Conference paper for the European Meeting of the International Microsimulation Association, Maastricht, 23-24 October 2014.

A microsimulation model for educational forecasting

Niels Erik Kaaber Rasmussen and Peter Stephensen, DREAM

A dynamic microsimulation model for forecasting educational patterns is presented. At the level of individuals the model simulates lifetime educational behavior, resulting in a long term forecast of the general educational level in Denmark. The model is a light-weight, dynamic, multithreaded and closed microsimulation model using discrete time.

Data on the full Danish population is used as the initial population. Each individual is characterized by age, gender, origin, educational attainment and current educational status. Future demographic events such as births, deaths, immigration and emigration are projected in a separate group-based model and given as input. In the model individuals lives their life's independently to decrease time-complexity and to utilize the potential of the multithreaded environment.

Transition probabilities are calculated from historical educational behavior using Danish register data. The historical observations are linked to a range of background variables (such as gender, age, origin, current participation in education, study length and educational attainment). Prior to running the model, transition probabilities are computed using conditional inference trees. This data-mining approach groups together observations with similar characteristics and responses based on statistical tests.

This paper describes the features of the model, briefly presents some results and points to the potential of the model in terms of policy analysis and already planned extensions to the model.

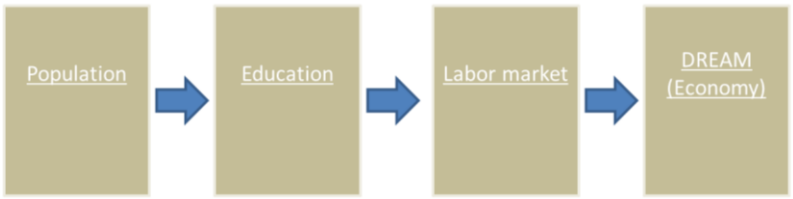
Keywords: microsimulation model, education, forecasting, education projection

Introduction

In the following we will present DREAMs education projection model a microsimulation model used for forecasting the long term general level of education in Denmark.

The model is developed as an integrated part of the DREAM system (Andersen et al. 2008). DREAM or the "Danish Rational Agent Model" is a dynamic computable general equilibrium (CGE) model with overlapping generations of households that plan their behavior in a manner consistent with rational expectations. The total DREAM system consists of four separate models: a population projection, an education projection, a socio-economic projection, and the DREAM economic model. In the DREAM system the education projection model takes the output from the population projection model, models educational behavior and its output serves as input for the socio-economic projection model which ends up in the economic model. However the educational projection model is a model in its own and can be run separately from the DREAM system. Actually the model is often used as a standalone model i.e. to investigate proposed policy changes within the educational sector.

Figure 1 - DREAMs education projection model is part of the DREAM system



Note: The education projection model is part of the DREAM system but is a model in itself and can be used independently.

The education projection model groups the full Danish population and future generations into 12 categories of education. These categories of education are here ranked with highest level mentioned first: Ph.d, university master degree, university bachelor degree, medium-cycle higher education, bachelor's degree, short-cycle higher education, vocational training, secondary education (business), secondary education (university-preparatory), 10th grade, primary school and unknown[1].

As result the model gives a rather detailed estimate on the number of people with certain degree of education and the number of students enrolled in educations in the years of simulation.

The model is based on transition probabilities calculated from rich Danish register data. It forecasts education levels by employing historical educational behavior. The model can therefore be used to predict future trends that can be attributed to the behavior of current students (or trends related to the future population composition). Examples of typical analyses include changes in dropout rates and in the age of students beginning a given level of education.

[1] Translated from Danish: Ph.d, universitets kandidat (hhv. delt og udelt), universitets bachelor, mellemlang videregående, professionsbachelor, kort videregående, erhvervsfaglig, erhvervsgymnasial, almengymnasial, 10. klasse, grundskole og ukendt.

Model characteristics

DREAMs education projection model is a dynamic, longitudinal, closed and lightweight microsimulation model simulating the full Danish population in discrete time.

The model is a *dynamic* microsimulation model because the educational status of every individual is updated for every year of simulation. This is in contrast to static microsimulation models where the behavior of individuals is constant (Li et al. 2013). An individual's educational history and state determines the following educational behavior. In addition the probability of different events depends on time. The probability that two individuals with identical demographic characteristics and educational history and state will attend next level of education can be different if they are not living in the same year of simulation.

However the probability of different events given the individuals characteristics does not change over time as a result of the overall progress of the simulation. In other words the individuals in the model does not interact neither at micro or macro level. Each individual makes its decision fully independently from the others. This feature is a prerequisite for unit-wise updating of the model, meaning that each life is fully simulated one at a time. Microsimulation models with unit-wise updating are called *longitudinal* models (http://www.tdymm.eu/sites/default/files/Dekkers_Belloni_unkn.pdf, p. 9). In other microsimulation models life is simulated one time period at the time for all units while the model continuously keeps track of the state of each unit. In this case is does not have any real impact on the result of the simulation but the unit-wise updating simplifies the implementation allows for smarter use of threading and underlines the independence of individuals in the model. The model operates in discrete time in one year intervals.

Making use of multithreading and having a lightweight setup enables the model to simulate the educational behavior of the full Danish population from 2014 to 2130 (18.8 million people in total) in approximately 2.5 minutes on a laptop PC[2].

DREAMs population projection which is also the official Danish population projection serves as input to DREAMs education projection model however other population data can be used as well. DREAMs population projection is a national demographic projection model which forecasts the Danish population across gender, age and origin. The projection is based on estimations of the age-, gender- and origin-dependent frequencies of birth, death and migration.

For each living person and for each future living person a person is created in the education projection model with similar demographic characteristics. 5.5 million people are alive in the model the first year of simulation corresponding to the current number of inhabitants in Denmark. Information on each living and future person's time of death or time of emigration from Denmark is extracted from the population projection data and thus not simulated within the model. Same goes for future immigration and births.

The model is a *closed* microsimulation model as the population including future growth is specified beforehand outside the model. In open microsimulation models key individuals can be added during simulation in order to simulate or highlight specific behavior.

Each person within the model has a range of demographic properties: gender, origin and year of birth. In addition time of death or emigration (ie. the year one leave the simulation) is

[2] Intel Core i5-240M CPU 2,5 GHz

treated as a property to a person within the model (while estimated outside the model). Simulation much similar to DREAMs education projection model is integrated in the larger SMILE model (Hansen et al. 2013) in which geographical location also has an impact on educational behavior.

The properties gender, origin and birth all have an impact on the behavior of a person during simulation while time of death/emigration only determines how for long a person stays alive in the simulation.

It is assumed that educational behavior does not have any impact on either fertility or death rates or vice versa. This follows from the fact that the full population is given as input to the model.

In addition to the demographic properties each person in the model has a history of education and a current state of education, which are updated whenever a person reaches an event in the model: undertakes a new level of education, dropout, studies one more period or fulfillment. The status is described by three variables:

- Highest achieved level of education meaning the highest ranked level of education the person has so far fulfilled regardless current participation in education and not necessary the most recent finished education.

- Current participation in education, meaning the education one undertakes a given year. If a person is not undertaking any education (for example if he/she is at the labor market) his/her status is set to “not undertaking education”.

- Duration/current study length is the number of years a person has been enrolled in the current education. This doesn’t necessarily correspond to the number of years successfully completed.

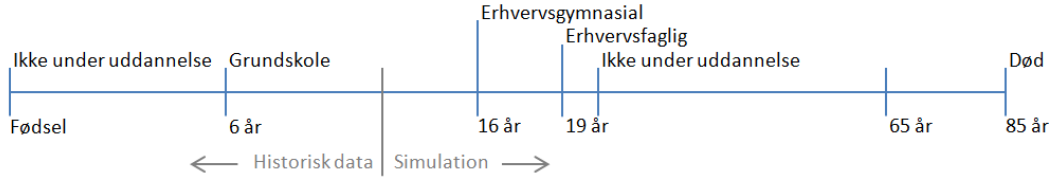
The yearly status of all individuals’ educational status all together constitutes the result of DREAMs education projection model.

The behavior of individuals in the model is simulated using transition probabilities based on the latest historical data using register data from Statistics Denmark. The exact transition probability used depends on the person’s demographic characteristics and current status of education.

Simulation of a single person

The above illustration shows how a specific person lives her life in the model. As an example a woman of Danish origin born in 2000 and estimated dead in 2085 is chosen. The demographic features are given by the population projection model.

Figure 2 Example: Simulation of a single person



The first years of the life of the woman are historical and thus not simulated in the model but read from register data. Far most likely she would have spent the first years not undertaking any education hereafter she would have attended primary school.

By start of simulation she is 14 years old and by 95 % chance she continues in primary school as a 15 year old. When 16 she has finished primary school and gets enrolled at a secondary school. Her status is now updated so that “highest level of education” is set to primary school and “current participation in education” is set to secondary school with the study length of 1 year. The most likely behavior for a 16 year old Danish woman having just finished primary school is to enroll at 10th grade (45%) however the behavior in the example is not unlikely (6%).

The women continues to live one year at a time in the model each with the chance of an event, either to enroll at an education, to fulfill an ongoing study, to drop out or to continue at current occupation. After the age of 65 (current pension age) everyone is by definition not undertaking any form of education.

The model is non-deterministic i.e. If we repeated the simulation with the exact same input of one person we would most likely not get the same result. At an aggregated level however because we have roughly 6 million persons in the model at a time the stochastic element is of minor importance to the overall educational level of the population. The smaller a group of people we are examining the more uncertainty however this problem is easily fixed by repeating the simulation a number of times and focusing at the average result. Because of the lightweight setup we have been able to run a simulation in 2.500 iterations within reasonable time.

Simulation of the full population

The patterns of education for the full population can be illustrated by looking at the observed transitions between different types of education.

The table below indicates the proportion of all students at a given education – regardless their demographic characteristics and educational history – who changes their status of current participation in education. It can be seen that 49 % of all pupils finishing primary school continues in 10th grade, while 24 % enrolls at secondary school and 9 % doesn’t undertake any education the following year.

Table 1 – Transition in current participation in education

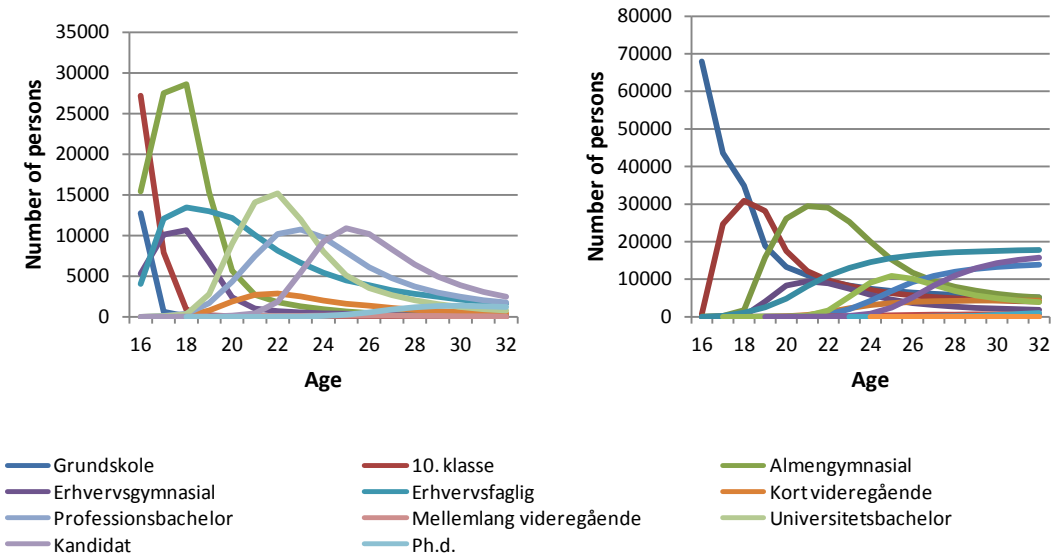
%	Primary school	10th grade	Secondary school (prepatory)	Secondary school (business)	Vocational training	Short-cycle education	Bachelor's degree	Medium-circle education	University bachelor	University candiadte	Ph.d.	Not undertaking education
Primary school		49	25	9	7	0	0	0	0	0	0	10
10th grade	0		41	15	24	0	0	0	0	0	0	19
Secondary school (prepatory)	0	0		2	5	2	6	0	9	0	0	77
Secondary school (business)	0	0	6		10	6	8	0	13	0	0	57
Vocational training	0	0	3	2		2	2	0	1	0	0	89
Short-cycle education	0	0	0	0	6		23	0	6	1	0	63
Bachelor's degree	0	0	0	0	3	2		0	3	8	0	83
Medium-circle education	0	0	0	0	0	1	9		8	8	0	74
University bachelor	0	0	0	0	2	3	6	0		62	0	27
University candidate	0	0	0	0	0	0	0	0	3		4	92
Ph.d.	0	0	0	0	0	0	0	0	0	7		92
Not undertaking education	1	1	6	1	35	7	19	1	20	6	3	

Note: The table lists the observed probabilities of transitions from one state (vertical) to another (horizontal). The percentage describes the proportion of persons changing status from being enrolled in the education named leftmost in the row to being enrolled in the education given by the column.
No distinction is made between students fulfilling and dropping out or between different study lengths. The numbers are based on simulated behavior of the full population. The sum of each row is 100.

Source: DREAMs education projection model 2013.

The figure below illustrates how on cohort works its way through different levels of education.

Figure 3 Current participation in education and highest level of education, one cohort.



Note: People with unknown educational status are not shown in the figure to the right. Age 16 corresponds to the year 2012, age 17 to 2013 and so on.

Source: DREAMs education projection model 2013.

The figure at left shows how 10th grade is almost solely for 16-17 year olds, while secondary school tops for the 17-18 year olds. Vocational training is attended by a much more diverse age range. Generally and not surprisingly it can be seen the highest levels of education starts at an older age primarily because one first has to graduate other levels of educations to qualify for the longer-term education. Few people are still undertaking education at the age of 30.

In the figure at right it can be seen that the full cohort starts out having completed primary school. Within the age of 17 and 18 roughly 30.000 has finished 10th grade and thereby upgraded their status of highest education. At the age of 20-22 many has completed a secondary school and in the following years many of those completes higher levels of education.

When the cohort reaches 32 years of age most are fully educated and will not attend further education.

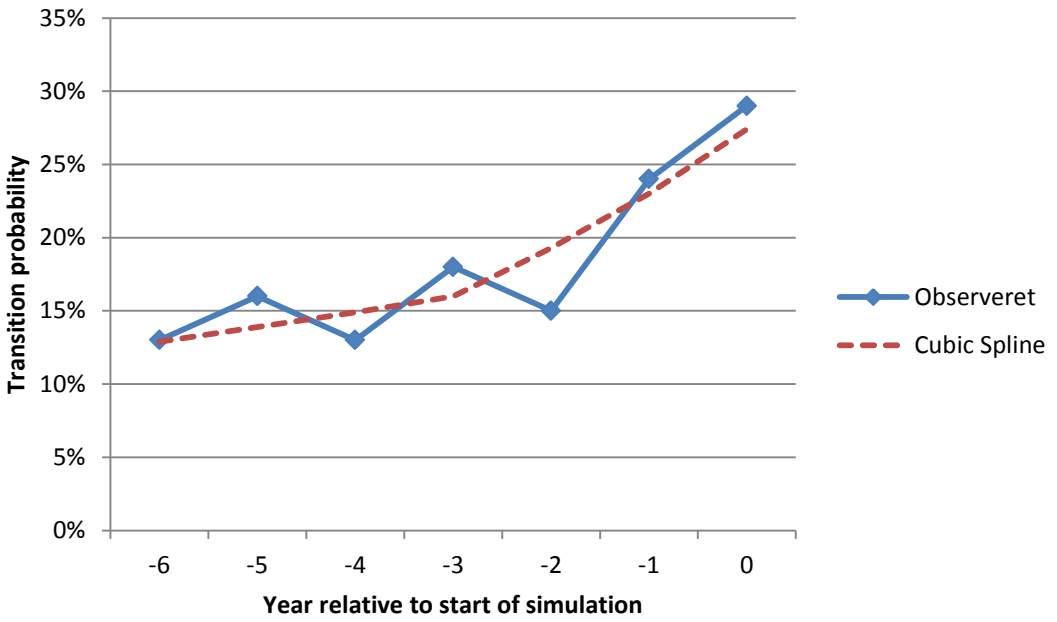
Transition probabilities

Determining transition probabilities is central to most dynamic microsimulation models. A transition probability describes the probability that a subject in a microsimulation model experience a transition from one state to another.

In DREAMs educational projection model transition probabilities constitute three key events: undertaking a new level of education, fulfilling an education and dropping out. The probabilities depend of the characteristics of the subject such as the age of the person or the number of children in the family.

Raw transition probabilities in the meaning observed historical frequencies are taken from Danish register data and covers a ten year period. Prior to being used in simulation transition probabilities are smoothed to minimize noise in data, extrapolated and processed by conditional interference trees[3]. The smoothing is done using the cubic spline method over the latest 10 years of data. The method ensures that the most recent data is particularly strongly weighted.

Figure 4. Smoothing of transition probabilities



Note: Transition probabilities are smoothed using cubic spline with cross validation using the most recent 10 years of data.

The actual transition probability used is the end point of the cubic spline function. All transition probabilities determining a person's choice of education is smoothed while transition probabilities determining whether a person finished, drops out or continues a current study is based on a simple average of the latest 3 years.

The model allows for extrapolation of transition probabilities. This is done by extending the cubic spline function a couple of years into the simulation using the slope in the end point of the cubic spline function. The idea is that certain trends in data can be expected to continue some years into the future however for our baseline projection we're not extrapolation transitions probabilities due to strong current trends in educational patterns that we have no strong reason to expect to continue.

[3] These are explained in greater details at the following pages.

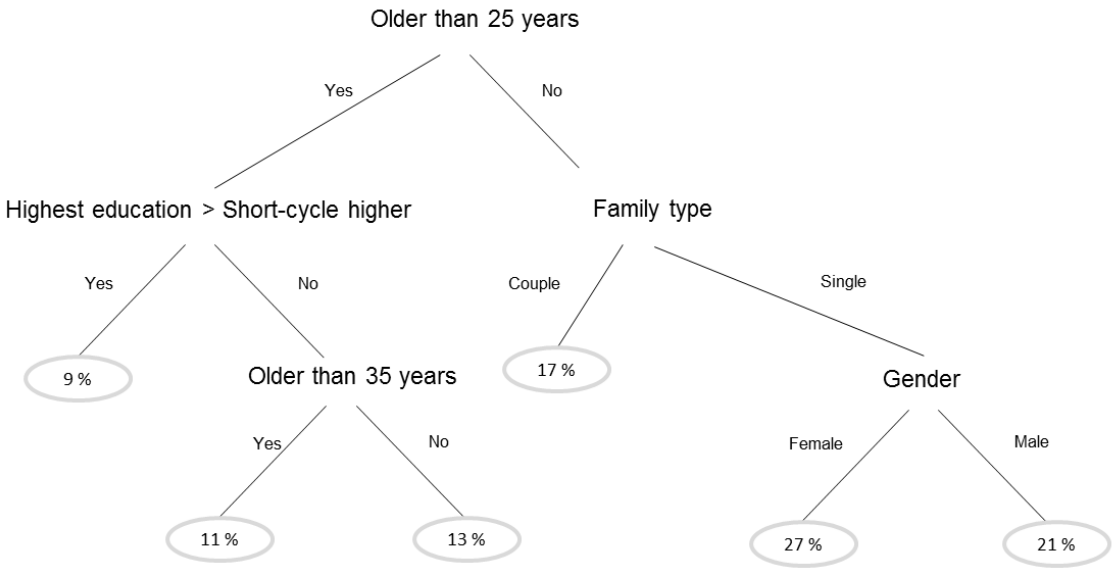
Conditional inference trees[4]

Transition probabilities are calculated for each combination of gender, origin, age, current participation in education, highest level of education and duration of current education. A large number of characteristics will therefore significantly decrease the number of historical observations available to calculate transition probabilities for a given combination of characteristics. Data will be sparse and accordingly the probabilities will be calculated using few (or even none) observations. This of course introduces a significant uncertainty to the transition probabilities. To solve this problem we're using conditional inference trees (CTREEs) a big data approach that is already successfully used in the SMILE model (Rasmussen et al 2013).

To decrease the number of possible combinations of explanatory variables one could choose to simply ignore selected variables. CTREEs provide a better solution: to group together observations which shares not all but many of the same explanatory variables and at the same time represents similar behavior. Ideally observations should be grouped in a way so that most characteristics are shared, the observed behavior is similar within the group of observation and the difference of observed behavior between groups is maximized (Hothorn, Hornik & Zeileis 2006).

The simple tree illustrated below is based on dummy data but illustrates well how different explanatory variables are used to determine which transition probability a subject can be paired to. In the example terminal nodes contains a transition probability representing a decision with a binary outcome (event vs. non-event).

Figure 5. Structure of a CTREE, dummy data.



Explanation: The structure of a CTREE. It can be seen by looking at the tree that the transition probability for a couple with an average age of 24 is 17 % while the transition probability for a single 20-year old male is 21 % and a family with an average age of 42 with a compulsory education has the transition probability 11 %.

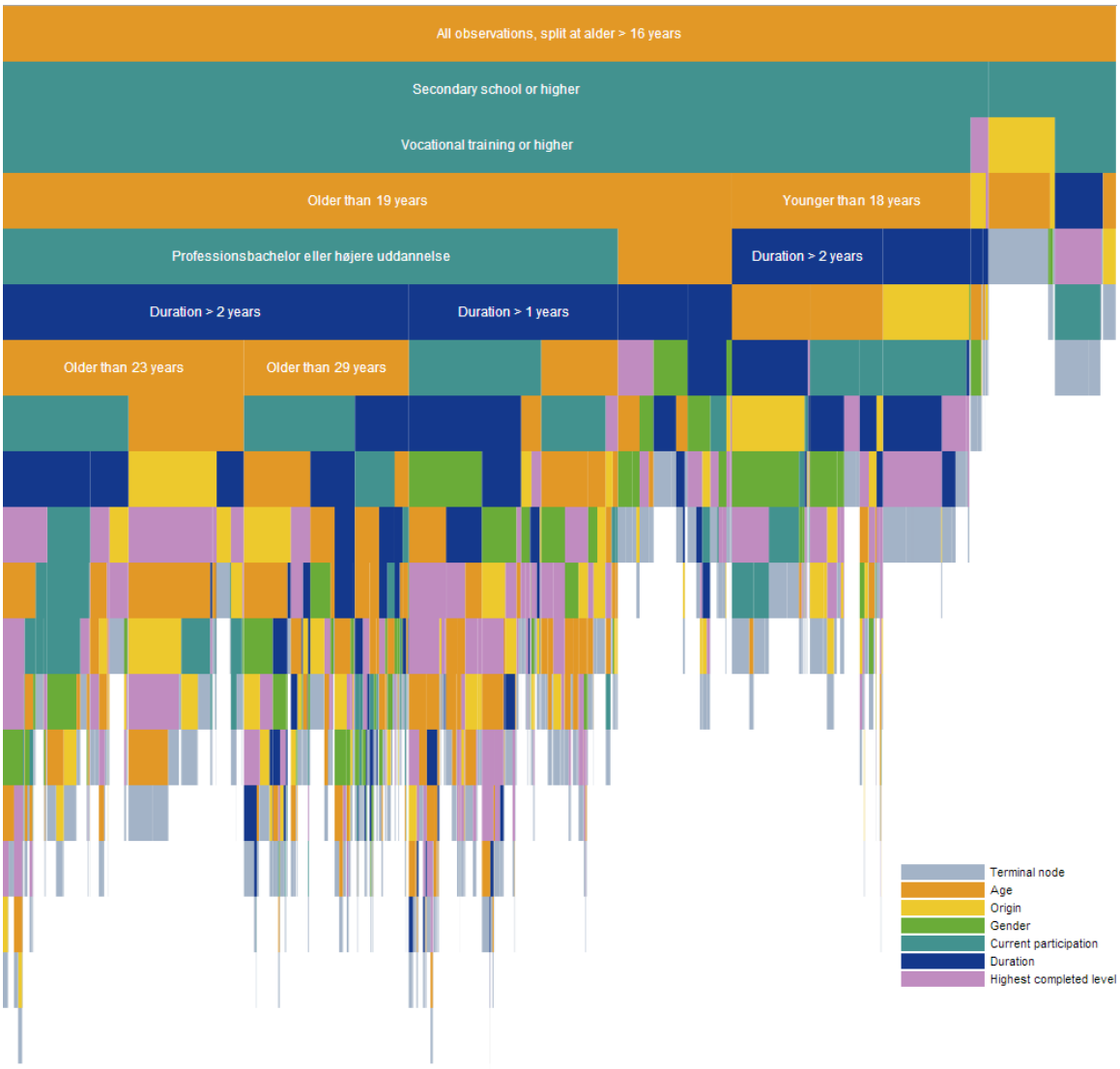
[4] This section is party based on the conference paper: "Conditional inference trees in dynamic microsimulation - modeling transition probabilities in the SMILE model", (Rasmussen et al. 2013).

Figure 6 illustrates the CTREE used in the model to represent transition probabilities for fulfillment of an educational level. More than 1.5 million observations are drawn from the latest three years of register data. The average of the three years is calculated reducing the number of observations to roughly 500.000, those are used as input for the CTREE.

The figure is read top down. Each row corresponds to a level in the CTREE. The area of each rectangle corresponds to the number of observations. Each explanatory variable is represented by a distinct color. Terminal nodes are colored light gray and contain the transition probability for the group with the given characteristics.

The top-most row contains all observations (root node) and is split into two (child nodes): persons older than 16 downwards to the left and persons at age 16 or younger downwards to the right. Most observations concerns persons who are 16 years. These are again split into two depending on the level of education they are currently undertaking and so the CTREE can be followed downwards until a terminal (leaf) node is reached.

Figure 6. Visualizing a CTREE for education related transition probabilities



Explanation: The visualization is read top-down. The top-most orange row represents all observations. In the CTREE the observations are split into two, a large part of persons older than 16 and a smaller part with the rest. The two resulting subgroups of observations are each split into two. For each row in the visualization the CTREE has performed splits.

Source: DREAMs education projection model 2013 and SMILE.

To reach a terminal node 4-17 inner nodes (splits) must be passed. The height of the tree is said to be 17 and the tree is said to be unbalanced. The tree is a full binary tree as every node in the tree has exactly 2 or 0 children, this follows from the fact that the CTREE-algorithm always splits in two.

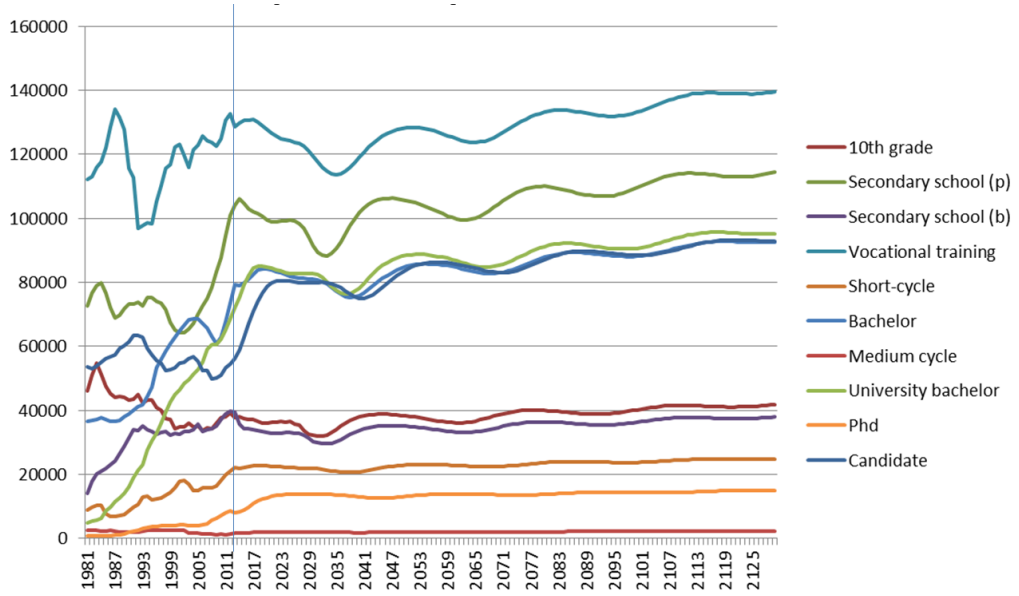
If a color covers a large area it means the explanatory variable is a strong indicator for the transition probability, on the other hand variables that are often ignored by the CTREE algorithm will cover a relative small area of the chart. In the chart above it can be seen that the cyan color covers a rather large area. Cyan represents current participation in education, so not surprisingly the chance of fulfilling an education depends strongly on the type of education one is undertaking. Gender and origin are less important indicators (one should be careful to conclude though as gender has only two responses and origin five while educational level has 12 and age has 50). If a variable is rarely represented in the chart one could consider omitting it.

Results

The education projection model gives detailed results on educational behavior and level of education. Results can be extracted for each cohort, by gender, origin, age and year. Below the total number of currently enrolled students is shown for each category of education.

Demography plays a major role in regards to the number os pupils in primary school. In the period of simulation it can be seen that besides the demographic effect there is an increase in the number of students enrolled in high levels of education. It the long term this effect stabilizes.

Figure 7. Current participation in education, historical and projected data.



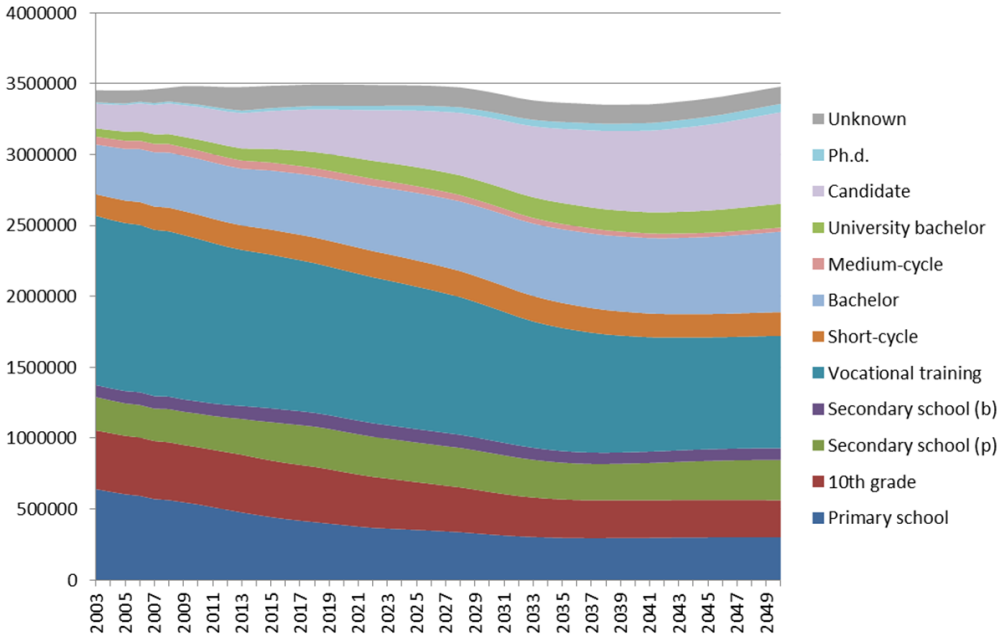
Note: The vertical gray line separates historical years from projected. Primary school covers only pupils in 8th and 9th grade.

Source: DREAMs education projection model 2013.

In regards to the number of university candidates it can be explained by the fact that there has been a recent increase in the number of students at secondary school. Many of those students will undertake and finish a university bachelor's degree and thereafter undertake an education as university candidates.

Since extrapolation of trends in our most recent projections is very limited, the general level of education stabilizes rather early in the simulation. The figure below shows the highest level of education for every 17-64 year old (the potential workforce if pension age is fixed). It can be seen that the population get better and better educated over time.

Figure 8. 17-64 year olds by highest level of education, number of persons.



Source: DREAMs education projection model 2013.

Most prevalent is the increase in the share of people with a long-term education. In contrast the share of people without a higher education decreases especially the share of vocational trained.

Conclusion

In this paper a model for education projection is presented, its characteristics are described and the main features are explained. DREAMs education projection model is a dynamic microsimulation model simulating the educational behavior of the full Danish population. It is a longitudinal, closed and lightweight microsimulation model using discrete time in a multithreaded environment.

Input data is a combination of data from the most recent version of the official Danish population projection and detailed Danish register data on recent year's educational behavior. Transition probabilities are smoothed using the cubic spline method to avoid noise in data. The model contains an option that makes it possible to extend trends in data into the simulation period. Transition probabilities are lastly processed using the technique of conditional inference trees. The CTREE-algorithm groups together explanatory variables for observations with similar outcomes based on statistical tests in order to avoid noise coming from transition probabilities being based on too few observations.

The results of our latest education projection show an overall increase in the number of enrolled students especially at higher educational levels and an increase in the general level of education in the Danish population; more university graduates and fewer with vocational training.

References

Andersen, Michael & Schou, Poul (2008): “DREAM Preliminary DREAM documentation”. http://www.dreammodel.dk/dwn_english.html

Hansen, Jonas Zangenberg; Stephensen, Peter & Kristensen, Joachim Borg (2013). “Modeling Household Formation and Housing Demand in Denmark”. Report (preliminary version). http://www.dreammodel.dk/SMILE/

Hothorn, T., K. Hornik & A. Zeileis (2006): Unbiased Recursive Partitioning: A Conditional Inference Framework, Journal of Computational and Graphical Statistics, Vol. 15, No. 3, page 651–74.

Hothorn, T., Hornik, K, Strobl, Carolin & Zeileis, A. ‘Party’ package for R: A Laboratory for Recursive Partytioning, October 2013. http://cran.r-project.org/web/packages/party/party.pdf

Jinjing Li & Cathal O’Donoghue (2013): “A survey of dynamic microsimulation models: uses, model structure and Methodology”. Internation Journal of Microsimulation 6(2) 3-55. http://www.microsimulation.org/IJM/V6_2/2_IJM_6_2_2013_Li_Odonoghue.pdf

Rasmussen, Niels Erik Kaaber; Hansen, Marianne Frank Hansen & Stephensen, Peter (2013): “Conditional inference trees in dynamic microsimulation - modeling transition probabilities in the SMILE model”. Conference paper for the 4th General Conference of the International Microsimulation Association. http://www.dreammodel.dk/SMILE/N2013_04.pdf

Rasmussen, Niels Erik Kaaber (2014): “DREAMs Uddannelsesfremskrivning 2013”. http://www.dreammodel.dk/pdf/Uddannelsesfremskrivning2013.pdf