Title: Estimating Net Migration by Level of Education in European NUTS3 Regions

Abstract: Although migration of people with various educational levels is essential for regional development all over Europe, very few European countries have available data at the regional NUTS3 level on migrations by age, sex and educational level.

In this study, we have developed a method for estimating net migration for all EU's NUTS3 regions by level of education. 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 start by estimating a model that predicts the population shares with high, medium and low education in each NUTS3 region, by sex and 5-year age groups. We use data from the Netherlands and Norway to estimate and test the model, where explanatory variables are the NUTS2 educational distribution and a regional economic index that includes Gross Regional Product and unemployment as indicators. Second, from these estimated population shares we use demographic accounting principles to calculate net migration from/to each of EU's NUTS3 regions, by age, sex and level of education. This is done for the years 2010, 2015 and 2020, and the model will also be used to make projections up until 2040.

Authors: Leo van Wissen [1], Becky Arnold [1], Marianne Tønnessen [2]

[1] NIDI - Netherlands Interdisciplinary Demographic Institute

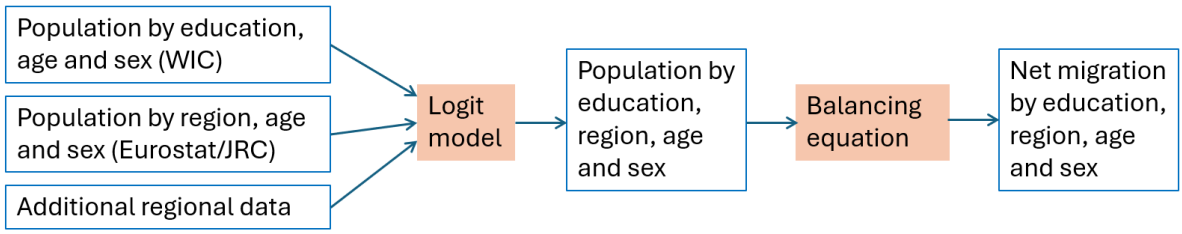
[2] NIBR, Oslo Metropolitan University, Norway

Extended abstract

Introduction

Regional development in various parts of Europe is highly dependent on the educational level of the people who live, move into or move out of each region. In European data, however, little information is available on the population's educational level at the NUTS3 regional level, and even less is known about the levels of education among those who move into and out of different regions.

In this study, we develop a method which combines various known data sources to estimate population and net migration by NUTS3 region, age, sex, and level of education. The study consists of two main tasks: 1) Estimating population in all NUTS3 regions in the EU by age, sex and education, and 2) using these estimates to calculate estimates of net migration for each NUTS3 region. A simplified version of the methods and data are illustrated below, before the two main tasks will be presented in detail.



First task: Estimating the regional population by age, sex and education

The first main task, estimation of population data by NUTS3 region, age (in 5 year age groups 0-4, 5-9, .., 85+) and sex, builds on partial information from several sources. From the REGIONS database of Eurostat, and harmonized by the Joint Research Centre JRC, time series exists for population by NUTS3 region, age and sex 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. However, 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, and detailed information of educational attainment by age and sex is provided at the national level, but not for regions within countries, by the Wittgenstein Centre for Demography and Human Capital WIC. Eurostat/JRC, based on European labor force surveys, also has some information about the population of working age (15-75 years) by sex and educational attainment, at the NUTS2 level. The available partial regional information is summarized in table 1.

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 1: Data resources for estimation

Model specification

The problem is specified using the GLIM (Generalized Linear Interactive Modelling) language. We have developed a predictive (logit) model with exogenous regional characteristics, where the probability *p* that a person living in region *i*, with sex *j* and age *k* will have educational attainment *l*, is

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

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 *p* can be either interpreted as the probability of having a certain educational level, or as population shares (which is more to the point in the current approach). The μ-terms describe differences in educational attainment between age- and sex groups, based on the national distribution from the WIC table SxAxE (Sex x Age x Education).

To estimate the model, we use educational attainment data, by region, sex and age (i.e. the observed table ExRxSxA) for the Netherlands, from the Statline database of Statistics Netherlands. We use the 2015 and 2020 data for the estimation. Table2 gives an overview of the explanatory variables used in the 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 3 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
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 by Arnold (2024); a composite index of regional product per capita and unemployment

Table 2: Explanatory variables

		2015		2020	
#	Model	Deviance	Df	Deviance	Df
1	ExRxSxA + ExSxA	298666	936	321639	936
2	Model 1 + Educ15-75	234582	933	240933	933
3	Model 2 + %Hightech	207490	931	232546	931
4	Model 3 + Econ_index	184016	929	183516	929

Tabel 3: Model fit of aggregated logit estimates of educational attainment in NUTS3 regions in the Netherlands, 2015 and 2020

Table 3 shows that the three explanatory variables have a substantial effect on the outcomes. Table 4 shows the parameter estimates of these three explanatory variables.

	2015			2020		
	Low	**Middle**	**High**	**Low**	**Middle**	**High**
Intercept	3,145	3,035	(ref)	3,176	3,430	(ref)
NUTS2 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 4: Parameter estimates of explanatory variables in 2015 and 2020

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 high tech sector is positive for the low, and mixed for the share of high educated: positive in 2015 but negative in 2020.

Figure 1 (see Figure Appendix) 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). Using these parameter estimates a prior distribution $\log \widehat{M}_{ijkl}$ can be estimated. This prior distribution contains the important {RxE} interaction term.

Next, we test if the estimated model for the Netherlands can be used to construct the regional educational distribution by age and gender for Norway. The estimated values can be compared with the observed counts, from Statistics Norway. Figure 2 shows the fit, each dot representing observed and expected shares of low, middle and high educated for each combination of region, age and gender. The fit is very satisfactory.

The variables Econ_index, %Hightech and NUTS2-Educ15-75, together with the estimated coefficients from the Dutch logit model would generate a prior distribution {RxE} for each country. Hence, for most countries in the EU (with sufficient data in the Eurostat Regions database) the educational distribution, by NUTS3 region, age and gender can be estimated, which we did for a total of 173 NUTS3 regions. These results can be partially tested using the NUTS2 level information from Eurostat's 2021 census. The test is partial because it only answers the question how well the model replicated the level of education at this geographical scale. Figures 3 and 4 provide evidence of the fit. Figure 3 provides a categorisation by level of education, and Figure 4 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 intercept 0.9 (i.e. less than 1 percentage point at educational shares close to zero), indicating only a very slight bias. The results for each of the educational categories are very similar, with R^2s of 0.93, 0.91 and 0.89 for Low, Middle and High educated respectively.

Figure 5 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 is partly linked to real differences in educational attainment (Portugal), but to some extent also to different educational systems and definitions (Italy).

Second main task: Estimating net migration for each region by age, sex and education

The method used to estimate net migration for each of the NUTS3 regions is based on demographic accounting principles and essentially starts with the balancing equation to arrive at net migration for a given region:

$$N_{t,t+5} = P_{t+5} - P_t - B_{t,t+5} + D_{t,t+5}$$

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. Our model works as well for estimates of the period 2010-2020 and for projections 2020-2040.

Several elements of the model complicate the straightforward application of the balancing equation. The population is defined by age, sex, and educational attainment. Moreover, educational attainment is a dynamic factor, governed by educational transition parameters. These complications are built into the balancing equation to arrive at a model of net migration by region, level of education, sex and age:

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

where $\Delta\mathbf{E}$ is the transition matrix of educational attainment, estimated based on the national educational attainment levels (High, Medium and Low) as given by WIC. With three educational levels, 18 age classes and 2 sexes the total number of rows and columns of $\Delta\mathbf{E}$ is 108 (3x18x2). $\Delta\mathbf{E}$ is a subdiagonal blockmatrix with age- and sex-specific 3x3 educational transition matrices in the subdiagonal.

With $\mathbf{P}_t$, and $\mathbf{P}_{t+5}$ estimated for t = 2010, 2015 and 2020 (in our first step), $\Delta\mathbf{E}$ 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.Figure 6 shows the geographical distribution of migration the estimate produces. (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).

Figure Appendix

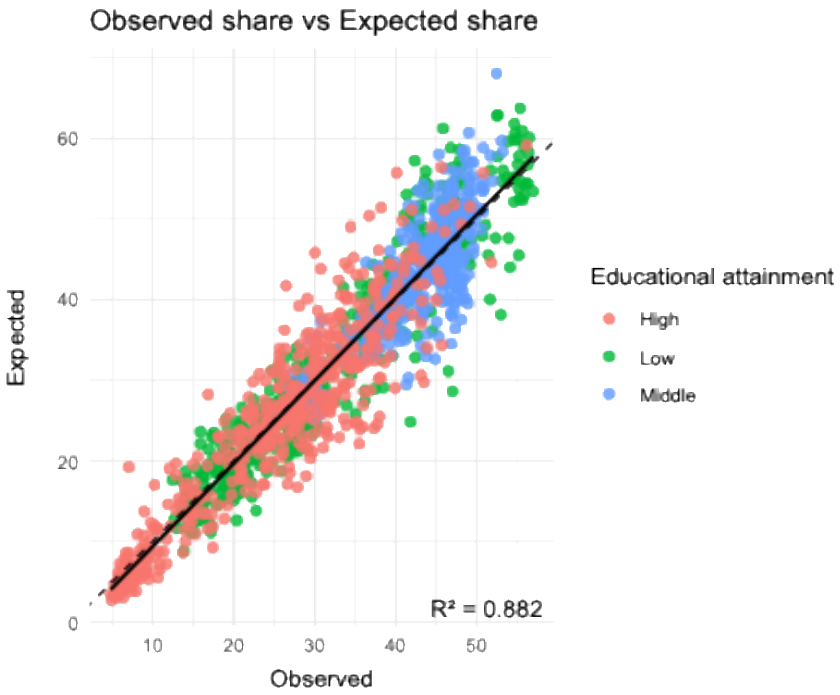


Figure 1 Observed and expected shares of levels of education for NUTS3 regions by age and sex, 2015, in the Netherlands

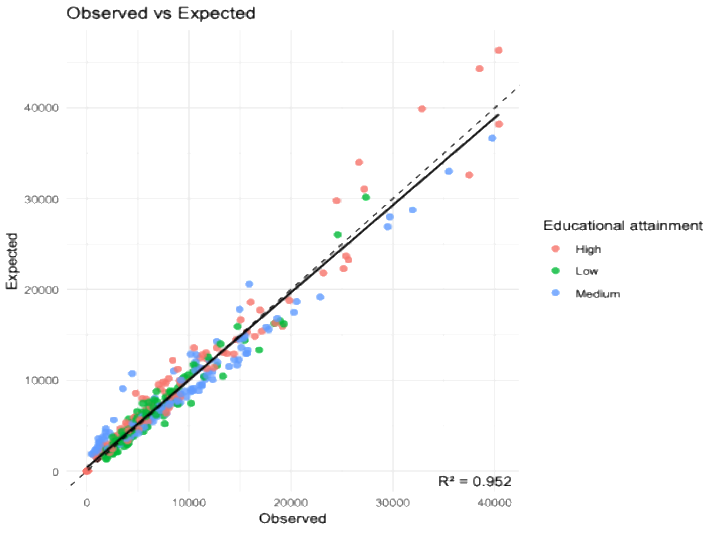


Figure 2 Observed and Expected counts of low, middle and high educated by age, sex and region, Norway, 2015

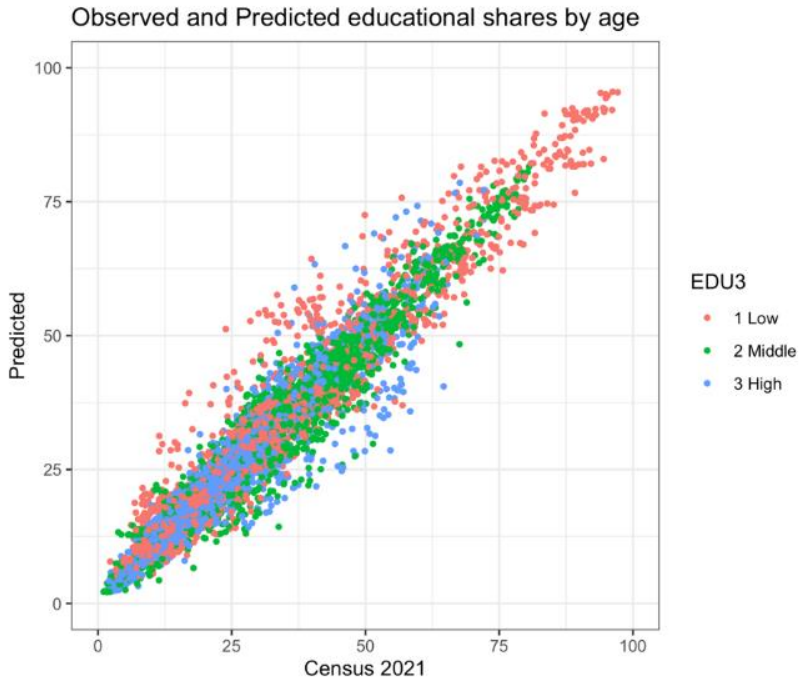


Figure 3 Observed and predicted educational attainment shares by NUTS2 regions, categorized by level of education

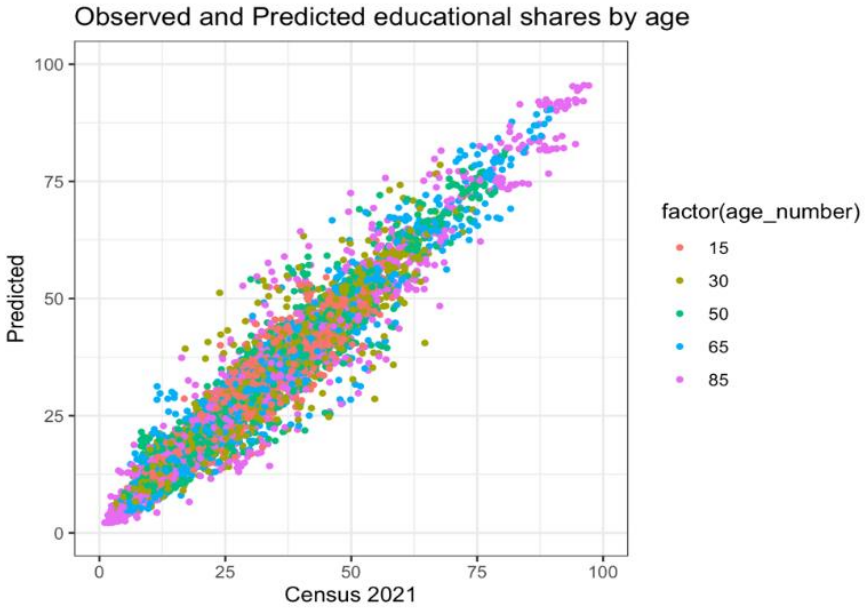
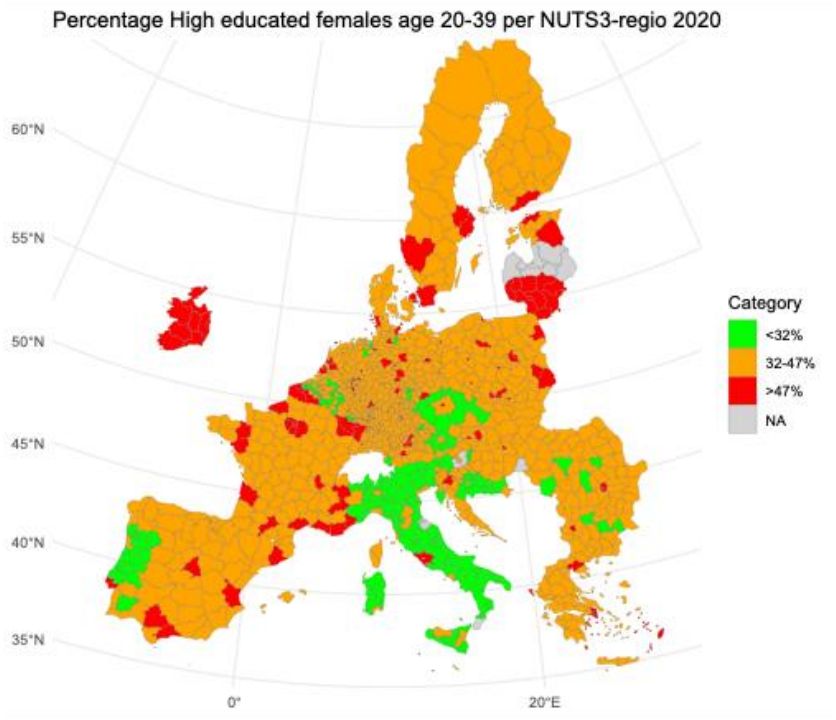


Figure 4 Observed and predicted educational attainment shares by NUTS2 region, categorized by age group



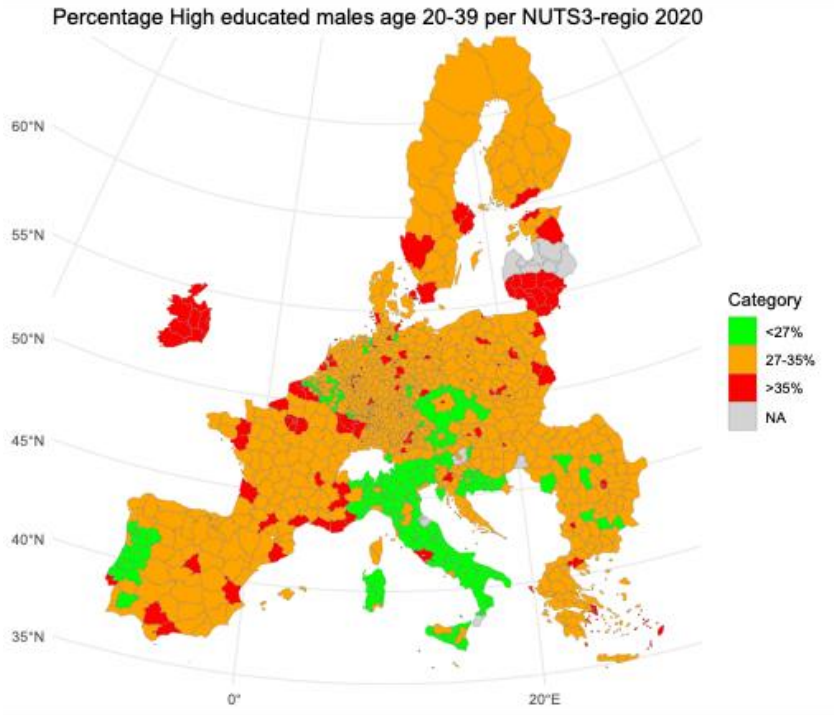


Figure 5 Estimated share of high educated females and males in NUTS3 regions in 2020

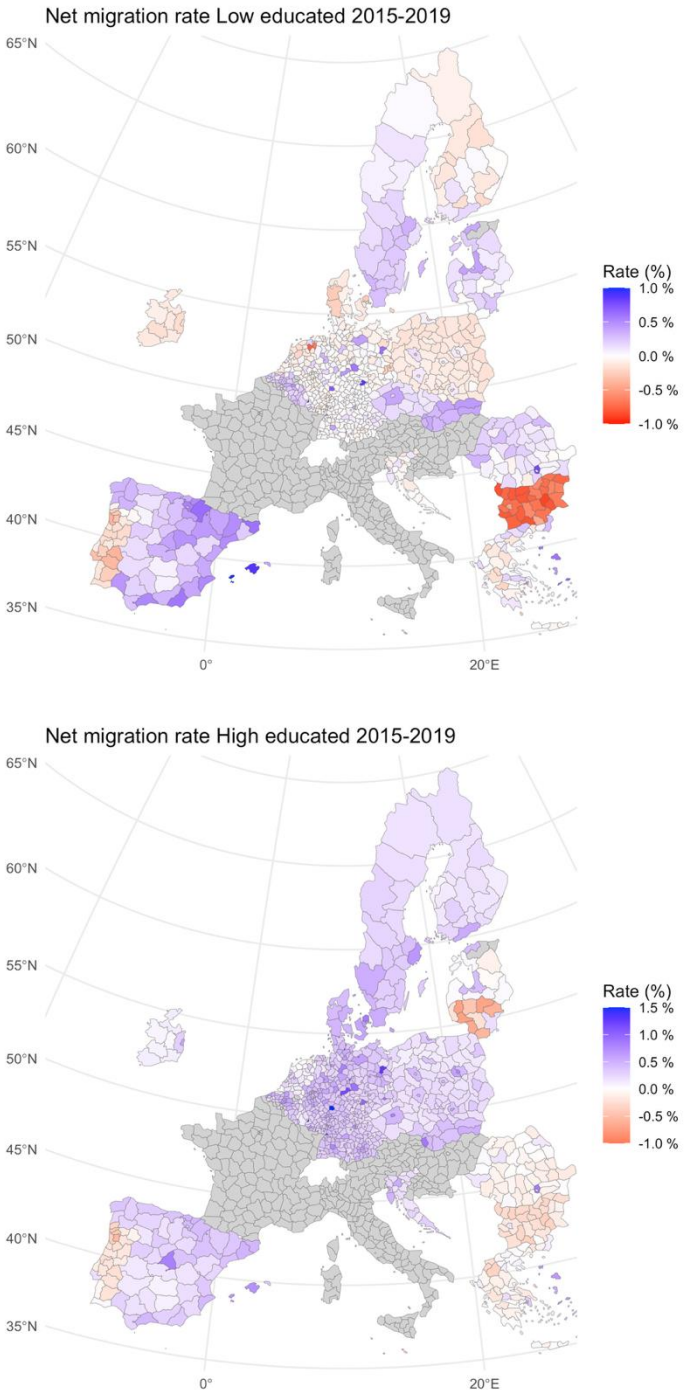


Figure 6 Estimated percentage Low and High educated in net migration 2015-2019