Project: PREMUM_EU

19.10.2025

Deliverable D3.2

Dataset of migration flows between and within European countries

PREMIUM_EU BENEFITS OF MOBILITY

Authors:
Dilek Yildiz, Leo van Wissen,
Marianne Tønnessen, Becky Arnold

PREMIUM_EU BENEFITS OF MOBILITY

Deliverable 3.2

Dataset of migration flows between and within European countries

Title: *Policy REcommendations to Maximise the beneficial Impact of Unexplored Mobilities in and beyond the European Union*

Project acronym: *PREMIUM_EU*

Grant Agreement number: *101094345*

Revision history:

Revision	Date	Contributor(s)	Description
1.0	24.06.2025	Dilek Yildiz	First draft CH1
1.1	20.08.2025	Leo van Wissen	First draft CH2
1.2	28.10.2025	Marianne Tønnessen, Becky Arnold	Second draft
2.0	04-11.2025	Final check	Final version (submitted)

Document type: *Report*

Submission date: *30-06-2025*

Lead parties for deliverable: *IIASA*

Dissemination level: CO *

Co-funded by the Horizon Europe of the European Union Research and Innovation Programme.

Consortium Members

Koninklijke Nederlandse Akademie van Wetenschappen (KNAW)- Coordinator.
Kloveniersburgwal 29 Het Trippenhuis, Amsterdam, Netherlands

Nordregio (NR)- Communications
Holmamiralens Vaeg 10, Stockholm, Sweden

Stichting Hanzehogeschool Groningen (HUAS)
Zernikeplein 7, Groningen, Netherlands

Danmarks Statistik (DST)
Sejrogade 11, Kobenhavn, Denmark

Oslomet- Storbyuniversitetet (OSLOMET).
Pilestredet 46, Oslo, Norway

Uniwersytet Ekonomiczny w Krakowie (UEK) .
Ulica Rakowicka 27, Krakow, Poland

Internationales Institut Fuer Angewandte Systemanalyse (IIASA) .
Schlossplatz 1, laxenburg, Austria

Max-Planck Gesellschaft Zur Forderung Der Wissenschaften EV (MPG)
Hofgartenstrasse 8, Munchen, Germany

Hacettepe Universitesi (HUIPS)
Universitesi Beytepe Kampusu Rektorluk Binasi, Cankaya Ankara, Turkey

Centro De Estudios Demograficos (CED)
Cerdanyola V Gr Universitat Aut Barcelona Edifici E, Barcalona, Spain

Acknowledgement

PREMIUM_EU is a Horizon EUROPE project funded by the European Commission under Grant Agreement no. *101094345*

Table of Contents

1. Integrated dataset of migration flows between European countries

This section presents the methodology behind the estimation of educational attainment composition of international migrants, and the bilateral migration flow dataset between European countries, categorized by age group, sex and educational attainment. Bilateral migration flows by age group and sex were estimated in Deliverable 2.2.

Our methodology aims to estimate the age and education composition of immigration and emigration flows globally. In this deliverable, we apply the education proportion within each age group and sex to the migration flows between European countries estimated in Deliverable 2.2.

1.1 Data

To predict the age and educational composition of immigration flows, we collected census samples from the Integrated Public Use Microdata Series International database (IPUMS, Ruggles et al. 2025). The age and education proportions of immigrant flows were calculated using 142 census samples from 73 countries. The proportions were calculated by dividing the size of each age-sex-education immigration flow by the total sex-specific immigration flow. Hence, the two sets of sex-specific proportions sum up to 1 for each country-year set of observations. Censuses with fewer than 100 observed immigrants or those with no observations in more than 20 age-education pairs in each set of sex-specific proportions were removed from the dataset.

1.2 Methodology

The age and education composition of bilateral migration flows are estimated in four steps as listed below.

1. Rogers-Castro migration age schedules for age smoothing
2. Super learning model for estimating education composition of immigrants
3. Estimating emigration education proportions
4. Iterative Proportional Fitting (IPF) for estimating bilateral migration education proportions

In the first step, the Rogers Castro (RC) migration models are used to smooth the age patterns in the immigration proportions derived from the IPUMS International database for each education group (Rogers and Castro 1981). We use a seven-parameter Rogers Castro model migration schedule, which includes pre-working and working age components as shown in in Equation 1:

$$m(x) = a_1 \exp(\alpha_1 x) + a_2 \exp\big(-\alpha_2(x-\mu_2) - exp\big(-\lambda_2(x-\mu_2)\big)\big) + c \quad (1)$$

Where $m(x)$ is the age-specific migration rate, x is age, 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 Rogers Castro age schedules in a probabilistic way (Yeung, Alexander, and Riffe 2022).

In the second step, the estimates from the first step and the predictor variables are used to build a Super Learner (SL) model. Super learning, also known as stacked generalisation, is an ensemble machine learning approach that combines multiple machine learning models (also interchangeably referred to as algorithms and learners) or the same model with different specialisations (Naimi and

Balzer 2018; Van Der Laan, Polley, and Hubbard 2007). Super learning utilises k-fold cross-validation to generate predictions from candidate (base) learners, which are then used as inputs to a meta-learner model that combines their predictions. The final predictions are the weighted combination of predictions (coefficients of the meta-learner listed in Table 1) from the selected base learners. Our Super Learner model utilizes eight different initial Random Forest (RF) learners listed in Table 1, and is used to make predictions for the missing countries and years. The initial RF learners use different mtry values which define the number of variables that can be chosen within each node of the model, and have been found to have the most significant influence on predictive performance. The higher the mtry value, the more chances that a tree can find a more useful variable, while low mtry values create less correlated trees.

Table 1 Base learners

Learner	Specification	Meta learner coefficients for immigration
Ranger	Fast(er) random forests with default mtry = 6	0.000
Ranger autotune	Automatically tuned Fast(er) random forests	0.089
Ranger10	Fast(er) random forests with mtry = 10	0.000
Ranger15	Fast(er) random forests with mtry = 15	0.415
Ranger20	Fast(er) random forests with mtry = 20	0.415
Ranger25	Fast(er) random forests with mtry = 25	0.000
Ranger30	Fast(er) random forests with mtry = 30	0.081
Ranger35	Fast(er) random forests with mtry = 35	0.000

The response variable of the SL algorithm is the log of the proportion of the immigrant stock at each age and education group within the total migrants in a country, categorised by five-year periods and separately for each gender. The last age group is an open-ended age group 75 years and older. The continuous predictor variables listed in Table 2 were standardised to have a mean of 0 and a standard deviation of 1, and the categorical predictor variables were one-hot encoded (the levels of categorical predictors were converted to a set of binary indicators without dropping any levels, similar to dummy variables).

In the third step, education composition of emigration flows within Europe are estimated using the information from the immigration flows following Yildiz and Abel (2023). It is assumed that the education distribution of emigrant flows are similar to the characteristics of all within-Europe immigrants except the immigrants in destination country for which the characteristics of immigrants are already predicted using the SL model. Since we are interested in within-Europe migration, the total immigration and emigration flows for males and females, at each five-year period, age group and educational attainment need to match at the European level. The aim of this step is to bring the education disaggregation to the emigration flows by country, year, age and sex from Deliverable 2.2 while ensuring this consistency at the European level. To achieve this aim, we follow a two-step approach.

For each origin country C we assume that the emigration rate to the rest of Europe in year *y*, for sex *s*, age *a* and education *e*, $m_{y,C,s,a,e}$, is proportional to the average immigration rate to all other European countries, which is the sum of the immigration flows $V_{y,c,s,a,e}$, into all other European countries *c*, by year *y* sex s, age group a, and education e, divided by the population $N_{y,c,s,a,e}$ of all other European countries, as shown in Equation 2:

$$m_{y,C,s,a,e} \sim \frac{\sum_{\substack{c=1,\\ c\neq C}}^{n} V_{y,c,s,a,e}}{\sum_{\substack{c=1,\\ c\neq C}}^{n} N_{y,c,s,a,e}} \quad (2)$$

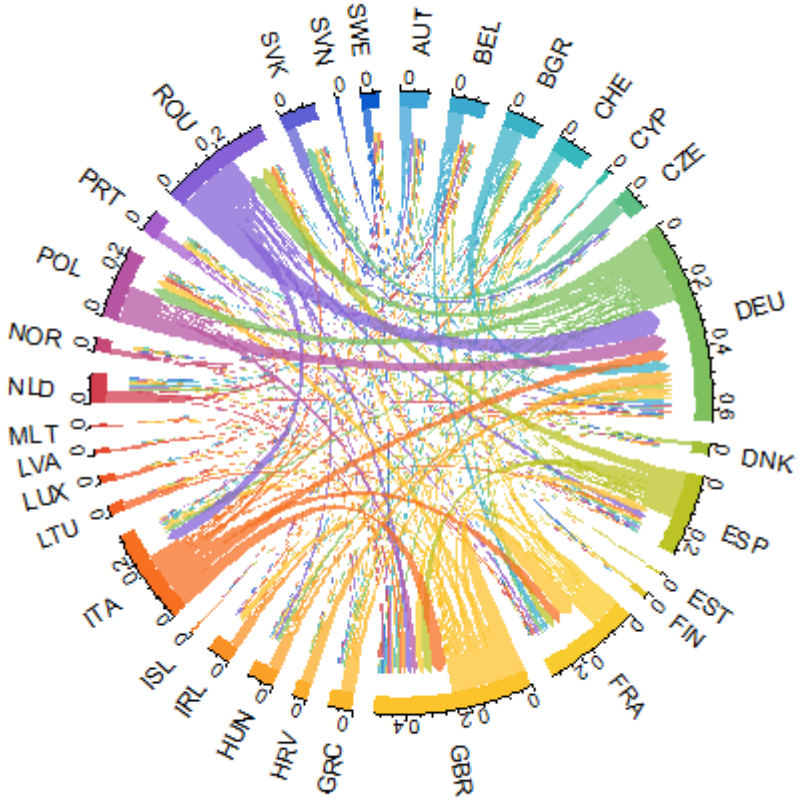
Emigration flows by year *y*, country *C*, sex *s*, age *a* and education *e*, calculated as $m_{y,C,s,a,e} \cdot N_{y,C,s,a,e}$, are adjusted to match emigration flows by year, country and sex in Deliverable 2.2.

In the final step, the educational composition of bilateral migration flows was estimated using the iterative proportional fitting algorithm (Raymer, de Beer, and van der Erf 2011) applied to immigration and emigration flows by age, sex, and education across European countries. For each five-year period and education level, the final model utilized the marginal totals of Origin-Age-Sex, Destination-Age-Sex, Origin-Age, Destination-Age, Origin-Sex, Age-Sex, and Destination-Sex to ensure consistency across origin and destination flows by demographic and educational categories. Figure 1 displays the resulting estimates of bilateral migration flows in 2015, by our five categories of educational attainment.

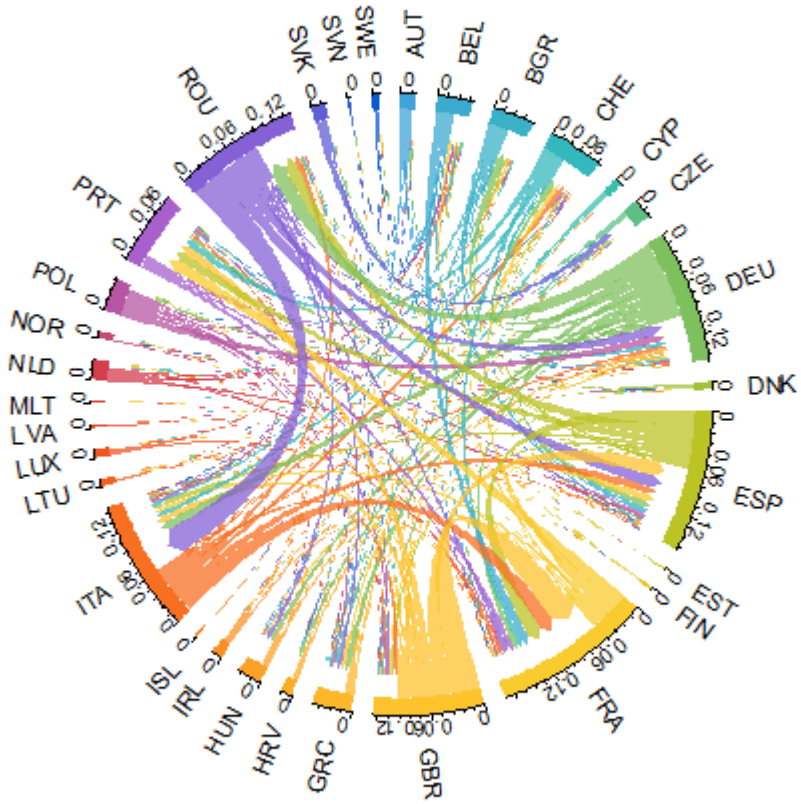
Table 2 Predictor variables

Predictor variables	Source
Period	IPUMS
Sex	IPUMS
Age group	IPUMS
Education	IPUMS
Migration interval	IPUMS
Population size by age group, education and sex	WIC
Population by previous age group, education and sex	WIC
Population by previous age group, education and sex	WIC
Proportion of the population by age group, education and sex	WIC
Proportion of the population by previous age group, education and sex	WIC
Proportion of the population by next age group, education and sex	WIC
Total population	WIC
Immigrant stock	UN
Size of the emigrant population	UN
GINI	Gap Minder
Sex ratio	UN WPP
Sex ratio at birth	UN WPP
Life expectancy at birth by sex	UN WPP
Life expectancy at age 80 by sex	UN WPP
Population density	UN WPP
Median age	UN WPP
Rate of natural change	UN WPP
Population growth rate	UN WPP
Births to 15-19 year old mothers	UN WPP
Crude birth rate	UN WPP
Total fertility rate	UN WPP
Mean age at childbearing	UN WPP
Crude death rate	UN WPP
Infant mortality rate	UN WPP
Under-five mortality rate	UN WPP
Net number of migrants	UN WPP
Net migration rate	UN WPP
Human Development Index	UNDP
Gross National Income per Capita	UNDP
Adolescent birth rate	UNDP
Carbon dioxide emissions per capita (production)	UNDP
Expected years of schooling	UNDP
World region	IPUMS

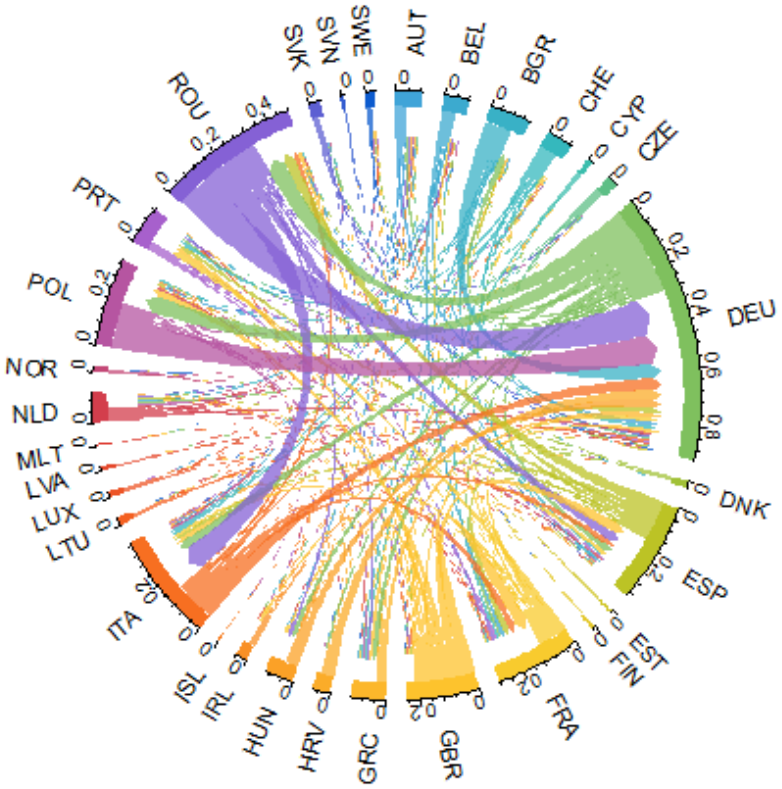
Figure 1: Bilateral migration flows by educational attainment in 2015.



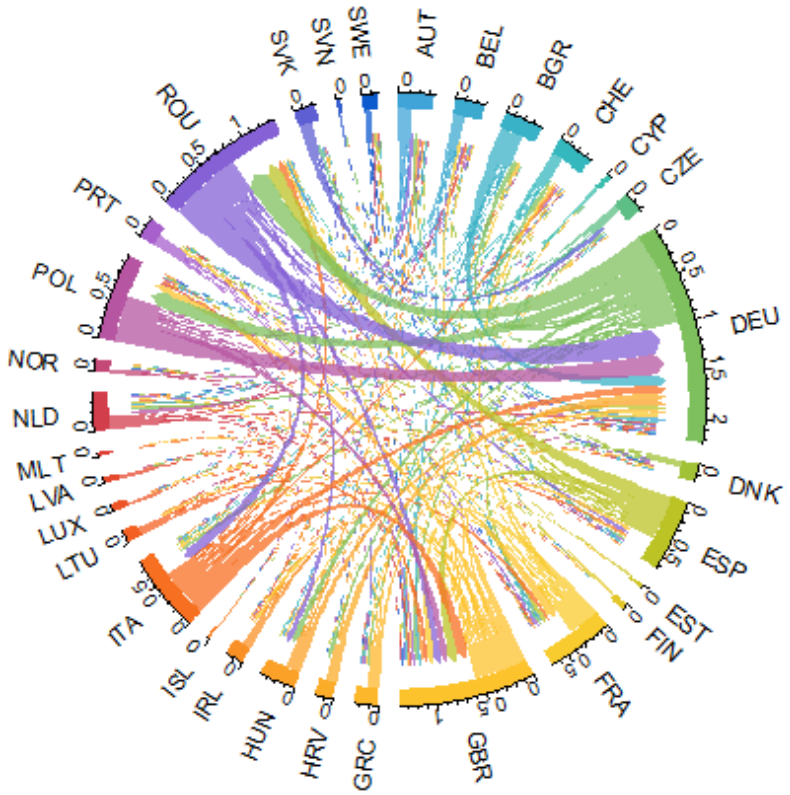
Under 15



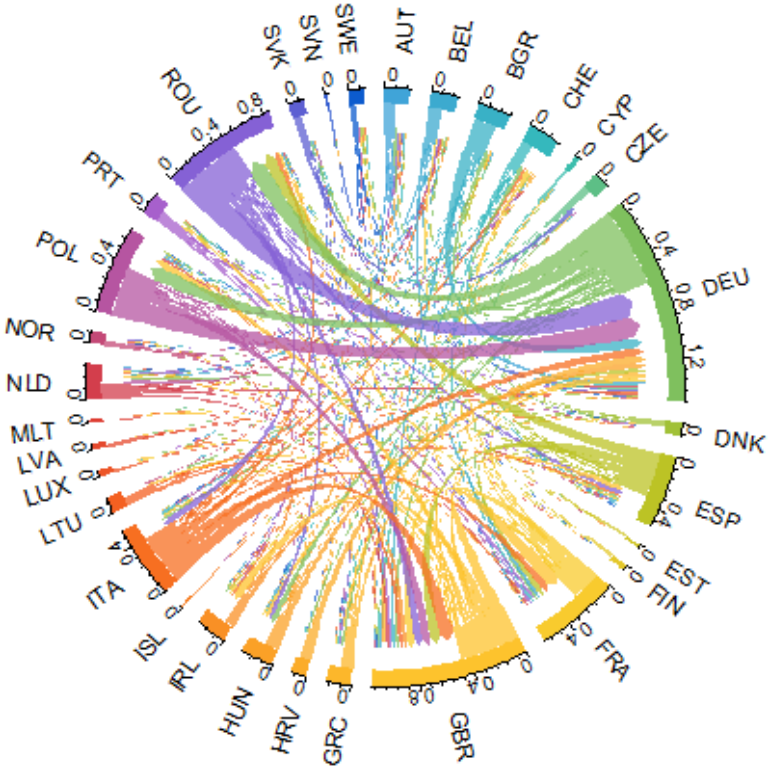
No Education



Primary



Secondary



Post Secondary

2 Migration flows within European countries

2.1 Introduction

In this chapter a method will be presented to include level of education into the estimates of net interregional migration by gender and age, using partial available information from various international statistical sources. This method is based on demographic accounting principles and essentially starts with the balancing equation to arrive at net migration for a given region:

$$N_{t,t+5} = P_{t+5} - P_t - B_{t,t+5} + D_{t,t+5} \quad (3)$$

Where $N_{t,t+5}$ is net migration between *t* and *t*+5, P_t is the population at time *t*, and $B_{t,t+5}$ and $D_{t,t+5}$ are births and deaths between *t* and *t*+5. A model will be presented that works as well for estimates of the period 2010-2020 and for projections 2020-2040 (where the period 2020-2025 is a projected period, since there is not yet sufficient information for the whole period to make estimates).

Several elements of the model complicate the straightforward application of the balancing equation. The population is defined by age-, gender, and educational attainment. Moreover, educational attainment is a dynamic factor, governed by educational transition parameters.

In the following sections, these complications are built into the balancing equation to arrive at a model of net migration by region, level of education, gender and age. In the next section the overall model structure is presented. Next, never mind the complications just mentioned, the basis of the balancing equation approach is to have estimates of the regional NUTS3 population by level of education, gender and age, which is not available for most countries in Europe. Section 2.3 explains in more detail how these data are estimated.

2.2 Model structure

The approach at the regional level in this project is to combine various known data sources into one integrated database, of estimates of population and regional migration by NUTS3 region, age, gender, and level of education, for the periods 2010- 2015 and 2015-2020. The method will be explained below based on the available information for the period 2010-2020.

For the period 2010-2020 we use the following data sources:

At the national level, from Wittgenstein Centre for Demography and Global Human Capital (WIC):

- Population by age (5 year categories 0, 5, .., 85+), gender and level of education for 2010, 2015 and 2020

At the regional (NUTS3) level, from Eurostat (and harmonized by the European Commission's Joint Research Centre JRC):

- Population by age (5 year age categories) and gender 2010, 2015, 2020
- Total births 2010-2019, further divided into births by gender using gender ratio's at birth
- Total deaths by age and gender 2010-2019. Harmonized data from JRC are only available by age, not gender. Eurostat has data by age and gender, but only since 2013. We use the gender shares of the Eurostat data for the total deaths as given by JRC.

As will be explained below, the available data does not cover information on level of education by NUTS3 region. We will use a predictive (logit) model, with exogenous regional characteristics to fill this information gap (see section 2.3).

2.2.1 Balancing equation by age and gender

Adding age and gender dimensions to the balancing equation (3) gives the following equation (which is called the forward survival method to estimate net migration,see Rowland (2003):

$$N_{s,a,t,t+5} = P_{s,a+5,t+5} - P_{s,a,t} + D_{s,a,t,t+5} \quad (4)$$

For all ages *a* = 0, 5, 10,.. 80 (where *a*= 0 means age group 0-4, etc.).

$N_{s,a,t,t+5}$ is the number of net migrants between *t* and *t*+5, who are in the age group 0-5 at time *t*; similarly for deaths $D_{s,a,t,t+5}$, for all ages 0, 5,.., 85+. For *a* = 0 we have:

$$N_{s,0,t,t+5} = P_{s,0,t+5} - B_{s,t,t+5} + D_{s,0,t,t+5} \quad (5)$$

For the highest open-ended age-group we have:

$$N_{s,85+,t,t+5} = P_{s,85+,t+5} - P_{s,85+,t} - P_{s,80,,t} + D_{s,80,t,t+5} + D_{s,85+,t,t+5} \quad (6)$$

The balancing equation (3), when expressed in vector formulation, takes these dimensions into account:

$$\mathbf{N}_{t,t+5} = \mathbf{P}_{t+5} - \mathbf{P}_t - \mathbf{B}_{t,t+5} + \mathbf{D}_{t,t+5} \quad (7)$$

Where **N**, **P**, **B** and **D** are appropriately sized vectors of all age and gender combinations.

2.2.2 Adding level of education to the balancing equation

We use three levels of education: low, middle and high (see Box 1 for details). The WIC approach for projecting level of education uses a cohort based approach, where for each birth cohort an age profile and an upper educational attainment level is assumed, based on the observed age profile and upper educational attainment levels of previous cohorts. The upper educational attainment levels are reached at age 30, and the age path starts at age 15 where a transition is made from no education (ages 0-14) to higher educational levels.
Here a different approach is taken but based on the educational attainment levels as given by WIC, for each age- and gender combination for each point in time (2010 up to 2040, although WIC makes projections up to 2100 for all countries in the world). We explicitly model transitions between educational levels by age and gender, whereas in the WIC approach these transitions remain implicit in the development of the cumulative educational achievement proportions by age. We use the transition approach since educational transitions and migration transitions interact, which we make explicit in the current approach.

Transitions between educational categories take place between the age categories 10-15 to 15-20, 15-20 to 20-25, 20-25 to 25-30, and 25-30 to 30-35. Beyond 35 years of age we assume no change of educational attainment occurs. For each gender we therefore have four 3x3 transition matrices

between educational levels. Because deaths below age 35 are rare in Europe, we can disregard the potential interaction between mortality and level of education and apply deaths proportionally to all educational categories. Since our input data from WIC are on the country level, all our transition matrices are country-specific.

The following example shows how the transition matrices are estimated. We use the example of Austria, with data for the period 2015-2020 for females. Table 3 shows the transition matrix for the transition between age groups 20-24 to 25-29. The WIC estimates of the distribution over the three educational categories for 2010 and 2015 are input to the table (bottom row for the 2010 distribution and right column for the 2015 distribution). The cells of the matrix are then derived as follows:

1. Changes from higher to lower educational levels are not possible. Therefore cells (1,2), (1,3) and (2,3) are zero by definition.
2. As a consequence of observation 1, cell (1,1) = rowtotal(1).
3. Similarly, cell (3,3) = columntotal(3).
4. Except for age 10-14 to 15-19, where small positive values have been observed, it is not possible to move up 2 levels of education within a 5-year period. Therefore, cell (3,1) is 0.
5. The other cells are then identified by subtraction. For instance, cell (3,2) = columntotal(3) – cell (3,3).

		2010			
		Low (1)	**Middle (2)**	**High (3)**	**2015 share**
2015	**Low (1)**	0.16	0	0	**0.16**
	Middle (2)	0.04	0.37	0	**0.41**
	High (3)	0	0.10	0.33	**0.43**
	2010 share	**0.20**	**0.47**	**0.33**	**1.00**

Table 3 Estimated educational attainment transition matrix in the period 2010-15, Austria, females, age group 20-24 to 25-29.

The four transition tables related to the four age group transitions determine the full matrix of educational transitions. These probabilities are transformed into conditional probabilities of having education level *f* at *t*+5, given that a person had education level *e* at time *t*. For instance, from table 3 we derive that the conditional probability of having a high education diploma at *t*+5, given that the person was middle educated at time *t* is 0.10/0.47 = 0.21. These conditional probabilities add up to 1.00 over each column. The two crucial transitions are from low to middle educated, and from middle to high educated. If you are still low educated in your early twenties, as Table 3 shows, the probability of moving from low to middle educated is not very high (0.04 / 0.20 =20%). In contrast, the result for the age group 10-14 to 15-19 is 54%.

The balancing equation (7) including educational dynamics of the resident population reads:

$$\mathbf{N}_{t,t+5} = \mathbf{P}_{t+5} - \Delta\mathbf{E} \cdot \mathbf{P}_t - \mathbf{B}_{t,t+5} + \mathbf{D}_{t,t+5} \qquad (8)$$

where $\Delta\mathbf{E}$ is the transition matrix of educational attainment. With three educational levels, 18 age classes and 2 genders the total number of rows and columns of $\Delta\mathbf{E}$ is 108 (3x18x2). $\Delta\mathbf{E}$ is a subdiagonal blockmatrix with age- and gender-specific 3x3 educational transition matrices in the subdiagonal.

Box 1: The classification into low, middle and high educated

The classification into low, middle and high education is not unambiguous across countries and institutions. Classifications are normally based on the ISCED2011 definitions, but there are slight differences between the way various organizations combine the ISCED97 categories (and also in the way they categorize persons with missing information about educational level). The table below gives an overview of the various definitions used by Eurostat, WIC, and a number of countries. The problem is in the delineation of middle and high educated. There is no ambiguity in the definition of low educated: everything up to the level of lower secondary education. The middle category is in any case upper secondary level, but post-secondary non-tertiary education is sometimes counted as middle (ISCED2011, Eurostat) and sometimes as higher educated. In any case, this is not a large category, but could lead to some distortions with national data.

In this project we follow the WIC classification. WIC classifies everyone <15 years of age in a separate category, which we classify as Low educated.

<table>
<tr><th>ISCED 97</th><th>ISCED2011</th><th>WIC 2023</th><th>Eurostat</th><th>Norway</th><th>Sweden</th><th>Finland</th></tr>
<tr><td rowspan="2">0 – Pre-primary edu</td><td>01 – Early childhood edu</td><td rowspan="2">No education (E1)</td><td rowspan="3">Primary & lower secondary</td><td rowspan="2">NUS0</td><td rowspan="3">Prim/sec edu, less than 9 years</td><td rowspan="4">9 Basic education</td></tr>
<tr><td>02 – Pre-primary edu</td></tr>
<tr><td>1 – Primary edu</td><td>1 – Primary edu</td><td>Incomplete (E2) and completed primary (E3)</td><td>NUS1</td></tr>
<tr><td>2 – Lower secondary edu</td><td>2 – Lower secondary edu</td><td>Lower secondary (E4)</td><td>NUS2</td><td>Prim/sec edu, 9-10 years</td></tr>
<tr><td>3 – Upper secondary edu</td><td>3 – Upper secondary edu</td><td>Upper secondary (E5)</td><td rowspan="2">Upper secondary & post-secondary non-tertiary</td><td>NUS3&4</td><td>Upper sec edu, 2 years or less + upper sec edu, 3 years</td><td>3 Upper secondary education</td></tr>
<tr><td>4 – Post-secondary non-tertiary</td><td>4 – Post-secondary non-tertiary</td><td rowspan="2">Short cycle (E6)</td><td rowspan="2">NUS5</td><td rowspan="2">Post-sec edu, less than 3 years</td><td>4 Post-secondary non-tertiary education</td></tr>
<tr><td>5B</td><td>5 – Short-cycle tertiary</td><td rowspan="4">Tertiary</td><td>5 Short-cycle tertiary education</td></tr>
<tr><td>5A First stage tertiary</td><td>6 – Bachelor</td><td>Bachelor (E6)</td><td>NUS6</td><td rowspan="2">Post-sec edu, 3years or more</td><td>6 Bachelor's</td></tr>
<tr><td>5A</td><td>7 – Master</td><td rowspan="2">Master (E6)</td><td>NUS7</td><td>7 Master</td></tr>
<tr><td>6 – Second stage tertiary</td><td>8 - Doctoral</td><td>NUS8</td><td>Post-graduate edu</td><td>8 Doctoral</td></tr>
</table>

Various agencies' categorization of educational attainment

The 3x3 educational transition matrices for all ages without educational change (all ages below 10 and above 30) are identity matrices. However, this above equation does not take into account the possible interaction between migration and educational dynamics. The two are clearly related. Educational dynamics is an important trigger for migration at young ages. For some migrants, the change in educational level takes place before the move (for instance a person migrates after graduating from university), for others it takes place after the move (for instance a person moves to a location to obtain a master degree). We deal with this interdependence by dividing net migration into two groups: half of the net migration takes place at the beginning of the period, and half of the net migration takes place at the end of the period (See also Preston et al., 2003, chapter 6). This means that half of the migrants are exposed to the educational transition probabilities in the origin region, and the other half are exposed to these probabilities in the destination region. In other words, half of the migrants who move from one educational category to the next will have this transition registered at the region of origin, and half will have it registered at the region of

destination. So we have that at the end of the period the net migrants are the sum of $\frac{1}{2}\mathbf{N}_{t,t+5}$ and $\frac{1}{2}\mathbf{\Delta E}\cdot\mathbf{N}_{t,t+5}$, or taken together: $\frac{1}{2}(\mathbf{I}+\mathbf{\Delta E})\cdot\mathbf{N}_{t,t+5}$, This means that the following identity holds:

$$\mathbf{P}_{t+5} = \Delta\mathbf{E}\cdot\mathbf{P}_t + \mathbf{B}_{t,t+5} - \mathbf{D}_{t,t+5} + \frac{1}{2}(\mathbf{I}+\Delta\mathbf{E})\mathbf{N}_{t,t+5} \quad (9)$$

Where $\mathbf{I}$ is the identity matrix. Reworking this equation results in the solution for net migration by age, gender and level of education:

$$\mathbf{N}_{t,t+5} = [\tfrac{1}{2}(\mathbf{I}+\Delta\mathbf{E})]^{-1}[\mathbf{P}_{t+5} - \Delta\mathbf{E}\cdot\mathbf{P}_t - \mathbf{B}_{t,t+5} + \mathbf{D}_{t,t+5}] \quad (10)$$

In the current model we do not assume regional differences between transition probabilities. In that case equation (9) reduces to

$$\mathbf{P}_{t+5} = \Delta\mathbf{E}\cdot\mathbf{P}_t + \mathbf{B}_{t,t+5} - \mathbf{D}_{t,t+5} + \Delta\mathbf{E}\cdot\mathbf{N}_{t,t+5} \quad (11)$$

And equation (10) reduces to

$$\mathbf{N}_{t,t+5} = \Delta\mathbf{E}^{-1}[\mathbf{P}_{t+5} - \Delta\mathbf{E}\cdot\mathbf{P}_t - \mathbf{B}_{t,t+5} + \mathbf{D}_{t,t+5}] \quad (12)$$

2.3 The estimation of the regional population by gender, age, and level of education

The terms $\mathbf{P}_t$ and $\mathbf{P}_{t+5,}$ the regional population at NUTS3 level by age, gender and level of education, are crucial for the estimation of net migration in equation (10). This information is not readily available at the European level. In this section a model is presented to estimate the distribution of educational attainment over the NUTS3 regions, by age and sex.

2.3.1 The estimation problem

The starting point for the estimation is the population data by NUTS3 region, age (in 5 year age groups 0-4, 5-9, .., 85+) and sex. This information is available in the REGIONS database of Eurostat, and harmonized by the Joint Research Centre JRC, to create a time series for each region back to 1990. The research problem that we try to solve here is to expand this table of three dimensions (region R, sex S, age A) with a fourth dimension: educational attainment E, in three categories: Low, Middle and High educated. For the Netherlands for example, having 40 NUTS3 regions (so called COROPs) this table contains 40x3x2x18 = 4320 cells. The level of education is not available for most European countries at the NUTS3 level. Eurostat provides information on level of education at the NUTS2 level but not broken down by age[1]. More detailed information of educational attainment is provided at the national level by the Wittgenstein Centre for Demography and Human Capital WIC. The available partial regional information is summarized in table 4. As already mentioned, Eurostat collects annual information about the population by age and gender for each NUTS3 region[2]: RxSxA. In the PREMIUM_EU project the years 2010, 2015 and 2020 are used. Educational attainment is a more problematic dimension. WIC has made estimates for all countries in the world of the population by sex, age and level of education, but not for regions within countries. There is some information at Eurostat/JRC, based on the ongoing European labor force

[1] Since the analysis for this contribution was finished, Eurostat released the data of the Census 2021, which contains a table at the NUTS2 level by broad age groups and sex. It was too late to include this table in the current estimation.
[2] This information is even available at the finer detailed regional level of the LAU (Local Area Units).

surveys: The population of working age (between 15 and 75 years) by sex and educational attainment, at the NUTS2 level. The notation in the first column of the table will be explained below.

Table	Source	Description
RxSxA	Eurostat/JRC	Population by Age (0-5,..,85+) and Gender for each NUTS3 Region
SxAxE	WIC	Educational attainment by Age (0-5,..,100+) and Gender
R2xSxE	Eurostat/JRC	Population 15-75 years for each Gender by Educational attainment for each NUTS2 (R2) Region

Table 4 Data resources for estimation

2.3.2 Model specification

The problem can be specified as follows, using the modelling language that was introduced in the programming language GLIM (an acronym for Generalized Linear Interactive Modelling) in the seventies (Aitkin et al., 2005). In this modelling language a contingency table with dimensions sex S and age A can be written as SxA. If we only have the marginal totals of age and sex, we can estimate a table S+A, assuming independence between both dimensions. The formulation of independence between the two dimensions in a loglinear formulation is given by:

$$\log \widehat{M}_{ij} = \mu + \mu_i^A + \mu_j^B \qquad (13)$$

Where $\widehat{M}_{ij}$ is the expected value of cell (*i,j*) in the two-way table under the assumption of independence of factors *A* and *B*, and the coefficients are given by:

$$\mu = \frac{\sum_{i,j} log\ \widehat{M_{ij}}}{IxJ},\ \mu_i^A = \frac{\sum_j log\ \widehat{M_{ij}}}{J} - \mu,\ \mu_j^B = \frac{\sum_i log\ \widehat{M_{ij}}}{I} - \mu \qquad (14)$$

If μ is determined, only $I-1$ coefficients of factor *A* can be estimated, and one is redundant, and similarly for factor *B*. Usually, the constraint that all coefficients sum to 0 is used, but other designs can be imposed as well, for instance that the coefficient of the first or last category of the factor is 0. Note that the parameters are estimated from the expected values $\widehat{M}_{ij}$. This means that the expected values under the model have to be estimated first, and from these expected values the parameters are derived. The sufficient statistics to estimate the expected values are the marginal totals implied by the specified main and interaction effects. For the model of independence (13) these are the row- and column totals of the AxB table.

By comparing the expected values $\widehat{M}_{ij}$ with observed values M_{ij}, the hypothesis of independence can be tested, using the well-known Chi-square test, with appropriate degrees of freedom: in this case $I * J - 1 - (I-1) - (J-1)$. However, many contingency tables of population data suffer from overdispersion, which heavily inflates the test statistic. Moreover, many tables are based on register or census data of the whole population, not samples. Therefore, the value of the test statistic is only indicative of the fit of the model.

If the hypothesis of independence is rejected, an interaction term is necessary. In a two-way table this leads to a saturated model, with the number of coefficients equal to the number of

observations (cells) in the table. The model of interdependence is AxB, or, in a loglinear formulation:

$$\log \widehat{M}_{ij} = \mu + \mu_i^A + \mu_j^B + \mu_{ij}^{AB} \quad (15)$$

and the interaction term in this model is estimated as:

$$\mu_{ij}^{AB} = log\ \widehat{M_{ij}} - \mu - \mu_i^A - \mu_j^B \quad (16)$$

This example readily extends to more than two dimensions. Our table to be estimated has four dimensions, with cell entries M_{ijkl}, where *i* is the index for region, *j* for sex, *k* for age and *l* for education. The available partial information allows only to estimate a model with restrictions on many parameters.

RxSxAxE = RxSxA + SxAxE + R2xSxE (17)

This is a hybrid loglinear model, since R2 (NUTS2) is an aggregate of NUTS3 regions and therefore not a standard factor in a loglinear model. In parametric form the model reads:

$$\log \widehat{M}_{ijkl} = \mu + \mu_i^R + \mu_j^S + \mu_k^A + \mu_l^E + \mu_{ij}^{RS} + \mu_{ik}^{RA} + \mu_{jk}^{SA} + \mu_{jl}^{SE} + \mu_{kl}^{AE} + \mu_{kl}^{R2E} + \mu_{ijk}^{RSA} + \mu_{jkl}^{SAE} + \mu_{jkl}^{R2SE} \quad (18)$$

Note that in this hierarchical model, if a higher dimensional interaction is included, the lower level interactions between these variables are also included. It is clear that the crucial interaction at the NUTS3 level is missing. Although R2xSxE provides some information between the regional dimension and the level of education, it does not differentiate between NUTS3 regions within a NUTS2 region. This is most clearly seen if we observe a single sex-age combination, say 25-30 Females. For this group we have the distribution over the regions R in a country, as well as the nation-wide level of education E. The interaction R2xSxE assigns values of educational attainment to each NUTS3 region based on the corresponding NUTS2 regional context, and for all ages 15-75. As a result, all 25-30 age Female categories in the NUTS3 regions of the same NUTS2 region will be assigned the same values. We therefore need an RxE interaction term (i.e. at the NUTS3 level), and possibly such interaction term could also be age- or sex-dependent. This information exists only for specific countries, but not European wide. To fill in this missing interaction we need a model that predicts the educational distribution of each NUTS3 region as a function of regional characteristics. This model, an aggregated logit model (a member of the family of loglinear models), is explained in the next section in some detail. The estimates of this model generate a distribution of level of education as a function of the regional characteristics. We include this term RxE, or, if it is age and sex-specific RxSxAxE, in the model as an *offset* in the model. An offset is a covariate in the model with a fixed parameter value of 1. This could be specified as:

RxExSxA = RxSxA + ExSxA +{ RxExSxA } (19)

where the notation {..} is used to denote the offset. This offset can be interpreted as prior information to be included in the estimation. The parametric form of this model is:

$$\log \widehat{M}_{ijkl} = \mu + \mu_i^R + \mu_j^S + \mu_k^A + \mu_l^E + \mu_{ij}^{RS} + \mu_{ik}^{RA} + \mu_{jk}^{SA} + \mu_{jl}^{SE} + \mu_{kl}^{AE} + \mu_{ijk}^{RSA} + \mu_{jkl}^{SAE} + log\widetilde{M}_{ijkl} \quad (20)$$

As will be explained in the next section, we include the R2xSxE table as a covariate in the estimation of the prior information $log\breve{M}_{ijkl}$. The formula has one intercept, four main effects, five two-way interactions, and two three-way interactions. This formulation makes clear that because of the partial information available, the two-way interaction μ_{il}^{RE}, nor the three-way interaction μ_{ijl}^{RSE} and μ_{ijl}^{RAE} are included (i.e. set to zero). Instead, the prior distribution $\breve{M}_{ijkl}$ contains an approximation of the interactions between these factors, derived from the logit predictions to be discussed below. Iterative Proportional Fitting IPF starts with the prior distribution {RxSxAxE}, which is scaled in a number of rounds to fit it to the marginal totals RxSxA and ExSxA, until convergence is reached.

2.3.3 The aggregated logit model to estimate the regional distribution over educational attainment

An aggregated logit model is equivalent to a loglinear model. In a loglinear model such as equation (15) the dependent variable is the cell count, but there is no dependent or independent variable among the factors that make up the table. The model estimates a multivariate discrete distribution. But we can designate one factor, say education, as the dependent variable, and estimate its value, conditional on the value of the other factors. For instance, we could model the probability that a person living in region *i*, with sex *j* and age *k* will have educational attainment *l*. This probability is

$$Prob_{l|ijk} = p_{l|ijk} = {M_{ijkl}}\Big/{\sum_{l'} M_{ijkl'}} \qquad (21)$$

Substitution of (20) in (21) gives:

$$p_{l|ijk} = \frac{\exp\left(\mu_{jl}^{SE}+\mu_{kl}^{AE}+\mu_{ijl}^{SAE}+\log\left(\breve{M}_{ijkl}\right)\right)}{\sum_{l'}[\exp\left(\mu_{jl'}^{SE}+\mu_{kl'}^{AE}+\mu_{ijl'}^{SAE}+\log\left(\breve{M}_{ijkl'}\right)\right)]} \qquad (22)$$

All terms not related to E cancel out. The *p* can be either interpreted as the probability of having a certain educational level, or as population shares. In the current approach the interpretation of shares is more to the point. The μ-terms describe differences in educational attainment between age- and sex categories, based on the national distribution from the WIC table SxAxE. By using data from countries that have all variables R, S, A and E, we can model the prior distribution $\log\left(\breve{M}_{ijkl}\right)$ with a set of region-specific explanatory variables. The estimated coefficients of such a model can then be applied to predict the value of the prior distribution for other countries without a complete RxSxAxE distribution, but with the explanatory variables. We present the results of the estimation for the Netherlands, as an example of a country with all variables available. We estimate the following model on Dutch data:

$$p_{l|ijk} = \frac{\exp\left(\mu_{jl}^{SE}+\mu_{kl}^{AE}+\mu_{ijl}^{SAE}+\sum_m \beta_{jkl,m}X_{i,m}\right)}{\sum_{l'}[\exp(\mu_{jl}^{SE}+\mu_{kl}^{AE}+\mu_{ijl}^{SAE}+\sum_m \beta_{jkl',m}X_{i',m})]} \qquad (23)$$

Where $\sum_m \beta_{jkl,m}X_{i,m}$ is a linear sum of *m* region-specific variables $X_{i,m}$, with weights $\beta_{jkl,m}$.

The educational attainment data, by region, sex and age (i.e. the observed table ExRxSxA) for the Netherlands come from the Statline database of Statistics Netherlands (Statistics Netherlands). Statistics Netherlands provides information for 10-year age categories (15-25, ..65-75). We use the

2015 and 2020 data for the estimation. Table 5 gives an overview of the explanatory variables used in the model. Note that the educational attainment at the NUTS2 level, R2_Educ15-75, is used as a predictor in this model.

The model was estimated using R function glm as a hybrid log-linear model including all available terms RxSxA and ExSxA, plus quantitative predictors of the regional educational attainment distribution. Table 6 shows the deviance fit for a number of nested models for 2015 and 2020. It gives an impression of the contribution of the explanatory variables to the overall fit between observed and predicted regional educational attainment shares.

Variable	Explanation
R2_Educ15-75	Share of the population in the age range 15-75 by educational attainment and sex, at NUTS2 level
%Hightech	Share of employment in professional, scientific and technical activities; administrative and support service activities.
Econ_index	The economic index as calculated in Work package 4 (Arnold, 2024). It is a composite index of regional product per capita, and unemployment

Table 5 Explanatory variables of the regional NUTS3 share of the population by educational attainment

		2015		2020	
#	Model	Deviance	Df	Deviance	Df
1	RxSxA + ExSxA	298666	936	321639	936
2	Model 1 + R2_Educ15-75	234582	933	240933	933
3	Model 2 + %Hightech	207490	931	232546	931
4	Model 3 + Econ_index	184016	929	183516	929

Tabel 6 Model fit of aggregated logit estimates of educational attainment in NUTS3 regions in the Netherlands, 2015 and 2020

Model 1 in table 6 denotes the baseline model, where all marginal effects are included that are available for all European countries, from Eurostat/JRC (i.e. the table RxSxA, which is the regional population by age and sex) and WIC (i.e. the table ExSxA, which is the table at the national level of the population by level of education, sex and age). In models 2, 3 and 4 the regional predictors of the level of education are added one by one. The three explanatory variables have a substantial effect on the model outcomes. The economic index, one of the dimensions of the regional development concept, is the strongest variable. Further interactions of these variables with age and sex do not add much to the fit. This means that the coefficients apply to all age- and sex categories simultaneously, or in other words: the β_{jkl}'s can be simplified to β_l.

The parameter estimates β_l of these three explanatory variables are given in table 7.

	2015			2020		
	Low	**Middle**	**High**	**Low**	**Middle**	**High**
Intercept	3,145	3,035	(ref)	3,176	3,430	(ref)
R2_Educ 15-75 (/ 10000)	-0,15	-0,12	(ref)	-0,25	-0,15	(ref)
% Hightech (/ 100)	0,255	(ref)	0,818	0,638	(ref)	-2,162
Econ_index	-0,178	(ref)	1,684	0,004	(ref)	2,760

Table 7 Parameter estimates of explanatory variables in 2015 and 2020

The most important variable is econ-index. The higher the economic index of a region the higher the share of high educated, as expected. In 2015 it also correlates negatively with the share of low educated. The percentage of employed in the Hightech sector is positive for the low, and mixed for the share of high educated: positive in 2015 but negative in 2020. The effect of the NUTS2 educational distribution is marginal and negative for low and middle educated shares, relative to the high educated shares at the NUTS3 level. This implies that on average higher shares of low or middle educated at the NUTS2 level correlate with lower shares at the NUTS3 level. Including interaction effects with either age or sex does not greatly improve the fit of the model, while making it substantially more complicated.

Using these parameter estimates a prior distribution $\log \widehat{M}_{ijkl}$ can be estimated. Figure 2 shows the expected and observed shares based on the 2015 data. The figure shows a decent fit (R^2 is 0.88 for 2015, and similarly 0.87 in 2020).

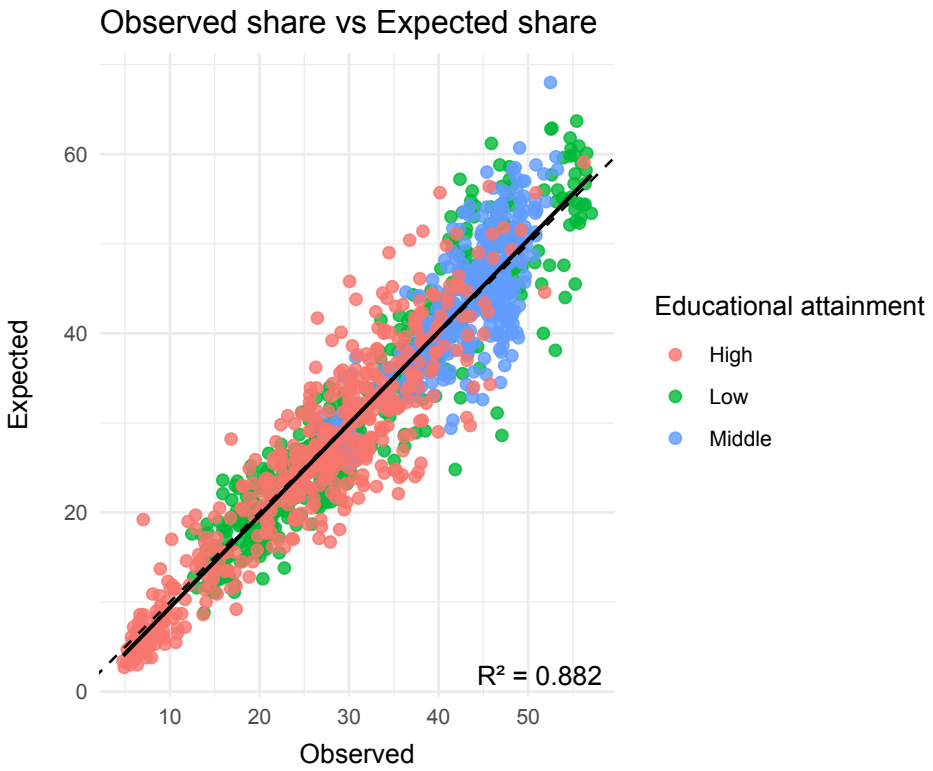


Figure 2 Observed and expected shares of levels of education for NUTS3 regions by age and sex, 2015, in the Netherlands

2.3.4 How good does this model work for other countries?

We use the estimated values of the logit model for the Netherlands to construct the {RxE} interaction term to be used as prior distribution in the estimation of the population by region, age, gender, and educational attainment for Norway. The model to be fitted using IPF is the RxSxA + ExSxA + {RxE}. The estimated values can be compared with the observed counts, which are available from Statistics Norway. Norway has (as of 2023) 11 NUTS3 regions, and Statistics Norway publishes data on regional educational attainment for 6 age categories. In total there are therefore 11x2x6x3=396 cell counts. Figure 3 shows the fit of the resulting model. Each dot represents observed and expected shares of low, middle and high educated for each combination of region, age and gender. The fit is very satisfactory.

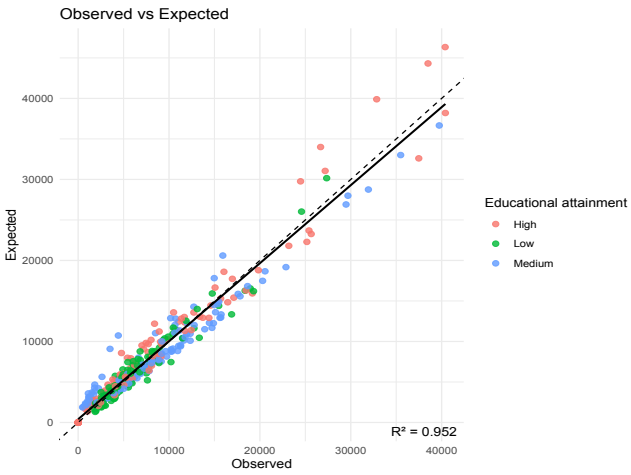


Figure 3 Observed and Expected counts of low, middle and high educated by age, gender and region, Norway, 2015

2.4 The population by age, gender and level of education at the NUTS3 level in Europe

The variables Econ_index, %Hightech and R2_Educ15-75 (the NUTS2 level of education), together with the estimated coefficients from the Dutch logit model would generate a prior distribution {RxE} for each country. Unfortunately, data for Austria, Germany and Spain are not available in the Eurostat Regions database[3]. For all other countries, the educational distribution, by NUTS3 region, age and gender can be estimated. We present here results for 173 NUTS3 regions. These data are included in the Regional Policy Dashboard[4].

There is no direct test of the results of the estimation of the population by sex, age, and level of education at the NUTS3 level, for the reasons mentioned above. At the NUTS2 level the 2021 census information of Eurostat enables a partial test at the European level which answers the question how good the model replicated the level of education at this geographical scale. The census 2021 provides a table of the population by NUTS2, sex, age in 6 groups (0-14, 15-29, 30-49, 50-65, 65-84 and 85+), and educational attainment in 11 categories, which we aggregate to Low, Middle and High educated. Figures 4 and 5 provide evidence of the fit.

[3] We hope to include these countries in an update of the database

[4] A prototype of the Dashboard is available at: https://nordregio.github.io/premium_eu/regional.html. The Dashboard will be launched in the spring of 2026 (Deliverable 7.8)

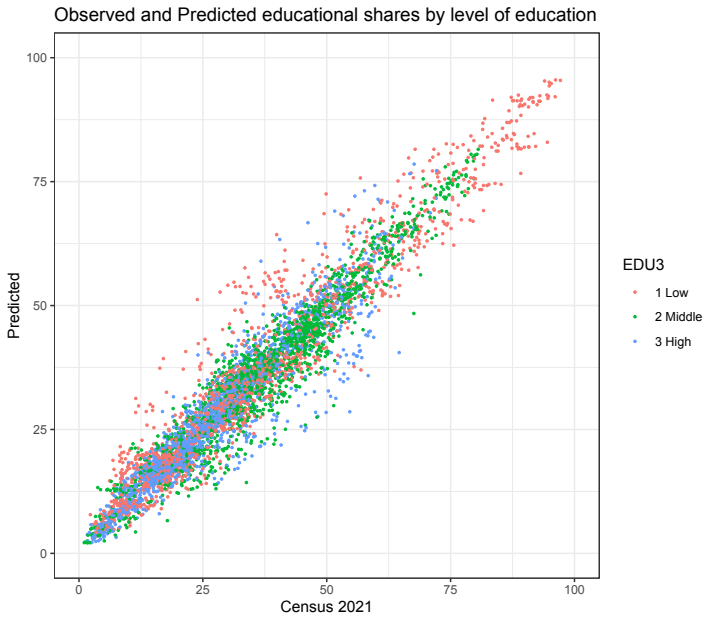


Figure 4 Observed and predicted educational attainment shares by NUTS2 regions, categorized by level of education

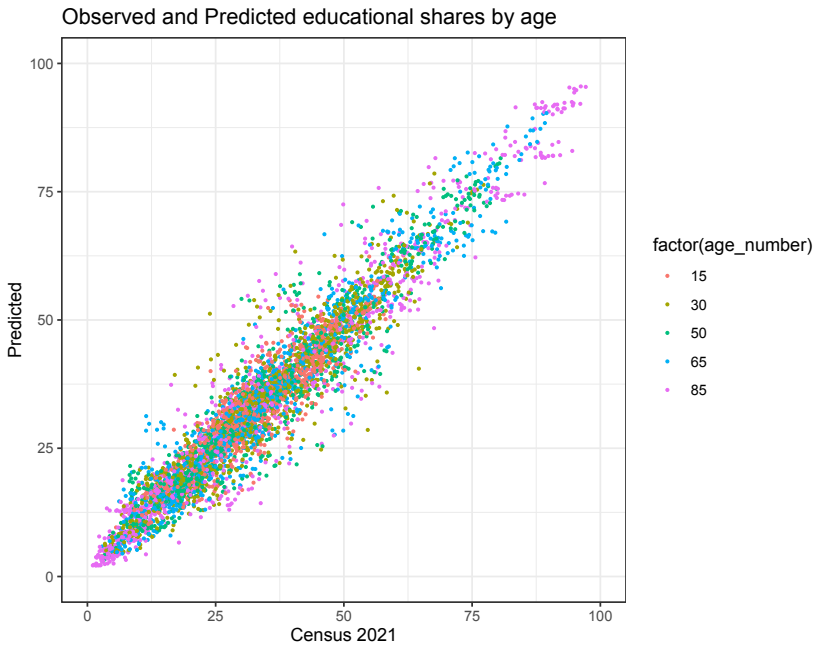


Figure 5 Observed and predicted educational attainment shares by NUTS2 region, categorized by age group

Figure 4 provides a categorisation by level of education, and Figure 5 by age group. The R^2 between observed educational shares and predicted shares is with 0.92 satisfactory. The slope of the regression line is 0.97, and the slope 0.9 (i.e. less than 1 percentage point at the lowest values), indicating only a very slight bias. The results for each of the educational categories is very similar, with R^2s of 0.93, 0.91 and 0.89 for Low, Middle and High educated respectively.

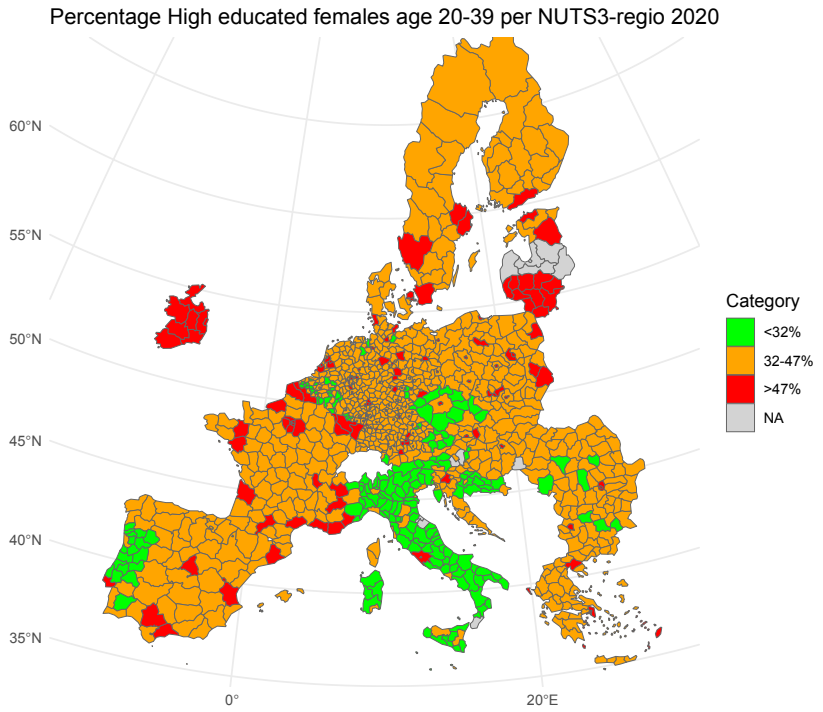
Figure 6 shows the resulting pattern for the high educated 20-39 year old females and males in 2020. Although the levels are different, with females on average higher educated (40 against 31 percent), the pattern is quite similar. Also visible are the country differences. This has partly to do with real differences in educational attainment (Portugal), but to some extent also with different educational systems and definitions (Italy).

2.5 Net migration by age, gender and level of education at the NUTS3 level in Europe

With $\mathbf{P}_t$, and $\mathbf{P}_{t+5}$ estimated, for t = 2010, 2015 and 2020, $\Delta\mathbf{E}$ estimated, and $\mathbf{B}_{t,t+5}$ and $\mathbf{D}_{t,t+5}$ observed, we have all required information to estimate $\mathbf{N}_{t,t+5}$ for the periods 2010-2015 and 2015 -2020, using equation (12). Figure 7 gives an idea of the result of the estimation. (Note that due to data limitations the results for a number of countries -France, Italy, Austria, Hungary- are not yet fully available. The data will be updated in the coming period however).

2.6 Conclusions of the estimation of net migration by level of education

In this chapter a method to estimate net migration by level of education has been developed. The method is essentially combining different sources of information together at the national and regional level, using iterative proportional fitting to find the minimum information distribution for regional populations, and demographic accounting principles to estimate the resulting net migration estimates from these population estimates. We could partially test the results. We used data from the Netherlands and Norway to estimate a model that predicts the regional educational attainment shares from exogenous information on the NUTS2 educational distribution and the regional economic index developed in Work Package 4, that includes Gross Regional Product and unemployment as indicators. These models fitted the data well. We used these models to predict the educational attainment shares for all European regions. That is clearly a strong assumption. We hypothesize that there is a clear relationship between regional educational attainment and economic indicators (which is the basis of the model estimated for the Netherlands and Norway), but the strength of the relationship may vary across countries. With more data available on regional educational attainment at the NUTS3 level we will investigate this relationship further, and include the results in the model estimates, whenever possible. We tested the results of the estimation of the population by level of education (and by sex and age) at the NUTS2 level, using the census 2021 data from Eurostat. The results were quite satisfactory, but this test does not say anything about the distribution of educational attainment at the NUTS3 level *within* NUTS2 regions. With more data at the NUTS3 level available our model results could be refined. This will only be possible for individual countries that have these data available. The Eurostat data are based on the European Labor Force Survey, which is not representative at the NUTS3 level.



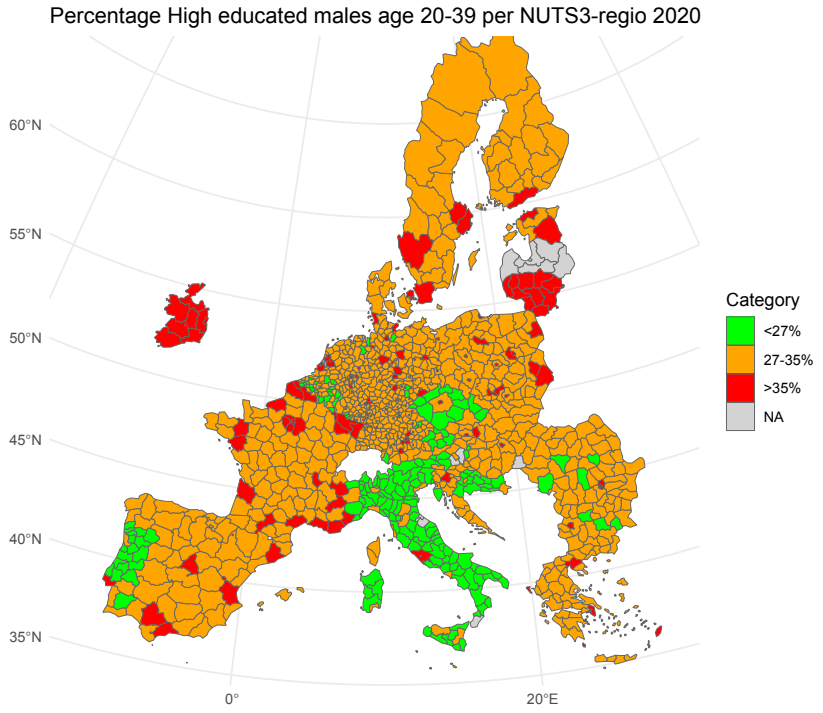


Figure 6 Estimated share of high educated females and males in NUTS3 regions in 2020

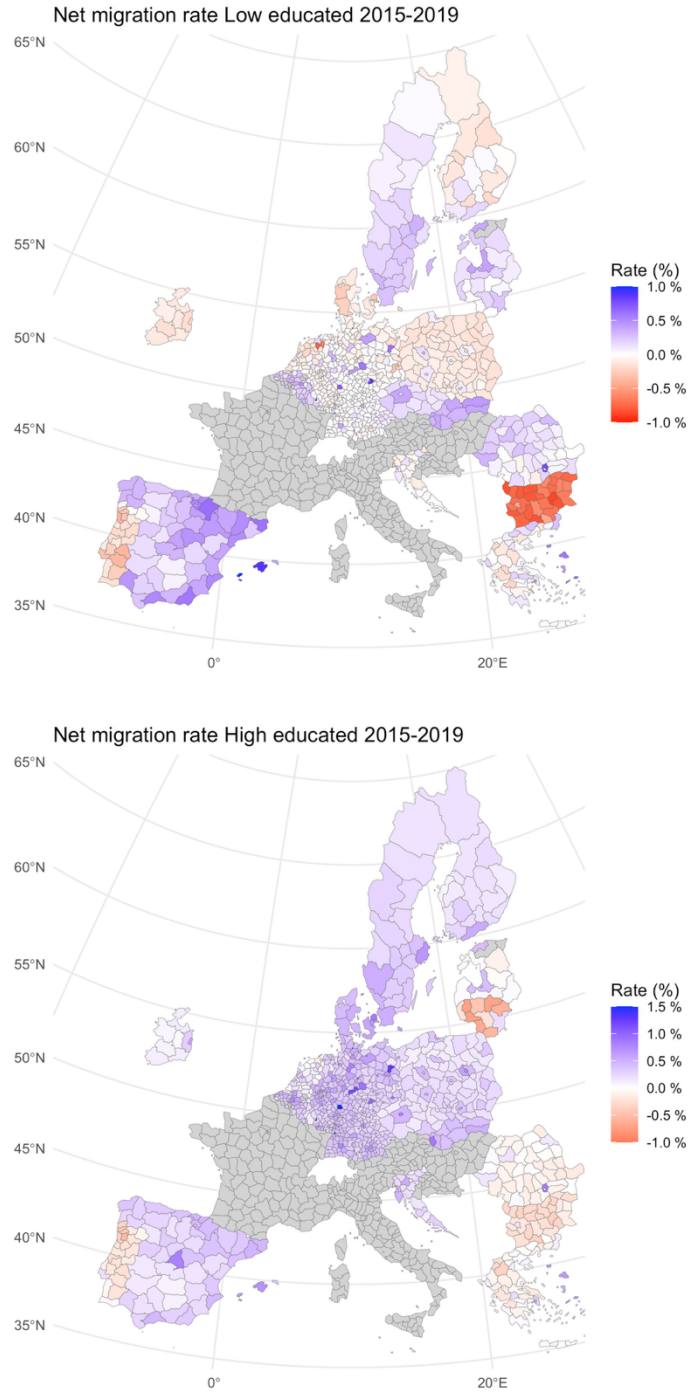


Figure 7 Estimated net migration rate of Low and High educated in net migration 2015-2019

References

Aitkin, M. A., Francis, B., & Hinde, J. (2005). *Statistical modelling in GLIM 4* (Vol. 32). Oxford University Press, USA.

Arnold, B. (2024) Deliverable 4.1 Workpackage 4, Horizon Europe. url….

Naimi, A. I. & Balzer, L. B. (2018). Stacked generalization: An introduction to super learning. *Eur. J. Epidemiol.* **33**, 459–464.

Preston, S. H., Heuveline, P., & Guillot, M. (2001). *Measuring and Modeling Population Processes*, Oxford: Blackwell Publishers, 2001, xv+ 291 pp.

Raymer, J., de Beer, J., & van der Erf, R. (2011). Putting the pieces of the puzzle together: Age and sex-specific estimates of migration amongst countries in the EU/EFTA, 2002-2007. European Journal of Population 27(2): 185-215. doi:10.1007/s10680-011-9230-5.

Rogers, A. & Castro, L. J. (1981). Model Migration Schedules. *Research Report RR-8 1-30*, International Institute for Applied Systems Analysis, Austria.

Rowland, D. T. (2003). *Demographic methods and concepts*. Oxford University Press, Oxford.

Ruggles, S. *et al.* (2025). Integrated Public Use Microdata Series, International: Version 7.6 [dataset]. Minnesota Population Center https://doi.org/10.18128/D020.V7.6.

Van Der Laan, M. J., Polley, E. C. & Hubbard, A. E. (2007). Super Learner. *Stat. Appl. Genet. Mol. Biol.* **6**.

Yeung, J., Alexander, M. & Riffe, T. (2023). Bayesian implementation of Rogers–Castro model migration schedules: An alternative technique for parameter estimation. *Demographic Research*. 49, 1201–1228.

Yildiz, Dilek, & G. J. Abel. (2024). Migration Flows by Age, Sex and Educational Attainment. IIASA Working Paper WP-24-001.