International Institute for Applied Systems Analysis
IIASA www.iiasa.ac.at

Address: IIASA, Schlossplatz 1, A-2361 Laxenburg, Austria
Email: permissions@iiasa.ac.at

Working paper

Migration flows by age, sex and educational attainment

Dilek Yildiz [1*] (yildiz@iiasa.ac.at)
Guy Abel [1,2] (abel@iiasa.ac.at)

[1] Population and Just Societies (POPJUS) Program, International Institute for Applied Systems Analysis (IIASA), Laxenburg, Austria.
[2] Department of Sociology, Faculty of Social Sciences, University of Hong Kong, Hong Kong.

*Corresponding author

WP-24-001

Approved by:

Anne Goujon
Program: Population and Just Societies (POPJUS) Program
Date: 9 January 2024

Table of contents

ZVR 524808900

Disclaimer, funding acknowledgment, and copyright information:
IIASA Working Papers report on research carried out at IIASA and have received only limited review. Views or opinions expressed herein do not necessarily represent those of the institute, its National Member Organizations, or other organizations supporting the work.

The authors gratefully acknowledge funding from European Union's H2020 Societal Challenges Research and Innovation Programme, for the project 'Future Migration Scenarios for Europe (FUME)' [grant agreement number 870649] and European Union's Horizon Europe Research and Innovation Programme for the project `Policy Recommendations to Maximise the beneficial Impact of Unexplored Mobilities in and beyond the European Union (Premium_EU)' [grant agreement number 101094345].

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License**.**
For any commercial use please contact permissions@iiasa.ac.at

Abstract

The 2013 and 2018 Wittgenstein Centre (WIC) global population projections by age, sex, and level of education considered differential fertility and mortality by educational attainment. However, the educational composition of migrants was not explicitly included in the projections. One of the main differences in the 2023 update is the projection of future immigration and emigration flows by age, sex, and educational attainment. In this paper, we outline the methodology used to estimate the proportions of immigration and emigration flows between 1990 and 2020 by four education levels and 16 age groups for 183 countries. These proportions are used to project age, sex and education specific migration flows to 2100.

Acknowledgments

The authors gratefully acknowledge funding from European Union's H2020 Societal Challenges Research and Innovation Programme, for the project 'Future Migration Scenarios for Europe (FUME)' [grant agreement number 870649] and European Union's Horizon Europe Research and Innovation Programme for the project `Policy Recommendations to Maximise the beneficial Impact of Unexplored Mobilities in and beyond the European Union (Premium_EU)' [grant agreement number 101094345].
The author would also like to express their gratitude to Samir KC for his contributions to this research.

Background

The first global population projections of educational attainment broken down by age and sex covering 195 countries were published in 2014 (WIC2013) (Lutz, Butz, and KC 2014). Since its publication, two major updates of the projections have taken place: Lutz et al. (2018) available in WIC2018 and WIC2023 (KC et al. 2023). WIC2023 follows a similar methodology as its predecessors with updated baseline structure and revised assumptions on demographic rates and educational progression to project future population projections under the Shared Socioeconomic Pathways (SSP). Similar to WIC2013 and WIC2018, in the new set of assumptions (WIC2023), country-specific total migration flow rates for the medium migration scenario are based on average historical migration rates and different future migration scenarios are employed for different SSP scenarios. One of the main differences in the methodology is the age, sex and education breakdown of future migration flows. In this paper, we outline the methodology used to estimate the proportions of immigration and emigration flows between 1990 and 2020 by four education levels, 16 age groups, and sex for 183 countries. We use these proportions to break down existing estimates of immigration and emigration flows by sex Abel and Cohen (2022). These flows are then used to calculate the age, sex and education specific immigration and emigration flows for the WIC2023 migration projections.

In the Global North/post-demographic transition countries, where fertility and mortality rates have been declining, and population has stabilised or is facing decline, migration plays a key role in the change of population size, structure, and characteristics. Especially, due to the 2015 refugee crisis following the civil war in Syria and concerns about climate migration, the past decade has seen a growing interest in quantitative research estimating and predicting migration flows. The education level of migrants holds significant importance due to its multifaceted impact on both the migrants themselves and the host society in terms of economic contributions, employment, ability to adapt to and integrate into the host society). However, there is a lack of global, comparative, and good quality migration flow data stratified by age and education. The majority of research that includes the characteristics of migrants has been restricted to migrant stocks, which are easier to collect than migration flows (Artuc et al. 2015). Availability of migration flows broken down by age is limited to high-income countries and only available for recent years (from 1990 onwards). Therefore, a formal modelling approach is required to produce global estimates of flows broken down by age, sex and education.

Statistical and economic models of migration typically draw upon the extensive literature on (international) migration theories, including push-pull models, neoclassical migration theory, dual labor market theory, and human capital theory (a detailed discussion on the determinants of migration and migration theories can be found in de Haas (2011) and de Haas et al. (2020)). Our approach is also grounded in theories and modelling of migration. In this paper, whenever available, we utilise the determinants of migration, such as the size of the diaspora in a destination country and its Gross Domestic Product (GDP), for our prediction models. However, the requirements of the WIC2023 global projections limit our selection of variables to the indicators available from global databases providing information for a large number of countries over time. Further, we utilise Rogers and Castro (1983) migration age schedules, a demographic method developed to handle missing data, smooth age distributions and generalise migration patterns for different population subgroups. It is important to note that the aim of our methodology is not to estimate the size or direction of migration, but to estimate the age and education distribution of sex specific immigration and emigration flows from Abel and Cohen (2022).

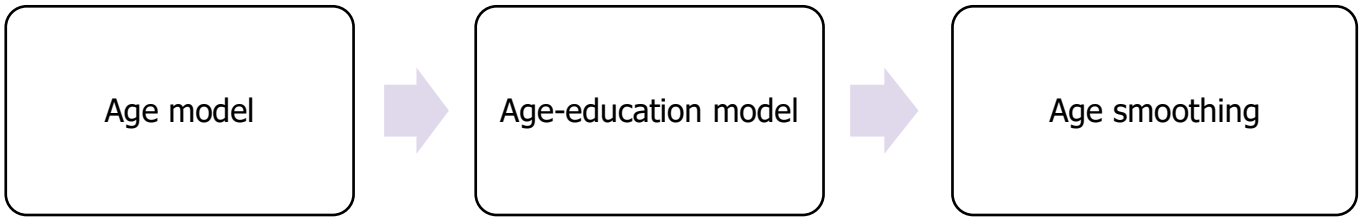
The relationship between age, education, and migration

The relationship between life course, hence age, and migration behaviour is well studied since Rogers and Castro (1981) published their seminal paper on migration age schedules (Preston, Heuveline, and Guillot 2005). The model migration age schedules are assumed to have four main components: labour force component, pre-labour force component, post-labour force component and constant component, where the highest migration rates are expected in the labour force and pre-labour force ages. The relationship between education and migration is more complicated. Human capital theory (Sjaastad 1962 in Haas, Castles, and Miller 2020) argues that the expected gain from migrating differs according to people's skills, abilities, age and gender. Haas, Castles, and Miller (2020) consider these differences as a potential reason of the 'selectivity' of migration based on skill and education levels that is evident in long established migration corridors in which low-skilled migrants move short distances compared to high-skilled migrants. This view is supported by Kerr et al. (2016) who outlined the "global talent mobility" and showed that high-skilled migrants are migrating to a small group of destination countries while their origin countries vary greatly.

In this paper, we explain how we estimate the proportions of male and female immigrants by age group (the age model), and age and educational attainment (the age-education model) at destination countries using random forest models (Breiman 2001). As shown in Figure 1, the estimation approach consists of three levels. The first level predicts male and female migration flow proportions by age groups. The second level further breaks down the estimates by four education levels. The last level involves smoothing the age distribution using Rogers-Castro migration age schedules (Rogers and Castro 1981). Male and female emigration proportions are derived from immigration estimates.

Details of the data sources and estimation process are outlined in the next and third section, respectively. The remaining part of the paper is organised in the following way. The results are presented in the fourth section, including a validation exercise; and the paper finishes with a conclusion and limitations section.

Figure 1: Modelling framework



Data Sources and variables

This paper uses several data sources to predict proportions of male and female immigration flows by age group and education. The main data source is the Integrated Public Use Microdata Series (IPUMS International, 2020) from which we obtain the observed proportions of immigration flows by age and education for each sex, the dependent variable. Other data sources include population size of the destination country by age, sex and education from WIC2018, several databases from the United Nations (UN)

organisations and estimates of bilateral migration flows by sex (Abel and Cohen 2022). The predictors for the models include age group (of migrants and of the population in destination country), sex (of migrants and of the population in destination country), period, educational attainment, migration interval (duration since migration, for more details see next section), country, share of migrants in the destination country, Human Development Index (HDI), Gross domestic product (GDP) per capita, share of illiterate population, life expectancy, old age dependency ratio, population size and proportion of population by age group and sex. The variables and their sources are explained in detail in the following subsections.

Our methodology consists of two random forest regression models, as mentioned above: one model to estimate the age-group proportions (age model) and one to break the estimates further down to education levels (age-education model). The age model excludes the education breakdown in the dependent and predictor variables and only focuses on age specific distributions. The details of data sources are explained in the following subsections.

Dependent variable

The IPUMS International database (2020) provides the world's largest archive of publicly available census microdata samples. Among demographic and socio-economic variables, censuses also collect information on internal and international migration. In terms of international migration, censuses can only directly measure immigration because it collects information in destination. Hence, in this paper we use characteristics of immigrants collected from the IPUMS International database from censuses conducted between 1970 and 2016 in 65 countries to construct the base data for the age and age-education prediction models.

One draw back of migration data in censuses is the inconsistencies in measuring migration. Countries use different migration intervals (duration since migration) to define migrants based on their administrative data needs. While many countries use the 12-month (one year) definition, some countries use five years, and some collect data on duration since migration. This inconsistency is also evident in IPUMS International data in which migration intervals range between 1 and 11 years. Therefore, the migration interval is also used as a predictor variable in our models.

Our dependent variable, $p_{t,c,s,a,e}$ ($p_{t,c,s,a}$in the age model) denotes the proportion of immigrants in destination country *c*, at each five-year period *t* for sex *s* measured by the migration interval *i* by age group *a*, and educational attainment level *e*. In other words, these proportions sum up to 1 in each five-year period, for each sex, country and for each migration interval as shown in Equation 1 for the age-model and in Equation 2 for the age-education model below.

$$\sum_{a=0-14}^{A=75+} p_{t=T,c=C,s=S,a} = 1 \quad (1)$$

$$\sum_{\substack{a=0-14,\\ e=NE}}^{A=75+,E=PSE} p_{t=T,c=C,s=S,a,e} = 1 \quad (2)$$

Below we provide a detailed explanation of the indices and their categories.

t: Five-year periods starting from 1970-1974 until 2015-2019. The last period only includes censuses conducted in 2015 and 2016.

c: ISO Country code.

s: Male, female.

a: Age groups, 0-14, 15-19, …, 75+. Immigrants younger than 15 years of age are grouped in one broad age group due to small number of observations at young ages, and then disaggregated to five-year age groups before age smoothing (see Methodology).

e: Information on education is collected in four categories as follows: less than primary completed, primary completed, secondary completed and university completed. We match these categories with the WIC educational attainment categories in predictor variables: no education (NE), primary (PE), secondary (SE) and post-secondary (PSE). Because a large proportion of the population are still at school we appoint educational attainment only to the population above age 14. For the younger age groups we use "Under 15" category. Similarly, we assume that post-secondary education is only completed from age group 20-24 onwards.

i: Migration interval in years: 1, 5, 6, 7, 8, 9, 10 and 11.

Predictor variables

Predictor variables for the random forest models were collected from several data sources and databases which provide estimates and indicators globally or for a large number of countries. The training and testing datasets (see Figure 2) use information starting from 1970 until 2020 for five-year periods to build a random forest model. Then employing this prediction model and predictor variables between 1990 and 2020 the immigration proportions are predicted for missing periods and countries. In case of missingness in predictor variables interpolation and extrapolation techniques were used. The sources for the predictor variables and data preparation are explained in detail below.

Age groups: The age model includes age groups as predictor variable. The age smoothing level further disaggregates the first 0-14 broad age group into three five-year age groups: 0-4, 5-9 and 10-14.

Share of migrants: To construct the prediction models, we use the share of migrants calculated using the IPUMS International data on total population and migrant size starting from 1970. In the prediction model,

international migrant stocks, as a percentage of the total population, are published by the United Nations in five-year periods between 1990 and 2020 are used (UN DESA 2020).

Human Development Index (HDI): HDI is a composite measure of average human development in three dimensions: health, education and standard of living (UNDP 2022). In this paper we use the datasets available in Our World in Data website (OWID 2023, Roser 2014). Two datasets, Human Development Index (HDI) by United Nations Development Programme (UNDP 2022) and Historical Index of Human Development (HIHD) by de la Escosura 2018 were downloaded using the *owidR* R package (York 2023).
The HDI dataset from UNDP includes single year estimates between 1990 and 2021. However, to build the model using IPUMS International data HDI estimates from 1970 are required. The HDID dataset goes back in time until 1870, however it does not include single year estimates. Therefore, the following imputation process was used to create a dataset for single years between 1970 and 2020.

- Whenever available HDI from UNDP (2022) are used.
- When HDI are missing a linear regression prediction model is employed incorporating HIHD and year as predictor variables.
- For the remaining missing values na.approx() function in R zoo package (Zeileis et al. 2023) is used for interpolation and extrapolation.

Gross Domestic Product (GDP): The GDP (constant 2010 US$) dataset was downloaded from UN DESA (2022) and World Bank (World Bank 2022). UN DESA provides GDP values starting from 1970 for single years. Similarly, World Bank provides GDP values starting from 1960 for single years until 2020. However, the data sets are not complete and some adjustments are required to build a prediction model. Therefore, we use the following data preparation steps.

- Whenever available we use UN DESA (2022) GDP.
- If UN DESA GDP values are missing, we use World Bank (2022) GDP.
- The remaining missing values are imputed by interpolation or extrapolation for each country.

Illiteracy rate by sex: Five-year period adult illiteracy rate by sex as a percentage of the total population for 185 countries was obtained from Reiter et al. (Reiter et al. 2021), who combined data from UNESCO Institute for Statistics (2020) and The World Bank (2020).

Life expectancy at birth by sex (e_0): Life expectancy at birth by sex and for 5-year period was extracted from the Wittgenstein Centre Data Explorer (Wittgenstein Centre for Demography and Global Human Capital 2018)[1]

Old age dependency ratio (ODR): The ratio of population above age 65 over the working age (15-64) population from Wittgenstein Centre Data Explorer (Wittgenstein Centre for Demography and Global Human Capital 2018)(see footnote 1).

[1] The historical data in the WIC data explorer relies on the UN world population prospects (2017), and projections for the 2015-2020 period (using the SSP2 – middle of the road scenario).

Population size: Logarithm of the size of population by age and education from Wittgenstein Centre Data Explorer (Wittgenstein Centre for Demography and Global Human Capital 2018). Additionally, population in the next and previous age group (in the same education level) are inputs to the model (see footnote 1).

Proportion of population: Proportions were calculated using size of population by age and education from Wittgenstein Centre Data Explorer (Wittgenstein Centre for Demography and Global Human Capital 2018). Additionally, proportion of population in the next and previous age group (within the same education level) are inputs to the model.

Methodology

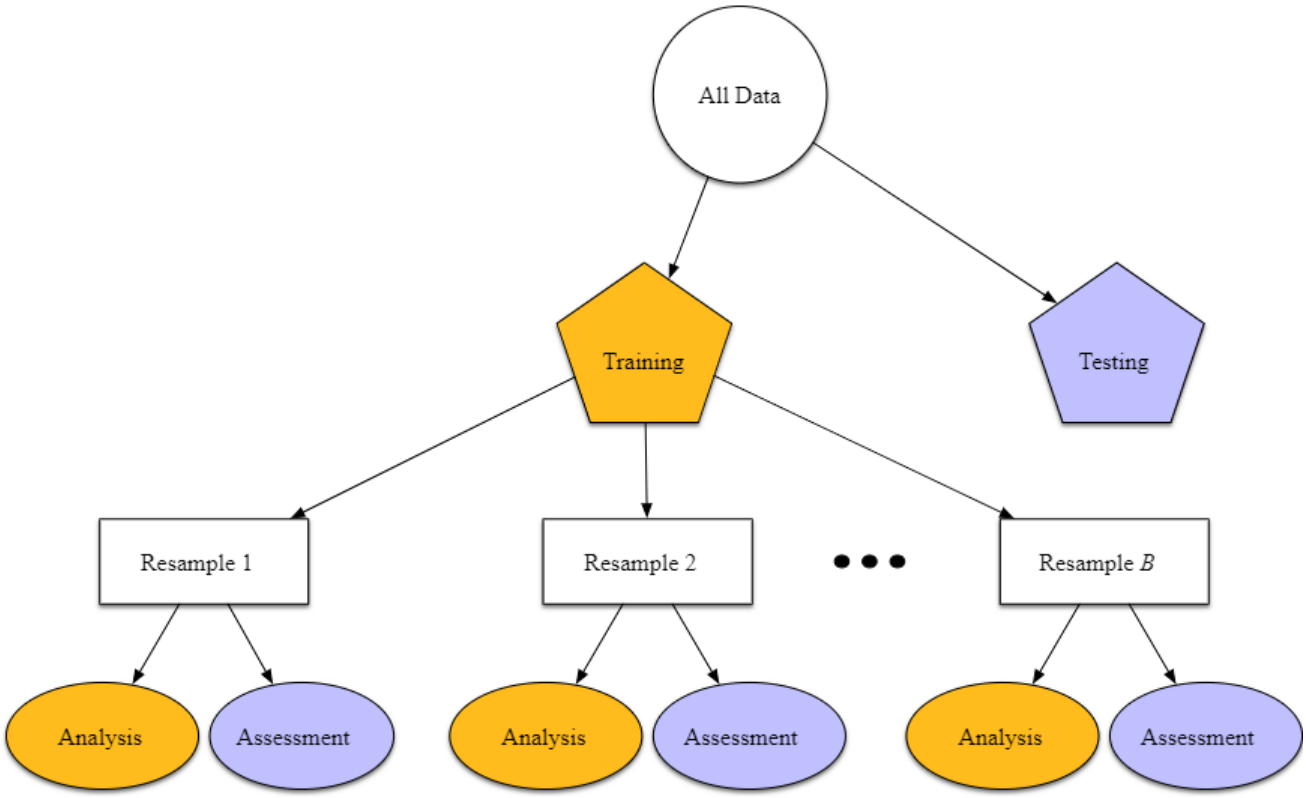
As mentioned in the Background section and shown in Figure 1 we designed a three-level modelling strategy to estimate proportions of immigration flows by age group and education between 1990 and 2019. In order to prevent issues that may arise from small sample sizes when disaggregated to education levels and calculate more reliable age distribution estimates, we start with the age model in which we estimate the proportions of immigration flows for each age group in each period and country for males and females. A random forest model is used to estimate the proportions for 183 countries. In the second step, we use a second random forest model to further break down the proportions at each age group by educational attainment. In the last step, we smooth the age distributions according to Roger-Castro age schedules. In the final step, the sex-specific emigration flows by Abel and Cohen (2022) are disaggregated to match corresponding estimated age and education proportions at the global level for both male and female migration flows.

Random forest is a machine learning technique that stems from bootstrap aggregation and similarly aims at reducing the variance while having a relatively lower bias (Hastie, Tibshirani, and Friedman 2009). It is often used to improve the accuracy and robustness of prediction models. More generally, machine learning is a collection of methods that the computer, "machine", automatically learns the patterns in data and generates a statistical model without a theoretical background. In this paper, random forest regression models are employed to make predictions by ensemble learning, i.e. combining predictions from multiple supervised machine learning algorithms. It is a supervised algorithm as dependent variables are defined and the models are used to predict them, rather than only focusing on identifying patterns in data using an unsupervised algorithm. In the random forest approach, each tree in the forest consists of random subsets of the data for prediction (see Figure 2). During the training phase, the algorithm builds an ensemble of multiple trees on training data, which are used to make predictions on new unseen data. The final prediction is obtained by averaging the predictions of all the individual trees. Using multiple trees instead of a single tree and dividing the data to training and test data reduces the overfitting and improves the model performance. It is argued that in complex analysis machine learning approaches may perform better in predictive models compared to standard regression models (Hindman 2015). Researchers showed that this advantage also holds for migration research (Best et al. 2021; Best et al. 2022). One drawback of machine learning practices in regression analysis is the complexity of models and their interpretation. However, the aim of this paper is to build a predictive model for missing countries and years with a limited number of covariates, rather than to build explanatory or causal relationships.

In our modelling framework, we first train the random forest models with the training dataset (75% of the IPUMS International data starting from 1970) and test the model performance with the test dataset (the

remaining 25% of the data of the IPUMS International data starting from 1970). Then we use a new dataset consisting of the predictor variables mentioned above to predict our dependent variable (starting from 1990). In other words, the regression models are built using IPUMS International data from 1970 and the predictions are made from 1990. The reason for this difference is that predictor variables such as migrant share are available for the countries in IPUMS International data from 1970s but globally they are only available from 1990. We use V-fold cross validation with five different resamples of the training data set (M. Kuhn and Johnson 2019). These five resamples are then divided into two data sets as shown in Figure 2, similar to the training and test data sets, to train the model (analysis data set) and assess the results (assessment data set). The results of model validation is presented in the next subsections. We use the *tidymodels* package (Max Kuhn and Wickham 2022) in R software to fit random forest models, and use similar terminology.

Figure 2: A diagram of resamples of the training data used to build random forest models



Source: Figure 3.5 in Kuhn and Johnson (2019)

Age model

As mentioned above, in the first level, we use a random forest model to predict the proportions of immigration flows at each age group starting from the 0-14 broad age group. The predictors for the age model are the variables listed in the previous section. It is important to note that, the predictor variables, population and proportion of population, in this model are only broken down by age and sex, not by educational attainment. In the input data processing stage, population counts are transformed to the log

scale and dummy variables are created from the categorical variables. Additionally, in this step we divide the input data into training and test data, 75% and 25% respectively.

After the data preparation, a random forest model with predictor variables as shown in Table 1, is fit to the training data set, and the metrics for the best model are evaluated using the test data set (Rsq = 0.884). Observed and predicted proportions of immigration flows by age group are compared in Figure 3. Overall, the figure shows a good model fit with most of the observed-predicted pairs lying around the 45-degree line. The highest difference was measured for the 0-14 age group in El Salvador in 1992, in which the observed proportions from IPUMS International (0.57) are unexpectedly high for this age group.

Unlike classical regression models random forest models do not have coefficients or a final model selection. Instead, measures of importance are used to identify the variables' predictive power. Figure 4 presents the visualisation of the variable importance scores for the top features in age model. Variable importance scores for each predictor are calculated based on their impact on the model improvement over the entire forest. In random forest regression models this impact is measured by the decrease a predictor is causing in the variance of the outcome (Hastie, Tibshirani, and Friedman 2009; K. Best et al. 2022). As expected, age groups and variables related to the age structure of the population are the most important variables. These variables are followed by share of migrant population (diaspora), illiteracy rate, GDP, and HDI. Size of the population, life expectancy at birth, period, old age dependency ratio and sex were found less important to predict the age distribution of migration flows. Once the model fit was checked using the training and the test datasets, proportions were estimated for 183 countries using the new dataset with predictor variables.

Table 1: Age model predictor variables

Predictor variables	
Period	t
Country	c
Sex	s
Age group	a
Population size of the destination country by age group, sex and period	$\log(Pop_{t,c,s,a})$
Population at the destination country by previous age group, sex and period	$\log(Pop_{t,c,s,a-1})$
Population at the destination country by previous age group, sex and period	$\log(Pop_{t,c,s,a+1})$
Proportion of the population in destination country by age group, sex and period	$prop_{t,c,s,a}^{POP}$
Proportion of the population in destination country by previous age group, sex and period	$prop_{t,c,s,a-1}^{POP}$
Proportion of the population in destination country by next age group, sex and period	$prop_{t,c,s,a+1}^{POP}$
Proportion of the migrant population in destination country by period	$prop_{t,c}^{MIG}$
Human Development Index of the destination country by period	$HDI_{t,c}$
Gross Domestic Product in destination country by period	$GDP_{t,c}$
Illiteracy rate in destination country by sex and period	$prop_{t,c,s}^{Illiterate}$
Life expectancy at birth in destination country by sex and period	$e_{0\ t,c,s}$
Old age dependency ratio in destination country by period	$ODR_{t,c}$

Figure 3 Predicted vs observed proportions for age model

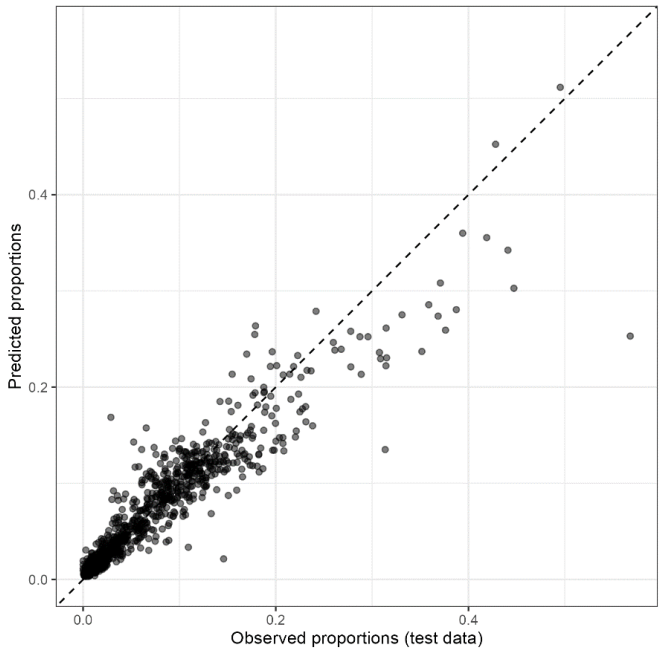
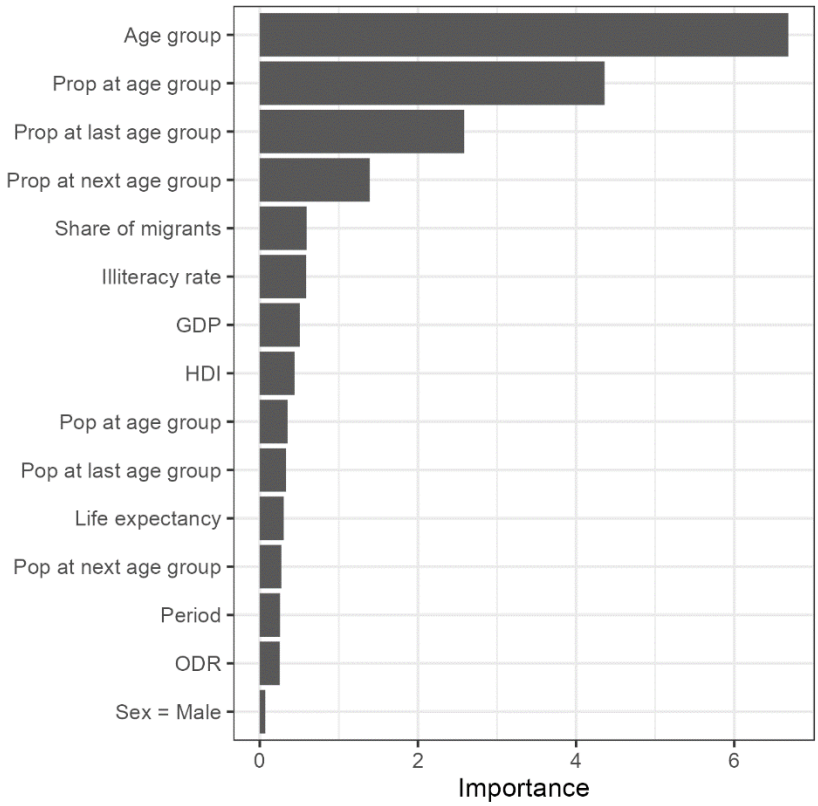


Figure 4 Variable importance scores for age model



The age categories for the predicted immigration proportions started with the broad 0-14 age group. To disaggregate this broad age group to three five-year age groups, we use the penalised composite link model (PCLM). PCLM is a model for count data that assumes that the counts in bins or in broad age groups follow a Poisson distribution of an ungrouped latent sequence. The age-model estimates the proportions; therefore,

we first calculate the immigration flows in each age group by applying the proportions obtained in the random forest age model to the total immigration flows in each destination country by sex, estimated in Abel and Cohen (2022). Secondly, the *pclm_graduate* function from the R *DemoTools* package (Riffe et al. 2019) is used to estimate flows in single ages. Finally, we group the immigration flow estimates in single ages to five-year age group estimates, and calculate the proportions.

Age-education model

The second level of our methodology uses a random forest model to further break down the age proportions of immigration flows into four education categories. This model is only applied to the population aged 15 and above as no education groups are used for the younger age groups (labelled as 'Under 15' education category in the final population projections). We use a similar model as the age model with additional education parameters and education specific population sizes and proportions as listed in Table 2 below.

Table 1: Age-education model predictor variables

Predictor variables	
Period	t
Country	c
Sex	s
Age group	a
Educational attainment	e
Population size of the destination country by age group, education, sex and period	$\log(Pop_{t,c,s,a,e})$
Population at the destination country by previous age group, education, sex and period	$\log(Pop_{t,c,s,a-1,e})$
Population at the destination country by previous age group, education, sex and period	$\log(Pop_{t,c,s,a+1,e})$
Proportion of the population in destination country by age group, education, sex and period	$prop^{POP}_{t,c,s,a,e}$
Proportion of the population in destination country by previous age group, education, sex and period	$prop^{POP}_{t,c,s,a-1,e}$
Proportion of the population in destination country by next age group, education, sex and period	$prop^{POP}_{t,c,s,a+1,e}$
Proportion of the migrant population in destination country by period	$prop^{MIG}_{t,c}$
Human Development Index of the destination country by period	$HDI_{t,c}$
Gross Domestic Product in destination country by period	$GDP_{t,c}$
Illiteracy rate in destination country by sex and period	$prop^{Illiterate}_{t,c,s}$
Life expectancy at birth in destination country by sex and period	$e_{0\ t,c,s}$
Old age dependency ratio in destination country by period	$ODR_{t,c}$

The age-education model is trained using the training data set and checking the model fit with the test data set. Then the new predictor dataset is used to predict the proportions of male and female immigrants in 183 countries at each five-year period by age group and educational attainment. The model has a R-squared value of 0.80. The predictions produced by the age-education model are plotted against the observed proportions from the IPUMS International dataset in Figure 5. The largest difference between two values is observed for the 15-19 year old female population in Nepal in 2001 in which the observed proportion of immigrants is above 0.30 while the predicted proportion for the same population is 0.1.

Figure 5 Predicted vs observed proportions for the age-education model

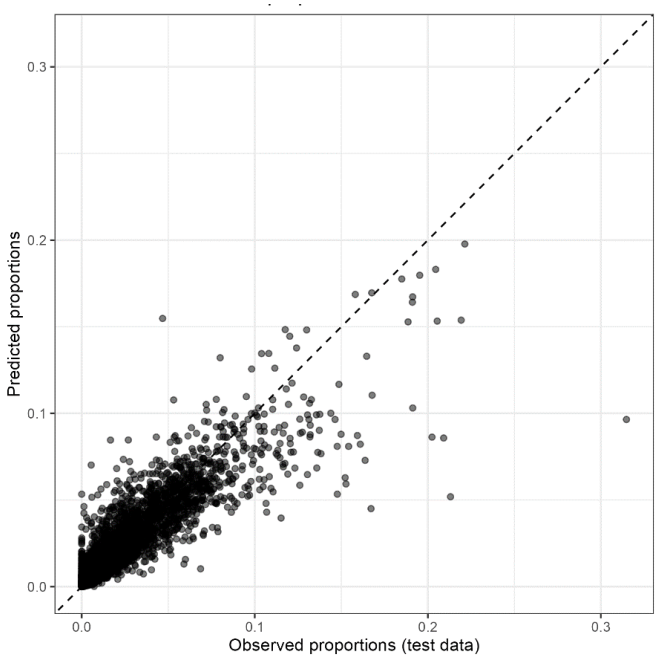
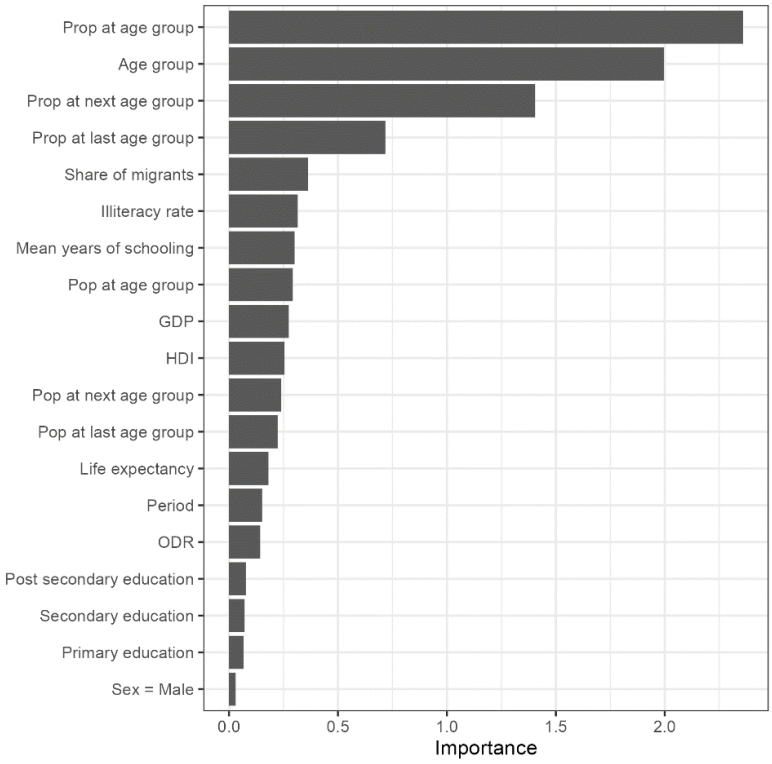


Figure 6 provides the variable importance score. Similar to the age model, age group, proportion of population and share of migrants variables are of great importance for the predictions. Dummy variables related to the educational attainment levels are found less important than other factors except for the sex dummy variable. This is probably because all population and life-course related predictor variables, such as age, were already included in the model as well as the illiteracy rate, the GDP and the HDI which are highly correlated to the education.

Finally, since the age model has the full range of age groups including under 15 and has a slightly better fit, we use the proportions from the age model as our basis and apply the proportions of flows at four education levels in each age group from the age-education model.

Figure 6 Variable importance scores for the age-education model



Estimating emigration flow proportions

WIC2018 used a biregional migration flow approach i.e. total emigration flows, summed over origin countries, are distributed to destination countries following the average age distribution. In WIC2023 an extended biregional migration flow approach is used which includes migrant characteristics and our starting point shifts to the proportions of immigration flows. In biregional approach, immigration and emigration rates are used instead of the origin-destination (bilateral) migration rates.

As characteristics of bilateral migrants are missing, at global level it is assumed that the age and education distribution of emigrants are similar to the characteristics of all global immigrants except the immigrants in destination country for which the characteristics of immigrants are already predicted. The immigration and emigration flows for males and females, at each five-year period, age group and educational attainment need to match at the global level. The aim of this step is to bring the age group and education disaggregation to the emigration flows by country, year and sex from Abel and Cohen (2022) while ensuring this consistency at the global level. To achieve this aim, we follow a two-step approach.

- For each origin country C we assume that the emigration rate by year, sex, age and education, $m_{y,C,s,a,e}$, is proportional to the average immigration rate by year, sex, age and education for the rest of the world (all countries except the origin country C), as shown in Equation 3:

$$m_{y,C,s,a,e} \sim \frac{\sum_{\substack{c=1,\\ c\neq C}}^{n} y_{t,c,s,a,e}}{N_{t,C,s,a,e}} . \quad (3)$$

where $m_{y,c,s,a,e}$ is the immigration flow by year, country, sex, age, and education; and $N_{y,C,s,a,e}$ is the rest of the world population - population of all countries except country C - by year, sex, age and education.

- Emigration flows by year, country, sex, age and education are calculated based on Equation 3, and adjusted to match Abel and Cohen (2022) emigration flows by year, country and sex.

Age smoothing

This step concerns smoothing of the age distribution of any irregular patterns in predicted migration flows.

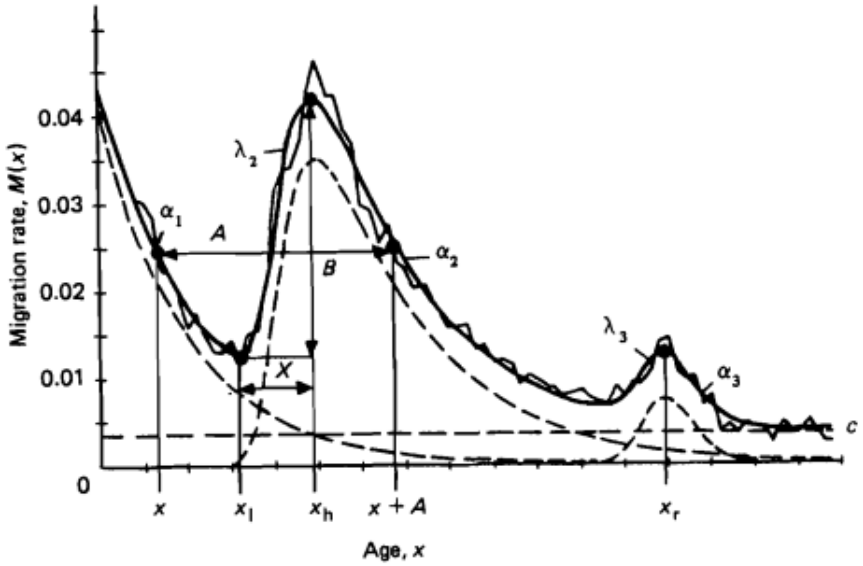
After obtaining immigration and emigration flows for all five-year age groups starting from the 0-4 age group, we continue with the age smoothing of age specific flows, and age- and education-specific flows separately. As mentioned before, for this step we rely on Rogers-Castro (RC) age schedules. Rogers and Castro (1981) developed a mathematical model to decompose the age schedule of migrants into four components. In a full model, i.e., including all four components, illustrated in Figure 7, the age schedule is explained by 13 parameters explaining the shape and intensity of migration (Rogers, Little, and Raymer 2010). For the purpose of this paper, we employed a 7-parameter RC model (Equation 4) that takes into account the pre-working and working age migration.

Equation 4: Rogers Castro model for pre-working and working age migration

$$m(x) = a_1 \exp(\alpha_1 x) + a_2 \exp\big(-\alpha_2(x-\mu_2) - exp\big(-\lambda_2(x-\mu_2)\big)\big) + c$$

where a_1 and a_2 represent the peaks of migration rates, α_1, α_2 and λ_2 represent the shape of the components (rate of change), μ_2 is the peak at labour force and c is the baseline level of migration. The R package, *rcbayes* is employed in the estimation of RC curves (Yeung, Alexander, and Riffe 2022; 2023), which allows inputs in the form of either age specific migration rates or age specific migration flows and population. Using the former option, we calculate the smoothed immigration rates by dividing the age specific immigration flows estimated in the first step by the age specific population from the WIC2018. As a last step, we transform the smooth migration rates back to proportions of migration flows for direct use in the WIC2023 population projection model.

Figure 7 Rogers Castro model migration age schedule



Source: Figure 4 in Rogers and Castro (1981)

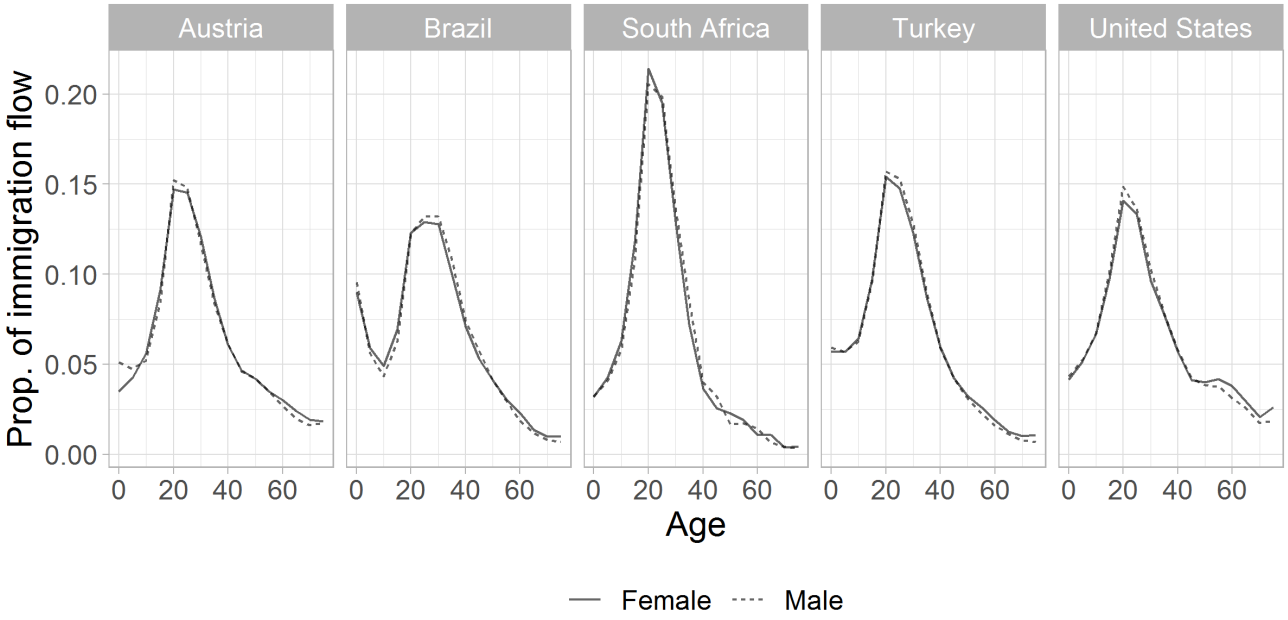
Results

Migration flow proportions by age and education

WIC2023 update uses 2015-2020 age- and education-specific proportions as a starting point for the proportions of future migration flows. The estimated proportions of immigration and emigration flows for all countries are presented in the Appendix. In this section we show the results for selected countries for the 2015-2020 period.

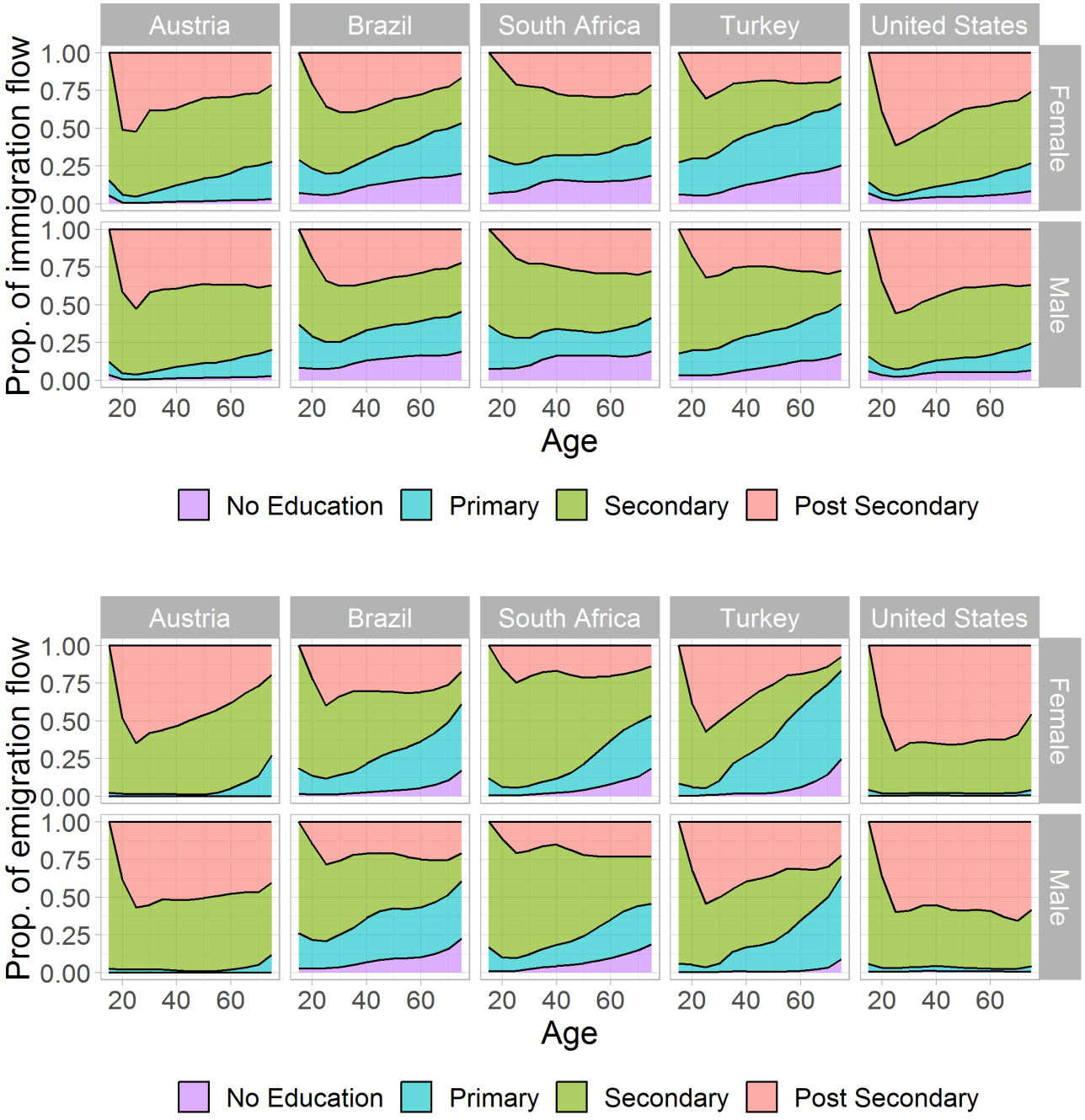
Figure 8 presents the smoothed age proportions in selected countries for females and males separately. In these figures age-specific proportions summing up to 1. The largest proportions are estimated at young working age groups, between 20 and 30. Countries show different age patterns – and to some degree different sex patterns. Among the selected countries, Brazil has the highest proportions at young age groups while South Africa has the highest proportions at 20-25 age group.

Figure 8 Age- and sex specific proportions of immigration flows for selected countries



The next figure presents the average education specific immigration and emigration proportions in selected countries for the 2015-2020 period. The top half of Figure 9 presents the proportions of immigration flows broken down by education levels at each age group There is a near zero-share of post-secondary education predicted in the age group 15-19, which only becomes non-zero in from age group 20-24 onwards. The share of immigrants with no education is almost always the lowest for all age groups, followed by immigrants with primary education, in the selected countries. At young working age groups immigrants with secondary and post-secondary education have a similar share of immigration flows. An exception to this can be found in the United States and Austria, where the proportion of post-secondary educated immigrants is equal or larger than the proportion of immigrants with a secondary education. At older age groups, the proportion of immigration flows with primary and secondary educated immigrants are higher. These results are similar to the age and education structures of the destination countries. The bottom part of the figure shows the proportions of emigration flows broken down by education levels at each age group. Emigration flows differ from immigration flows in terms of education structure. For example, in South Africa, a higher proportion of people with secondary education is estimated to emigrate compared to their counterparts who are immigrating to the country. Similarly, in United States proportion of emigrants with post-secondary education is high at all age groups starting from age 30.

Figure 9 Age- (15+), sex- and education-specific immigration and emigration flow proportions for selected countries, 2015-2019 period

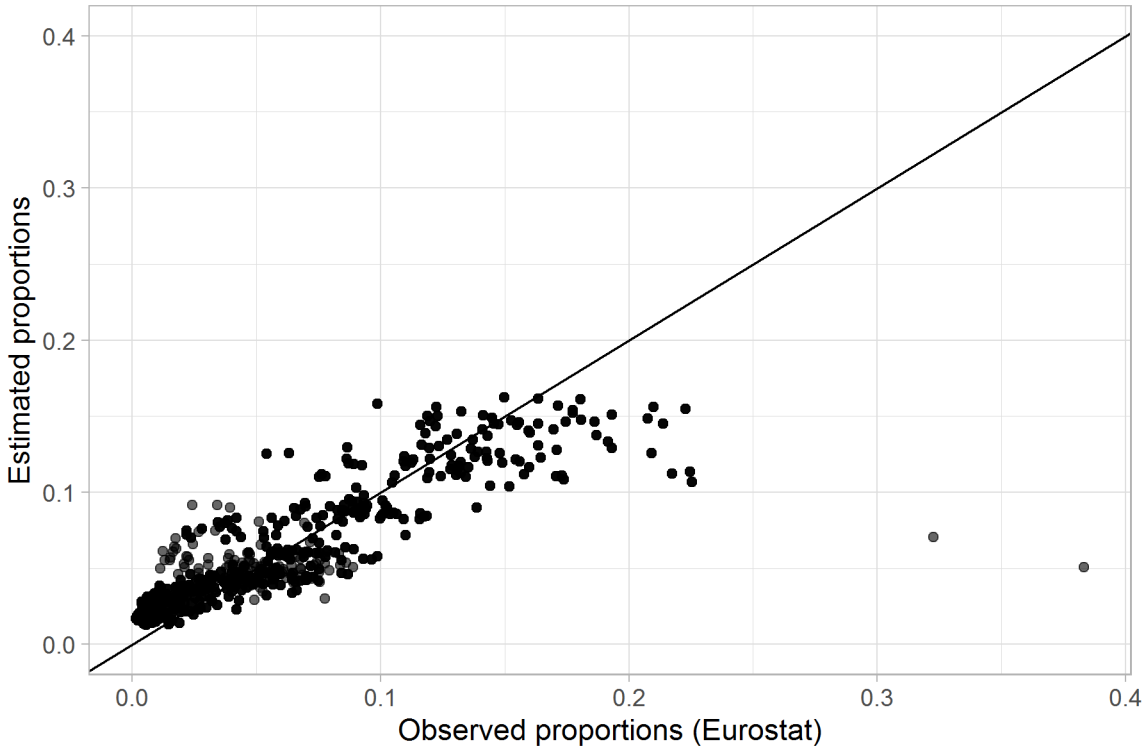


Validation

We finalise our methodology section with a validation exercise on the proportions of immigration flows. Eurostat publishes within-EU bilateral migration flows broken down by age and sex but not by educational attainment. Figure 10 compares the average age specific immigration proportions for the period 2015-2019 from Eurostat and compare them with our estimates. The largest differences between Eurostat and our estimates were found for the 0-4 age group immigrants in Hungary, Slovakia and Portugal. In these countries

proportion of immigrants in 0-4 age group were reported as 0.65 (Hungary, males), 0.56 (Hungary, females), 0.34 (Slovakia, females), 0.32 (Slovakia, females) and 0.00 (Portugal, females) in Eurostat. The Pearson correlation coefficient between the reported and estimated proportions for the same population 0.89 when data for Hungary and Portugal are ignored in the calculation.

Figure 10 Observed (Eurostat) vs estimated immigration proportions



Source: (Eurostat 2023)

Conclusion and limitations

This paper presents the methodology used to estimate the proportion of female and male immigration flows by age group and education level. To our knowledge, this is a first-of-a-kind estimation of migration flows by the three key demographic characteristics across a wide range of countries. Our estimates cover migration between 183 countries, covering 99 per cent of the global population in 2020. The estimates serve as base data for migration measures in the WIC2023, considerably improving on previous assumptions on migration levels by age and education that were constructed based on naive assumptions without reference to empirical measures of immigration flows available in census samples.

Similar to all prediction models, our methodology is based on assumptions and limited by data availability. There are a number of key limitations that we are considering in future work. First, there are alternative ways to estimate age-education proportions such as modeling education distribution before the age distribution or modeling age-education distribution at the same time. In this paper we chose to model age distribution first and rely on migration age schedules to smooth any irregular patterns found in the estimated age-specific proportions predicted by the model. We hope to test alternative specifications of the steps and additional predictor variables in future version to further optimize our predictive approach. Second, emigration

proportions are derived from the age and education distribution of immigration flows. Future work may include a similar methodology to estimate emigration proportions separately, which might be obtainable from census records on questions related to country of previous residence. However, the quality and availability of previous country data is unknown. In many countries census forms only allow respondents to choose from a limited number of options on their previous country and may include aggregated regions (such as Other Asia) that might be inconsistent over many countries and limit the use of indirect emigration measures from census data over multiple countries Third, we did not predict age and education shares of migration for all countries. We were restricted by the availability of predictor variables. However, in some cases the predictor variables did not provide much predictive power and could be removed to allow a greater coverage of countries.

References

Abel, Guy J., and Joel E. Cohen. 2022. "Bilateral International Migration Flow Estimates Updated and Refined by Sex." *Scientific Data* 9 (1): 173. https://doi.org/10.1038/s41597-022-01271-z.

Best, Kelsea B., Jonathan M. Gilligan, Hiba Baroud, Amanda R. Carrico, Katharine M. Donato, Brooke A. Ackerly, and Bishawjit Mallick. 2021. "Random Forest Analysis of Two Household Surveys Can Identify Important Predictors of Migration in Bangladesh." *Journal of Computational Social Science*. https://doi.org/10.1007/s42001-020-00066-9.

Best, Kelsea, Jonathan Gilligan, Hiba Baroud, Amanda Carrico, Katharine Donato, and Bishawjit Mallick. 2022. "Applying Machine Learning to Social Datasets: A Study of Migration in Southwestern Bangladesh Using Random Forests." *Regional Environmental Change* 22 (2): 52. https://doi.org/10.1007/s10113-022-01915-1.

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45 (1): 5–32. https://doi.org/10.1023/A:1010933404324.

Escosura, L.P. de la. 2018. "Historical Index of Human Development." https://espacioinvestiga.org/home-hihd/?lang=en.

Eurostat. 2023. "Immigration by Age Group, Sex and Country of Previous Residence."

Haas, Hein de. 2011. *The Determinants of International Migration: Conceptualizing Policy, Origin and Destination Effects*. Oxford: International Migration Institute, IMI Working Paper.

Haas, Hein de, Stephen Castles, and Mark J. Miller. 2020. *The Age of Migration: International Population Movements in the Modern World*. 6. Auflage, Reprinted by Bloomsbury Academic. London: Bloomsbury Academic.

Hastie, T., R. Tibshirani, and J.H. Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. Springer New York.

Hindman, Matthew. 2015. "Building Better Models: Prediction, Replication, and Machine Learning in the Social Sciences." *The ANNALS of the American Academy of Political and Social Science* 659 (1): 48–62. https://doi.org/10.1177/0002716215570279.

"Integrated Public Use Microdata Series, International: Version 7.3 [Dataset]." 2020. Minnesota Population Center. https://international.ipums.org/.

KC, Samir, Moradhvaj, Michaela Potancokova, Saroja Adhikari, Dilek Yildiz, Maria Mamolo, Tomas Sobotka, et al. 2023. "Wittgenstein Center (WIC) Population and Human Capital Projections - 2023." Zenodo. https://doi.org/10.5281/ZENODO.7921989.

Kerr, Sari Pekkala, William Kerr, Çağlar Özden, and Christopher Parsons. 2016. "Global Talent Flows." *Journal of Economic Perspectives* 30 (4): 83–106. https://doi.org/10.1257/jep.30.4.83.

Kuhn, M., and K. Johnson. 2019. *Feature Engineering and Selection: A Practical Approach for Predictive Models*. Online version. https://bookdown.org/max/FES/.

Kuhn, Max, and Hadley Wickham. 2022. "Tidymodels: Easily Install and Load the 'Tidymodels' Packages." https://cran.r-project.org/web/packages/tidymodels/index.html.

Lutz, Wolfgang, William P. Butz, and Samir KC, eds. 2014. *World Population and Human Capital in the Twenty-First Century*. First edition. Oxford, United Kingdom ; New York, NY: Oxford University Press.

Lutz, Wolfgang, Anne Valia Goujon, Samir KC, Marcin Stonawski, and Nikolaos Stilianakis. 2018. "Demographic and Human Capital Scenarios for the 21st Century: 2018 Assessment for 201 Countries." Luxembourg: Publications Office of the European Union. https://doi.org/10.2760/41776.

Preston, Samuel H., Patrick Heuveline, and Michel Guillot. 2005. *Demography: Measuring and Modeling Population Processes*. Reprint. Oxford: Blackwell.

Reiter, Claudia, Caner Özdemir, Dilek Yildiz, Anne Goujon, Raquel Guimaraes, and Wolfgang Lutz. 2021. "The Demography of Skills-Adjusted Human Capital." WorkingPaper. WP-20-006. Vol. IIASA Working Paper.

Rogers, Andrei, and Luis J. Castro. 1981. "Model Migration Schedules." RR-81-30. Laxenburg, Austria: International Institute for Applied Systems Analysis. http://webarchive.iiasa.ac.at/Admin/PUB/Documents/RR-81-030.pdf.

Rogers, Andrei, Jani Little, and James Raymer. 2010. *The Indirect Estimation of Migration: Methods for Dealing with Irregular, Inadequate, and Missing Data*. THE SPRINGER SERIES ON DEMOGRAPHIC METHODS AND POPULATION ANALYSIS. Dordrecht: Springer.

Roser, Max. 2014. "Human Development Index (HDI) [Dataset]." Our World in Data. https://ourworldindata.org/human-development-index.

Sjaastad, Larry A. 1962. "The Costs and Returns of Human Migration." *Journal of Political Economy* 70 (5).
The World Bank. 2020. "Literacy Rate (%)." https://genderdata.worldbank.org/indicators/se-adt/.
UN DESA. 2020. "International Migrant Stock 2020." United Nations Department of Economic and Social Affairs, Population Division.
———. 2022. "GDP, Per Capita GDP at Constant 2010 Prices - US Dollars." https://unstats.un.org/unsd/snaama/basic.
UNDP. 2022. "Human Development Report 2021/2022 Overview." UNDP.
UNESCO Institute for Statistics. 2020. "Literacy."
Wittgenstein Centre for Demography and Global Human Capital. 2018. "Wittgenstein Centre Data Explorer Version 2.0 (Beta)." http://www.wittgensteincentre.org/dataexplorer.
World Bank. 2022. "GDP, Per Capita (Constant 2010 USD)." https://data.worldbank.org/indicator/NY.GDP.PCAP.KD.
Yeung, Jessie, Monica Alexander, and Tim Riffe. 2022. "Rcbayes: Estimate Rogers-Castro Migration Age Schedules with Bayesian Models." https://CRAN.R-project.org/package=rcbayes.
———. 2023. "Bayesian Implementation of Rogers–Castro Model Migration Schedules: An Alternative Technique for Parameter Estimation." *Demographic Research* 49 (December): 1201–28. https://doi.org/10.4054/DemRes.2023.49.42.
York, P. 2023. "OwidR." R package. https://github.com/piersyork/owidR.
Zeileis, A., G. Grothendieck, J.A. Ryan, J. Ulrich, and F. Andrews. 2023. "S3 Infrastructure for Regular and Irregular Time Series (Z's Ordered Observations)." R. https://zoo.R-Forge.R-project.org/.