

VIENNA INSTITUTE OF DEMOGRAPHY

WORKING PAPERS

05/2023

BAYESIAN MULTI-DIMENSIONAL MORTALITY RECONSTRUCTION

ANDREA TAMBURINI, ARKADIUSZ WIŚNIEWSKI AND
DILEK YILDIZ

ABSTRACT

Even though mortality differentials by socio-economic status and educational attainment level have been widely examined, this research is often limited to developed countries and recent years. This is primarily due to the absence of consistently good-quality inherent data. Systematic studies with a broad geographical and temporal spectrum that engage with the link between educational attainment and mortality are lacking. In this paper, we propose a mortality rates reconstruction model based on multiple patchy data sources, and provide mortality rates by level of education. The proposed model is a hierarchical Bayesian model that combines the strengths of multiple sources in order to disaggregate mortality rates by time periods, age groups, sex and educational attainment. 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 five-year age groups starting at age 15 for the 1980-2015 time period. Furthermore, we evaluate the model's performance relying on standard convergence indicators and trace plots, and validate our estimates via posterior predictive checks. This study contributes to the literature by proposing a novel methodology to enhance the research on the relationship between education and adult mortality. It addresses the lack of education-specific mortality differentials by providing a flexible method for their estimation.

KEYWORDS

Bayesian Reconstruction, Mortality, Education.

AUTHORS

Andrea Tamburini, Wittgenstein Centre for Demography and Global Human Capital (IIASA, OeAW, University of Vienna), Vienna Institute of Demography/Austrian Academy of Sciences. Email: andrea.tamburini@oeaw.ac.at

Arkadiusz Wiśniowski, Department of Social Statistics, School of Social Sciences, University of Manchester, United Kingdom. Email: a.wisniowski@manchester.ac.uk

Dilek Yıldız, IIASA, Wittgenstein Centre for Demography and Global Human Capital (IIASA, OeAW, University of Vienna). Email: yildiz@iiasa.ac.at

ACKNOWLEDGEMENTS

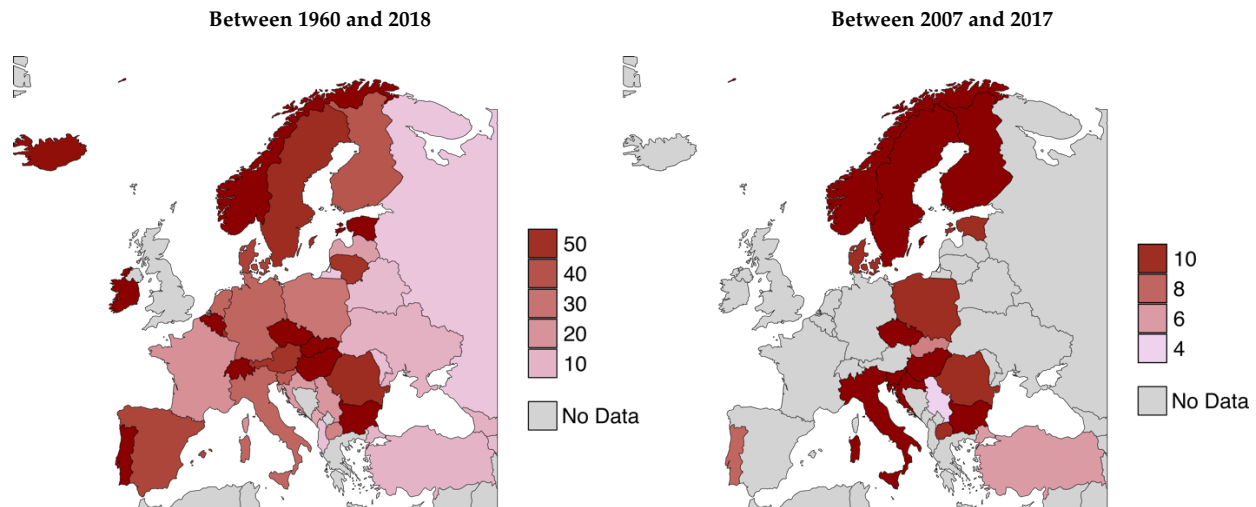
This paper partially results from the work carried out in connection with the BayesEdu project at the Vienna Institute of Demography, which received funding from the “Innovation Fund Research, Science and Society” established by the Austrian Academy of Sciences (ÖAW).

1 INTRODUCTION

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 sub-nationally (Bora, Raushan, and Lutz 2018; Sasson and Hayward 2019), the research to date has focused on specific sub-populations only (e.g., sub-groups of the adult population, infants), and has not addressed the systematic reconstruction of age-specific mortality rates for adults. Moreover, most 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, at aggregate levels only. The primary obstacle that has constrained the growth of 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 data sets that provide mortality rates or counts of deaths by educational attainment for a large group of countries (including developing countries) and over a long time period (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 of education and mortality evolved for sub-populations in different countries; (2) to extend our knowledge of socio-economic disparities in mortality to a broader geography and to longer time periods; and (3) to provide more accurate baseline estimates to project multidimensional populations.

Mortality data broken down by educational attainment have been collected for recent periods only, and typically for a few high-income countries in the Global North. This pattern is obvious in Europe, where this information is only available from a few national statistical offices (see Figure 1, right panel), in addition to from the recent Eurostat data collection (Eurostat 2022). High-quality data are rarely available even for broad age groups. For countries in the Global South, which often lack valid civil registration systems, the main sources of demographic data are nationally representative surveys such as the Demographic and Health Surveys (DHS) (USAID 2022). However, those surveys rarely collect information on adult mortality. Existing estimates rely on indirect estimates such as life tables and the sisterhood method for maternal deaths (Graham, Brass, and Snow 1989; United Nations 1983). Globally, the main source of comparable mortality data is the United Nations World Population Prospects (UN WPP) (United Nations 2022). It provides population counts, vital rates estimates and projections between 1950 and 2100 for 235 countries or areas. However, these estimates are not broken down by levels of education. The most comprehensive systematically verified source of information on population counts and consistent demographic rates (e.g., total fertility rate, age-specific survival ratio) disaggregated by educational attainment is the Data Explorer of the Wittgenstein Centre for Demography and Human Capital (Wittgenstein Centre Data Explorer 2018). While data concerning survival rates based on educational achievement levels are available for the reference period (2015-2020) and for future predictions under various Shared Socioeconomic Pathways (SSPs) scenarios, the information on survival ratios by educational attainment is limited to assumptions for future projections under different Shared Socioeconomic Pathways (SSPs) scenarios (Wittgenstein Centre Data Explorer 2018).

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:

Own calculations based on Eurostat data (Eurostat 2022).

In this paper, 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, DHS and UN WPP. Our contribution is two-fold. First, we propose a method that fills the gap in the literature on reconstructing multi-dimensional 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.

2 BACKGROUND

Education is primarily acquired at younger ages, and is a fundamental determinant of individual and inter-generational social mobility that is closely linked to people's health (Avison 2005). For this reason, the level of education has often been used as an indicator of socio-economic status, occupation (Davey Smith et al. 1998; Luy et al. 2019) or both (Luy, Giulio, and Caselli 2011). A number of studies have examined the connection between educational attainment, health outcomes and mortality (see Baker et al. (2011) for a detailed review). All of these studies, irrespective of their geographical and temporal scale, found that higher educated individuals live longer and generally healthier lives. Previous studies have also identified connections between educational attainment and health risks, such as alcohol consumption (Murakami and Hashimoto 2019; Rosoff et al. 2019), smoking (Assari and Mistry 2018; Tomioka, Kurumatani, and Saeki 2020) and an unbalanced diet (Fard et al. 2021). In addition, several studies have postulated a connection between cause of death and level of education in numerous countries (Malamud, Mitrut, and Pop-Eleches 2018; Clark and Royer 2013; Tjepkema, Wilkins, and Long 2012; Gavurova, Vagasova, and Grof 2017), or for specific age groups (Gakidou et al. 2010). Other studies have evaluated the association between education and health and mortality, and the causal relationship between them (e.g., Zimmerman and Woolf 2014; Avison 2005).

Furthermore, Luy et al. (2019) examined the effects of structural changes in populations due to increasing educational levels. This investigation uncovered strong associations between education and the overall health of a population, which suggests that educational policies might even be regarded as indirect health policies. Moreover, Lutz and Kebede (2018) demonstrated a strong and consistent link between educational attainment and life expectancy improvements as well as reductions in child mortality, with the beneficial effect of education being even more significant than that of GDP per capita. The analysis of populations and their characteristics broken down by various attributes (level of education, marital status, etc.) is often undertaken via multi-state analysis (Keyfitz 1980; Rogers 1980). In demographic data reconstructions, the methodological approaches used in multi-state analysis have focused primarily on the study of population sizes and compositions (Lutz et al. 2007; Goujon et al. 2016; Wheldon et al. 2013a), rather than of demographic rates. As was mentioned above, the most comprehensive multi-state analysis that has addressed the relationship between vital rates, population sizes and educational levels stemmed from Lutz and colleagues (2018), but it did not estimate past mortality rates by education, and it focused only on future scenarios. Moreover, this study relied on several scenarios that did not include any assessment of uncertainty regarding mortality differences by educational attainment. Thus, although previous research has established and investigated the connection between education and mortality from numerous perspectives, the data describing this relationship in the past are still limited.

Population and vital rates reconstruction is a key research topic in demography (see Wheldon et al. (2013a) for an overview). Methods of reconstruction have been developed mainly in two directions, which are distinguished by whether they move backwards or forwards in time. The first approach, demographic back projection, attempts to revert the relationships between population size and composition and mortality, fertility and migration rates based on the Cohort Components Method for Population Projections (CCMPP). Pioneering work on this subject was carried out by Wrigley and Schofield (1983). More recently, this approach was employed systematically for multi-state population reconstructions by Lutz and colleagues (2007) and by Goujon and colleagues (2016). In the latest work using this approach, back projections are available for the 1950-2015 period in five-year steps for 201 countries and six levels of education (Lutz et al. 2018; Springer et al. 2021).

The second approach, inverse projection, was first introduced by Lee (1974), with subsequent work (Lee 1985), addressing certain technical inconsistencies associated with the CCMPP inversion used in back projection. Although the results obtained with both these techniques were validated using historical data, they are plainly deterministic; therefore, uncertainties related to data scarcity and quality and the underlying assumptions are not included in the modelling design, and are not embodied in the results.

This issue was explored by another stream of research using Bayesian inference to simultaneously reconstruct population sizes and demographic rates (mortality and fertility rates and net migration flows) by combining incomplete data sources. Measurement errors were incorporated in a method developed by Wheldon et al. (2013b) to estimate missing population counts using fragmentary data. In that paper, the model was employed in a case study that aimed to reconstruct the female population of Burkina Faso from 1960 to 2005. This reconstruction approach was tested in different data quality environments, and was extended to countries that do not have regular censuses (Wheldon et al. 2016). It was subsequently shown that this approach can be employed for two-sex populations as well, and that probabilistic estimates of various sex ratio measures can be obtained (Wheldon et al. 2015). While providing results that take into account the possible uncertainties in the modelling process, this method is limited to analyses of age and sex structures. It has not been applied to multi-state populations, e.g., to populations disaggregated by the level of education.

Additionally, Bayesian hierarchical models have been utilised to independently reconstruct fertility rates, distinct from other population and vital rates. For instance, Alkema et al. (2012) accounted for deficiencies in data sources to estimate total fertility rates (TFR) by combining and adequately weighting observations from DHS, World Fertility Surveys and other surveys. They produced estimates for West Africa, and thus for a context characterised by data scarcity. In an alternative approach, Schmertmann and Hauer (2019) combined information regarding the age-sex population structure and the child-to-woman ratio to infer the TFR.

In the study of migration, the absence of data or the fragmented nature of the data is a more pronounced problem (see Willekens et al. (2016) for a detailed review). Recently, Bayesian hierarchical models have been developed that aim to integrate various types of migration data (Raymer et al. 2013; Wiśniowski 2017; Gendronneau et al. 2019; Wiśniowski 2021). The statistical framework proposed in these works relies on correcting measurement errors and imputing missing information, and it permits the inclusion of information derived from social media.

Finally, Bayesian approaches have been employed to estimate mortality rates. Alkema and New (2014) and Alexander and Alkema (2018) dealt with limited data availability and data quality issues in developing countries to estimate under-five and neonatal mortality rates, respectively. In these works, the authors used Bayesian regression spline models. They took into account data quality issues and various sources of error, as well as the considerable differences in data availability across various countries. Their Bayesian hierarchical models permit the borrowing of information from multiple countries and over time, and are 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 based on borrowing from other data sources was proposed in Alexander, Zagheni and Barbieri (2017). They addressed the problem of sample sizes in sub-populations (the population of the U.S. split up at the county level) by sharing information across different geographical levels. They did so by using the state-level mortality profiles to inform their estimates of the mortality rates in counties (small areas) via singular value decomposition (SVD). The SVD approach allows for the imputation of missing observations and the correction of measurement irregularities in small-area data.

The aim of this research is to develop a methodology to estimate age- and education-specific mortality rates that integrates data from various sources and produces estimates with measures of uncertainty that take into account variability in data quality. Building upon the above-described literature, we propose a multi-dimensional hierarchical Bayesian model. It integrates the 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. The model generates age-specific mortality rates for various 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 model is similar to that developed by Alexander, Zagheni and Barbieri (2017), as it also uses 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. These are the estimated age-year-country-education-specific mortality rates for which the estimation requires the interaction of several data sources.

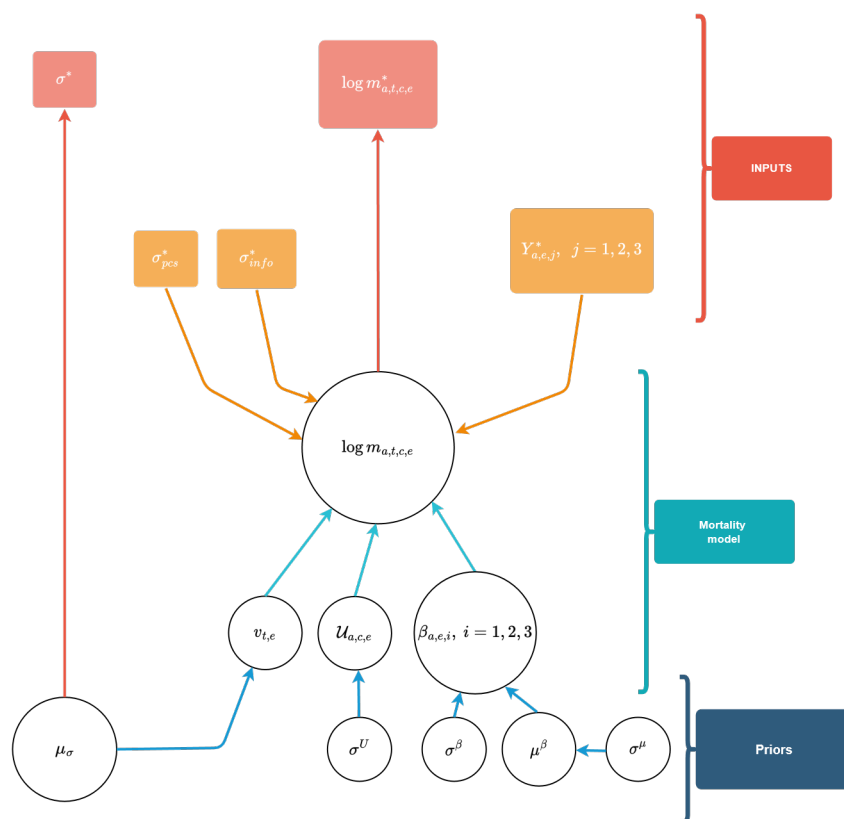
The general reconstruction model specification is outlined in section 2. The case study application is presented in section 3. In particular, the preparation of two sets of input data – namely, the age- and education-specific principal components and the initial age, year- and education-specific log-mortality rates – is described sub-section 3.2.2. This process, which requires the combining of information from different data sources, is explained in detail in sub-section 3.2, and is illustrated in Figure 4. While the proposed methodology can be extended to different geographical regions and periods, the input preparation step is specific to our case study, and may be different for other applications. The model validation, performance analysis and results are summarised in section 4. Finally, our conclusions are presented in section 5.

3 THE RECONSTRUCTION MODEL SETUP

In this section, we introduce the general modelling framework. It is presented in Figure 2, which shows that we are reconstructing unknown (latent) mortality rates, $m_{a,t,c,e}$, 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 prior distributions.

In the context of Bayesian modelling, priors are our initial beliefs about the model parameters before the data have been observed. They help us to incorporate prior knowledge into the analysis by influencing the posterior distribution, which represents our updated beliefs after the data observation. Priors can be informative, meaning that they can strongly guide inference by capturing substantial prior knowledge. Alternatively, priors can weakly guide inference or be non-informative, allowing the data to dominate, and resulting in less biased parameter estimates. Priors can play a crucial role in striking a balance between relying on prior information and on observed data, and can thus allow for a coherent and flexible approach to statistical inference in scientific research. The general formulation and the formulation used in the case study are not sex-specific, and have been developed for the female population. However, the same modelling technique can be used to produce estimates for the male population 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.

FIGURE 2: THE MODEL'S GRAPHICAL REPRESENTATION



Graphical notation:

Red Squares: quantities estimated outside of the model that are used as data; *Orange Squares*: quantities estimated outside of the model that are used as hyper-parameters; *Circles*: random variables.

In our model, we first assume that the externally estimated log-mortality rates, $\log m_{a,t,c,e}^*$, are normally distributed:

$$\log m_{a,t,c,e}^* \sim \mathcal{N}(\log m_{a,t,c,e}, \sigma_{info}^*) \quad (1)$$

with an expectation being a key quantity of interest, that is, unobserved (reconstructed) log-mortality rates $\log m_{a,t,c,e}$. Throughout the paper, asterisk * is used in the superscript to denote a fixed quantity rather than a model parameter. 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 needs to be taken into account. In equation 1, parameter σ_{info}^* denotes the standard deviation reflecting the uncertainty around $\log(m_{a,t,c,e})$ that can be derived from one or more data sources.

Next, the unobserved mortality rates are reconstructed by using information derived from various data sources (Eq. 2). In this reconstruction, we assume that the reconstructed mortality rates are informed by three age- and education-specific principal components ($Y_{a,e,j}, j \in \{1,2,3\}$) that provide a time-independent basic structure of the mortality curves together with their time-dependent loads ($\beta_{a,t,e,j}$), and a set of random effects:

$$\log m_{a,t,c,e} \sim \mathcal{N}\left(\sum_{j=1}^3 \beta_{a,t,e,j} * Y_{a,e,j} + u_{a,c,e} + v_{t,e}, \sigma_{pcs}^*\right) \quad (2)$$

The random effects denoted as $u_{a,c,e}$ capture deviations from the education-specific profiles described by the principal components for each country (Eq. 3), and are informed by the data through weakly informative uniformly distributed hierarchical priors for their variance $\sigma_{a,e}^u$ (Eq. 4):

$$u_{a,c,e} \sim \mathcal{N}(0, \sigma_{a,e}^u) \quad (3)$$

$$\sigma_{a,e}^u \sim \mathcal{U}[0,40] \quad (4)$$

Standard deviation σ_{pcs}^* accounts for the potential variation resulting from the selection of the curves¹ employed in the SVD, which are used to derive principal components. Furthermore in Equation 2, we assume that random effects $v_{t,e}$ depend on the available data through a hyperparameter μ_{σ} ,

$$v_{t,e} \sim \mathcal{N}(0, \mu_{\sigma} v_{t,e}) \quad (5)$$

The mean of the normal distribution in Equation 2 is derived from an expansion of the principal components structure outlined in Alexander, Zagheni and Barbieri (2017). The most notable difference from the original formulation is the education-specific formulation of the principal components. We have chosen to use this specification because the mortality profiles differ for various educational attainments. Aside from that, we follow the specification of the hierarchical prior distribution for the factor loading (Eq. 5), as in Alexander, Zagheni and Barbieri (2017) (Eqs. 6-9):

¹ Hereinafter, a mortality curve refers to an age-specific mortality profile.

$$\beta_{a,t,e,j} \sim \mathcal{N}(\mu_{t,e,j}^\beta, \sigma_{t,e,j}^\beta) \quad (6)$$

$$\sigma_{t,e,j}^\beta \sim \mathcal{U}[0,40] \quad (7)$$

$$\begin{cases} \mu_{t,e,j}^\beta \sim \mathcal{N}(0, \sigma_{e,j}^\mu) & (8) \\ \mu_{t,e,j}^\beta \sim \mathcal{N}(2 * \mu_{t-1,e,j}^\beta - \mu_{t-2,e,j}^\beta, \sigma_{e,j}^\mu) & (9) \end{cases}$$

$$\sigma_{e,j}^\mu \sim \mathcal{U}[0,40] \quad (10)$$

Where $2 * \mu_{t-1,e,j}^\beta - \mu_{t-2,e,j}^\beta$ denotes a random walk specification of the time effects over time. The priors for the variance parameters are weakly informative, leaving the posteriors unconstrained and leveraging the data to shape the posterior distribution.

4 CASE STUDY AND DATA

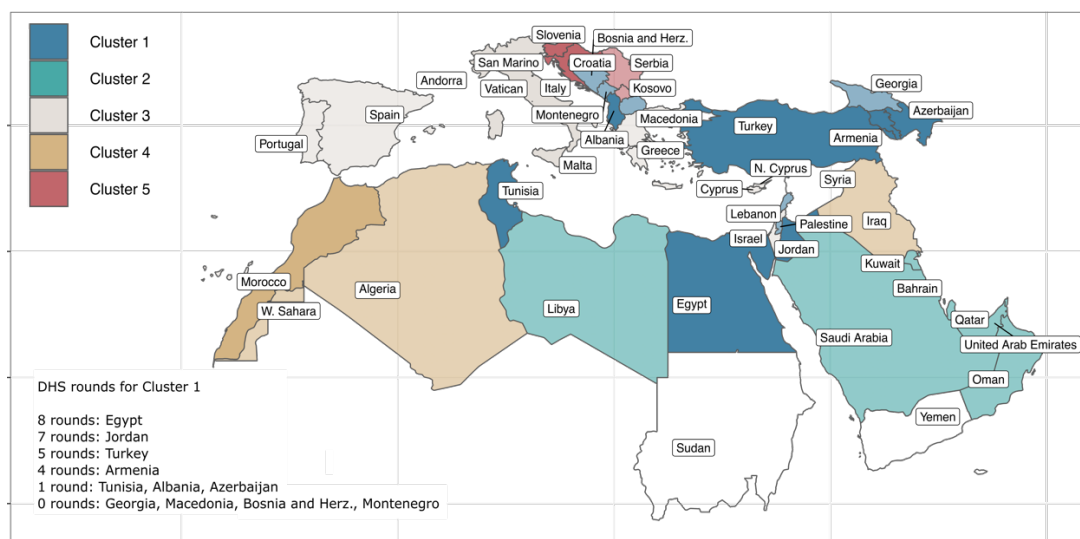
4.1 CASE STUDY SETTING

In our case study, we apply our model to a group of countries, which have been selected to represent a wide range of geographical locations, socio-economic development levels, and levels of data quality and availability. The countries have been chosen in a supplementary step of the reconstruction procedure, which ensures the efficient borrowing of information across countries and over time. The case study is exclusively focused on the female population. However, it is important to note that the same methodology can be applied to the male population or to the total population of the included countries.

In order to demonstrate the borrowing of information between Eurostat and DHS, we have selected a macro-region comprising countries in 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 socio-economic status, mortality and schooling trends. Details of the geographical setting and the clustering procedure are presented in Appendix A-1. 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), State of Palestine (PSE), Tunisia (TUN) and Turkey (TUR). These countries have a noteworthy range of data availability in relation to 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, as well as for the male population.

FIGURE 3: CLUSTERING RESULTS AND INFORMATION ABOUT DHS DATA AVAILABILITY FOR CLUSTER 1



Source:

Authors' own calculations and DHS.

Note:

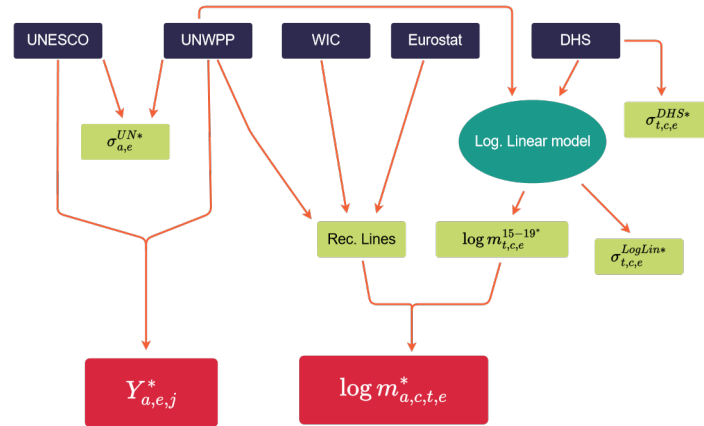
Solid colours represent the availability of DHS rounds or Eurostat data. More information about the DHS data is available in Appendix A-6.

4.2 DATA SOURCES AND MODEL INPUTS

Given the scarcity of data on mortality disaggregated by level of education, such as death counts or mortality rates, we borrowed and combined information from various sources. Considering the systematic nature of our approach and our desire to ensure replicability in different country clusters, 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 possible concerns about their quality and the uncertainty generated by their integration.

The two main data inputs for our model are the age- and education-specific principal components and the age-, year- and education-specific log-mortality rates. We schematically describe the construction of inputs, starting from the data sources until the principal components, $Y_{a,e,j}$, and the multi-dimensional 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



Note:

This scheme presents a schematic workflow for the construction of the model's inputs. 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.

4.2.1 DATA SOURCES

For our case study, the data sources and the information taken from them include:

1. Eurostat Database: life expectancy by age, sex and level of education for 19 countries², between 2007 and 2017 (Eurostat 2021).
2. United Nations World Population Prospects (UN WPP): mortality rates by age, sex, period and country. These are collected for the Cluster 1 countries, and are available for five-year intervals from 1980 to 2015 (United Nations 2022).
3. Demographic and Health Surveys (DHS): infant mortality rates by mother's level of education (USAID 2022). A detailed description of the DHS data is presented in Appendix A-6.
4. Wittgenstein Centre for Demography and Global Human Capital (WIC): population counts by five-year age group, sex, country, five-year period and educational attainment, and mean years of schooling for the population aged 15+ by sex, country and period (Wittgenstein Centre Data Explorer (WCDE) 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 (UIS) (2023)).

² Bulgaria, Croatia, Czechia, Denmark, Estonia, Finland, Greece, Hungary, Italy, Malta, Norway, Poland, Portugal, Romania, Serbia, Slovakia, Slovenia, Sweden, Turkey.

6. World Bank Data Base: multiple indicators regarding education, mortality and health at the national level. These indicators have been employed to cluster countries in the initial step (see section 3.1) (The World Bank 2022).

4.2.2 MODEL INPUTS

Considering the temporal coverage of the DHS waves and the recall period of 10 years³ before each survey date, we focus on the 1980-2015 period. Here we explain the key assumptions and methodological steps for the construction of the inputs: that is, the variables with an asterisk * in Figures 2 and 4, which 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 – i.e., the reconstruction curves and the mortality rates for the 15-19 age group – can be found in Appendices A-2 and A-3, and the inputs are described in more detail in Appendix A-3. Consistent with the notation introduced in section 2 in this section, we introduce an additional notation. Superscript UN is used for quantities that stem from the UN WPP and UNESCO data, while *LogLin* denotes the outputs from the Bayesian log-linear model that is used to estimate the log-mortality rates for the 15-19 age group. Superscript DHS marks the quantities obtained from the DHS data. The inputs are summarised as follows:

1. $\log m_{a,c,t,e}$: log-mortality rates resulting from the application of the region-specific reconstruction curves to the estimated starting points $\log m_{t,c,e}^{15-19*}$, which is the log-mortality rate for 15-19 age group resulting from the log-linear model estimation (Appendix A-2). The reconstruction curves are obtained following a procedure developed by Sauerberg (2021). More specifically, $\log m_{a,c,t,e}^*$ are based on data stemming from 18 European countries, which we have grouped into four regions⁴. In our case study, we use the reconstruction curves for Central-Southern Europe (see Figure A-5 and Appendix A3.2), given the geographical location of the countries in cluster 1. These profiles are applied to the log-mortality rates for the 15-19 age group by mother's level of education, which are obtained via Bayesian log-linear modelling (for details, see Appendix A-2). Then, using the WIC and UN WPP data, we ensure that the results are coherent with the aggregated mortality rates. We obtain log-mortality rates for all 13 countries in cluster 1. 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 (see Figure 4).

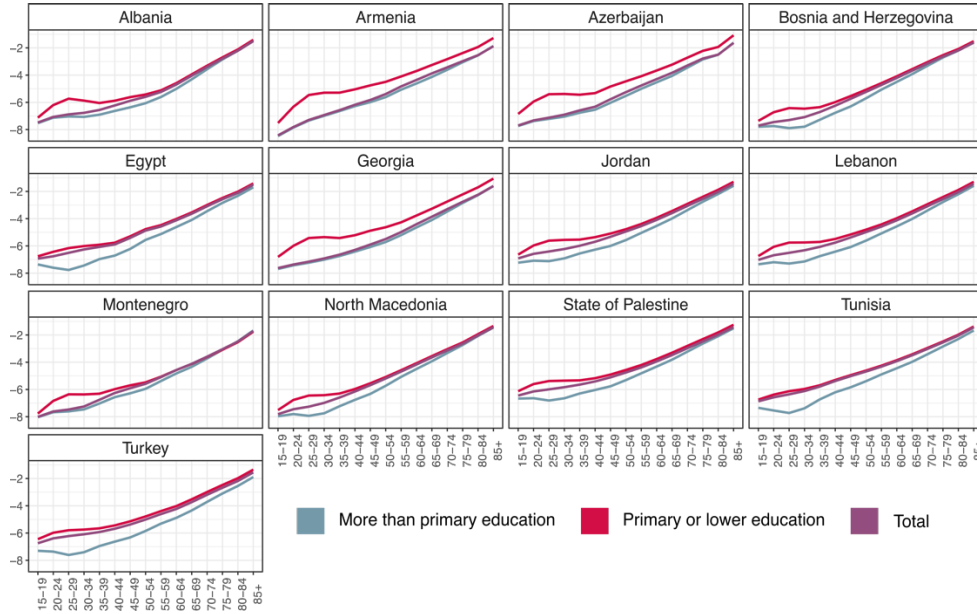
In Figure 5, we present the inputs for the year 1980 for all countries in our case study. As shown in the figure, our reconstruction method defines sets of mortality profiles that are country- and year-specific. Utilising the complete mortality schedules, we can harness the information provided by UN WPP as the foundational framework for our reconstruction. By displaying the total mortality rates in one figure, we demonstrate that our method effectively exploits the information on population size by period and level of education. As expected, the mortality profiles referring to the most represented level of education in the population are those that are most similar to the total profile (for population composition, see Figure A-11 in Appendix A-7). 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. We present details of the construction of $m_{a,t,c,e}^*$ in Appendix A-3.

³ 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 administration.

⁴ North: DNK, EST, FIN, NOR, SWE. South: ITA, GRC, PRT, MLT. Central East: BGR, HUN, POL, ROU, SVN, SVK. Central South: SRB, HRV, TUR.

2. $Y_{a,e,j}^*$: principal components extracted from the collection of mortality rates referring to the relevant time span and region. We utilise female population life tables from the UN WPP database for the case study countries for the 1980-2015 period. We combine these data with information from the WIC Data Explorer on mean years of schooling and primary education duration in years from UNESCO to separate the mortality curves into two groups. One group comprises year-country combinations in which the average years of schooling exceed the duration of primary school (country-specific), and the other 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 distinct 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 (Cluster 1) over a specified time interval (1980-2015) according to the estimated average level of education. Further details of their derivation can be found in Appendix A-4.
3. $\sigma_{a,e}^{UN*}$: standard deviations derived from the estimated distribution of age- and education-specific log-mortality rates obtained through the estimation steps performed for the $Y_{a,e,j}^*$ (Appendix A-4).
4. $\sigma_{t,c,e}^{DHS*}$: standard deviations obtained from the confidence intervals published by DHS concerning estimates of infant mortality by mother's level of education.
5. $\sigma_{t,c,e}^{LogLin*}$: standard deviations retrieved from the log-linear component of our model. This is used to derive starting points, i.e., mortality rates for the 15-19 age group by level of education and over time, and for the reconstruction over all years and countries in our case study (Cluster 1). The model is estimated within the Bayesian inferential framework, which permits us to learn about the uncertainty of the resulting estimates. Details on the model specification can be found in Appendix A-2.

FIGURE 5: EDUCATION-SPECIFIC LOG-MORTALITY RATES, CLUSTER 1, 1980, FEMALE POPULATION



Source:

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

4.2.3 CASE STUDY MODEL SPECIFICATION

In this chapter, we describe in detail the specification of the case study model, according to the model outlined in chapter 2. A graphical representation of the model, enriched with a full set of uncertainty measures derived from the inputs' reconstruction, is provided in Figure A-9 in the Appendix. In our case study, the age-, year-, country- and education-specific log-mortality rates $\log m_{a,c,t,e}^*$, that rely on information from other countries and various sources are constructed by applying the mortality profiles to the 15-19 log-mortality estimates, which are the result of a log-linear model implemented outside of the principal model. That is, we assume they are normally distributed (as in equation 1):

$$\log m_{a,t,c,e}^* \sim \mathcal{N}(\log m_{a,t,c,e}, \sigma_{t,c,e}^{LogLin*}) \quad (10)$$

where $\sigma_{t,c,e}^{LogLin*}$ is the standard deviation referring to the credible interval estimated with the aforementioned log-linear model. The inclusion of this uncertainty is necessary given the role played by the starting point in the reconstruction steps. The result of the log-linear model influences the development of the entire schedule.

The reconstructed mortality rates are then assumed to follow a normal distribution that borrows information from the countries that have reliable data through principal components (β and Y^*), random effects (u) capturing country-specific deviations from age-education profiles and period-education effects based on infant mortality by mother's education (v) derived from DHS:

$$\log m_{a,t,c,e} \sim \mathcal{N}\left(\sum_{j=1}^3 \beta_{a,t,e,j} * Y_{a,e,j}^* + u_{a,c,e} + v_{t,e}, \sigma_{a,e}^{UN*}\right) \quad (11)$$

Principal components have been derived for two education-specific mortality profiles separately and independently. Thus, the σ_{pcs}^* becomes $\sigma_{a,e}^{UN*}$. Random effects $v_{t,e} \sim \mathcal{N}(0, \mu_{\sigma_{t,e}}^{DHS})$, with $\mu_{\sigma_{t,e}}^{DHS}$ being the standard deviation associated with the DHS mortality estimates. These are informed by the standard deviations extracted from the DHS confidence intervals (Figure A-2) sourced from the STATcompiler website⁵:

$$\sigma_{t,c,e}^{DHS*} \sim \mathcal{N}\left(\mu_{\sigma_{t,e}}^{DHS}, \sigma_U^{DHS}\right) \quad (12)$$

The inclusion of $\mu_{\sigma_{t,e}}^{DHS}$ captures the uncertainty of the survey-based estimates (such as sampling or recall period errors). The hyperparameters of the prior in our case study ensure that practically, the resulting prior is positive. Priors for the parameters that capture variation in the DHS are weakly informative:

$$\mu_{\sigma_{t,e}}^{DHS} \sim \mathcal{U}[0,0.5] \quad (13)$$

$$\sigma_U^{DHS} \sim \mathcal{U}[0,1] \quad (14)$$

The rest of the model is structured in the same way as described in section 2. We sample from the posterior distributions of the model parameters 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.

5 RESULTS

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, State of Palestine, Tunisia and Turkey, which are the countries in our cluster 1 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 educational composition of the population (Figure A-12 in Appendix A-7). In greater detail, the coherence between the results and the overall mortality rates at the country level is achieved by incorporating the known information regarding the age-education composition of the population and the age-specific total mortality rates. These data are utilised to refine the estimated log-mortality rates employed as inputs in the model.

⁵ <https://www.statcompiler.com/en/>

FIGURE 6: LOG-MORTALITY RATES BY LEVEL OF EDUCATION, AGE GROUP AND TIME, ALBANIA, FEMALE POPULATION

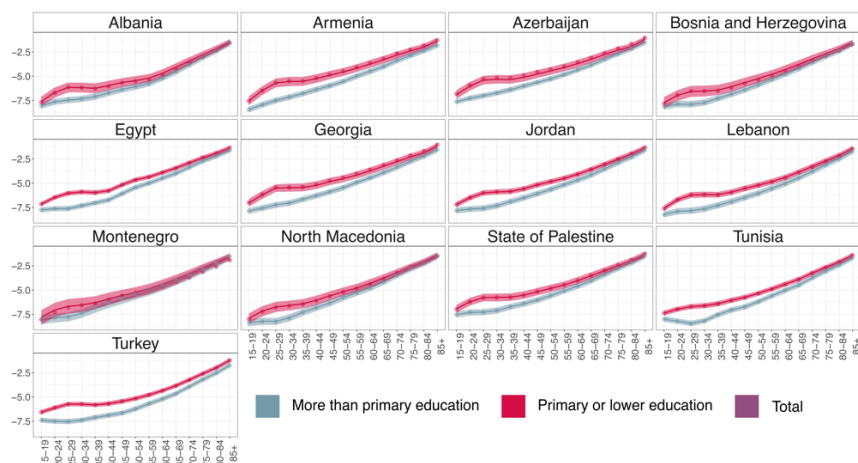


Remark:

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.

In Figure 7, we present the estimated mortality rates for all countries in cluster 1 for the year 2000. While the common features derived from the principal components are maintained, the mortality rates are differentiated for each country under consideration. We observe common features: 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 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 socio-economic 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.

FIGURE 7: LOG-MORTALITY RATES BY LEVEL OF EDUCATION, AGE GROUP AND TIME, 2000, FEMALE POPULATION



Remark:

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.

5.1 MODEL PERFORMANCE

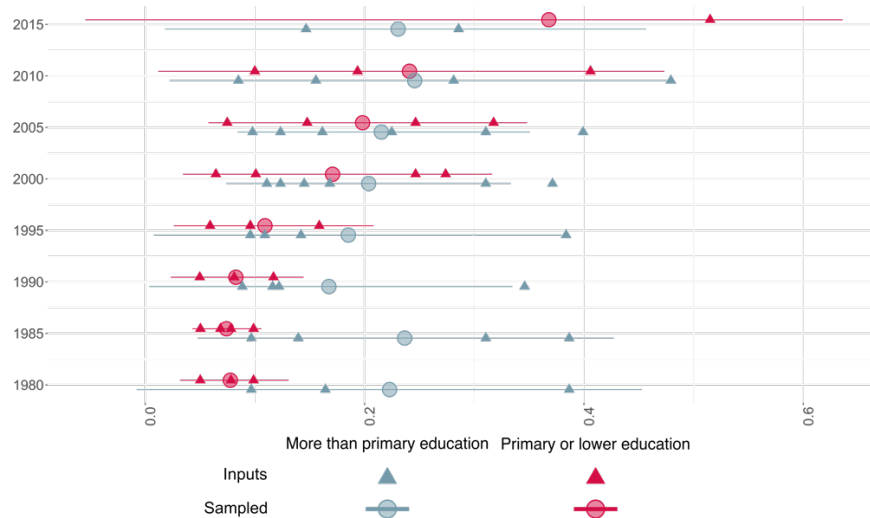
Due to the aforementioned lack of data, we do not have a gold standard against which we can evaluate our estimates. Furthermore, the measurement of goodness-of-fit is complicated by the fact that the inputs to the model are derived from a variety of data sources. Hence, we assess the performance of the 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 UN WPP.

5.1.1 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 a posterior predictive distribution (analogous to fitted values). For instance, in Figure 8 we present the PPDs for the $\mu_{\sigma_{t,e}}^{DHS}$ as defined in equation 12 for each year along the inputs (i.e., observed data). In the plot, the circular markers represent the median values of the posterior predictive distribution, 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, and subsequently affecting standard deviation. Conversely, higher education levels exhibit less susceptibility to this widening of values.

We observe that only 8% of the total available observations fall outside of the aggregated PPDs or a given year, all of them for the more than primary education level. This suggests that our model reproduces the inputs reasonably well.

FIGURE 8: STANDARD DEVIATION ASSOCIATED WITH THE DHS MORTALITY ESTIMATES ($\mu_{\sigma t,e}^{DHS}$)



Remark:

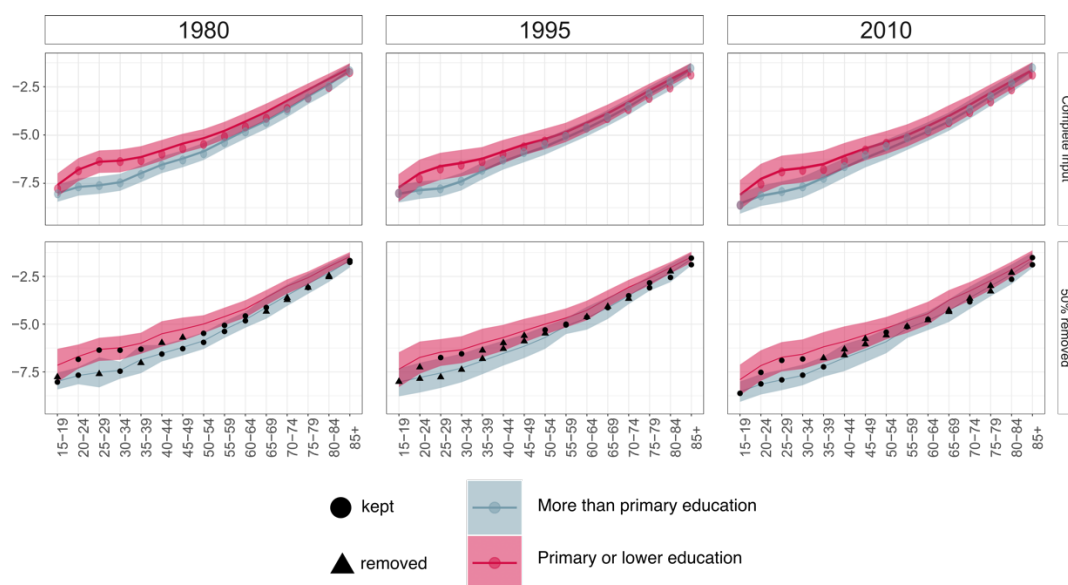
The solid lines present the 60% credible intervals. The circles represent the medians of the posterior predictive distributions. The triangles represent the inputs to our model.

5.1.2 RANDOMISED INPUT REDUCTION

Next, we randomly remove a portion of the data used to inform the age- and education-specific mortality rates. That is, we progressively remove 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 the imputation of missing information.

Figure 9 shows the mortality rates obtained for Montenegro when all the input data 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. ESTIMATES BASED ON 50% OF THE $\log m_{a,t,c,e}^*$ INPUTS REMOVED



Remark:

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.

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 1, we present the percentage of inputs contained in the 50% credible interval⁶ (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 97%, which suggests that our model underfits the inputs (i.e., the uncertainty is relatively large). However, the resulting uncertainty is based on the uncertainty about the inputs, such as confidence intervals of the DHS data. When the inputs are reduced at random, the width of the CIs increases but the coverage decreases, reaching 79% when three-quarters of the inputs are removed. This decrease is reasonably small, which suggests that the estimated mortality profiles are stable, and are not overly susceptible to even dramatic (removal of 75% of observations) changes in the inputs. This decline is relatively modest, which indicates that the estimated mortality profiles are stable, and are resilient to substantial changes in the inputs, even when as much as 75% of the observations are removed.

⁶ In the Bayesian context, utilising a narrower credible interval presents a more stringent and challenging assessment of the model's performance.

TABLE 1 ASSESSMENT OF MODEL PERFORMANCE: PERCENTAGE OF INPUTS FALLING IN THE 50% CREDIBLE INTERVAL (CI)

% Reduced inputs	% Contained in the 50% CI
0 (i.e., full model inputs)	96.9%
20	95.0%
50	87.9%
75	79.3%

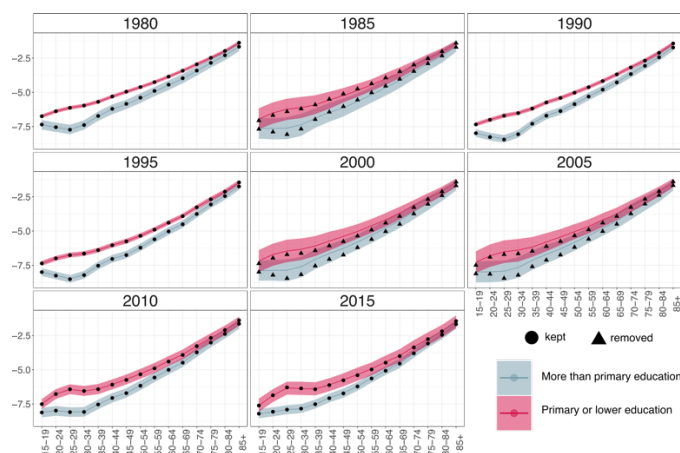
Note: The performance is calculated as the percentage of data falling into the 50% CI according to the amount of inputs employed.

5.1.3 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 Tunisia from a model with 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, TUNISIA, FEMALE POPULATION



Remark:

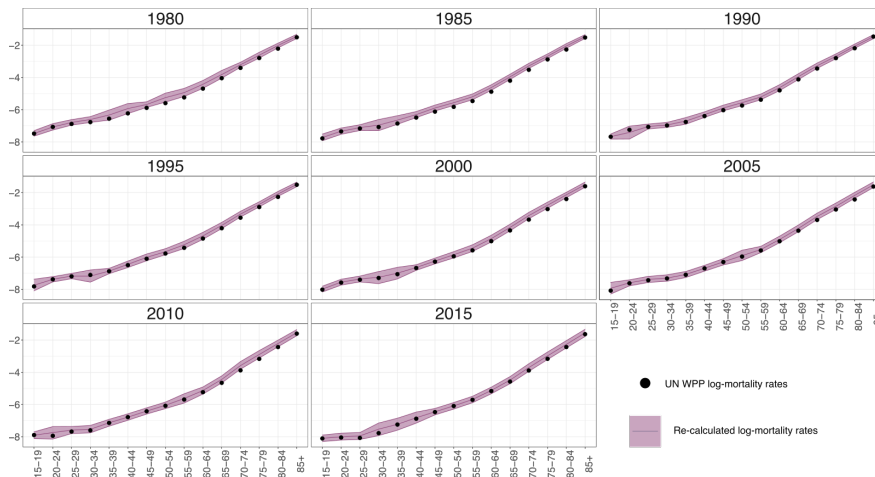
The solid lines present the estimated medians, while the 80% credible intervals are visualised via shaded areas. The dots represent the inputs to our model.

Note: Inputs for the years 1985, 2000 and 2005 were removed for Azerbaijan, Georgia, North Macedonia and Tunisia.

5.1.4 COHERENCY WITH TOTAL MORTALITY RATES

Another check addresses the coherency of the results with the (only) available data: namely, the overall mortality rates (i.e., not differentiating by education). These comparisons assess the consistency of our estimates with total mortality 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 compare the resulting total mortality rates with those published by 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 UNWPP



Remark:

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.


6 CONCLUSIONS

Our work makes a methodological contribution by proposing a modelling framework that includes a Bayesian hierarchical model and a mechanism for constructing inputs into it. 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 differentials in mortality by other characteristics, such as socio-economic 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 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.

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 DHS at irregular intervals. For other countries (like Bosnia and Herzegovina, Georgia, Lebanon, North Macedonia, Montenegro and 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 fill a gap in the data.

The results we presented reaffirm, consistent with the established literature, a well-documented relationship between education and mortality. Specifically, our analysis underscores that higher educational attainment is closely linked to reduced mortality rates. Notably, the differential 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 socio-economic 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. 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).

Remarkably, these differentials lessen in older age categories. 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 UN WPP. Second, 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 socio-economic 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.

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. https://EconPapers.repec.org/RePEc:spr:demogr:v:54:y:2017:i:6:d: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). doi:10.1214/14aoas768.
- 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.
- 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.
- Avison, W. (2005). Education, social status, and health by John Mirowsky and Catherine E. Ross: Education, social status, and health. *American Journal of Sociology* 110(5): 1511–1513. doi:10.1086/431614.
- Baker, D., Leon, J., Greenaway, E., Collins, J., and Movit, M. (2011). The education' effect on population health: A reassessment. *Population and Development Review* 37: 307–32. 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 Institute of Demography Working Papers* 03/2018, Vienna. 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(1): 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). doi:10.18637/jss.v076.i01
- Clark, D. and Royer, H. (2013). The effect of education on adult mortality and health: evidence from Britain. *American Economic Review* 103(6): 2087–2120. doi:10.1257/aer.103.6.2087.
- Cutler, D., Lange, F., Meara, E., Richards-Shubik, S., and Ruhm, C. (2011). Rising educational gradients in mortality: The role of behavioural risk factors. *Journal of Health Economics* 30: 1174–87. doi:10.1016/j.jhealeco.2011.06.009.

- Davey Smith, G., Hart, C., Hole, D., MacKinnon, P., Gillis, C., Watt, G., Blane, D., and Hawthorne, V. (1998). Education and occupational social class: which is the more important indicator of mortality risk? *Journal of Epidemiology & Community Health* 52(3): 153–160. doi:10.1136/jech.52.3.153.
- 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* (Wayne State University. Press) 55: 224–249. doi:10.1353/mpq.0.0030.
- 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: 191–204. doi:10.1080/14748460500372309.
- Eurostat (2021). Life expectancy by age, sex and educational attainment level. https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=demo_mlexpecedu&lang=en.
- Eurostat (2022). Deaths by age, sex and educational attainment level [demo maeduc]. https://ec.europa.eu/eurostat/web/products-datasets/-/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(1). 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. *Lancet* 376: 959–74. doi:10.1016/S0140-6736(10)61257-3.
- Gavurova, B., Vagasova, T., and Grof, M. (2017). Educational attainment and cardiovascular disease mortality in the Slovak republic. *Economics & Sociology* 10: 232–245. doi:10.14254/2071-789X.2017/10-1/17.
- 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.
- Gendronneau, C., Wiśniowski, A., Yildiz, D., Zagheni, E., Fiorio, L., Hsiao, Y., Stepanek, M., Weber, I., Abel, G., and Hoorens, S. (2019). Measuring labour mobility and migration using big data: exploring the potential of social-media data for measuring EU mobility flows and stocks of EU movers. Publications Office of the European Union.
- Goujon, A., K.C., S., Springer, M., Barakat, B., Potančoková, 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. <http://www.jstor.org/stable/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.
- Karlsen, S., Say, L., Souza, JP. et al. (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

- Keyfitz, N. (1980). Multistate demography and its data: A comment. *Environment and Planning A: Economy and Space* 12(5): 615–622. doi:10.1068/a120615.
- 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 metanalysis. *PLOS ONE* 14: 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): 1–13. doi:10.1371/journal.pone.0131809.
- Lee, R. (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. PMID: 11620664.
- 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. PMID: 11630559.
- Lee, R.D. and Carter, L.R. (1992). Modelling and forecasting u. s. mortality. *Journal of the American Statistical Association* 87(419): 659–671. <http://www.jstor.org/stable/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: 899–908. doi:10.2105/AJPH.2009.186916.
- Ludeke, S., Gensowski, M., Junge, S., Kirkpatrick, R., John, O., and Andersen, S. (2020). 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. 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.
- Lutz, W., Goujon, A., KC, S., Stonawski, M. and Stilianakis, N., (2018) Demographic and human capital scenarios for the 21st century: 2018 assessment for 201 countries. Publications Office of the European Union, Luxembourg. doi:10.2760/835878.
- Lutz, W. and Kebede, E. (2018). Education and health: Redrawing the Preston curve: Education and health. *Population and Development Review* 44. doi:10.1111/padr.12141.
- Luy, M., Giulio, P.D., and Caselli, G. (2011). Differences in life expectancy by education and occupation in Italy, 1980–94: Indirect estimates from maternal and paternal orphanhood. *Population Studies* 65(2): 137–155. doi:10.1080/00324728.2011.568192.
- 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/s41118019-0055-0.
- Malamud, O., Mitrut, A., and Pop-Eleches, C. (2018). The effect of education on mortality and health: Evidence from a schooling expansion in Romania. *Working Paper* 24341, National Bureau of Economic Research. doi:10.3386/w24341.
- 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* 0(0): 1–7. doi:10.1080/07481187.2019.1692970.

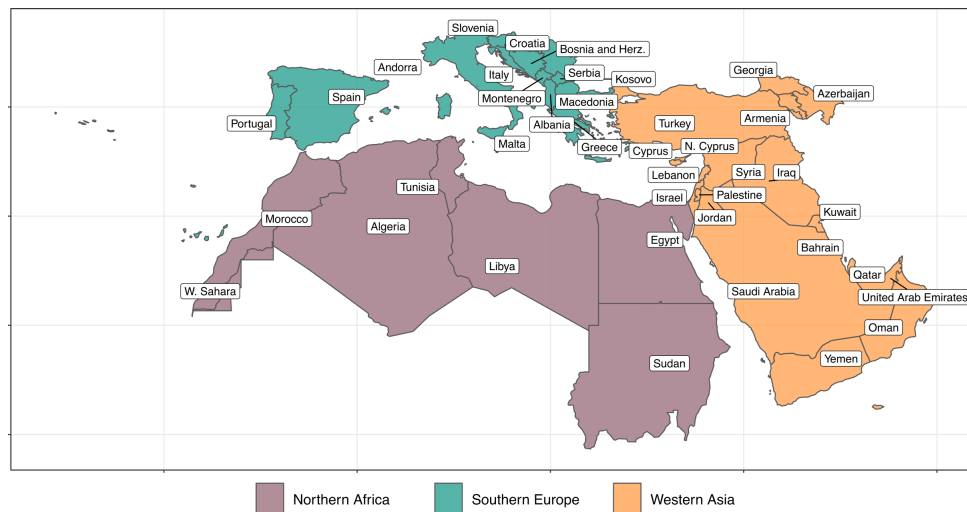
- Montez, J.K., Hummer, R.A., and Hayward, M.D. (2012). Educational Attainment and Adult Mortality in the United States: A Systematic Analysis of Functional Form. *Demography* 49(1): 315–336. doi:10.1007/s13524-011-0082-8.
- Murakami, K. and Hashimoto, H. (2019). Associations of education and income with heavy drinking and problem drinking among men: evidence from a population-based study in Japan. *BMC Public Health* 19(1). doi:10.1186/s12889-019-6790-5.
- Murtagh, F. and Legendre, P. (2011). Ward’s hierarchical clustering method: Clustering criterion and agglomerative algorithm. *arXiv*. doi: 10.48550/arXiv.1111.6285
- Nielsen, F. (2016). Introduction to HPC with MPI for Data Science. Springer.
- Plummer, M. (2003). Jags: A program for analysis of Bayesian graphical models using gibbs sampling. 3rd International Workshop on Distributed Statistical Computing (DSC 2003); Vienna, Austria 124.
- Pradhan, E., Suzuki, E., Martinez, S., Schaferhoff, M., and Jamison, D. (2017). The Effects of Education Quantity and Quality on Child and Adult Mortality: Their Magnitude and Their Value. Washington (DC): The International Bank for Reconstruction and Development / The World Bank, 423–440. doi:10.1596/978-1-4648-0423-6 ch30.
- Raghupathi, V. and Raghupathi, W. (2020). The influence of education on health: an empirical assessment of OECD countries for the period 1995–2015. *Archives of Public Health* 78: 20. doi:10.1186/s13690-020-00402-5.
- Raymer, J., Wiśniowski, A., Forster, J.J., Smith, P.W.F., and Bijak, J. (2013). Integrated modelling of European migration. *Journal of the American Statistical Association* 108(503): 801–819. doi:10.1080/01621459.2013.789435.
- Rogers, A. (1980). Introduction to multistate mathematical demography. *Environment and Planning A* 12(5): 489–498.
- Rosoff, D.B., Clarke, T.K., Adams, M.J., McIntosh, A.M., Smith, G.D., Jung, J., and Lohoff, F.W. (2019). Educational attainment impacts drinking behaviours and risk for alcohol dependence: results from a two-sample mendelian randomization study with ~780,000 participants. *Molecular Psychiatry* 26(4): 1119–1132. doi:10.1038/s41380019-0535-9.
- 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. (2019). Global reconstruction of educational attainment, 1950 to 2015: Methodology and assessment. *Vienna Institute of Demography Working Papers* 02/2019, Vienna. doi:10.1553/0x003cb434.
- Stan Development Team (2018a). RStan: the R interface to Stan. <http://mc-stan.org/9>. R package version 2.17.3.
- Stan Development Team (2018b). The Stan Core Library. <http://mc-stan.org/9>. Version 2.18.0.
- The World Bank (2022). Databank. <https://databank.worldbank.org/>.

- Tjepkema, M., Wilkins, R., and Long, A. (2012). Cause-specific mortality by education in Canada: A 16-year follow-up study. *Health reports / Statistics Canada, Canadian Centre for Health Information = Rapports sur la santé / Statistique Canada, Centre canadien d'information sur la santé* 23: 23–31.
- Tomioka, K., Kurumatani, N., and Saeki, K. (2020). The association between education and smoking prevalence, independent of occupation: A nationally representative survey in Japan. *Journal of Epidemiology* 30(3): 136–142. doi:10.2188/jea.je20180195.
- UNESCO Institute for Statistics (UIS) (2023). Sdg global and thematic indicators. <http://data.uis.unesco.org/>.
- United Nations (2022). World Population Prospects 2022. United Nations. <https://www.un-ilibrary.org/content/books/9789210014380>.
- United Nations, N.R.C.U. (1983). Indirect techniques for demographic estimation. United Nations New York.
- USAID (2022). Statcompiler: the DHS program. <https://www.statcompiler.com/en/>.
- Wheldon, M., Raftery, A., Clark, S., and Gerland, P. (2013). Reconstructing past populations with uncertainty from fragmentary data. *Journal of the American Statistical Association* 108: 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.
- 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.
- 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)* 185–202.
- Wiśniowski, A. (2021). Migration forecasting using new technology and methods. In: McAuliffe, M. (ed.). *Research Handbook on International Migration and Digital Technology*. Edward Elgar Publishing: 376–392.
- Wittgenstein Centre Data Explorer (WCDE) (2018). Wittgenstein centre for demography and global human capital (WIC). <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. <http://www.jstor.org/stable/2173980>.
- Zimmerman, E. and Woolf, S. (2014). Understanding the relationship between education and health. *NAM Perspectives* 4. doi:10.31478/201406a.

© 2006 The Authors
Journal compilation © 2006 Blackwell Publishing Ltd

A-1.1 THE GEOGRAPHICAL SETTING

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 socio-economic 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.



A-1.2 COUNTRIES CLUSTERING

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 socio-economic differences between the sub-regions.

To cluster the countries, we selected variables representing mortality, education and socio-economic 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 countries⁷ because of the excessive amount of missing data.

Mortality variables:

1. Cause of death, by injury (% of total);
2. Cause of death, by non-communicable 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); and
6. Survival rate from age 15-60.

Education and socio-economic 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 aged 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, ages 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 (%); and
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.

⁷ Syria, Gibraltar, San Marino, Andorra, Sudan, 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

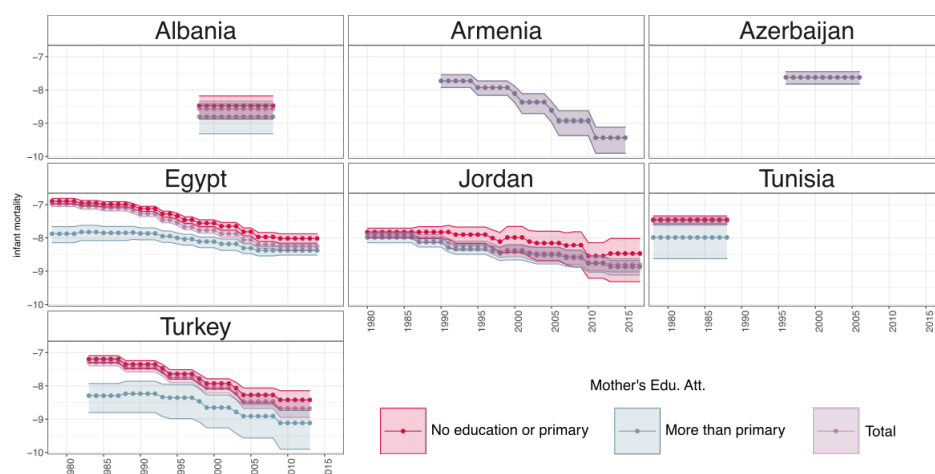
A-2 THE LOG-LINEAR MODEL

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 DHS, 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.

A-2.1 THE 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-4: 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).

A-2.2 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 DHS is linked to a region made up of countries that are covered by 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, 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

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

All models were analysed in terms of residual deviance and differences between fitted and observed values. Then, the best-performing models was:

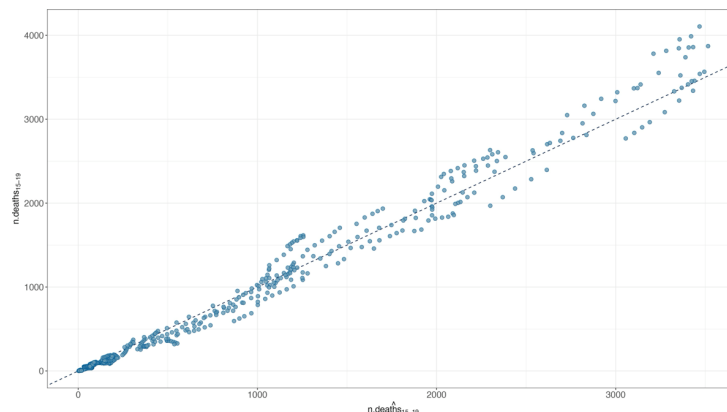
$$n.deaths \sim edu.att + region + sex + year + edu.att * region + region * sex \quad (15)$$

2. Bayesian framing of the best fitting model resulting from the model selection

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 UN WPP before they were used as starting values of the reconstruction.

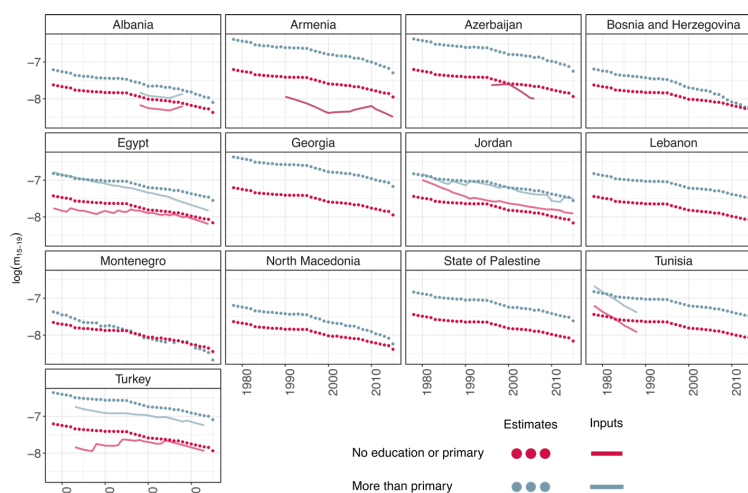
We implemented the model in R software, package *rstan* (Stan Development Team 2018a; Carpenter et al. 2017; Stan Development Team 2018b), which we used to sample from the posterior distribution. We used weakly informative normal priors. To check convergence, we relied on the Gelman and Rubin diagnostic (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 5000 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-2: LOGLIN 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-3: THE LOG-LINEAR MODEL RESULTS



Source: Own calculations based on DHS data.

A-3 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 Eurostat⁸, 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 A-2). By applying these profiles to the starting values and then correcting for possible discrepancies from 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.

A-3.1 15-19 MORTALITY RATES

As described in Appendix A-2, 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.

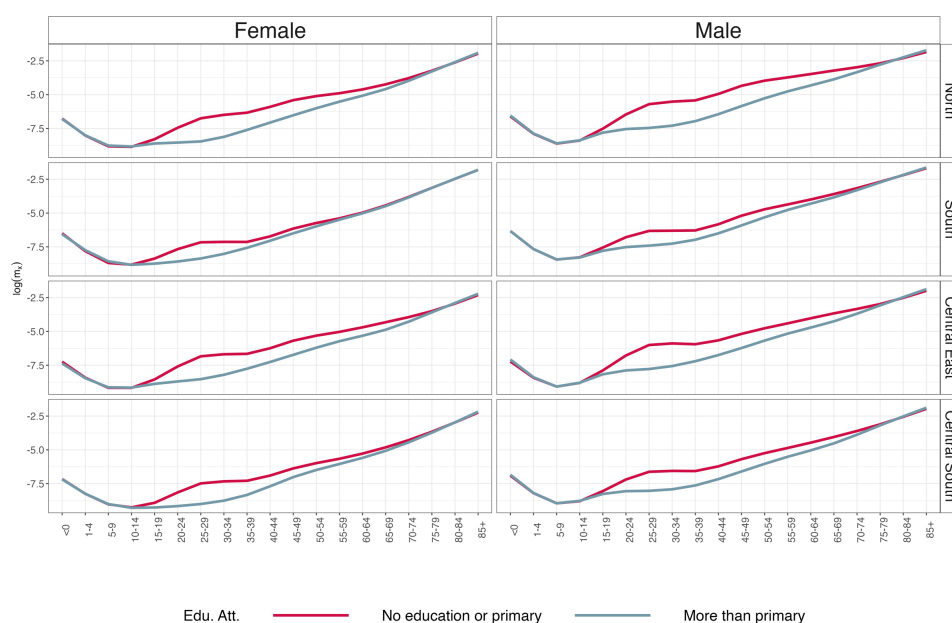
⁸ DNK, EST, FIN, NOR, SWE, ITA, GRC, PRT, MLT, BGR, HUN, POL, ROU, SVN, SVK, SRB, HRV, TUR

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. 2020; 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).

A-3.2 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 years⁹. By grouping the levels of education and the countries into four groups (Figure A-5), we identified profiles for European sub-regions over the 2007-2017 period for the two levels of education under consideration.

FIGURE A-5: THE GROUPED EUROPEAN SUB-REGIONAL MORTALITY PROFILES



⁹ 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.

Source: Own calculations based on Eurostat data (Eurostat 2021).

Note: 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; and
- Central South: SRB, HRV, TUR.

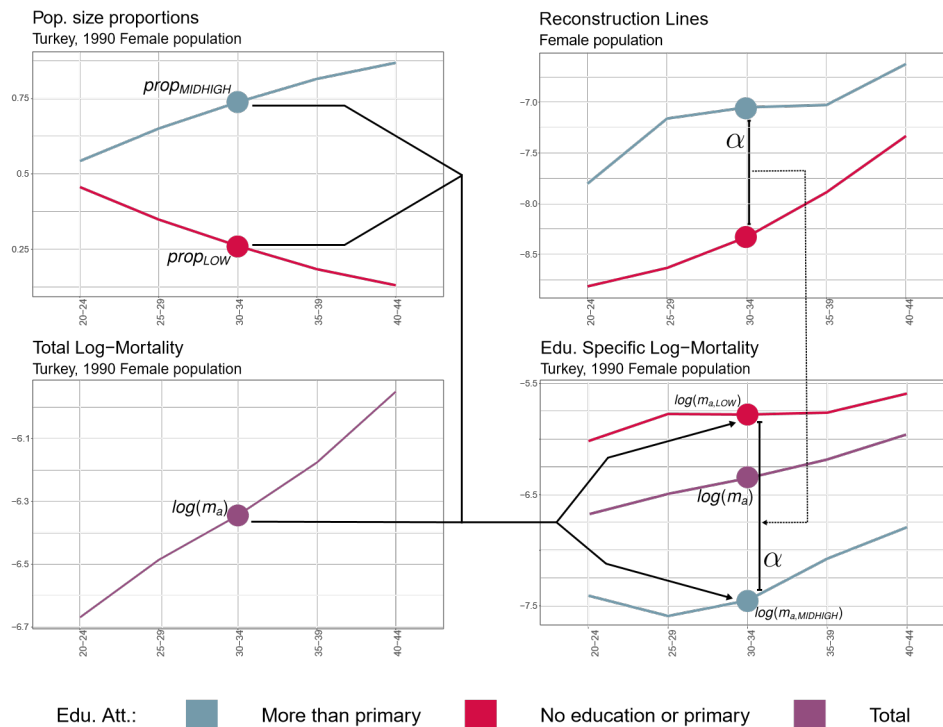
We then used ratios of the education-specific mortality profiles and education-specific population sizes to split the total mortality rates profile from UN WPP. The splitting operation ensures consistency with the total mortality rate and with the difference between the two different mortality levels.

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:

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:

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

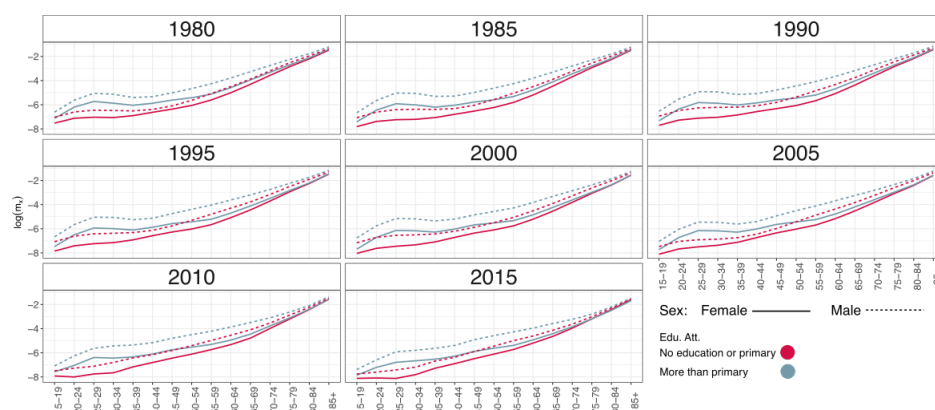
FIGURE A-6: THE CONSTRUCTION PROCEDURE



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

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 A-2) that allows for estimating mortality for countries not covered by the DHS.

FIGURE A-7: THE LOG-MORTALITY PROFILES ESTIMATED FOR ALBANIA



A-4 AGE-, SEX-, EDUCATION-SPECIFIC PRINCIPAL COMPONENTS

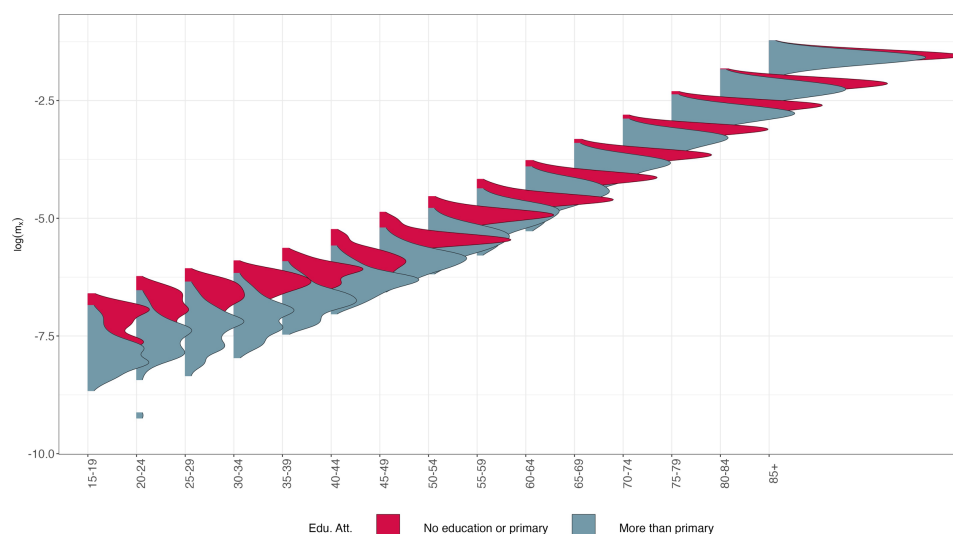
In this section, we describe in detail the construction of the sex-, age- and education-specific principal components. As in the work of Alexander, Zagheni and Barbieri (2017), these components are employed to represent the key characteristics, in terms of variation, of a family of mortality curves. 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:

- 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 the five-year period for the countries under consideration;
- 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
- 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 Centre 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 “less than primary” or “primary or more” 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 integer).

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

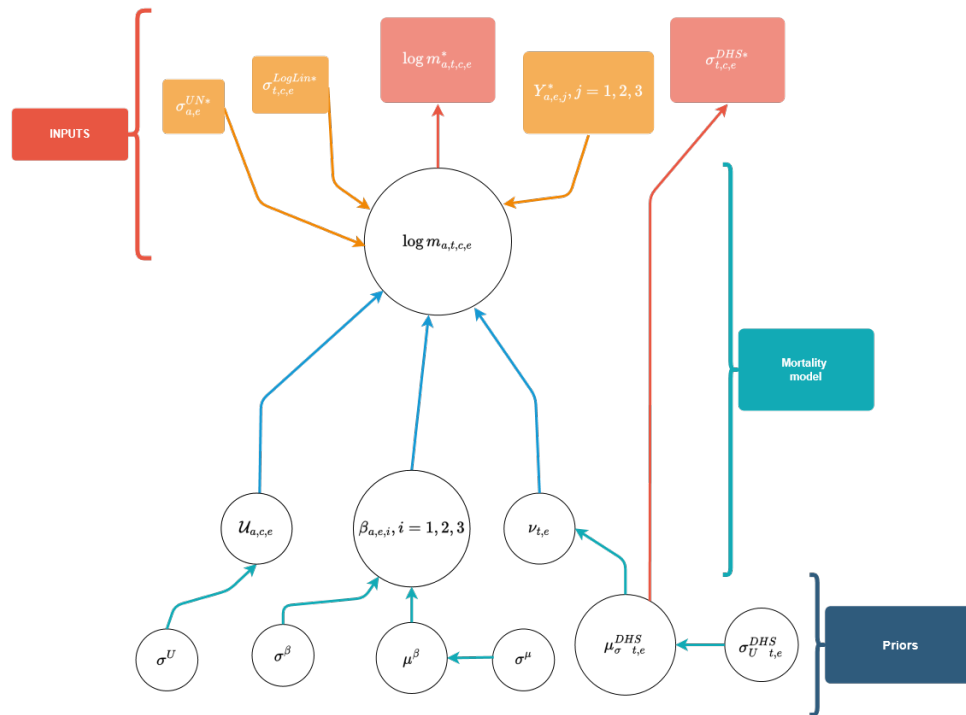


Source: Own calculations based on UN WPP, WIC and UNESCO data.

Note: The mortality curves referring to the countries in cluster 1 for the 1980-2015 period are related to the female population.

A-5 THE CASE STUDY MODEL FORMULATION

FIGURE A-9: THE MODEL: GRAPHICAL REPRESENTATION



Note:

Graphical notation

Circle: objects with a distribution.

Orange Square: quantities estimated outside of the model used as hyper-parameters.

Red Square: quantities estimated outside of the model used as data.

A-6 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, 2022, 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

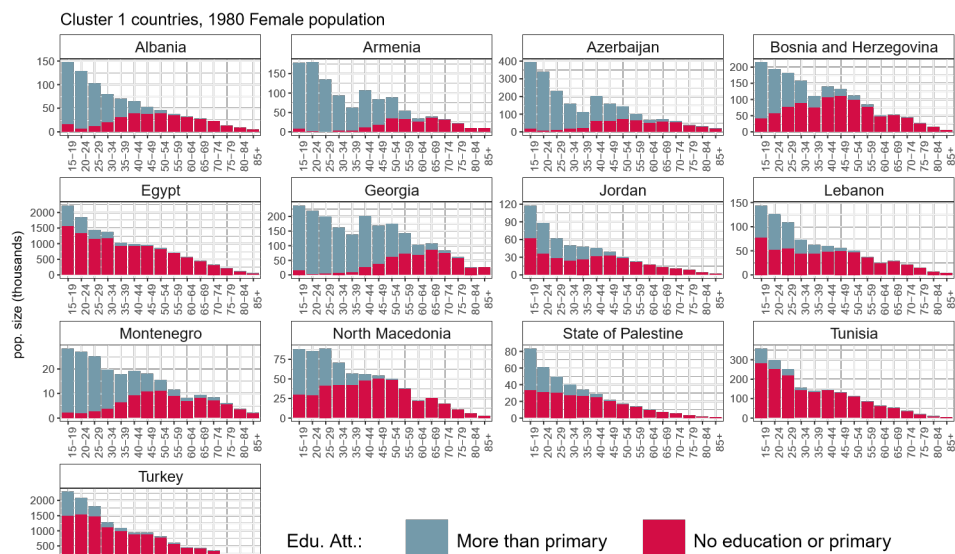
A-7 ADDITIONAL FIGURES

FIGURE A-10: LOG-MORATLITY RATES (80% C.I.). GEORGIA, FEMALE POPULATION



Note: Reporting results for the selected years 1985, 2000 and 2015. In the first row, we report the results from the model with full inputs. In the second row, we report the results from the model with 50% of the inputs removed (of the total amount of inputs), and in the last row, we report the results from the model for which all the inputs for Azerbaijan, Georgia, North Macedonia and Tunisia were removed.

FIGURE A-11: POP. SIZES



Source: WIC Data Explorer.

FIGURE A-12: POP. SIZES



Source: WIC Data Explorer.



Wittgenstein Centre

FOR DEMOGRAPHY AND
GLOBAL HUMAN CAPITAL



universität
wien

The Wittgenstein Centre is a collaboration among the Austrian Academy of Sciences (ÖAW), the International Institute for Applied Systems Analysis (IIASA) and the University of Vienna.

www.wittgensteincentre.org