EU migration flow proportions by age, sex and educational attainment

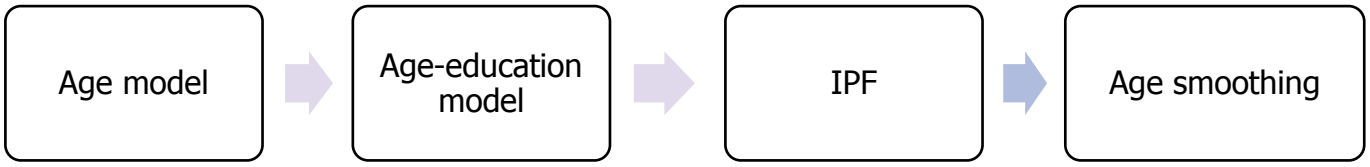
Introduction

The migration statistics for European Union (EU) member states are more detailed and frequent, and of better quality than most of the countries worldwide. However, while Eurostat provides information on migrant stocks and migration flows between member states, the age, sex and educational attainment distributions of migration flows over time are still not complete. In this paper, we propose a methodology to estimate the migration flow proportions by age, sex and educational attainment. This ongoing work is based on previous research which estimated immigration and emigration flow proportions globally. The previous work employed random forest models to estimate the sex specific proportions by age and education. In this paper, we introduce the methodology for estimating immigration and emigration flow proportions using random forest models, and present how we envisage to extend our methodology to estimate bilateral flow proportions by age, sex and education.

We estimate the proportions of male and migration flows by age group (the age model), and age and educational attainment (the age-education model) using a machine learning tool, Random Forest models (Breiman 2001). As shown in Figure 1, our estimation approach consists of four levels. The first level predicts male and female immigration and emigration flow proportions by age groups. The second level further breaks down the estimates by four education levels. The third level, uses the iterative proportional fitting algorithm (IPF) to estimate bilateral immigration and emigration flows by age, sex and education for the European Union member states. The fourth level involves smoothing the age distribution using Rogers-Castro migration age schedules (Rogers and Castro 1981).

In this extended abstract we present the methodology and results to estimate age and education proportions of immigration and emigration flows for 183 countries (Yildiz and Abel, forthcoming). The methodology will be extended to estimate proportions of within EU bilateral flows using the IPF algorithm and the Rogers-Castro age schedules.

Figure 1: 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) in which we obtain the observed proportions of immigration and emigration flows by age and education for each sex, the dependent variable. Other data sources include population size by age, sex and education from Wittgenstein Centre Data Explorer (WIC, 2018), multiple databases from the United Nations (UN) organisations, and estimates of bilateral migration flows by sex (Abel and Cohen 2022). The predictor variables of the random forest models include age group, sex, period, educational attainment, migration definition, country, share of migrants in the country (UNDESA, 2020), Human Development Index (HDI; UNDP, 2022), Gross domestic product (GDP) per capita (World Bank, 2022), share of illiterate population (UNESCO, 2020), life expectancy, old age dependency ratio, population size and proportion of population by age group and sex (and educational attainment; WIC, 2018).

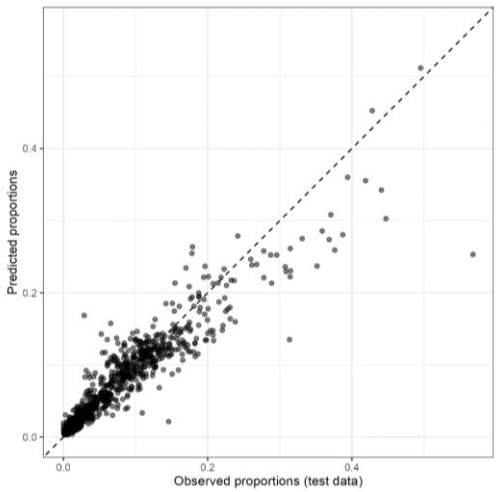
Methodology for immigration and emigration flow proportions by age and education

After the data preparation, two random forest models (one for immigration and one for emigration), as shown in Equation 1, are 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 2. 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 which the observed proportions from IPUMS are higher than 0.5 for this age group.

Equation 1: Age model

$$prop^{IMM}_{y,c,s,a} \sim Period + Country + Sex + Age\ group + \log\left(Pop_{y,c,s,a}\right) + \log\left(Pop_{y,c,s,a-1}\right) + \log\left(Pop_{y,c,s,a+1}\right) + prop^{POP}_{y,c,s,a} + prop^{POP}_{y,c,s,a-1} + prop^{POP}_{y,c,s,a+1} + prop^{MIG}_{y,c} + HDI_{\ y,c} + GDP_{\ y,c} + prop^{Illiterate}_{y,c,s} + e_{0\ y,c,s} + ODR_{y,c}$$

Figure 2 Predicted vs observed proportions for age model



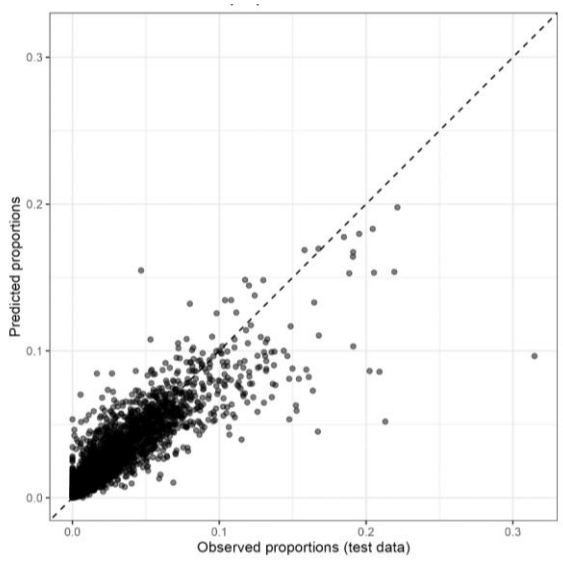
The second level of our methodology uses two random forest model to further break down the age proportions of immigration and emigration flows into four education categories. These models are only applied to the population above age 14 as all of the younger age groups are in the 'Under 15' education category. We use similar models as the age model with additional education parameters and education specific population sizes and proportions as shown in Equation 2 below.

Equation 2: Age-education model

$$prop^{IMM}_{y,c,s,a,e} \sim Period + Country + Sex + Age\ group + Education + \log(Pop_{y,c,s,a,e}) + \log(Pop_{y,c,s,a-1,e}) + \log(Pop_{y,c,s,a+1,e}) + prop^{POP}_{y,c,s,a,e} + prop^{POP}_{y,c,s,a-1,e} + prop^{POP}_{y,c,s,a+1,e} + prop^{MIG}_{y,c} + HDI_{y,c} + GDP_{y,c} + prop^{Illiterate}_{y,c,s} + e_{0\ y,c,s} + ODR_{y,c}$$

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 and emigrants in 183 countries at each five-year period by age group and educational attainment. The immigration 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 dataset in Figure 3. The largest difference between two values is observed for the 15-19 year old female population in Nepal in which the observed proportion of immigrants is above 0.30 while the predicted proportion for the same population is 0.1.

Figure 3 Predicted vs observed proportions for the age-education model



Bilateral flow proportions

The third level is still in progress and aims to estimate the bilateral migration flow proportions by age and education. It assumes the within EU immigration and emigration flows at each age and education should be equal. In this level, we will investigate the possibility of using the IPF algorithm in estimating the bilateral flow proportions.

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.

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

Eurostat. 2023. "Immigration by Age Group, Sex and Country of Previous Residence." https://doi.org/10.1177/0002716215570279.

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

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.

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.

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.

Yildiz, D and Abel, G. (Forthcoming) Migration flows by age, sex and educational attainment. IIASA Working Paper. Laxenburg, Austria: International Institute for Applied Systems Analysis.