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 level on migrations by age, sex and educational level.

In this study, we have developed methods for estimating net migration for all EU's NUTS3 regions by level of education. The methods essentially combine different sources of information together at the national and regional level, using iterative proportional fitting to estimate the educational distribution for regional populations, and demographic accounting principles to estimate the resulting net migration estimates from these population estimates.

We start by estimating the population shares with high, medium and low education in each NUTS3 region, by sex and 5-year age groups. Data from the Netherlands and certain other European countries are used to estimate and test a logit model where explanatory variables are the NUTS2 educational distribution and a regional economic index that includes Gross Regional Product and unemployment as indicators. The model provides prior distributions of educational shares, which through iterative proportional fitting are aligned with existing, partial data. Second, from these population estimates 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 for 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

Introduction

Regional development in all parts of Europe is closely linked to migration of people with various educational levels. For regional policy makers, projections of the regions' future net migration – broken down by educational level – could constitute significant input in decision making about industrial development, service provision and strategies to counteract labor shortages.

In European statistics, however, very little is known about the levels of education among those who move into and out of different regions and how this can be expected to develop in the decades to come.

In this study, we have developed methods which combines various known data sources to estimate population and net migration by NUTS3 region, age, sex, and level of education for the EU and Norway for 2020 and projected until 2040. The educational dimension at such a low geographical level, with estimates for both population and net migration, for current and future periods, and for all EU countries, is unprecedented in previous research.

Previous work

Recent decades have seen large progress in the estimation and projection of populations and migrations by educational attainment, at the national level (see, for instance, Lutz et al. 2014, Speringer et al. 2019). Most European countries, as well as the EU's statistical office Eurostat, now provide data on the educational attainment of the population, sometimes also by age and sex. A focus on the educational attainment of populations and migrants is not surprising, since education is seen as a major engine of the demographic transition and the demographic dividend, and given the well-established link between education and development (Murtin 2013, Lutz et al. 2019, Psacharopoulos & Patrinos 2018, Hanushek & Woessmann 2023).

At the sub-national level, however, considerably less data exist – although previous research has shown that also at such lower geographical levels the population's educational attainment is closely linked to regional development, with multiple interlinkages between regional development and migration selectivity (Faggian&McCann 2009, Bell et al. 2015, van Wissen & Schutjens 2015, Fratesi & Percoco 2014). The educational selectivity is large among migrants, with the highly educated having much higher migration rates (Bernard and Bell 2018, González-Leonardo et al. 2022).

Given the relevance of the population's educational attainment on regional development, with some regions winning and other reasons loosing vital human capital, it is remarkable that available data on regional populations and migrants' educational attainment are so scarce. Some national statistical offices in Europe provide sub-national data on the population's educational attainment (not necessarily by age and sex),[1] and Eurostat has data on educational attainment of the population stock at NUTS2 level,[2] but when it comes to the educational attainment of those moving into and out of Europe's regions, very little data is available.[3] Moreover, when it comes to population projections, many subnational projections exist for European countries, including by Eurostat (see also Rees et al. 2012), but even those statistical offices with the best register data do not make sub-national projections by educational attainment, neither of population stocks nor of (net) migration.

In this study, we aim at filling some of this gap, building on advances made in estimation and projections of migration at the national level,[4] but with a focus on the NUTS3 regions,[5] of which there are currently 1,165 across Europe.

Methodology

In short, our contribution consists of two main tasks: 1) Estimating population in all NUTS3 regions by age, sex and education, and 2) using these estimates to calculate estimates of net migration for each NUTS3 region by education.

Figure 1 shows the methods and the data used in each of these two main tasks. Below, the tasks will be presented in detail.

[1] See for instance for Spain: Population 16 years of age and over by level of education attained, sex and Autonomous Community. Absolute values(66104), France: Population des 16 ans ou plus selon le niveau de diplôme, le sexe et l'âge de 1968 à 2022 (1990 à 2022 pour les DOM) | Insee, Italy: Popolazione per titolo di studio e regioni, Hungary: KSH Statinfo v40, Ireland: https://data.cso.ie/table/EDQ03, Finland: Population aged 15 or over by level of education, municipality, gender and age by Year, Area, Age, Gender, Level of education and Information. PxWeb, Sweden: Population 16-74 years of age by region, highest level of education, age and sex. Year 1985 - 2024. PxWeb, Norway: 09429: Educational attainment, by municipality and sex (M) 1970-2024 Denmark: https://www.statistikbanken.dk/statbank5a/default.asp?w=1280

[3] Hence, the internal migration variable used by for instance González-Leonardo et al. (2022) is deducted from changing NUTS-2 region of residence.

[4] See for instance the MIMOSA IMEM QUANTMIG projects

[5] NUTS = Nomenclature of territorial units for statistics, developed by the EU. NUTS3 is the most detailed level. https://ec.europa.eu/eurostat/web/nuts/

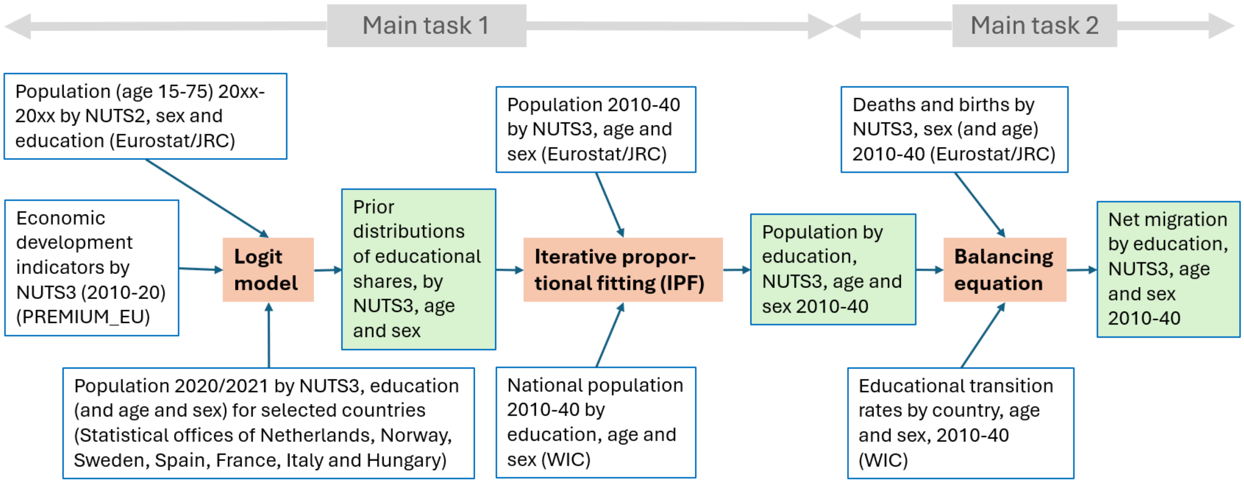

Figure 1 Main tasks of this study, with available data used (white boxes), models applied (orange boxes) and data sets produced (green boxes)

Estimating the regional population by age, sex and education

The first main task, estimation of population data by NUTS3 region, sex, and age (in 5 year age groups 0-4, 5-9, .., 85+), will provide population data which is crucial for the estimation of net migration in our second main task.

This task builds on partial information from several sources (see also Figure 1): Population data by NUTS3 region, age (in 5 year age groups 0-4, 5-9, .., 85+) and sex is available from 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 (see Appendix 1 for details). For the Netherlands, for example, having 40 NUTS3 regions (so called COROPs) this table would contain 40x3x2x18 = 4320 cells. 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 but not broken down by age. More detailed information of educational attainment is provided at the national level by the Wittgenstein Centre for Demography and Human Capital WIC. This available partial regional information is summarized in table 1. The notation in the first column of the table will be explained below.

Table 1 Existing partial data sources on population by region (R), gender (S), age (A) and education (E)

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

As can be seen from the table, 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. The predictive logit model provides prior distributions of educational shares which will be aligned with the existing WIC SxAxE and Eurostat/JRC RxSxA data (table 1) through iterative proportional fitting.

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:

$$\log \widehat{M}_{ij} = \mu + \mu_i^A + \mu_j^B \qquad (1)$$

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:

$$\mu = \frac{\sum_{i,j} log\, \widehat{M_{ij}}}{IxJ},\ \mu_i^A = \frac{\sum_{,j} log\, \widehat{M_{ij}}}{J} - \mu,\ \mu_j^B = \frac{\sum_{,i} log\, \widehat{M_{ij}}}{I} - \mu \qquad (2)$$

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 (1) 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:

$$\log \widehat{M}_{ij} = \mu + \mu_i^A + \mu_j^B + \mu_{ij}^{AB} \tag{3}$$

and the interaction term in this model is estimated as:

$$\mu_{ij}^{AB} = log\ \widehat{M_{ij}} - \mu - \mu_i^A - \mu_j^B \tag{4}$$

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 (5)

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:

$$\log \widehat{M}_{ijkl} = \mu + \mu_i^R + \mu_j^S + \mu_k^A + \mu_l^E + \mu_{ij}^{RS} + \mu_{ik}^{RA} + \mu_{jk}^{SA} + \mu_{jl}^{SE} + \mu_{kl}^{AE} + \mu_{kl}^{R2E} + \mu_{ijk}^{RSA} + \mu_{jkl}^{SAE} + \mu_{jkl}^{R2SE} \tag{6}$$

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 } (7)

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:

$$\log \widehat{M}_{ijkl} = \mu + \mu_i^R + \mu_j^S + \mu_k^A + \mu_l^E + \mu_{ij}^{RS} + \mu_{ik}^{RA} + \mu_{jk}^{SA} + \mu_{jl}^{SE} + \mu_{kl}^{AE} + \mu_{ijk}^{RSA} + \mu_{jkl}^{SAE} + log\breve{M}_{ijkl} \qquad (8)$$

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.

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 (3) 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

$$Prob_{l|ijk} = p_{l|ijk} = {M_{ijkl}}\Big/{\sum_{l'} M_{ijkl'}} \tag{9}$$

Substitution of (8) in (9) gives:

$$p_{l|ijk} = \frac{\exp\left(\mu_{jl}^{SE}+\mu_{kl}^{AE}+\mu_{ijl}^{SAE}+\log\left(\breve{M}_{ijkl}\right)\right)}{\sum_{l'}\left[\exp\left(\mu_{jl'}^{SE}+\mu_{kl'}^{AE}+\mu_{ijl'}^{SAE}+\log\left(\breve{M}_{ijkl'}\right)\right)\right]} \tag{10}$$

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:

$$p_{l|ijk} = \frac{\exp\left(\mu_{jl}^{SE}+\mu_{kl}^{AE}+\mu_{ijl}^{SAE}+\sum_m \beta_{jkl,m} X_{i,m}\right)}{\sum_{l'}\left[\exp\left(\mu_{jl}^{SE}+\mu_{kl}^{AE}+\mu_{ijl}^{SAE}+\sum_m \beta_{jkl',m} X_{i',m}\right)\right]} \tag{11}$$

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 2 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 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.

Table 2 Explanatory variables of the regional NUTS3 share of the population by educational attainment

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 by Arnold (2024). It is a composite index of regional product per capita, and unemployment.

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

		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

Model 1 in table 3 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 4.

Table 4 Parameter estimates of explanatory variables in 2015 and 2020

	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

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 $\log \widehat{\widehat{M}}_{ijkl}$ 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).

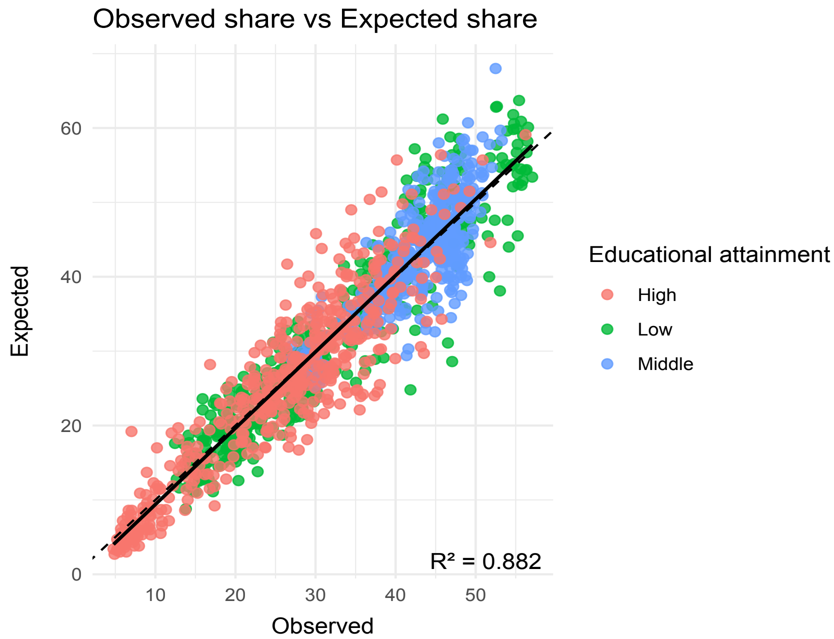

Figure 2 Observed and expected shares of levels of education for NUTS3 regions by age and sex, 2015, in the Netherlands

How well 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.

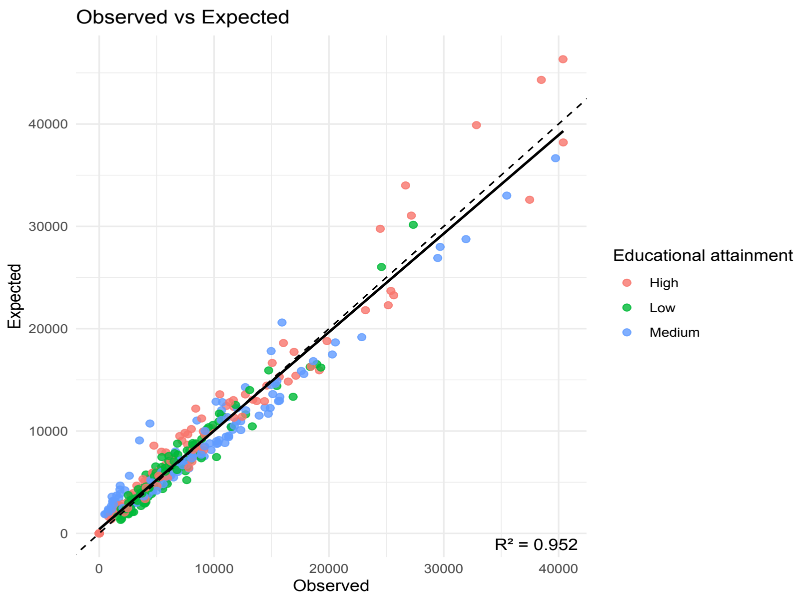

Figure 3 Observed and Expected counts of low, middle and high educated by age, sex and region, Norway, 2015

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. 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 from the PREMIUM_EU project.[6]

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 provide evidence of the fit.

[6] The Dashboard is available at https://nordregio.github.io/premium_eu/regional.html

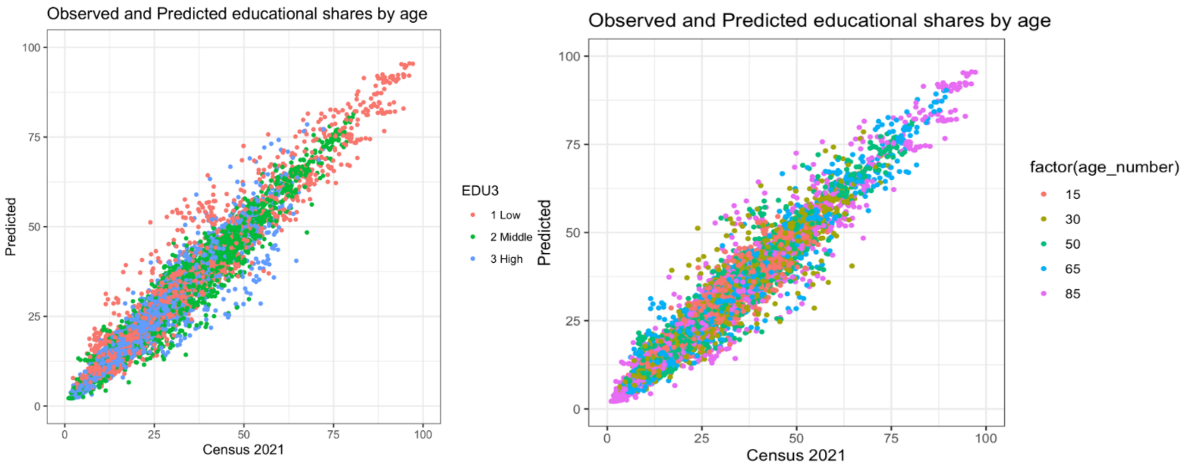

Figure 4 Observed and predicted educational attainment shares by NUTS2 regions, categorized by level of education (left) and age groups (right)

Figure 4's left panel provides a categorisation by level of education, and the right panel 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 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 has partly to do with real differences in educational attainment (Portugal), but to some extent also with different educational systems and definitions (Italy).

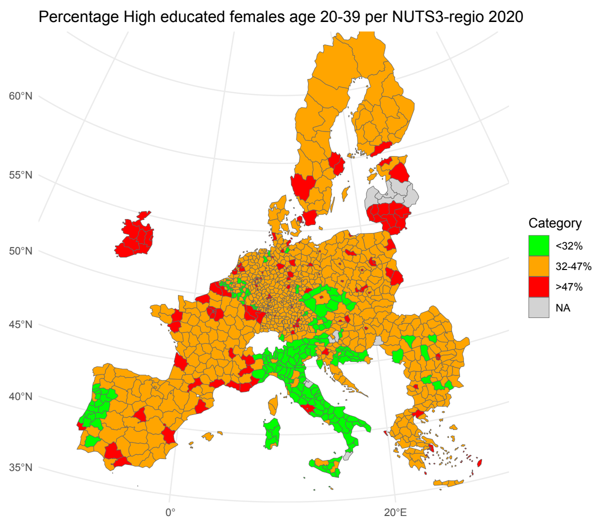

Figure 5 Figure 5 Estimated share of high educated females (left) and males (right) in NUTS3 regions in 2020

Estimating net migration for each NUTS 3 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:

$$N_{t,t+5} = P_{t+5} - P_t - B_{t,t+5} + D_{t,t+5} \tag{12}$$

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.

Model structure

Balancing equation by age and gender

Adding age and gender dimensions to the balancing equation (12) gives the following equation (which is called the forward survival method to estimate net migration,see Rowland (2003):

$$N_{s,a,t,t+5} = P_{s,a+5,t+5} - P_{s,a,t} + D_{s,a,t,t+5} \tag{13}$$

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:

$$N_{s,0,t,t+5} = P_{s,0,t+5} - B_{s,t,t+5} + D_{s,0,t,t+5} \tag{14}$$

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

$$N_{s,85+,t,t+5} = P_{s,85+,t+5} - P_{s,85+,t} - P_{s,80,,t} + D_{s,80,t,t+5} + D_{s,85+,t,t+5} \tag{15}$$

The balancing equation (12), when expressed in vector formulation, takes these dimensions into account:

$$\mathbf{N}_{t,t+5} = \mathbf{P}_{t+5} - \mathbf{P}_{t} - \mathbf{B}_{t,t+5} + \mathbf{D}_{t,t+5} \tag{16}$$

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

In these calculations, some additional data sources from Eurostat/JRC are used: Total births 2010-2019, further divided into births by gender using gender ratio's at birth, and 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. Additionally, data on educational transitions (table 5) are provided WIC, see the next section.

Adding level of education to the balancing equation

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).

Table 5 Estimated educational attainment transition matrix in the period 2010-15, Austria, females, age group 20-24 to 25-29.

		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**

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 (16) including educational dynamics of the resident population reads:

$$\mathbf{N}_{t,t+5} = \mathbf{P}_{t+5} - \Delta\mathbf{E} \cdot \mathbf{P}_t - \mathbf{B}_{t,t+5} + \mathbf{D}_{t,t+5} \tag{17}$$

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.

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:

$$\mathbf{P}_{t+5} = \Delta\mathbf{E} \cdot \mathbf{P}_t + \mathbf{B}_{t,t+5} - \mathbf{D}_{t,t+5} + \frac{1}{2}(\mathbf{I} + \Delta\mathbf{E})\mathbf{N}_{t,t+5} \tag{18}$$

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

$$\mathbf{N}_{t,t+5} = [\tfrac{1}{2}(\mathbf{I} + \Delta\mathbf{E})]^{-1}[\mathbf{P}_{t+5} - \Delta\mathbf{E} \cdot \mathbf{P}_t - \mathbf{B}_{t,t+5} + \mathbf{D}_{t,t+5}] \tag{19}$$

In the current model we do not assume regional differences between transition probabilities. In that case equation (18) reduces to

$$\mathbf{P}_{t+5} = \Delta\mathbf{E} \cdot \mathbf{P}_t + \mathbf{B}_{t,t+5} - \mathbf{D}_{t,t+5} + \Delta\mathbf{E} \cdot \mathbf{N}_{t,t+5} \tag{20}$$

And equation (19) reduces to

$$\mathbf{N}_{t,t+5} = \Delta\mathbf{E}^{-1}[\mathbf{P}_{t+5} - \Delta\mathbf{E} \cdot \mathbf{P}_t - \mathbf{B}_{t,t+5} + \mathbf{D}_{t,t+5}] \tag{21}$$

Estimation of net migration by sex, age, NUTS3 and educational level

With $\mathbf{P}_t$, and $\mathbf{P}_{t+5}$ estimated for t = 2010, 2015 and 2020 (and more??), $\Delta\mathbf{E}$ 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 (21). Figure 6 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).

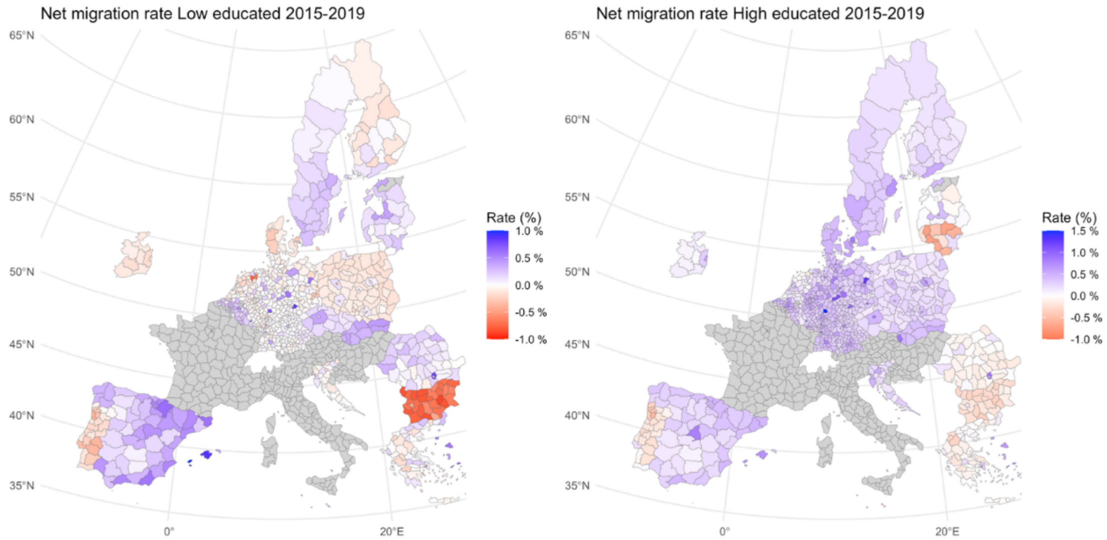

Figure 6 Estimated percentage Low and High educated in net migration 2015-2019

Summary and discussion

This paper shows how net migration by level of education at the regional level can be estimated based on partial, existing data. 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 a regional economic index which includes Gross Regional Product and unemployment as indicators. These models fitted the data well. 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 available at the NUTS3 level, our model results could be further refined. For instance, we used the results from the Netherlands 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 it would be possible to investigate this relationship further. Moreover, it could be possible to expand our estimations to include more European countries, such as Switzerland, Iceland, more Eastern European countries.

In general, more detailed and comparable data provided by international agencies such as Eurostat as well as national statistical offices could strengthen the knowledge base on regional migration by educational level. For instance, different educational groupings between agencies (see Appendix 1) makes it challenging to estimate cross-European models, and lack of data on the RxExSxA level makes it necessary to make stronger assumptions than otherwise needed. Ideally, if sufficiently detailed and comparable data were to be available, large parts of our model would be superfluous. Until then, it may be a useful tool for providing European regional policy makers with estimates of migration trends that may be crucial for their regional development.

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….

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

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

Van Wissen, L., & Rima, A. (1988). Modelling urban housing market dynamics: evolutionary patterns of households and housing in Amsterdam. North-Holland Publishing Company.

Willekens, F. J. (1977). The recovery of detailed migration patterns from aggregate data: An entropy maximizing approach. Research Memoranda 77-58, International Institute for Applied Systems Analysis, Laxenburg.

Willekens, F. J. (1980). Entropy, multiproportional adjustment and the analysis of contingency tables. Systemi Urbani, 2, 171-201.

Willekens F. J. (1982) Multidimensional population analysis with incomplete data. In Multidimensional mathematical demography, Land KC and Rogers A, Eds, pp. 43-111. Academic Press, New York.

Willekens, F. J. (1983). Log-linear modelling of spatial interaction. Papers in Regional Science, 52(1), 187-205.

Appendix 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 ISCED 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 some European statistical offices. 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 (e.g. Eurostat) and sometimes as higher educated (e.g. WIC). In any case this is not a large category but could lead to some distortions.

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

<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="4">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 rowspan="3">5B
5A First stage tertiary
5A</td><td>5 – Short-cycle tertiary</td><td rowspan="4">Tertiary</td><td>5 Short-cycle tertiary education</td></tr>
<tr><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>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>