4th General Conference of the International Microsimulation Association
Canberra, Wednesday 11th to Friday 13th December 2013

Conditional inference trees in dynamic microsimulation - modelling transition probabilities in the SMILE model

Niels Erik Kaaber Rasmussen, Marianne Frank Hansen and Peter Philip Stephensen
DREAM

Abstract

Determining transition probabilities is a vital part of dynamic microsimulation models. Modelling individual behaviour by a large number of covariates reduces the number of observations with identical characteristics. This challenges determination of the response structure. Data mining using conditional inference trees (CTREEs) is found to be a useful tool to quantify a discrete response variable conditional on multiple individual characteristics and is generally believed to provide better covariate interactions than traditional parametric discrete choice models, i.e. logit and probit models.

Deriving transition probabilities from conditional inference trees is a core method used in the SMILE microsimulation model forecasting household demand for dwellings. The properties of CTREEs are investigated through an empirical application aiming to describe the household decision of moving from a number of covariates representing various demographic and dwelling characteristics.

Using recursive binary partitioning, decision trees group individuals' responses according to a selected number of conditioning covariates. Recursively splitting the population by characteristics results in smaller groups consisting of individuals with identical behaviour. Classification is induced by recognized statistical procedures evaluating heterogeneity and the number of observations within the group exposed to a potential split. If a split is statistically validated, binary partitioning results in two new tree nodes, each of which potentially can split further after the next evaluation. The recursion stops when indicated by the statistical test procedures. Nodes caused by the final split are called terminal nodes. The final tree is characterized by a minimum of variation between observations within a terminal node and maximum variation across terminal nodes. For each terminal node a transitional probability is calculated and used to describe the response of individuals with the same covariate structure as characterizing the given terminal node. That is, if a terminal node consists of single males aged 50 and above living in rental housing, individuals with such characteristics are assumed to behave identically with respect to moving when transitioning from one state to another.

Keywords: transition probabilities, conditional inference trees, data mining, microsimulation, classification tree.

Conditional inference trees in dynamic microsimulation - modelling transition probabilities in the SMILE model

1. Introduction

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. As in most other microsimulation models the transitions in SMILE constitute events such as death, fulfillment of an education or moving. The probabilities depend of the characteristics of the subject such as the age of the person or the number of children in the family. In the SMILE model transition probabilities are used to determine demographic, socioeconomic and housing-related events (Hansen, J. Z. & Stephensen, P., 2013).

One straight forward approach when dealing with transition probabilities is to apply the historically observed frequencies of events to the subjects in the microsimulation model. To this one could add a method of smoothing the historical transition probabilities and extending trends into the forecast period. In SMILE the probabilities of demographic events such as birth and death are calculated based on observed frequencies. Other well-known approaches to transition probabilities in microsimulation models are logit and probit regression models.

For each subject in the microsimulation model the probability for a given event depends on the characteristic of the subject. A male usually dies at a younger age than a female, a couple living together will have higher fertility than two singles, the chance of a 15-year old moving away from her parents to find an expensive house on the country side is small and so on. The larger the number of relevant characteristics used to describe the subjects in the microsimulation model, the more accurate transition probabilities can be predicted and ultimately the better the model will be. This holds true however only as long as the number of historical observations describing a subject with characteristics corresponding to the characteristics of the subject in the model is large enough. Transition probabilities will normally be calculated for each combination of characteristics (in the following the terms explanatory variables and input variables will be used to interchangeably). For each explanatory variables used the number of combinations will be multiplied. 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.

In SMILE moving patterns depends on a number of characteristics of the household. The characteristics are age (18-99 year), gender, three family types, six possible levels of education, five possible origins, labor market status, eleven regions and an indicator of whether there's children in the household or not. The combination of these explanatory variables sums up to roughly 1.6 million combinations. For families with two adults some of these variables will apply to both adults and further increase the number of possible combinations. If both adults educational background, origin and labor market status is taken into account the number of combinations will

reach 97.4 million. With a population of 5.5 million in Denmark it is easy to see that data is too sparse to reasonable cover all possible combinations of explanatory variables.

To decrease the number of possible combinations of explanatory variables one could choose to simply ignore selected variables or better to group together observations which shares not all but many of the same explanatory variables and at the same time represents a 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. Conditional inference trees (CTREEs) provides an algorithm which uses statistical tests to split a large group of observations into such groups. CTREE is one of the newer algorithms for this purpose (Hothorn, Hornik & Zeileis 2006).

The CTREE is based on recursive binary partitioning meaning that the full group of observations will be split into two recursively until a stop criterion is reached. The result is a binary decision three representing information on which explanatory variables that shall be grouped together and the transition probability for each group of observations.

The CTREE-algorithm can be described in three steps (for details see Hothorn, Hornik, Strobl & Zeileis 2013):

1) Test the global null hypothesis of independence between any of the explanatory variables and the behavior/the response.
    a. Stop if this hypothesis cannot be rejected ($p>0.05$).
    b. Otherwise select the input variable with strongest association to the response. This association is measured by a p-value corresponding to a test for the partial null hypothesis of a single input variable and the response. For SMILE we choose to use the default implementation with c_{quad}-type test statistics and Bonferroni-adjusted p-values to avoid overfitting.
2) Implement a binary split in the selected input variable. To find the optimal binary split in the selected input variable the algorithm uses a permutation test. A stopping criterion makes sure that groups containing less than 20 observations will not be split and that the resulting groups after a split will contain at least 7 observations[1].
3) Recursively repeat steps 1) and 2) until a stop criterion is reached.

This approach of data mining the transition probabilities makes it possible to take into account a large number of relevant characteristics of a subject in the microsimulation model and reduces the need for detailed domain specific knowledge otherwise required to decide which explanatory variables to group together. Instead the statistical test in the algorithm does the work.

Unlike similar recursive fitting algorithms CTREEs are not biased towards selecting input variables with many missing values and many possible splits.

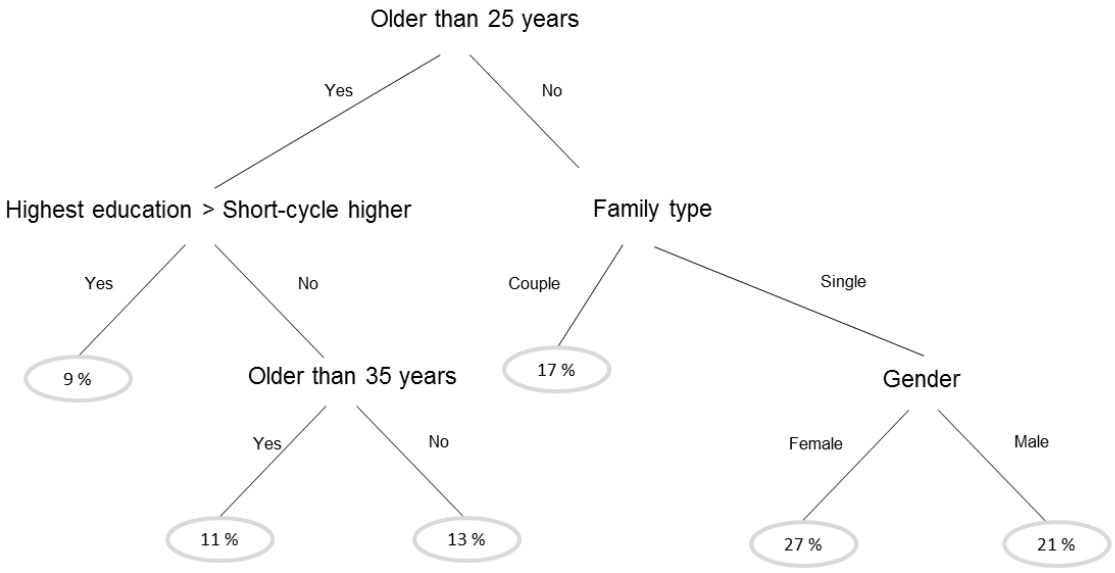
[1] These stopping criterias as well as the p-value can be adjusted.

The CTREE-framework is applicable to a wide range of regression problems where both response and covariates can be measured at arbitrary scales, including nominal, ordinal, discrete and continuous as well as censored and multivariate variables.

2. Using CTREEs in SMILE

Figure 1 below shows an example of the structure of a very simple CTREE. The CTREE is based on dummy data but illustrates well how different explanatory variables are used to determine which transition probability a subject in the model will be paired to.

Figure 1. Structure of a CTREE, dummy data.



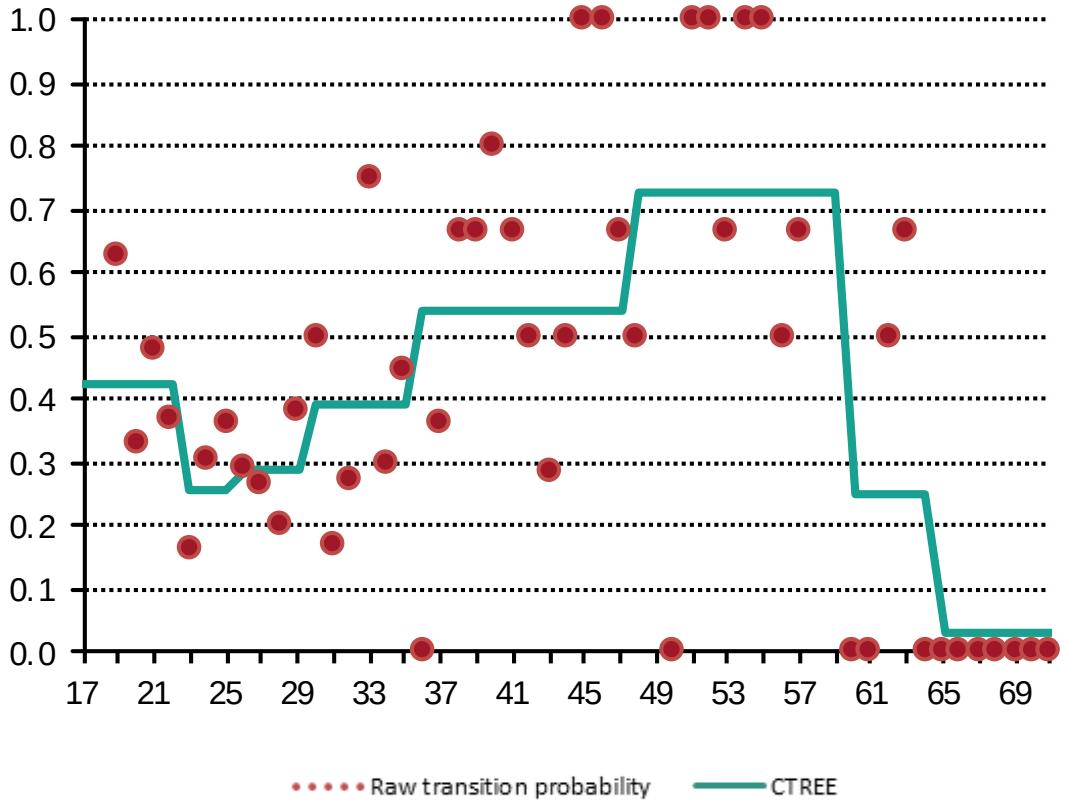
Explanation: The structure of a CTREE. It can be seen by looking at the tree that the transition probability for a couple with an average age of 24 is 17 % while the transition probability for a single 20-year old male is 21 % and a family with an average age of 42 with a compulsory education has the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

To look up a transition probability for a subject in the microsimulation model the CTREE is used as a decision-tree. In this example the characteristics of a household determines the probability for an event. First split concerns the average age of the adults in the family while the second split depends on the answer to the first split. If the average age is above 25 the next split is dependent on educational level while second split depends on family type if the average age is below 26. To find a transition probability the tree structure shall be followed until a terminal node is reached. The tree is unbalanced (height varies according to the characteristics of the subject) and can perform splits on both continuous and categorical data. By specifying explanatory variables as continuous the numerical order of the input variable will be respected. In the example above age is a continuous variable. The algorithm can decide to group families older than 25 together but cannot make a split that places 16- and 18-year olds together while at the same time 15- and 17-year olds are placed in a separate group.

In the example terminal nodes contains a transition probability representing a decision with a binary outcome (event vs. non-event). In SMILE CTREEs are used on decisions with binary

outcome as well as with multiple possible outcomes (i.e. To decide which region a person is moving to or which level of education a person is undertaking).

Figure 2. Age dependent employment frequencies for immigrants from more developed countries, men with short-cycle higher education.



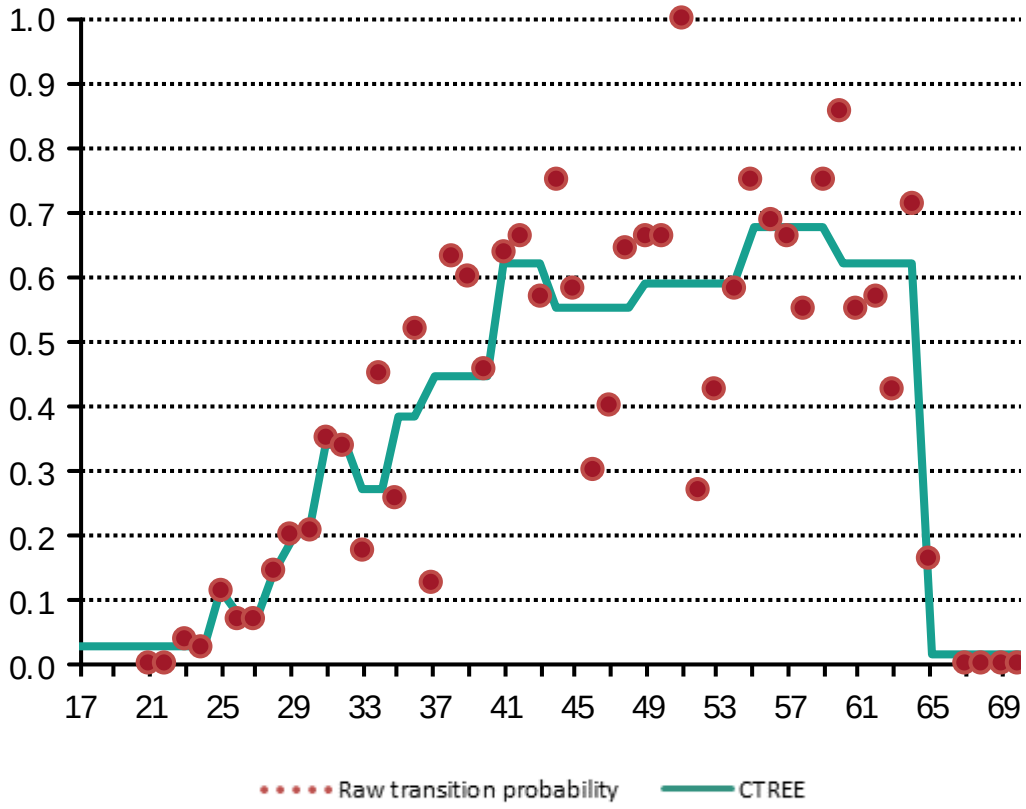
Source: Register data, Statistics Denmark, own calculations.

Figure 2 above shows the observed employment frequencies (red dots) for men with short-cycle higher education that are immigrants from more developed countries. Since this is a rather specific group there are few observations in historical data and consequently the uncertainty is high. For 50-year olds the probability is 0 while it's 1 for 51-year olds but obviously this does not describe a general pattern rather it is a result of sparse data and we would not like to reproduce the pattern in the forecasting model.

The cyan line shows the transition probabilities given by the CTREE. The transition probabilities calculated by the CTREE-algorithm seems reasonable given the variation of the observed frequencies.

Figure 3 below shows a related example of age dependent employment frequencies for female immigrants from more developed countries with a medium-cycle higher education.

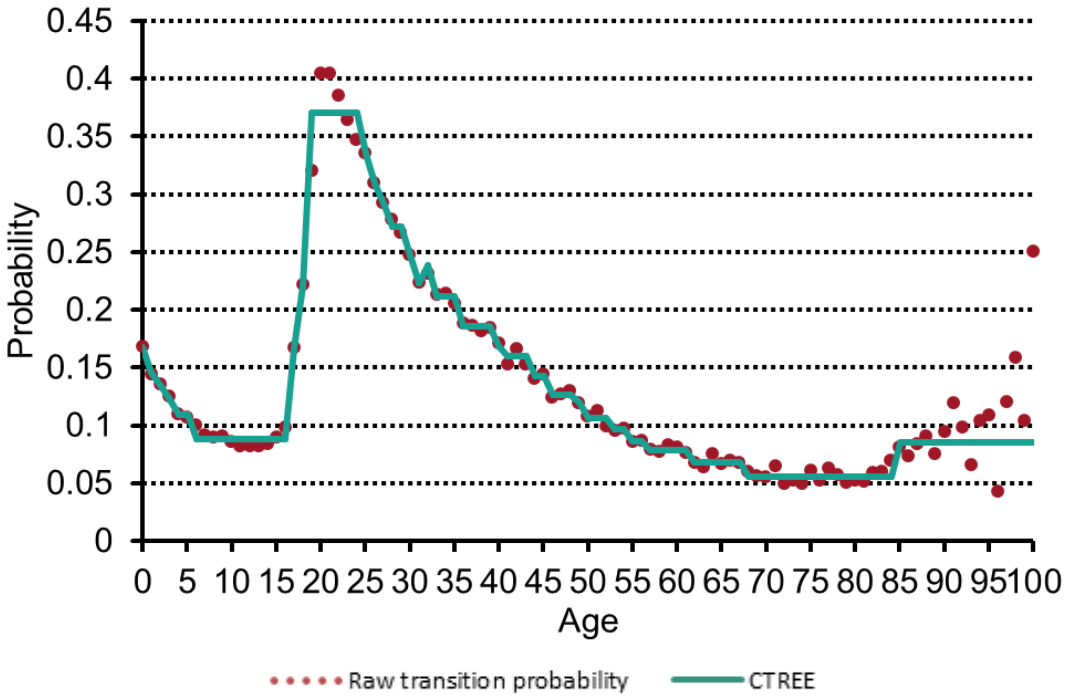
Figure 3. Age dependent employment frequencies for immigrants from more developed countries, women with medium-cycle higher education.



Source: Register data, Statistics Denmark, own calculations.

The cyan curve representing the transition probabilities given by the CTREE follows the raw data well until 30 years. For the years 30-64 much of the disturbing variation in the raw data is gone. The variation seen for these years can be explained by the low number of observations. By the age of 64-65 we expect to see a decrease in the employment frequency as this is the current retirement age. This decrease it is desirable to keep and it is cleverly maintained by the CTREE.

Figure 1. Probability of moving, single men.



Source: Register data, Statistics Denmark, own calculations.

Figure 1 shows the probability of moving for a single man. Moving is rare in primary school years and at an old age (60-85). In Denmark it is normal to move away from ones parents in the late teenage years or early twenties. During adulthood the probability of moving decreases. At a very old age many are moving to a retirement home thus the increase in moving frequencies for very old people (85+).

The CTREE-curve follows the observed frequencies closely for most ages. For the very old people there a few observations and the leveling done by the CTREE seems fine in that perspective. The raw transition probabilities top around the age of 20-22 this is a time where many young men are moving out from their parents. In other words there's a reasonable explanation for why data tops. However the CTREE-curve cuts the top of the raw transition probabilities. The CTREE is constructed around multiple dimensions (input variables) that could help explain the decision made by the algorithm. Fx. It could be that moving patterns for young men depend very much on the region they live in or the education they are undertaking. Also the fact that there for most men will be a change of family-type (going from children living at home to start a household of his own) can affect the overall probability.

To avoid the cut it could be considered to adjust the p-value used in the statistical test in the CTREE.

In DREAMs educational projection model as well as in the forthcoming education module in SMILE educational events are simulated at micro level. Educational events are events such as beginning at a new level of education, dropout and fulfilment.

The educational projection is based on transition probabilities calculated from Danish register data, and thus it forecasts education levels by employing historical educational behavior. The historical observations are linked to background variables such as gender, age, origin, current participation in education, study length and highest level of education. In SMILE educational behavior also depends on the geographical region the person lives in and comes from.

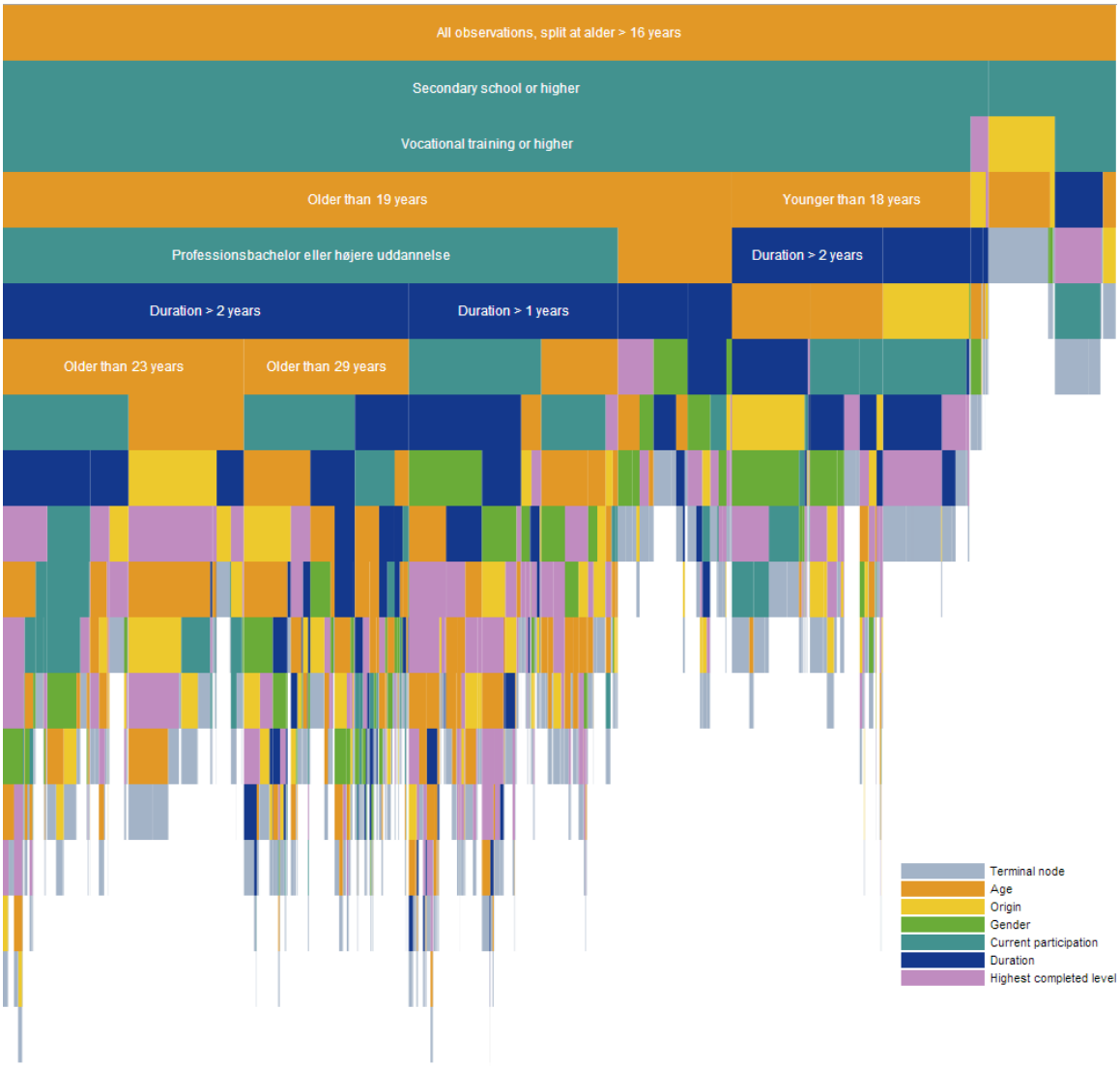
As with the housing-related events the possible combinations of these background variables are many. For certain combinations no data exists at all (fx. 16 year old girls with a compulsory education and foreign origin who's studying for a Ph.D. on her third year), for other combinations there will only be one observation or observations will simply be rare. In order not to base the transition probabilities on too few observations it is desired to group the observations across the different background variables using CTREEs.

Figure 4 illustrates the CTREE used in the model to represent transition probabilities for fulfilment 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 4. 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.

3. Usage

Usage of CTREE in SMILE can be explained in 5 simple steps: 1. Extract raw data with transition frequencies, 2. Create CTREE with transition probabilities, 3. Export CTREE to text format, 4. Import CTREE into the SMILE-model, 5. Use in CTREE SIMLE to look up transition probabilities.

Firstly transition frequencies are drawn at micro level from register data. The raw data serves as input to the CTREE-algorithm. The CTREE algorithm used in SMILE is implemented in the R-package "party: A Laboratory for Recursive Partytioning" by Hothorn, T., K. Hornik & A. Zeileis (2013).

DREAM has developed R-code to print the constructed CTREEs to plain text-files thereby exporting the results from R. As SMILE is developed in C# next step is to import the CTREE into C# and create a data structure to represent the threes in the model, DREAM has developed code to do this as well. The sample code below shows how a CTREE can be easily imported into C# and used to look up a transition probability within the model.

```
Partytree _probFinish = new PartyTree<double>(ctreeFile); //import tree

//lookup transition probability
double probabilityToFinish = _probFinish[_age, _origin, _gender,
_ongoingEducation, _duration, _highestCompletedEdu];


if (_random.NextDouble() < probabilityToFinish)
  // person just finished education
else
  //person doesn't finish education this year
```

Both the R-code used to export CTREEs and the C# library capable of importing CTREEs into a C# project and look up values are published as open source software under the MIT license. Code can be found at https://github.com/DREAMmodel/.

4. Conclusions

The computational complexity of constructing a CTREE depends on the number of background variables and their types (ordered variables are considerable less time consuming than categorical variables). Adding additional explanatory variables might introduce a bottleneck in the microsimulation model. However the problem can be solved by the use of a Monto Carlo sample of random permutations.

CTREEs makes it possible to take into account a large number of relevant characteristics of a subject in the microsimulation model and reduces the need for detailed domain specific knowledge otherwise required to decide which explanatory variables to group together. However CTREEs cannot be used blindly. The events described by a CTREE shall be similar in nature and the input variables shall be interpreted in a similar manner across observations.

In the SMILE model CTREEs are successfully used to manage transition probabilities. The CTREE-algorithm groups together explanatory variables for observations with similar outcomes based on statistical tests. The data mining approach is found to be a useful tool to quantify a discrete response variable conditional on multiple individual characteristics and is generally believed to provide better covariate interactions than traditional parametric discrete choice models, i.e. logit and probit models.

5. References

Hansen, J. Z. & Stephensen, P. (2013): *Modeling Household Formation and Housing Demand in Denmark using the Dynamic Microsimulation Model SMILE*, DREAM Conference Paper, December 2013. The paper can be downloaded from www.dreammodel.dk/SMILE

Hansen, J. Z., Stephensen, P. & Kristensen, J. B. (2013): *Household Formation and Housing Demand Forecasts*, DREAM Report, December 2013. The report can be downloaded from 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.

Stephensen, P. (2013): *The Danish microsimulation model SMILE - An overview*, DREAM Conference Paper, December 2013. The paper can be downloaded from www.dreammodel.dk/SMILE

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