States. *PLOS One* 10(7): e0131809. doi:10.1371/journal.pone.0131809.

Lee, R. (1974). Estimating series of vital rates and age structures from baptisms and burials: A new technique, with applications to pre-industrial England. *Population Studies* 28(3): 495–512. doi:10.1080/00324728.1974.10405195.

Lee, R.D. (1985). Inverse projection and back projection: A critical appraisal, and comparative results for England, 1539 to 1871. *Population Studies* 39(2): 233–248. doi:10.1080/0032472031000141466.

Lee, R.D. and Carter, L.R. (1992). Modeling and forecasting U.S. mortality. *Journal of the American Statistical Association* 87(419): 659–671. doi:10.2307/2290201.

Li, Q. and Keith, L. (2010). The differential association between education and infant modality by nativity status of Chinese American mothers: A life-course perspective. *American Journal of Public Health* 101(5): 899–908. doi:10.2105/AJPH.2009.186916.

Ludeke, S.G., Gensowski, M., Junge, S.Y., Kirkpatrick, R.M., John, O.P., and Andersen, S.C. (2021). Does parental education influence child educational outcomes? A

developmental analysis in a full-population sample and adoptee design. *Journal of Personality and Social Psychology* 120(4): 1074–1090. doi:10.1037/pspp0000314.

Lutz, W., Goujon, A., KC, S., and Sanderson, W. (2007). Reconstruction of population by age, sex and level of educational attainment of 120 countries for 1970–2000. *Vienna Yearbook of Population Research* 5: 193–235. doi:10.1553/populationyearbook2007s193.

Lutz, W., Goujon, A., KC, S., Stonawski, M., and Stilianakis, N. (eds.) (2018). *Demographic and Human Capital Scenarios for the 21st Century: 2018 assessment for 201 countries.* Luxembourg (Luxembourg): Publications Office of the European Union. doi:10.2760/835878.

Luy, M., Zannella, M., Wegner-Siegmundt, C., Minagawa, Y., Lutz, W., and Caselli, G. (2019). The impact of increasing education levels on rising life expectancy: A decomposition analysis for Italy, Denmark, and the USA. *Genus* 75(11). doi:10.1186/s41118-019-0055-0.

Mandal, S., Paul, P., and Chouhan, P. (2019). Impact of maternal education on under-five mortality of children in India: Insights from the National Family Health Survey, 2005–2006 and 2015–2016. *Death Studies* 45(10): 788–794. doi:10.1080/07481187.2019.1692970.

Montez, J.K., Hummer, R.A., and Hayward, M.D. (2012). Educational attainment and adult mortality in the United States: A systematic analysis of functional form. *Demography* 49(1): 315–336. doi:10.1007/s13524-011-0082-8.

Murakami, K. and Hashimoto, H. (2019). Associations of education and income with heavy drinking and problem drinking among men: Evidence from a population-based study in Japan. *BMC Public Health* 19(420). doi:10.1186/s12889-019-6790-5.

Murtagh, F. and Legendre, P. (2011). Ward's hierarchical clustering method: Clustering criterion and agglomerative algorithm (arxiv preprint). doi:10.48550/arXiv.1111.6285.

Nielsen, F. (2016). *Introduction to HPC with MPI for data science*. Cham: Springer. doi:10.1007/978-3-319-21903-5.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In: *Proceedings of the Third International Workshop on Distributed Statistical Computing (DSC 2003) March 20–23*. Vienna: Technische Universität Wien: 1–9.

Pradhan, E., Suzuki, E.M., Martinez, S., Schäferhoff, M., and Jamison, D.T. (2017). The effects of education quantity and quality on child and adult mortality: Their magnitude and their value. In: Bundy, D.A.P., Silva, N.d., Horton, S., Jamison, D.T., and Patton,

G.C. (eds.). *Child and adolescent: Health and development*. Washington D.C.: World Bank Group: 423–440. doi:10.1596/978-1-4648-0423-6_ch30.

Raymer, J., Wiśniowski, A., Forster, J.J., Smith, P.W.F., and Bijak, J. (2013). Integrated modeling of European migration. *Journal of the American Statistical Association* 108(503): 801–819. doi:10.1080/01621459.2013.789435.

Sasson, I. and Hayward, M.D. (2019). Association between educational attainment and causes of death among white and black US adults, 2010–2017. *JAMA* 322(8): 756–763. doi:10.1001/jama.2019.11330.

Sauerberg, M. (2021). The impact of population's educational composition on Healthy Life Years: An empirical illustration of 16 European countries. *SSM – Population Health* 15: 100857. doi:10.1016/j.ssmph.2021.100857.

Schmertmann, C.P. and Hauer, M.E. (2019). Bayesian estimation of total fertility from a population's age–sex structure. *Statistical Modelling* 19(3): 225–247. doi:10.1177/1471082X18801450.

Speringer, M., Goujon, A., KC, S., Potančoková, M., Reiter, C., Jurasszovich, S., and Eder, J. (2021). Global reconstruction of educational attainment, 1950 to 2015: Methodology and assessment – annex tables and data documentation. Vienna: Vienna Institute of Demography (VID Working Paper 02/2019). doi:10.1553/0x003cb434.

Stan Development Team (2018a). RStan: The R interface to Stan. R package version 2.17.3. http://mc-stan.org/.

Stan Development Team (2018b). The Stan Core Library. Version 2.18.0. http://mc-stan.org/.

UNESCO Institute for Statistics (2023). SDG global and thematic indicators. Montreal: UNESCO Institute for Statistics (UIS). http://data.uis.unesco.org.

United Nations (1983). *Manual X. Indirect techniques for demographic estimation*. New York: United Nations. Department of International Economic and Social Affairs. Population Division.

United Nations (2022). World population prospects 2022. New York: United Nations. https://www.un-ilibrary.org/content/books/9789210014380.

Wheldon, M.C., Raftery, A.E., Clark, S.J., and Gerland, P. (2013). Reconstructing past populations with uncertainty from fragmentary data. *Journal of the American Statistical Association* 108(501): 96–110. doi:10.1080/01621459.2012.737729.

Wheldon, M.C., Raftery, A.E., Clark, S.J., and Gerland, P. (2015). Bayesian reconstruction of two-sex populations by age: Estimating sex ratios at birth and sex ratios

of mortality. *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 178(4): 977–1007. http://www.jstor.org/stable/43965780.

Wheldon, M.C., Raftery, A.E., Clark, S.J., and Gerland, P. (2016). Bayesian population reconstruction of female populations for less developed and more developed countries. *Population Studies* 70(1): 21–37. doi:10.1080/00324728.2016.1139164.

Willekens, F., Massey, D., Raymer, J., and Beauchemin, C. (2016). International migration under the microscope. *Science* 352(6288): 897–899. doi:10.1126/science.aaf6545.

Wiśniowski, A. (2017). Combining labour force survey data to estimate migration flows: The case of migration from Poland to the UK. *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 180(1): 185–202. doi:10.1111/rssa.12189.

Wiśniowski, A. (2021). Migration forecasting using new technology and methods. In: McAuliffe, M. (ed.). *Research handbook on international migration and digital technology*. Cheltenham (UK): Edward Elgar Publishing: 376–392.

Wittgenstein Centre for Demography and Global Human Capital (2018). Wittgenstein Centre Data Explorer version 2.0. [online]. http://dataexplorer.wittgensteincentre.org/wcde-v2/.

Wrigley, E.A. and Schofield, R.S. (1983). English population history from family reconstitution: Summary results 1600–1799. *Population Studies* 37(2): 157–184. doi:10.2307/2173980.

Yildiz, D., Wiśniowski, A., Abel, G.J., Weber, I., Zagheni, E., Gendronneau, C., and Hoorens, S. (2024). Integrating traditional and social media data to predict bilateral migrant stocks in the European Union. *International Migration Review* 59(1): 90–118. doi:10.1177/01979183241249969.

Yildiz, D., Wiśniowski, A., Brzozowska, Z., and Durowaa-Boateng, A. (2023). A flexible model to reconstruct education-specific fertility rates: Sub-saharan Africa case study. Vienna: Vienna Institute of Demography (VID Working Paper 02/2023). doi:10.1553/0x003e65e0.

Appendix A: Full model specification

$$\log m^{*}_{a,t,c,e} \sim \mathcal{N}\Big(\log m_{a,t,c,e},\ \sigma^{info}_{t,c,e}\Big), \tag{14}$$

$$\log m_{a,t,c,e} \sim \mathcal{N}\Big(\sum_{i=1}^{3} \beta_{a,t,e,i} \cdot Y^{*}_{a,e,i} + u_{a,c,e} + \nu_{t,e},\ \sigma^{pcs}_{a,e}\Big). \tag{15}$$

$$\beta_{a,t,e,i} \sim \mathcal{N}\Big(\mu^{\beta}_{t,e,i},\ \sigma^{\beta}_{t,e,i}\Big), \tag{16}$$

$$\begin{cases} \mu^{\beta}_{t,e,i} \sim \mathcal{N}(0, \sigma^{\mu}_{e,i}) & t \in \{1,2\} \quad (17) \\ \mu^{\beta}_{t,e,i} \sim \mathcal{N}(2 \times \mu^{\beta}_{t-1,e,i} - \mu^{\beta}_{t-2,e,i}, \sigma^{\mu}_{e,i}) & t > 2, \quad (18) \end{cases}$$

$$u_{a,c,e} \sim \mathcal{N}\Big(0,\ \sigma^{u}_{a,e}\Big) \tag{19}$$

$$\nu_{t,e} \sim \mathcal{N}\Big(0,\ {\mu^{data}_{\sigma}}_{t,e}\Big) \tag{20}$$

$$\begin{aligned} \log \sigma^{data*}_{t,c,e} &\sim \mathcal{N}\Big(\log {\mu^{data}_{\sigma}}_{t,e}, {\sigma^{data}_{u}}_{t,e}\Big) \\ {\mu^{data}_{\sigma}}_{t,e} &= \exp\Big(\log {\mu^{data}_{\sigma}}_{t,e}\Big) \end{aligned} \tag{21}$$

Priors

$$\begin{aligned} \tau^{info}_{t,c,e} &\sim \Gamma(0.1, 0.1) \\ \sigma^{info}_{t,c,e} &= 1/\sqrt{\tau^{info}_{t,c,e}} \end{aligned} \tag{22}$$

$$\begin{aligned} \tau^{pcs}_{a,e} &\sim \Gamma(0.1, 0.1) \\ \sigma^{pcs}_{a,e} &= 1/\sqrt{\tau^{pcs}_{a,e}} \end{aligned} \tag{23}$$

$$\sigma_{u}^{data}{}_{t,e} \sim \mathcal{U}[0, 40] \tag{24}$$

$$\log \mu_{\sigma}^{data}{}_{t,e} \sim \mathcal{N}(0, 1) \tag{25}$$

$$\sigma_{t,e,i}^{\beta} \sim \mathcal{U}[0, 40], \tag{26}$$

$$\sigma_{e,i}^{\mu} \sim \mathcal{U}[0, 40], \tag{27}$$

$$\sigma_{a,e}^{u} \sim \mathcal{U}[0, 40]. \tag{28}$$

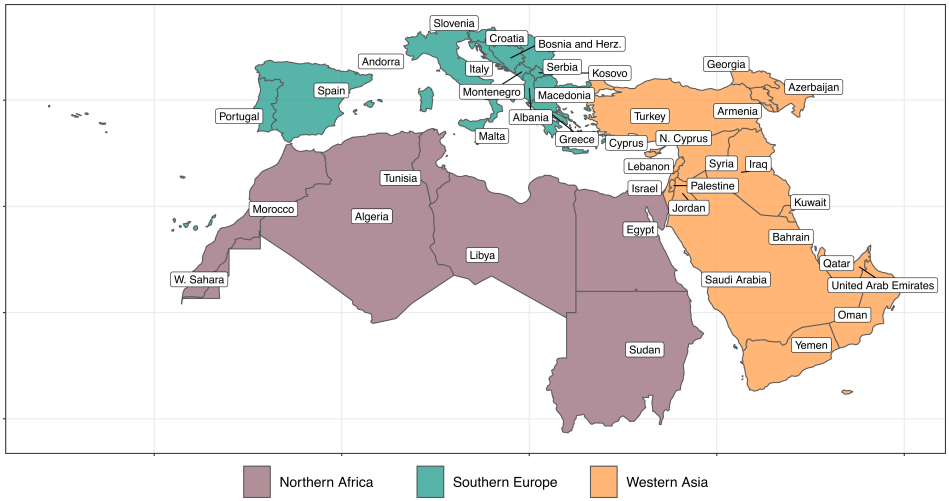
Appendix B: Case study setting and countries clustering

Geographical setting

The selection of the geographical setting for our case study was heavily influenced by the availability of data. In accordance with our adopted strategy and the need to utilise data from multiple sources, we chose a region consisting of countries with varying levels of data availability. This was done in order to evaluate the generalisability of our methodology, and to provide a means of validating our results against established figures.

Initially, we narrowed down the potential countries to a macro region comprising Southern Europe, Western Asia, and Northern Africa (see Figure A-1). This selection provided us with a group of countries that are vastly different in terms of their socioeconomic development and average levels of education. Furthermore, variables directly related to mortality profiles, such as life expectancy or total mortality by age and sex, also vary greatly across these regions.

Figure A-1: The macro region



Significant differences between the countries make it challenging for the model to accurately distinguish the data.

Many of the countries in this area have also been included in DHS, which can be used to obtain indirect information about the differentials in mortality by level of education.

Clustering of countries

One key feature of the proposed model is its capacity to share information in order to optimise the use of limited data. This means relying first on borrowing information across countries for which a similar mortality development is most likely, and second on sharing a mortality structure through principal components analysis. For this purpose, a primary grouping based on a hierarchical clustering algorithm was performed to identify clusters of countries with similar education-specific mortality schedules. While this initial clustering is not a structural necessity, it is a significant alternative to relying on a simple geographical categorisation, given the profound socioeconomic differences between the subregions.

To cluster the countries, we selected variables representing mortality, education, and socioeconomic status macro characteristics for 2010 (which provided a good trade-off

between the available variables and the historical focus of our study). We had to drop some countries because of the excessive amount of missing data.[9]

Mortality variables:

1. Cause of death, by injury (% of total)
2. Cause of death, by noncommunicable diseases (% of total)
3. Lifetime risk of maternal death (%)
4. Life expectancy at birth, total (years)
5. Mortality rate, neonatal (per 1,000 live births)
6. Survival rate from age 15–60

Education and socioeconomic variables:

1. Access to electricity (% of population)
2. Adjusted net enrolment rate, primary (% of primary-school-age children)
3. Adjusted net national income per capita (annual % growth)
4. Adolescents out of school (% of lower secondary school age)
5. Bank capital to assets ratio (%)
6. Educational attainment, at least bachelor's or equivalent, population 25+, total (%) (cumulative)
7. Female share of employment in senior and middle management (%)
8. Literacy rate, adult total (% of people ages 15 and above)
9. Literacy rate, youth (ages 15–24), gender parity index (GPI)
10. Progression to secondary school (%)
11. Barro–Lee: Average years of primary schooling, age 15–19, total
12. Barro–Lee: Average years of primary schooling, age 50–54, total
13. Barro–Lee: Average years of secondary schooling, age 30–34, total
14. Government expenditure on education as % of GDP (%)
15. Human Capital Index (HCI) (scale 0–1)

Considering these variables, the Ward's hierarchical clustering (Murtagh and Legendre 2011) resulted in the clusters shown in Table A-1 and Figure 3. We selected cluster 1 for the case study because it is the largest group, has a substantial number of available DHS datasets, and is geographically and socioeconomically closer to the countries from which the mortality schedules were derived.

[9] Syria, Gibraltar, San Marino, Andorra, Sudan, and Yemen.

Table A-1: Clusters composition

Cluster	Countries
1	ALB, ARM, AZE, BIH, EGY, GEO, JOR, LBN, MKD, MNE, PSE, TUN, TUR
2	ARE, BHR, KWT, LBY, OMN, QAT, SAU
3	CYP, ESP, GRC, ISR, ITA, MLT, PRT
4	DZA, IRQ, MAR
5	HRV, SRB , SVN

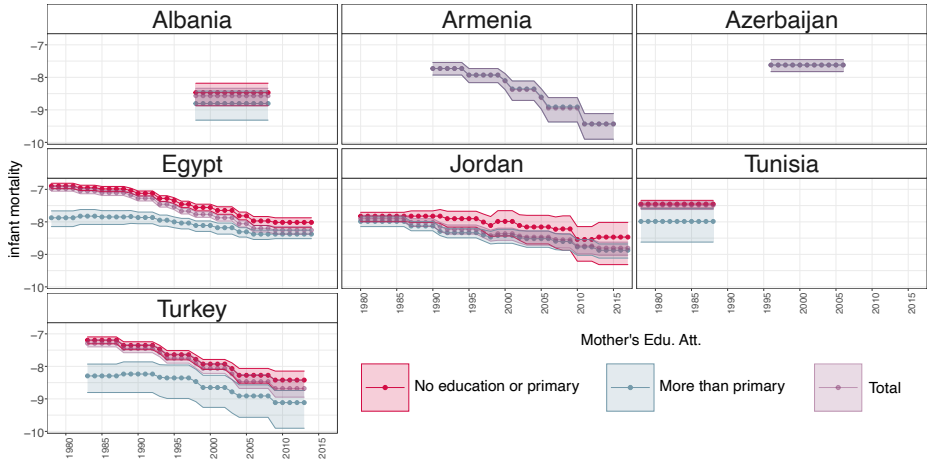
Appendix C: Log-linear model

This modelling step is required to improve the availability of the data on the 15–19 mortality rates. Since the infant mortality by mother's education is available just for the countries and the years represented in the DHS surveys, we would have been otherwise able to apply the proportional splitting of the 15–19 mortality rates just for these country-year combinations. This would have reduced the amount of available information, and it would have prevented us from coherently sharing the information between the countries. Therefore, we applied a log-linear models to estimate the 15–19 log-mortality rates. Log-linear models belong to a family of generalised linear models (GLM) that are often applied to analyse contingency tables.

Data

For the purposes of our case study, we used the DHS data on infant mortality by mother's education, as well as the UN WPP sex-, country-, and year-specific mortality rates interpolated over time for the 15–19 age group. The infant mortality rates by mother's level of education were acquired from the DHS STATcompiler database. These values are based on a recall over nine years preceding the year of data collection (see Figure A-2). Two levels of education are reported: less than primary education and primary education or more. These two categories were satisfactory to analyse the role of education (in this case, the completion of at least one course of study) in the determination and development of mortality differentials over time. When there were multiple estimates for the same period, we used the average rates.

Figure A-2: DHS infant mortality rates by mother's education



Note: The data are extrapolated and averaged to account for the recall period in the DHS.

We first applied the procedure shown in Figure A-6 to split the 15–19 mortality by education (by using infant mortality by mother's education from DHS) only for countries for which the DHS data were available. Then, we predicted the mortality rates for all countries belonging to cluster 1 (Table A-1).

The modelling approach

To impute missing information on 15–19 mortality, we first introduced an additional geographical layer (dimension), based on proximity, so that each country not represented in the DHS is linked to a region made up of countries that are covered by the DHS for each relevant year. This addresses the necessity to have observations for combinations of variables. The countries are grouped as follows:

- Region 1: Albania, Bosnia and Herzegovina, North Macedonia, Montenegro
- Region 2: Armenia, Azerbaijan, Georgia, Turkey
- Region 3: Egypt, Jordan, Lebanon, the State of Palestine, Tunisia

Next, to utilise the log-linear modelling setting, the log-mortality rates were transformed by using the population sizes to create deaths counts (Poisson model family) and using the logarithmic transformation of the education-, sex-, year-, and country-specific population sizes as offsets. Then, we proceeded in two steps.

1. Selection of the best-fitting model (frequentist approach): First, we identified the best-fitting model from a pool of possible model formulations. These were generated by the combinations of the available variables and the pairwise interactions of them (i.e., the models contained only main and two-way interaction terms). We tested the fitting of all the possible combinations of the elements of the set

$$edu.att,\ region,\ year,\ sex,\ edu.att * region,\\ edu.att * sex,\ region * sex.$$

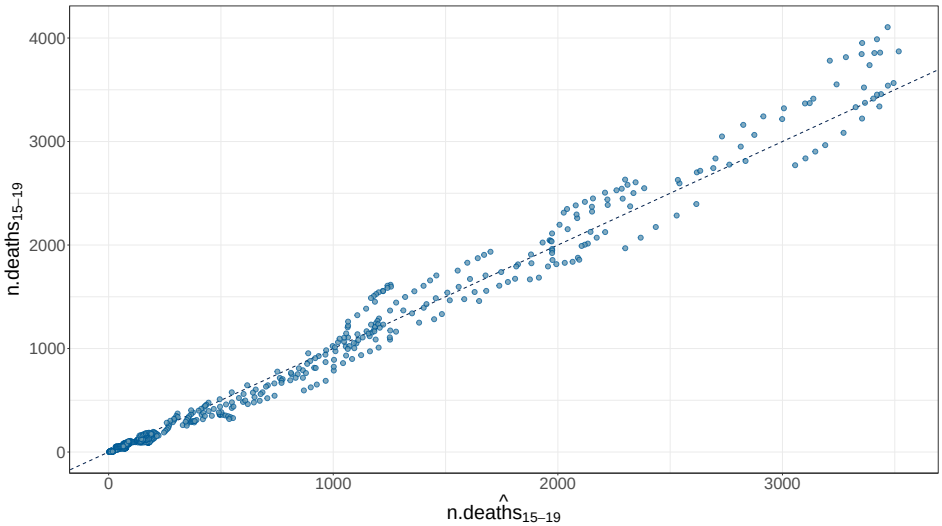
All models were analysed in terms of Bayesian information criterion (BIC) and differences between fitted and observed values. Then, the best-performing models was

$$n.deaths \sim edu.att + region + year + edu.att * region + \\ region * sex + edu.att * sex. \tag{29}$$

2. Bayesian estimation of the best-fitting model: The best-fitting model was estimated within Bayesian inference. This was done to better reflect the uncertainties deriving from the combination of different data sources. It also allowed us to include the predictive uncertainty in the hierarchical structure of our main model (where we indicate with the index $LogLin$ the uncertainty that originated from this step). These estimated rates were additionally corrected via the total mortality data from the UN WPP before they were used as starting values of the reconstruction.

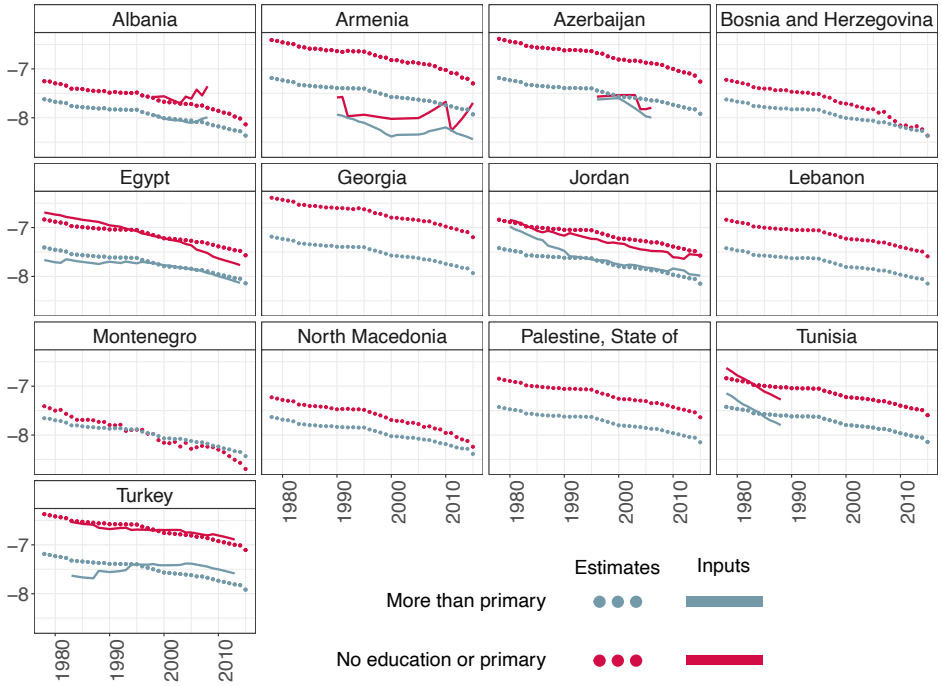
We implemented the model in `R` software, package `rstan` (Stan Development Team 2018a; Carpenter et al. 2017; Stan Development Team 2018b), which we used to sample from the posterior distribution. We used weakly informative normal priors. To check convergence we relied on the Gelman and Rubin diagnostic (Gelman and Rubin 1992), the effective sample size, and a visual inspection of trace plots analysis and posterior predictive checks. In Figure A-3, we report the scatter plot resulting from the comparison of 5,000 posterior predictive draws of the predicted number of deaths (x-axis) and the input data (y-axis). We observe that the final model predicts the data well.

Figure A-3: Estimations scatter plot



This model was employed to produce annual estimates of the (log-)mortality rates for the 15–19 age group differentiated by the level of education. In Figure A-4, the posterior medians (dots) are reported along with the data (lines). By using a limited number of inputs, we can gather a set of values for all countries that exhibit a more uniform and refined pattern compared to the raw data provided by DHS. Moreover, by using Bayesian inference, the results are accompanied by measures of uncertainty. These values, together with those supplied by the DHS database, are integrated into our hierarchical reconstruction model.

Figure A-4: The log-linear model results



Source: Own calculations based on DHS data.

Appendix D: Construction of model inputs

The procedure was based on applying country-, education-, and time-specific mortality profiles to time- and education-specific log-mortality rates for the 15–19 age group as starting points. The profiles were extrapolated by applying the mortality differentials between the levels of education (based on data obtained from Eurostat,[10] see Sauerberg 2021) to the remaining age groups (from 20–24 to 85+) in the UN WPP total mortality schedules. Their consistency with total mortality was ensured by population size weights. The starting values were obtained by exploiting the differences in infant mortality by mother's education, which are available in DHS and are estimated for the countries without DHS data with a Bayesian log-linear model (Appendix C). By applying these profiles to the starting values and then correcting for possible discrepancies from

[10] DNK, EST, FIN, NOR, SWE, ITA, GRC, PRT, MLT, BGR, HUN, POL, ROU, SVN, SVK, SRB, HRV, TUR.

total log-mortality, we obtained the age-, time-, country-, and age-specific log-mortality schedules $\log m^*_{a,t,c,e}$ to be used as model inputs.

Two quantities are necessary for the reconstruction: (1) 15–19 log-mortality rates, and (2) education-specific reconstruction curves. We describe details of their construction below.

15–19 mortality rates

As described in Appendix C, the mortality rates for the 15–19 age group were adopted as starting points of the reconstruction for the rest of the age groups. These 15–19 mortality rates by education were estimated by a Bayesian log-linear model informed by the DHS data on infant mortality by mother's education.

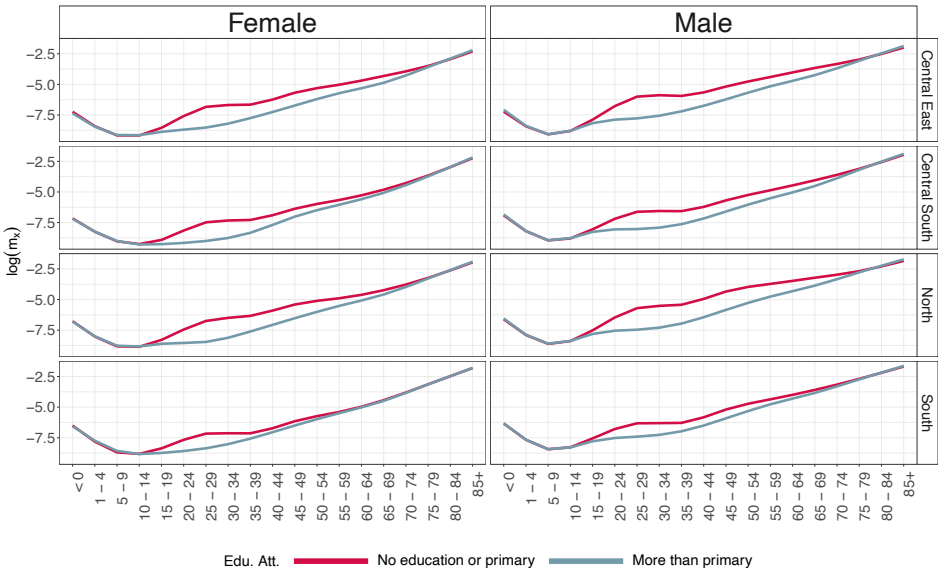
The procedure for disaggregating the 15–19 mortality by education relies on the DHS data on infant mortality by mother's education. From these data, we calculated the ratio of the mother's education-specific infant mortality rates to the total infant mortality rates. By using these ratios, we disaggregated the sex-, period-, and country-specific mortality rate for the 15–19 age group. We also used population size weighting such that the resulting average total rate equals the total value available in UN WPP for the analogous period. This procedure rests on the assumption that the level of education of the mother suffices to differentiate the education-specific mortality for the 15–19 age group. This is supported by its coherence with the procedure followed by Eurostat for the definition of life expectancy by age, sex, and educational attainment, and by the consistent evidence indicating that parents' education efficiently predicts the educational outcomes of their children (Eccles 2005; Ludeke et al. 2021; Dubow, Boxer, and Huesmann 2009), and that maternal schooling plays a key role in determining children's chances of survival (Kiross et al. 2019; Li and Keith 2010; Green and Hamilton 2019; Caldwell and McDonald 1982; Mandal, Paul, and Chouhan 2019).

Education-specific reconstruction curves

The 15–19 mortality by education was a starting point for reconstructing mortality profiles by education for the older age groups. They were obtained by exploiting the information from the Eurostat database and from estimates by Sauerberg (2021). These were combined with the period-, sex- and country-specific (log-)mortality rates published by UN WPP and the period-, sex-, education-, and country-specific population sizes from the WIC database to provide the reconstruction curves, which were then used as inputs to the model. The methods presented in Sauerberg (2021) were employed to obtain a collection

of mortality curves for 18 European countries in different years.[11] By grouping the levels of education and the countries into four groups (Figure A-5), we identified profiles for European subregions over the 2007–2017 period for the two levels of education under consideration.

Figure A-5: The grouped European subregional mortality profiles



Source: Own calculations based on Eurostat data (Eurostat 2023).
Notes: The countries were grouped as follows: North: DNK, EST, FIN, NOR, SWE; South: ITA, GRC, PRT, MLT; Central East: BGR, HUN, POL, ROU, SVN, SVK; Central South: SRB, HRV, TUR.

We then used ratios of the education-specific mortality profiles and education-specific population sizes to split the total mortality rates profile from UN WPP. The splitting operation ensures consistency with the total mortality rate and with the difference between the two different mortality levels.

Figure A-6 illustrates the steps used to achieve an age-, country-, and period-specific log-mortality rate for the 30–34 age group for Turkey. We applied the same procedure to all other countries and age groups. To explain the procedure, we introduce the following notation:

[11] Represented countries: BGR, DNK, EST, GRC, HRV, ITA, HUN, MLT, POL, PRT, ROU, SVN, SVK, FIN, SWE, NOR, SRB, TUR. Time span (maximum): 2007–2017.

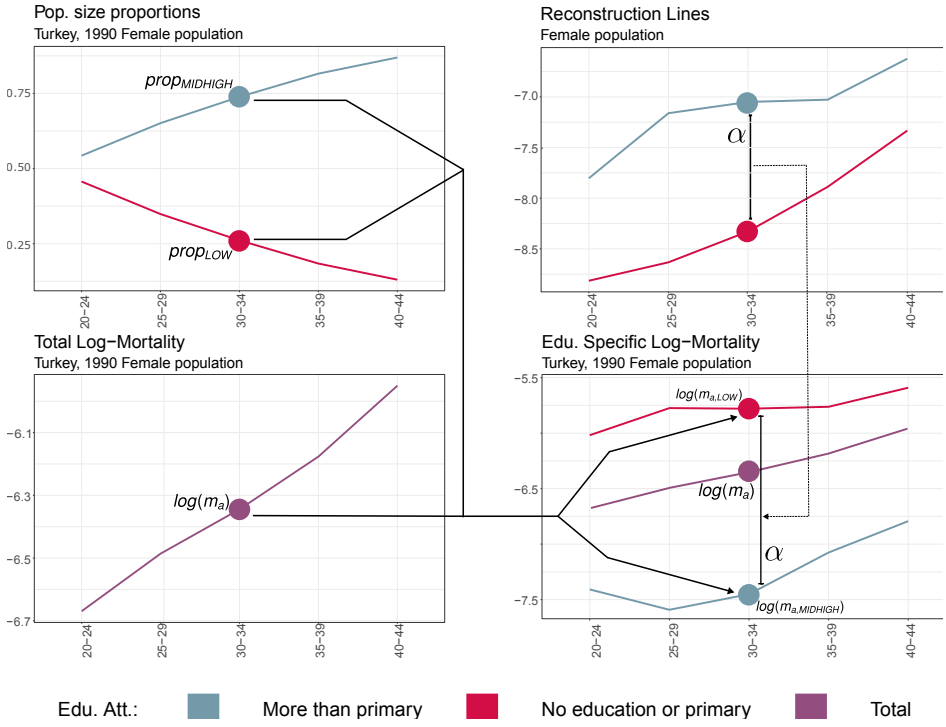
1. $prop_{LOW}$ and $prop_{MIDHIGH}$: proportions of population in category up to primary education (LOW) and more than primary educated (MIDHIGH) that are attained in a specific country in a specific period (subscripts dropped for the clarity of presentation).
2. $log(m_a)$: period-, country-, and age-group-specific log-mortality rate as published by UN WPP.
3. $log(m_{a,LOW})$ and $log(m_{a,MIDHIGH})$: age-specific log-mortality rates for the two levels of education.
4. α: ratio of lower and mid-higher education log-mortality rates. The ratio was calculated based on the collection of mortality curves derived from the Eurostat data.

The disaggregation of the total values into education-specific mortality is obtained by solving a two equations system with two unknowns:

$$\begin{cases} \frac{log(m_{a,LOW})}{log(m_{a,MIDHIGH})} = \alpha, \\ \\ prop_{LOW} * log(m_{a,LOW}) + prop_{MIDHIGH} * log(m_{a,MIDHIGH}) = log(m_a). \end{cases}$$

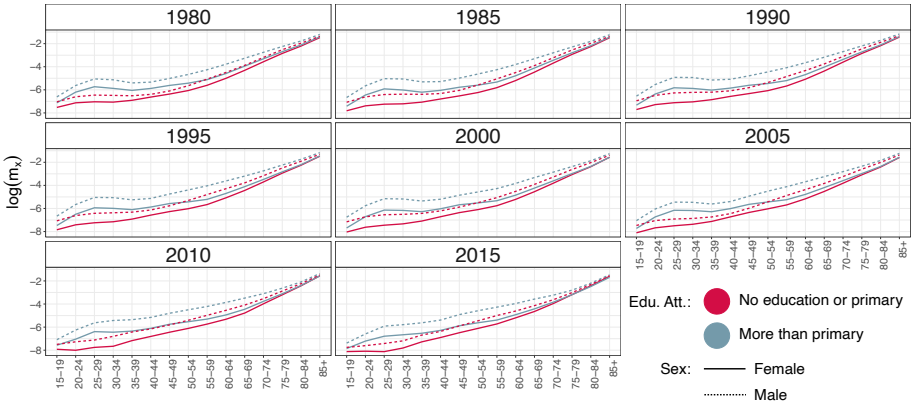
As a concluding step, for the inputs construction, the values derived from the above procedure were then corrected to ensure consistency with the total mortality rates. We thus obtained a collection of country-, period-, age-, and education-specific mortality rates that are consistent with the total mortality rates, when education-specific rates are weighted with the population size of a given level of education. An example of outputs for Albania is presented in Figure A-7. The approach is easily generalisable to other countries, regions, and periods, especially in combination with the log-linear modelling approach (Appendix C) that allows for estimating mortality for countries not covered by the DHS.

Figure A-6: The construction procedure



Source: Authors' own calculations based on WIC, Eurostat, and UN WPP data.

Figure A-7: The log-mortality profiles estimated for Albania



Appendix E: Age-, sex-, and education-specific principal components

In this section, we describe in detail the construction of the sex-, age-, and education-specific principal components. As in the work of Alexander, Zagheni, and Barbieri (2017), these components are employed to represent the key characteristics, in terms of variation, of a family of mortality curves. Their use is conceptually comparable to the Lee–Carter approach (Lee and Carter 1992), and it is based on the representation of a set of mortality curves as a combination (weighted by loadings) of principal components. Principal components analysis (PCA) is a widely known method for dimension reduction and the summarising of variability of the data. Principal components can be obtained through a single value decomposition (SVD) method. In our case, the decomposed matrices are those containing information on how the mortality curves develop in a given space-time region for a given average level of education of the different age groups. In particular, we considered the countries belonging to cluster 1 and the 1980–2015 time period. To obtain the principal components, we made use of three data sources:

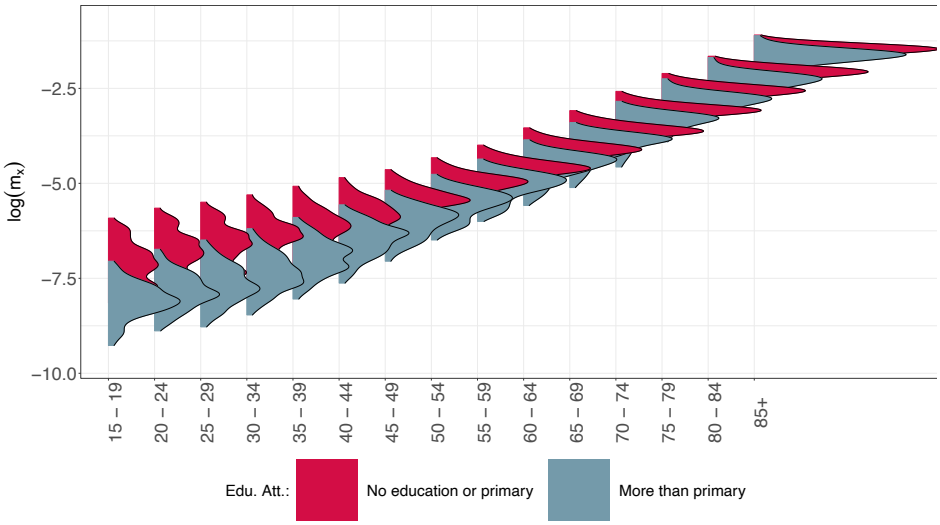
a) The WIC Data Explorer, from which we acquired data regarding the average number of years of schooling for the five-year age groups by sex and five-years period for the countries under consideration;
b) The UNESCO DataBase, from which we obtained, for the countries and for the period of our interest, the average duration of the study cycles to finalise the primary schooling; and

c) The UN WPP database, from which we obtained the age-, sex-, and period-specific mortality tables (and in particular the mortality rates), which we then used to populate the two different matrices.

By using the 15+ age group in the Wittgenstein Center database as a reference age group, we obtained the average years of schooling of the population aged 15+ (specified by sex, country, and period). By cross-referencing this information with the precise duration of the different cycles of study in the countries and periods considered, we assigned the labels "no education or primary" or "more than primary" to all sex-period-country combinations under study. The labels refer to the estimated average level of education of the population in that specific year and country in the 15+ age group. After doing so, we assigned mortality curves obtained from the UN WPP database to these labels for each country-period. Then we performed PCA for two matrices representing two levels of education and obtained two separate sets of principal components vectors, specific to the approximate average level of education. As was already mentioned in the main body of the paper (Section 3.2.2), since the time intervals for which the data were available did not coincide, we performed a yearly interpolation of the values before crossing the values (school duration was kept as an integer). This procedure relies on the assumption of associating the average years of schooling with a representative label for the entire country's mortality curve. While this is a simplification, it is a reasonable approach given that the curves are ultimately summarized using SVD. The primary purpose of this step is to construct collections of curves that approximate the general mortality patterns based on education levels within the specified region and time period.

The visualisation of the labelling outcome is presented in Figure A-8. The density plots for each age-education group depict mortality rates for the female population in line with the case study, across all countries falling within cluster 1, during the specified time period of interest (1980–2015). In this plot, the reference period is from 1980 to 2015, and the countries are those we studied in the case study. It is immediately apparent that, for all age groups, the mortality rate of the lower educated is higher than that of individuals who have at least completed primary education. It is also interesting to note that the differences in mortality (and the reduction of variability of the densities) decrease with age, as does the distance between the modes of the distribution.

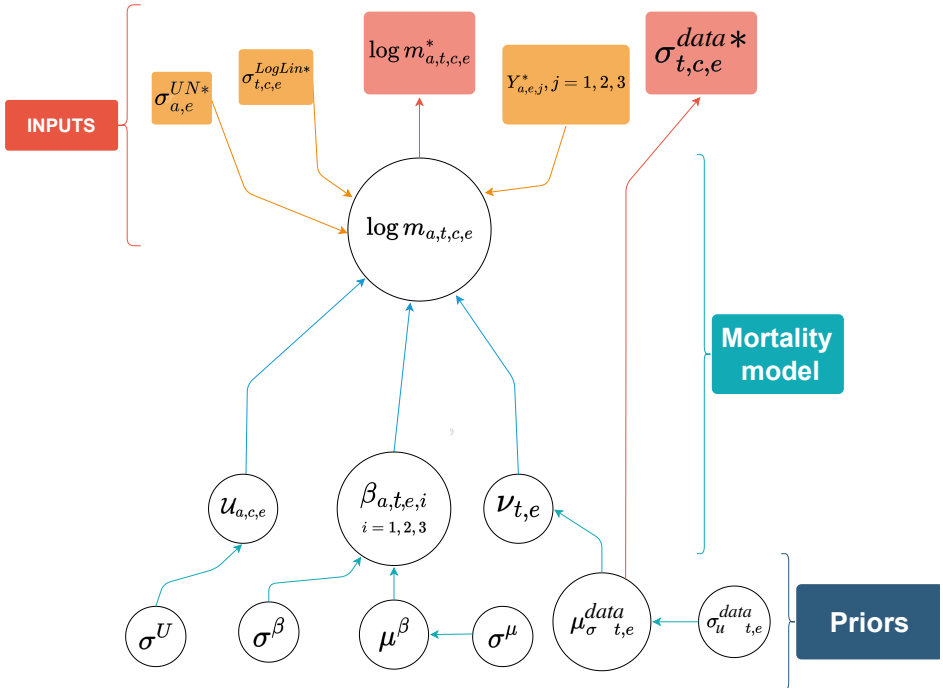
Figure A-8: Approximated education-specific log-mortality distributions



Source: Own calculations based on UN WPP, WIC and UNESCO data.
Note: The mortality curves referring to the countries in cluster 1 for the 1980–2015 period are related to the female population.

Appendix F: The case study model formulation, graphical representation

Figure A-9: The model: Graphical representation



Notes: The different components are visualised as follow: circle: objects with a distribution; orange square: quantities estimated outside of the model used as hyperparameters; red square: quantities estimated outside of the model used as data.

Appendix G: Additional tables

Table A-2: Available DHS rounds for cluster 1

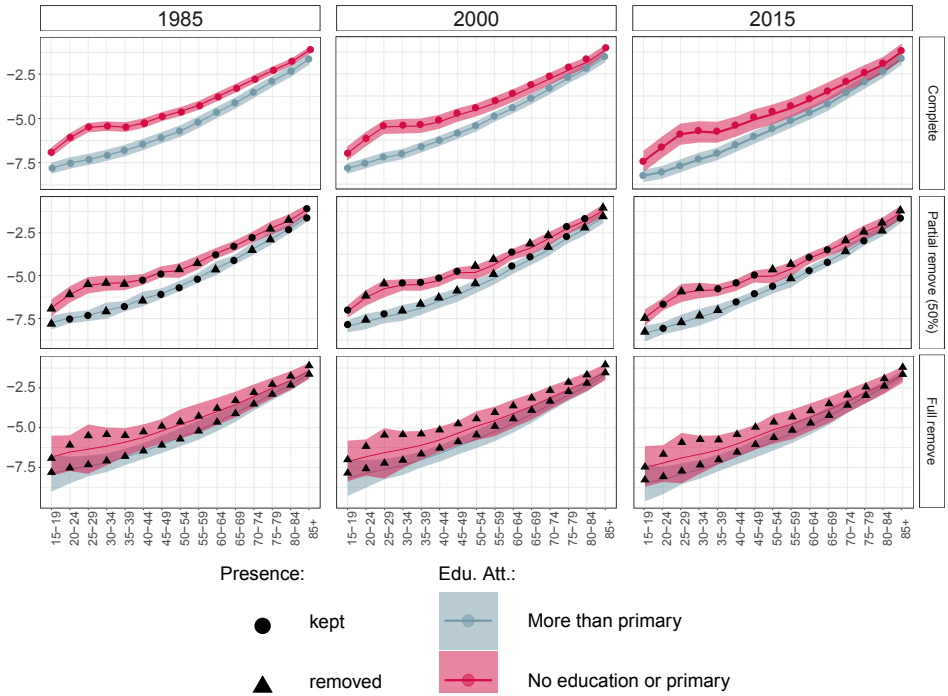
Country	DHS rounds
Albania	2008-09 (1)
Armenia	2000, 2005, 2010, 2015-16 (4)
Azerbaijan	2006 (1)
Egypt	1988, 1992, 1995, 2000, 2003, 2005, 2008, 2014 (8)
Jordan	1990, 1997, 2002, 2007, 2009 (5)
Tunisia	1988 (1)
Turkey	1993, 1998, 2003, 2008, 2013 (5)

Table A-3: Educational attainments conversions

ISCED (Eurostat)	WIC Explorer
ISCED 0–2: Early childhood education Primary education Lower secondary education	No education, incomplete primary, primary, lower secondary
ISCED 3–4: Upper secondary education Post-secondary non-tertiary education	Upper secondary, post-secondary, short post-secondary
ISCED 5–8: Short-cycle tertiary education Bachelor's degree or equivalent tertiary education level Master's degree or equivalent tertiary education level Doctoral degree or equivalent tertiary education level	Bachelor's, master's and higher

Appendix H: Additional figures

Figure A-10: Log-mortality rates by level of education, age group, and time. Georgia, female population



Note: Reporting results for selected years 1985, 2000, and 2015. In the first row we report the results from the model with full input. In the second row the ones for the model with 50% of the inputs removed (of the total inputs amount), and in the last row for the one with the model for which all the inputs for Azerbaijan, Georgia, North Macedonia, and Tunisia were removed. The solid lines present the estimated medians, while the 80% credible intervals are visualised via shaded areas. The dots represent the inputs to our model, the triangles the inputs which were removed.

Figure A-11: All the results, 95% CI

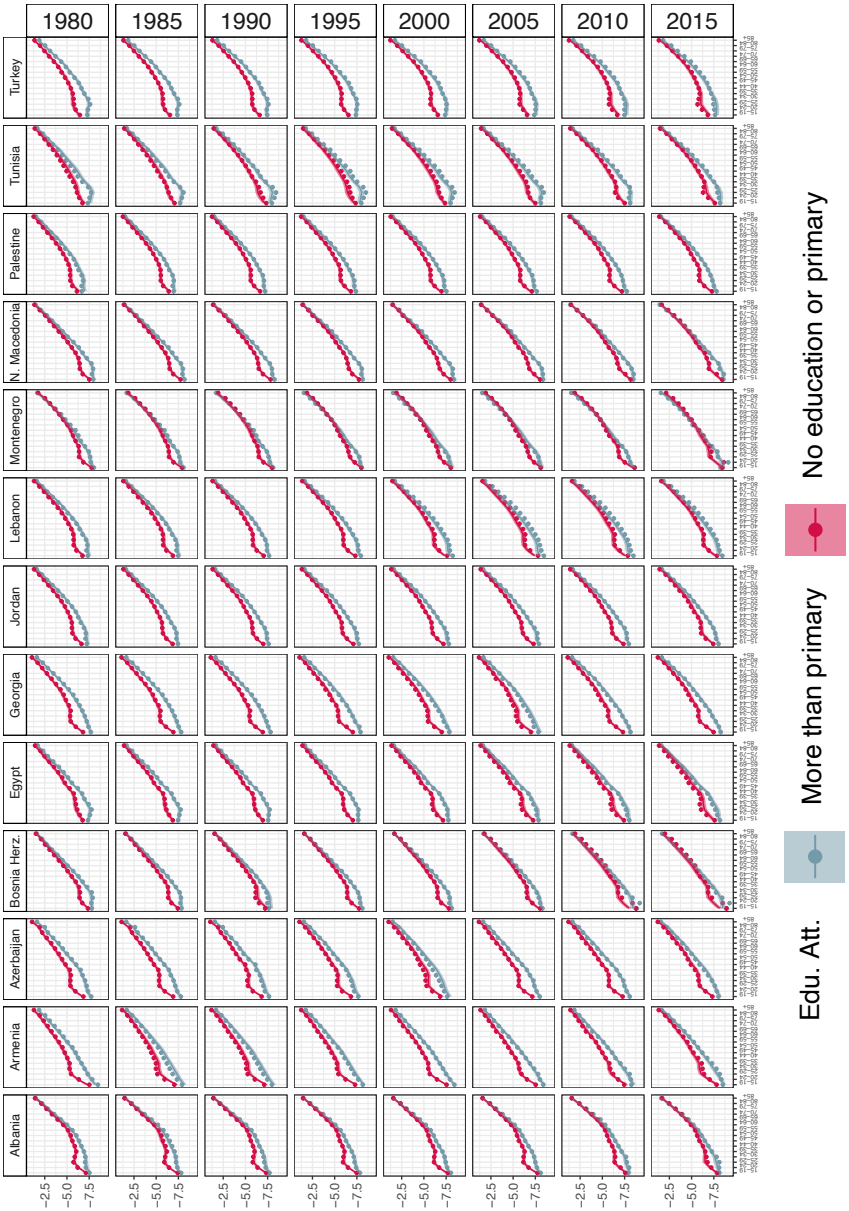


Figure A-12: Trace plots examples

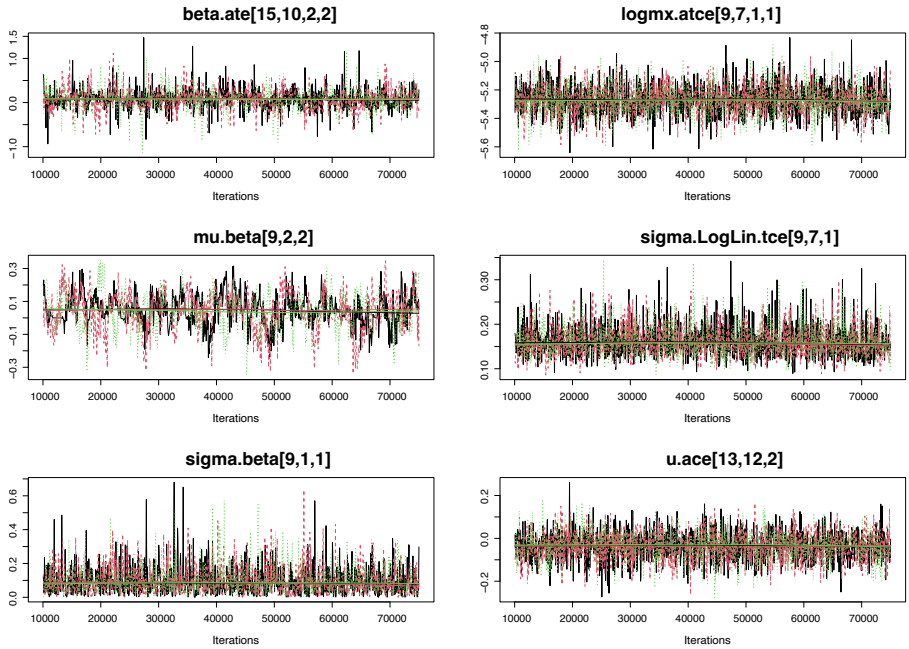
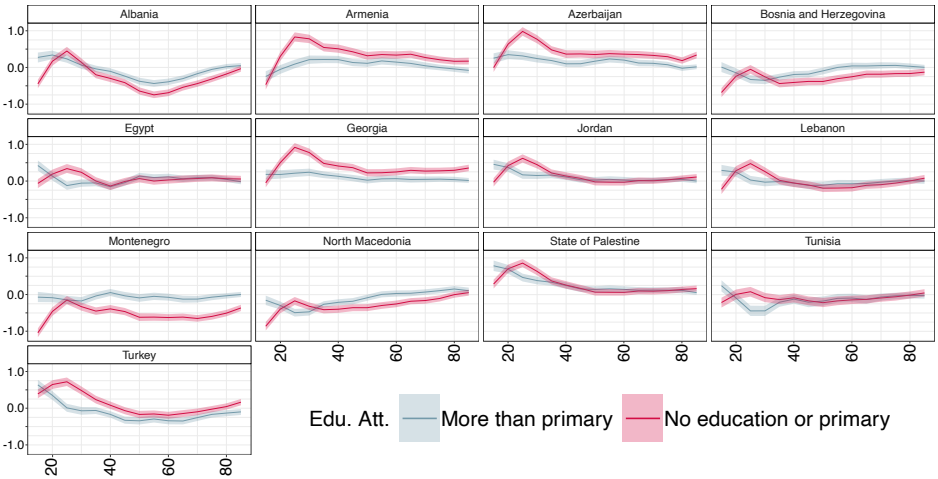
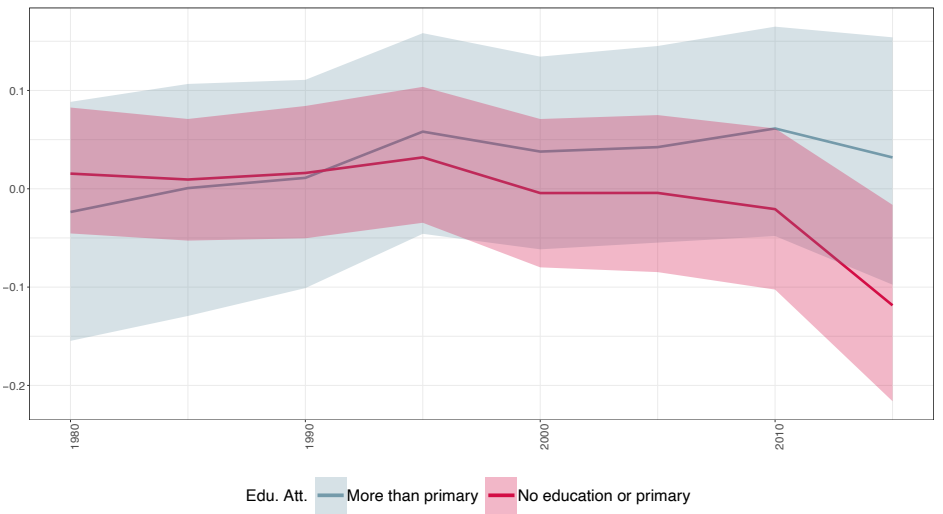


Figure A-13: $u_{a,c,e}$ random effects



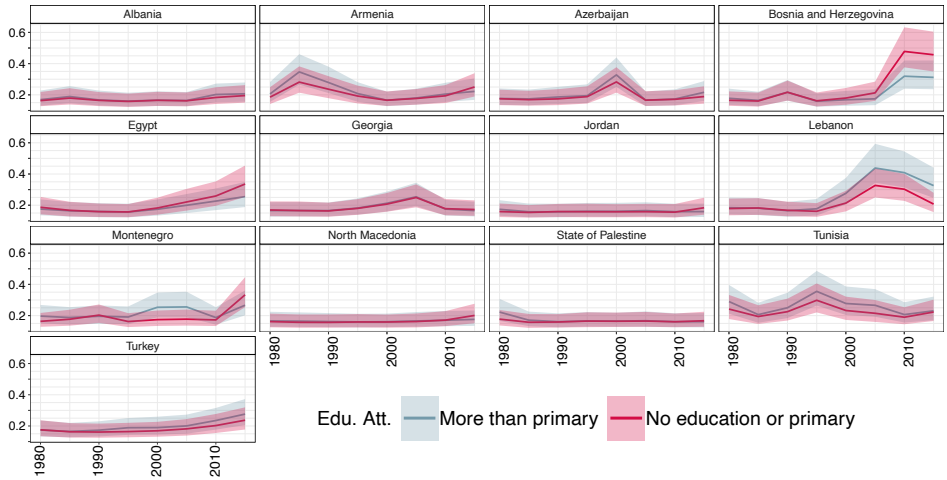
Note: The solid lines present the estimated medians, while the 80% credible intervals are visualised via shaded areas.

Figure A-14: $\nu_{t,e}$ random effects



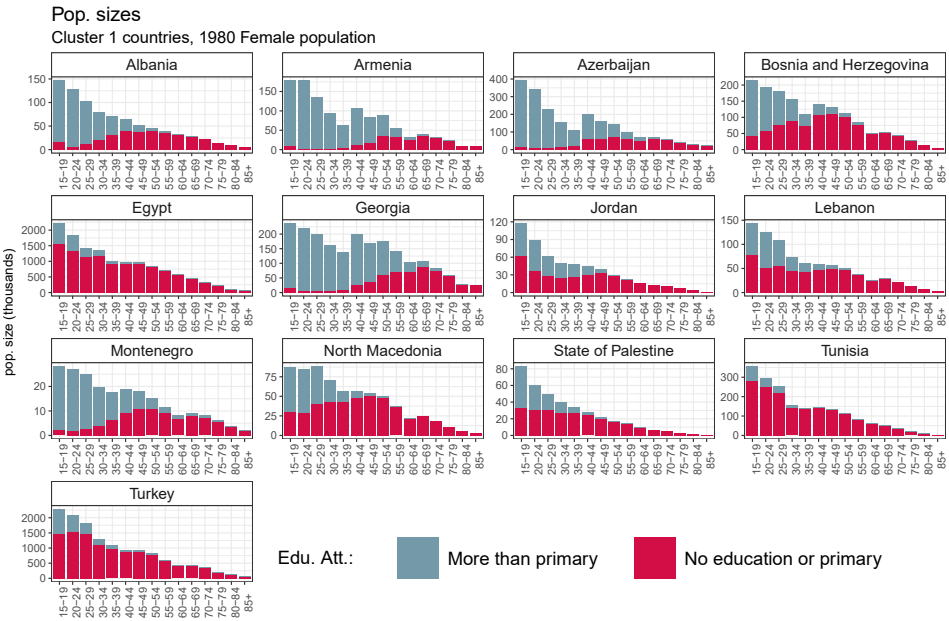
Note: The solid lines present the estimated medians, while the 80% credible intervals are visualised via shaded areas.

Figure A-15: $\sigma_{t,c,e}^{info}$



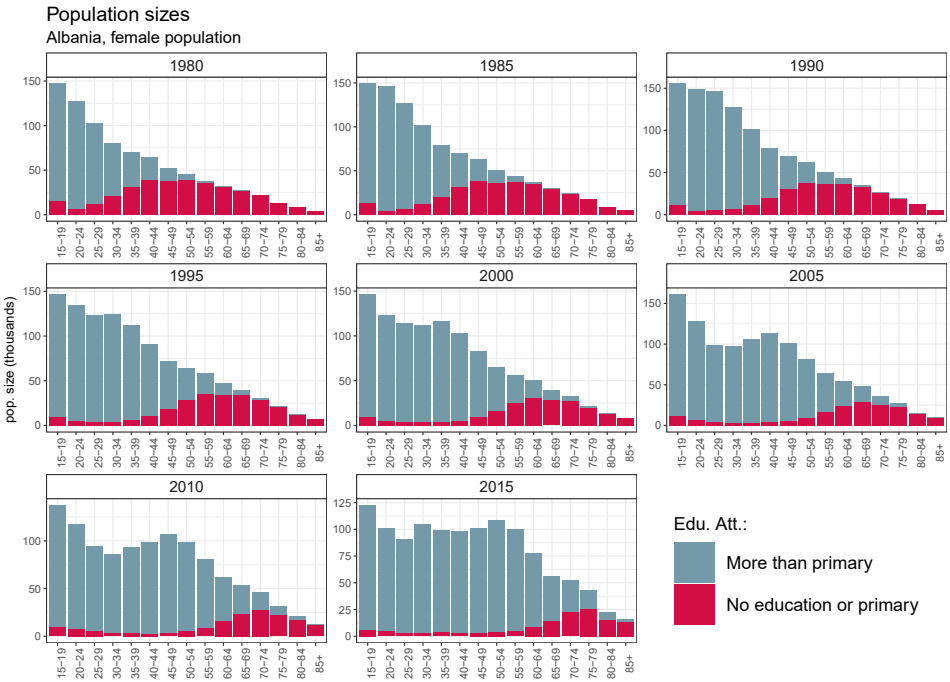
Note: The solid lines present the estimated medians, while the 80% credible intervals are visualised via shaded areas.

Figure A-16



Source: WIC Data Explorer.

Figure A-17



Source: WIC Data Explorer.