aging·health·labor
CHRISTIAN DOPPLER LABORATORY

Predicting educational attainment of the Austrian population using data from the Austrian Social Security Institutions

by

Christina NEUWIRTH

Working Paper No. 1601

February 2016

Thanks to Bettina Grün for support in writing this paper and to the Christian Doppler Society

Christian Doppler Laboratory
Aging, Health and the Labor Market
cdecon.jku.at

Johannes Kepler University
Department of Economics
Altenberger Strasse 69
4040 Linz, Austria

Statutory Declaration

I hereby declare under oath that the submitted Master thesis has been written solely by me without any third-party assistance, information other than provided sources or aids have not been used and those used have been fully documented. Sources for literal, paraphrased and cited quotes have been accurately credited. The submitted document here present is identical to the electronically submitted text document.

....	....
date	signature

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The highest education completed of a person is an important variable for microeconomic research, such as for the analysis of the effect of education on income or of the relation between education and health. Unfortunately educational attainment is rarely recorded by the Austrian institutions.

Missing values regarding the educational attainment are thus a big problem in the research of the Department of Economics of Linz and have to be imputed in a lot of datasets, such as in the data of the National Research Network (NRN) Labor & Welfare State (`www.labornrn.at`). At the moment only for 40% of the people recorded educational attainment is known, for 60% there is no information regarding the highest level of education available in this dataset.

The purpose of my master thesis is therefore an imputation of educational attainment for the Austrian population.

In general, there exist a variety of methods that help to handle or to predict missing values. Missing values may be deleted, replaced with the median or the mean or they may be predicted with the help of statistical learning methods. The idea for the imputation of this thesis is based on two methods: first, on random forests and second, on association rules. Two different datasets will be used to predict educational attainment: Austrian census data and data of the NRN Labor & Welfare State.

This master thesis is organized as follows: After the introduction (Chapter 1), Chapter 2 defines educational attainment. In addition, it gives a short overview of the Austrian education system, based on ISCED 2011 and describes and categorizes the variable educational attainment. Chapter 2 also describes the already known information of the highest level of education of the Austrians and compares this information with data from Statistik Austria.

In Chapter 3 the general problem of missing values and the methods how to deal with this problem are described. For this purpose statistical learning is presented and three specific methods: classification trees, random forests and association rules are explained in detail.

Chapter 4 is the practical part of the thesis and the imputation of educational attainment. In this chapter the datasets which are used are first presented and descriptive statistics are carried out. Next, the statistical learning methods (random forests and association rules) are applied to predict educational attainment. Some of these results (the results for the year 2001) are presented and then the set-up for the final imputation is explained. Finally, the imputation is carried out and educational attainment is predicted for all Austrians. In the end, the results are compared with data from Statistik Austria. A brief summary concludes the thesis in Chapter 5.

Chapter 2

Educational attainment

This chapter first defines educational attainment and the possibilities how to categorize educational attainment. In addition, it gives a short overview over the Austrian education system. Next, this chapter summarizes the already known information about the highest education completed of the Austrians and a first descriptive analysis is carried out. Moreover, this information is compared with data from Statistik Austria and at the end of this chapter the missing values regarding educational attainment are analysed briefly.

2.1 Definition of educational attainment

Education is a complex phenomenon within a society that considers many aspects. The International Standard Classification of Education (ISCED), which was developed by the UNESCO, defines education as *"Processes by which societies deliberately transmit their accumulated information, knowledge, understanding, attitudes, values, skills, competencies and behaviours across generations. It involves communication designed to bring about learning"* (ISCED, 2011, p. 79).

Education can be divided into formal and non-formal education and covers in total a variety of education programmes, such as initial education, regular education, second chance programmes, literacy programmes, adult education, continuing education, open and distance education, apprenticeships, technical or vocational education, training, or special needs education (ISCED, 2011, p. 11).

2.1.1 Levels of education

Levels of education are a construct, that are represented by an ordered set and which group education programmes in relation to gradations of learning experiences in a set of categories. *"These categories represent broad steps of educational progression in terms of the complexity of educational content. The more advanced the programme, the higher the level of education"* (ISCED, 2011, p. 12).

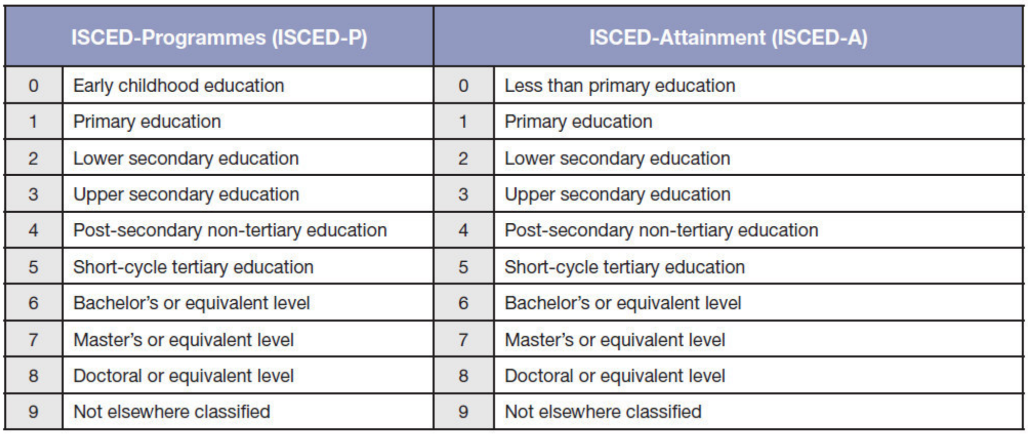
The highest level of education of a person is called educational attainment. *"Educational attainment refers to the highest level of education completed by a person, shown as a percentage of all persons in that age group"* (OECD, 2015).

In general, there exist various methods how to structure educational attainment. In this section the international definition of the UNESCO, the structure of the Micro-Census and "Bildungsstandregister" of Statistik Austria and the structure of the already generated variable education are described in detail.

2.1.1.1 ISCED 2011

The UNESCO Institute for Statistics has developed an International Standard Classification of Education (ISCED) which should help to compare performance in the education systems across countries and over time. Its current version that was adopted in November 2011 is ISCED 2011 (ISCED, 2011, p. iii).

The ISCED coding scheme consists of a system of nine different levels, starting from “Early childhood education” to “Doctoral or equivalent level” and a further decomposition into categories and subcategories (see Figure 2.1) (ISCED, 2011, p. 21).



ISCED-Programmes (ISCED-P)		ISCED-Attainment (ISCED-A)	
0	Early childhood education	0	Less than primary education
1	Primary education	1	Primary education
2	Lower secondary education	2	Lower secondary education
3	Upper secondary education	3	Upper secondary education
4	Post-secondary non-tertiary education	4	Post-secondary non-tertiary education
5	Short-cycle tertiary education	5	Short-cycle tertiary education
6	Bachelor’s or equivalent level	6	Bachelor’s or equivalent level
7	Master’s or equivalent level	7	Master’s or equivalent level
8	Doctoral or equivalent level	8	Doctoral or equivalent level
9	Not elsewhere classified	9	Not elsewhere classified

Figure 2.1: ISCED coding of level (first digit) (Source: ISCED 2011)

2.1.1.2 Austrian education system

The Austrian education system may also be structured into the nine ISCED levels. The “Institut für Bildungsforschung der Wirtschaft” provides a graphical overview of the Austrian education system structured with the ISCED classification (see Figure 2.2, IBW, 2015).

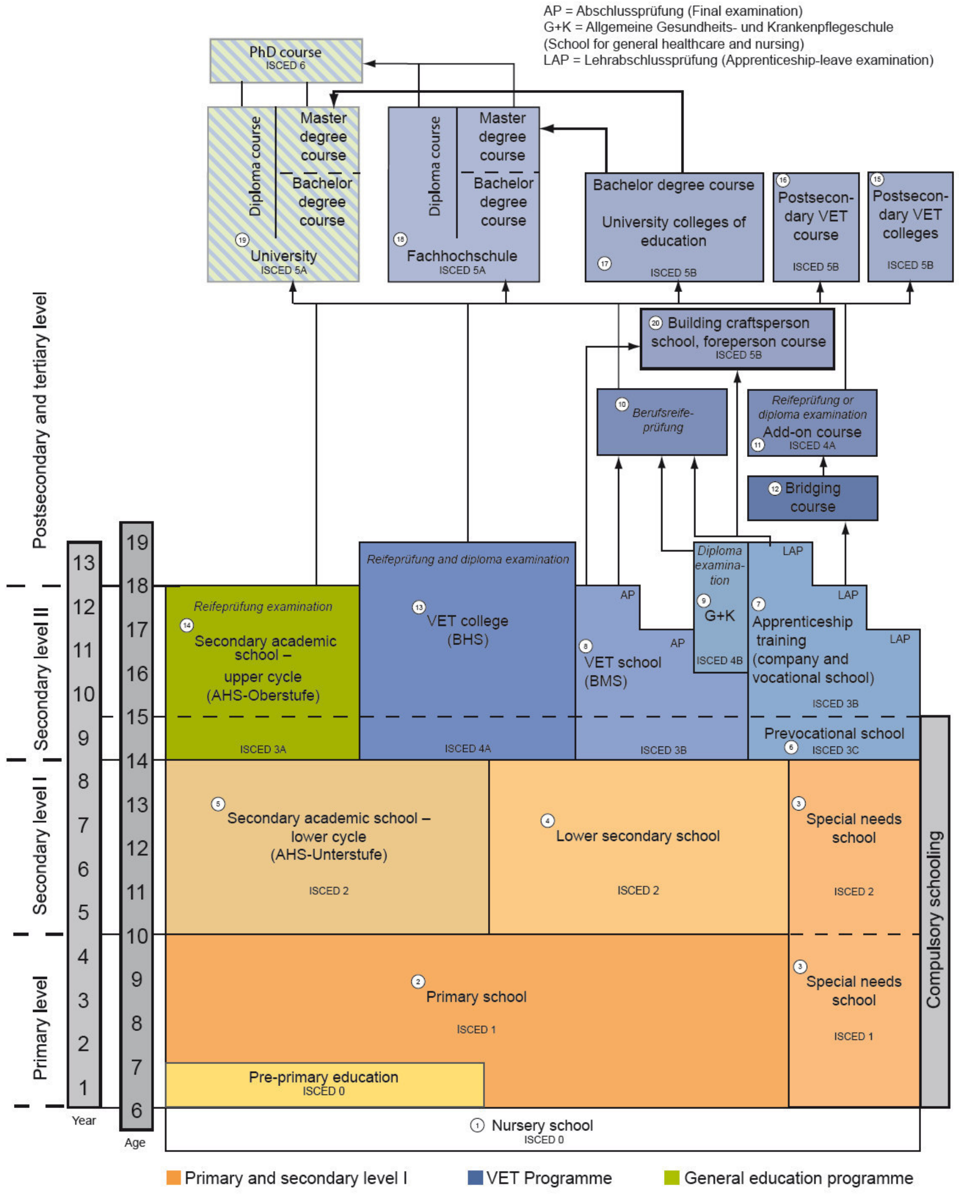

AP = Abschlussprüfung (Final examination)
G+K = Allgemeine Gesundheits- und Krankenpflegeschule
(School for general healthcare and nursing)
LAP = Lehrabschlussprüfung (Apprenticeship-leave examination)
PhD course ISCED 6
Diploma course
Master degree course
Bachelor degree course
19 University ISCED 5A
18 Fachhochschule ISCED 5A
Bachelor degree course
University colleges of education
17 ISCED 5B
16 Postsecondary VET course ISCED 5B
15 Postsecondary VET colleges ISCED 5B
20 Building craftsperson school, foreperson course ISCED 5B
10 Berufsreifeprüfung
Reifeprüfung or diploma examination
11 Add-on course ISCED 4A
12 Bridging course
Postsecondary and tertiary level
Secondary level II
Secondary level I
Primary level
Year
Age
Reifeprüfung examination
14 Secondary academic school – upper cycle (AHS-Oberstufe)
ISCED 3A
Reifeprüfung and diploma examination
13 VET college (BHS)
ISCED 4A
AP
8 VET school (BMS)
ISCED 3B
Diploma examination
9 G+K
ISCED 4B
LAP
7 Apprenticeship training (company and vocational school)
ISCED 3B
6 Prevocational school ISCED 3C
5 Secondary academic school – lower cycle (AHS-Unterstufe)
ISCED 2
4 Lower secondary school
ISCED 2
3 Special needs school
ISCED 2
2 Primary school
ISCED 1
3 Special needs school
ISCED 1
Pre-primary education ISCED 0
1 Nursery school ISCED 0
Compulsory schooling
Primary and secondary level I
VET Programme
General education programme

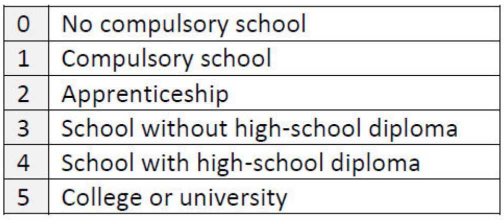
Figure 2.2: The Austrian education system (Source: IBW, 2015)

Statistik Austria provides also often in addition to this international structure, a national structure where educational attainment is structured in another way. Especially in the publications of the results of the Micro-Census the following levels can be found (Statistik Austria, 2012a, p. 39):

- University
- School with diploma
- School without diploma
- Apprenticeship training
- Compulsory school

2.1.1.3 Structure of educational attainment in this thesis

In this thesis educational attainment is also an ordinal variable which has six levels, starting in principle from “No compulsory school” to “College or university”. As there are, however, only very few people with “No compulsory school”, level 0 and level 1 will be combined subsequently in this thesis.
In addition, the imputation methods in Chapter 4 will also show that for the data analysis it is hard to distinguish between “School without diploma” and “School with diploma”. For this reason these two levels will be also combined in the final results, so that in the final imputation the variable educational attainment has only four levels: “Compulsory school”, “Apprenticeship training”, “School with or without diploma”, “College or university”.



0	No compulsory school
1	Compulsory school
2	Apprenticeship
3	School without high-school diploma
4	School with high-school diploma
5	College or university

Table 2.1: Structure of the variable education

2.2 What do we already know?

As already mentioned in Chapter 1 educational attainment is rarely recorded by the Austrian Institutions, such as by the Austrian Social Security Institutions or the Ministry of Finance. It is, however, essential for a lot of economic research questions.

There are several institutions that partly collect information about educational attainment of the Austrians, such as the Austrian Social Security Institutions, the Public Employment Service Austria or the Ministry of Finance.

The only reliable source in this context is the Public Employment Service Austria, which always asks educational attainment of the unemployed. Therefore, if a person is unemployed or has been unemployed at least once, his or her highest level of education at this time point is known for sure. The other institutions, e.g., the Ministry of Finance, only sometimes collect information about education.

In total, there are more than seven sources that may collect educational attainment of the Austrians, such as data sources about

- Apprenticeship training
- Training period
- Subsidies
- AFDC (aid to families with dependent children)
- Free transport for pupils
- Register of births
- Income tax

The Department of Economics in Linz has already combined these different data sources and has created a variable “educ”, which is an ordinal variable that has six levels, starting from “No compulsory school” and ending with “College or university”.

Due to the combination of the different sources information is available for in total 5,407,538 persons. As the dataset consists, however, of more than 11 million observations this variable has a coverage of 39%; in 61% of all cases there is no information about educational attainment available. For this 61% educational attainment should be thus predicted in the course of this thesis.

2.2.1 Descriptive analysis

Table 2.2 shows that 37.64% of those 39% whose highest education completed is known, completed an apprenticeship training, 16.76% have a college or university degree and 16.29% finished a school with diploma. 13.09% finished a school without diploma and for 15.21% the highest education completed was compulsory school.

Highest education completed	absolut	percentage
No compulsory school	54,608	1.01%
Compulsory school	822,788	15.21%
Apprenticeship	2,035,826	37.64%
School without diploma	707,730	13.09%
School with diploma	881,161	16.29%
College or university	906,425	16.76%
Total	5,408,538	100.00%

Table 2.2: Descriptive analysis of education

2.2.1.1 Comparison with data from Statistik Austria

For about 39% of the population the level of education is already known. In order to check the quality of the known information, the generated variable education is compared with data from Statistik Austria. Statistik Austria collects information about education with two different methods. On the one hand, with the "Bildungsstandregister" and on the other hand, with the "Micro-Census – Arbeitskräfte- und Wohnungserhebung", whose main concept is the “Labour Force-Concept”.

2.2.1.2 Micro-Census – Labour-Force-Concept

The Labour-Force Concept (LFC) was developed by the International Labour Organisation and the “Micro-Census – Arbeitskräfte- und Wohnungserhebung” is a continuous primary sample survey of the Austrian households (Statistik Austria, 2014a, p. 4ff).

Statistik Austria categorizes educational attainment with the national concept into five levels: “Compulsory school”, “Apprenticeship training”, “School without diploma”, “School with diploma” and “College or university”.

Table 2.3 shows a comparison of the generated variable education with data from the LFC/ Micro-Census of 2011 and 2013. As Statistik Austria combines the persons who have “No compulsory school” with the level “Compulsory school” this was also done for the variable “educ”.

In addition, as both groups have to have the same composition to be comparable, they include both the whole Austrian population, except for the retired and unemployed people. It is obvious that especially the results of 2013 are similar to those of the generated variable “educ”. The largest difference that may be found is 1.19 percentage points for the level “Apprenticeship training”.

	educ	**2011**	**2013**
Compulsory school	13.38%	15.03%	13.83%
Apprenticeship	37.80%	38.95%	38.99%
School without diploma	13.90%	13.95%	13.17%
School with diploma	17.44%	17.01%	17.35%
College or university	17.49%	15.06%	16.66%

Table 2.3: Comaprison with data from the Micro-Census

A comparison of all Austrians, except for the retired shows similar results (see Table 2.4). The differences are, however, a little bit larger compared to the previous Table 2.3.

	educ	2011	2013
Compulsory school	21.64%	21.98%	20.88%
Apprenticeship	39.82%	34.89%	34.95%
School without diploma	12.60%	12.59%	11.87%
School with diploma	13.96%	17.50%	17.76%
College or university	11.98%	13.04%	14.54%

Table 2.4: Comparison with data from the Micro-Census 2

2.2.1.3 Register – "Bildungsstandregister"

An additional source, apart from the Micro-Census, is the "Bildungsstandregister" which provides also information about educational attainment of the Austrians at the age 15+. The main data in the register is based on the results of the national census from 2001. In the following years it was updated yearly with the information from schools, universities, the Economic Chamber (for the number of finished apprenticeship trainings), etc. (see Statistik Austria, 2014b).

Data is available for the Austrian population at the age 25 to 64 years and in this case Statistik Austria structures educational attainment into three levels: "Primary school", "Secondary school" and "Tertiary school".

Table 2.5 indicates a comparison of education with results of the "Bildungsstandregister" 2011. To provide a valid comparison with the generated variable "educ", data from the "Bildungsstandregister" will again concentrate on the Austrian population, except for the retired and unemployed. Also Table 2.5 shows that the generated variable "educ" seems to display educational attainment of the Austrians quite well.

The two comparisons with data from the Micro-Census and the "Bildungsstandregister" showed that the generated variable represents educational attainment of the Austrians quite well. Therefore, the already generated vari-

	educ	“Registerzählung” 2011
Primary school	13.38%	17.81%
Secondary school	69.14%	66.77%
Tertiary school	17.49%	15.42%

Table 2.5: Comparison with data from the “Bildungsstandregister”

able may be used as a training set for the imputation model in the further thesis.

2.2.1.4 Analysis of the missing values

Table 2.6 shows an analysis of the missing values of the variable “educ” in reference to the birth decades of the Austrians. It may be seen that information about educational attainment is available especially for those who were born between 1960 and 1980. For the youngest and oldest people in the sample, the data contains nearly no information about educational attainment. For those persons it may be difficult to predict educational attainment. Therefore, the imputation will only concentrate on the Austrians who were born between 1930 and 1990.

birthyear	missing values	information
x <1900	99.00%	1.00%
1900≤ x <1910	98.37%	1.63%
1910≤ x <1920	97.29%	2.71%
1920≤ x <1930	93.75%	6.25%
1930≤ x <1940	78.89%	21.11%
1940≤ x <1950	64.29%	35.71%
1950≤ x <1960	42.68%	57.32%
1960≤ x <1970	25.31%	74.69%
1970≤ x <1980	31.67%	68.33%
1980≤ x <1990	42.32%	57.68%
1990≤ x <2000	66.06%	33.94%
2000≤ x <2010	99.95%	0.05%

Table 2.6: Missing values in reference to the birthyear

Chapter 3

Missing values

This chapter now describes the problem of missing values. Therefore, it analyses the consequences of missing values in a general way and describes the methods how to deal with these values. In detail, statistical learning methods are presented and especially random forests and association rules are described.

3.1 Problem of missing values

Missing values are values that we wanted to obtain during data collection, but which we did not get due to different reasons. This problem of missigness might appear because of different reasons: the respondents did not answer all questions, there might have been problems during the manual data entry process, data might be censored, the measurement may be incorrect, etc. (see Kaiser, 2014, p. 42).

Barnard and Meng find three main problems that occur as a result of missing values (see Barnard/Meng, 1999, p. 17):

- loss of information or power;
- complication in data handling, computation and analysis due to irregularities in the data patterns and non applicability of standard software;

- potentially very serious bias due to systematic differences between the observed data and the unobserved data.

3.1.1 Mechanisms of missing values

Mechanisms of missingness describe the relationship between the missing values and the observed units (see Göthlich, 2009, p. 120). In general, three different mechanisms of missing values exist: Missing Completely at Random (MCAR), Missing at Random (MAR) and Missing Not at Random (MNAR) (Rubin, 1976).

The following description of the missing data mechanisms and the standard methods how to handle missing data are based on the book *Statistical Analysis with Missing Data*, written by Little & Rubin (2002).

If we define the complete data $Y = (y_{ij})$ and the missing data indicator matrix $M = (M_{ij})$. The missing data mechanism is defined by the conditinal distribution of M given Y: $f(M|Y, \Phi)$, where Φ are the unknown parameters. Y may be split up into Y_{obs}, which denotes the observed components and Y_{miss} the missing components.

Missing completely at random

Missing completely at random (MCAR) occurs when there is no relationship between the missingness and the data record, which means that the missing values occur totally at random. Therefore, $f(M|Y, \Phi) = f(M|\Phi)$ for all Y, Φ.

Missing at random

Missing at random means, that given the observed data, data are missing independently of the unobserved data. Thus, $f(M|Y, \Phi) = f(M|Y_{obs}, \Phi)$ for all Y_{miss}, Φ.

Missing not at random

If data is missing not at random the missing observations are related to the values of the unobserved data.

As in this thesis educational attainment is known for all people who are unemployed or have been unemployed at least once, the data for the imputation is missing not at random.

3.2 Methods

In general, there exists a wide range of different methods that might be used if missing values occur (see Little & Rubin, 2002).

3.2.1 Standard methods

Little and Rubin (2002) distinguish between four methods to handle missing data: complete case analysis, weighting procedures, imputation methods and model-based methods.

The first and simplest method is to delete the incomplete units and only use the complete recorded units. The second method is to use weighting procedures, where first the incomplete units are deleted. Then the observed units are weighted by their design weights, which are inversely proportional to their probability of selection. The third method are imputation based methods, where the missing values are filled in. Then the complete data record can be analysed with standard methods. Examples for this kind of method, is the hot deck imputation, where the recorded units in the sample are used to substitute the missing values, the mean imputation, where missing values are replaced with the means of the variable or regression imputation, where the missing values are predicted by a regression model.

The fourth type of method are model-based methods. These models are generated by defining a model, which is based on the observed data, and basing inferences on the likelihood or posterior distribution under that model. The parameters are estimated by procedures as for instance maximum likelihood. (see Little & Rubin, 2002).

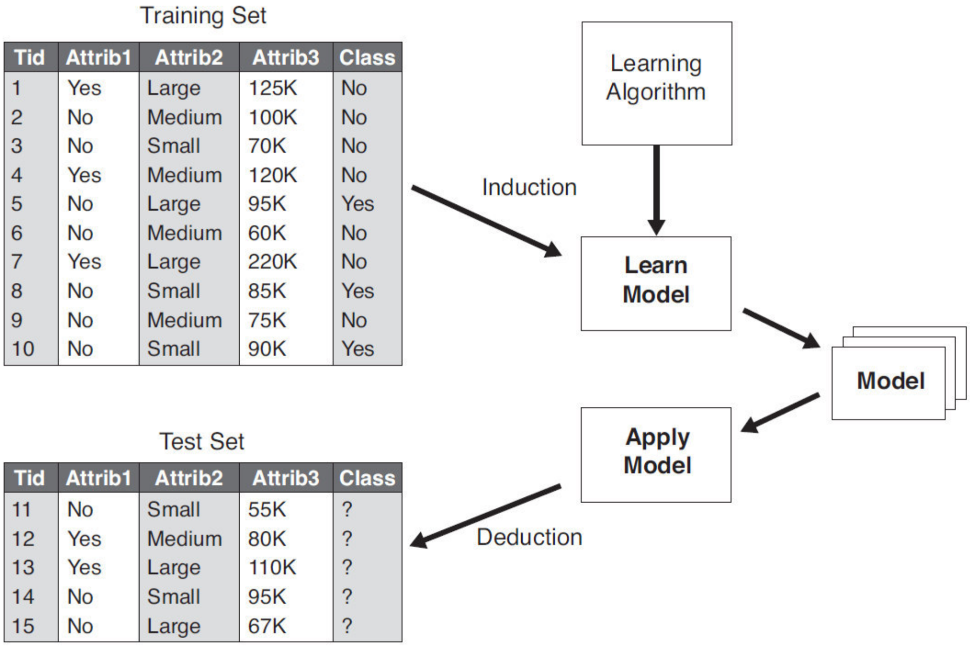
As there are about 40% missing values regarding educational attainment and the data is not MCAR complete case analysis is not an appropriate method for the imputation. However, as there exist variables that can explain educational attainment, such as income or the age at the first job, the further analysis will concentrate on imputation based methods. In the next subsection statistical learning will be described in detail.

3.2.2 Statistical learning

"Statistical learning refers to a vast set of tools for understanding data" (James et al., 2013, p. 1).

With statistical learning we want to learn from data. Statistical learning plays an important role in many fields of statistics, data mining and artificial intelligence and is even intersecting with areas of engineering and other disciplines (see Hastie, Tibshirani & Friedman, 2011, p. 1).

Statistical learning may be classified into *supervised* or *unsupervised* learning (see James et al, 2013, p. 1). The aim of *supervised* learning is to predict the value of an outcome measure with a number of input variables/features. With the help of a training set which contains the outcome variable, as well as the features, a prediction model (learner) is built. This prediction model enables then to predict the outcome for new objects (see Hastie, Tibshirani & Friedman, 2011, p. 1f). The output may be either quantitative or categorical, which leads to two different prediction types: regression or classification (see Hastie, Tibshirani & Friedman, 2011, p. 10).
In *unsupervised* learning there is no outcome measure and the goal is to describe the associations and patterns among the variables (see Hastie, Tibshirani, & Friedman, 2011, p. xi).



Training Set

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Test Set

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Figure 3.1: Classification model (Source: Tan et al., 2005, p. 148)

Classification is one supervised learning type. It is the task of assigning objects to one of several predefined classes, where the input data is a collection of records (see Tan et al., 2005, p. 145). It is *“the task of learning a target function f that maps each attribute x to one of the predefined class labels y.”* (Tan et al., 2005, p. 146). Classification models can be distinguished in two different types: descriptive models, that serve as an explanatory tool and predictive models that help to predict the class of unknown labels. A general example of a predictive classification model can be seen in Figure 3.1.

As the imputed variable education will be used for further research and to avoid bias and problems in further estimations, the thesis will focus on two simple and non-parametric statistical learning methods: Random Forests (RF) and association rules, that are explained in the following subsections. As classification trees are the basis for Random Forests, tree-based methods are described first.

3.2.3 Tree-based methods

With tree-based methods the input space is partitioned into a set of rectangles (R), where in each rectangle ($R_1, .., R_m$) a simple model (e.g. a constant) is fit to the data. Figure 3.2 shows a two-dimensional example with two variables X_1 and X_2, where the square input is first split at $X1 = t1$, then the rectangle $X1 < t1$ is split at $X2 = t2$. After that the region $X1 > t1$ is split at $X1 = t3$ and then $X1 > t3$ is split at $X2 = t4$, so that there are five regions in the end. In the corresponding model Y is predicted with a constant c_m in region R_m: $\hat{f}(x) = \sum_{m=1}^{5} c_m I\{(X_1, X_2) \in R_m\}$ (see Hastie et al., 2006, p. 306).

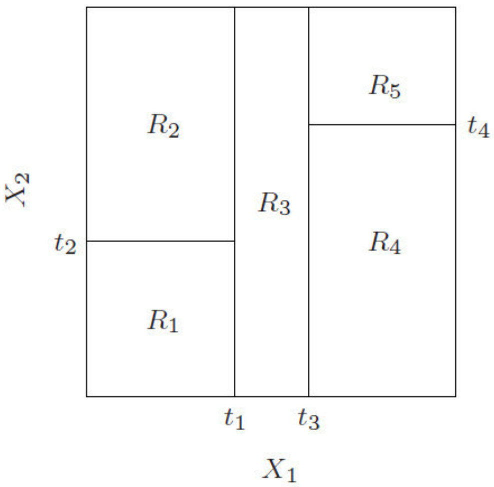


Figure 3.2: Recursive binary splitting (Source: Hastie et al., 2006, p. 306)

Figure 3.3 shows the same model, represented as a binary tree (see Hastie et al., 2006, p. 306).

If the output of the tree is continuous we talk about regression trees; with categorical output we have classification trees. A decision tree has a hierarchical structure and consists of several nodes. In general, there are three types of nodes: root nodes, internal nodes and leaf or terminal nodes. In the leaf or terminal nodes the different classes of the variable that should be predicted can be found, the root nodes and internal nodes contain the ex-

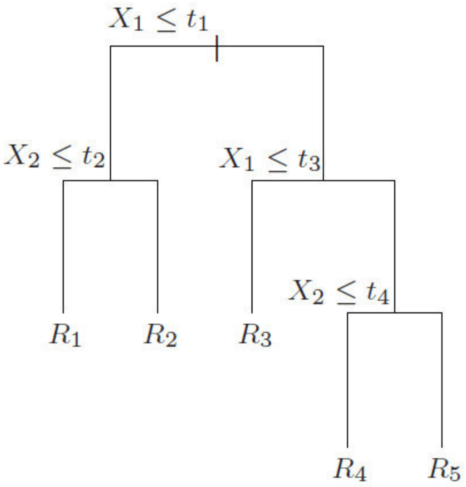


Figure 3.3: Example of a binary tree (Source: Hastie et al., 2006, p. 306)

planatory attributes. If a new object should be classified, the starting point is the root node, then the object is pulled down the tree until a final class in a terminal node is reached. The construction of a classification tree may be based on several different algorithms (see Tan et al., 2005, p. 150f).

In order to explain the construction of a tree regression trees are described first. Then, classification trees will be explained.

3.2.3.1 Regression trees

The algorithm for the tree construction needs to automatically choose the best splitting variable and split points. The following description of the construction is based on *Elements of Statistical Learning* (Hastie et al., 2011). The data, which consists of p inputs and a response for each observation is first partitioned into M regions $R_1, R_2, .., R_M$ where the response is modelled as a constant in each region R_m:

$$f(x) = \sum_{m=1}^{M} c_m I(x \in R_m). \tag{3.1}$$

If the criterion minimization of the sum of squares is chosen $(\sum(y_i - f(x_i))^2)$

the best $\hat{c}_m$ is the average of y_i in region R_m:

$$\hat{c}_m = ave(y_i | x_i \in R_m). \tag{3.2}$$

In order to find the best binary partition regarding the minimum sum of squares a greedy algorithm is applied: Starting with all of the data, with the splitting variable j and the split point s the pair of half-planes can be defined as follows:

$$R_1(j,s) = \{X | X_j \leq s\} \quad \text{and} \quad R_2(j,s) = \{X | X_j > s\} \tag{3.3}$$

The splitting variable j and the split point s which solves following equation is searched:

$$\min_{j,s} [\min_{c_1} \sum_{x_i \in R_1(j,s)} (y_i - c_1)^2 + \min_{c_2} \sum_{x_i \in R_2(j,s)} (y_i - c_2)^2] \tag{3.4}$$

For any j and s the inner minimization can be solved by

$$\hat{c}_1 = ave(y_i | x_i \in R_1(j,s)) \quad \text{and} \quad \hat{c}_2 = ave(y_i | x_i \in R_2(j,s)) \tag{3.5}$$

For each splitting variable the split point s can be found by scanning through all the inputs and determining the best pair (j, s).
If the best split is found, the data is split into the two resulting regions and the splitting process is repeated on each of the two regions. This process is then repeated on all of the resulting regions.

Normally a large tree T_0 is grown, which is stopped when some minimum node size (e.g. 5) is reached. Then this large tree is pruned using the *cost-complexity pruning*, which works as follows:
We define the subtree $T \subset T_0$, which is any tree that can be created by pruning T_0. Terminal nodes are indexed by m, with node m representing the region R_m and $|T|$ is the number of terminal nodes in T.
Then $N_m = \#\{x_i \in R_m\}, \hat{c}_m = \frac{1}{N_m} \sum_{x_i \in R_m} y_i, Q_m(T) = \frac{1}{N_m} \sum_{x_i \in R_m} (y_i - \hat{c}_m)^2$

and the cost complexity criterion is defined as:

$$C_\alpha(T) = \sum_{m=1}^{|T|} N_m Q_m(T) + \alpha|T| \tag{3.6}$$

For each α a subtree $T_\alpha \subseteq T_0$ that minimizes $C_\alpha(T)$ is found. $\alpha \geq 0$ is the tuning parameter that governs the tradeoff between the tree size and the goodness of fit to the data. T_α is found by weakest link pruning, where the internal node that produces the smallest per-node increase in $\sum_m N_m Q_m(T)$ is collapsed until the single-node tree is created. α is found by five- or tenfold cross-validation (see Hastie et al., 2011, p. 306f.).
If the target variable is not metric but categorical we have classification trees.

3.2.3.2 Classification trees

"Classification trees are used to classify an object or an instance (such as insurant) to a predefined set of classes (such as risky/non-risky) based on their attributes values (such as age or gender)" (Rokach & Maimon, 2008, p. 5f).

To choose the best splitting variables in classification trees different metrics exist. In regression trees the squared error node impurity measure $Q_m(T)$ was used. For classification trees an important feature is the proportion of class k observations in node m: $\hat{p}_{mk} = \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i = k)$.
The observations in node m is classified to class $k(m)$= arg$max_k \hat{p}_{mk}$, the majority class in node m.
Measures $Q_m(T)$ of node impurity include the following (see Hastie et al., 2011, p. 309):

Missclassification error: $\frac{1}{N_m} \sum_{i \in R_m} I(y_i \neq k(m)) = 1 - \hat{p}_{mk(m)}$

Gini index: $\sum_{k \neq k'} \hat{p}_{mk} \hat{p}_{mk'} = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$

Cross-entropy or deviance: $-\sum_{k=1}^{K} \hat{p}_{mk} log(\hat{p}_{mk})$

The node impurity is 0 when all patterns at the node are of the same category and it becomes maximum when all the classes at node m are equally likely (see Tan et al., 2005, p. 158).

Figure 3.4 (see Tan et al., 2005, p. 151) plots an example classification tree where animals should be classified into “Mammals” and “Non-mammals”. If a new animal will be classified, the starting point (the root node) is the first decision criterion where the body temperature is asked. If the animal is a cold blood animal, the leaf node “Non-mammal” is already reached and the animal is classified as “Non-mammal”. If the answer is “warm” the next internal node and the question if the animal gives birth is asked. If this question is answered with “Yes” the animal is classified as “Mammals”, if not it is a “Non-mammal”.

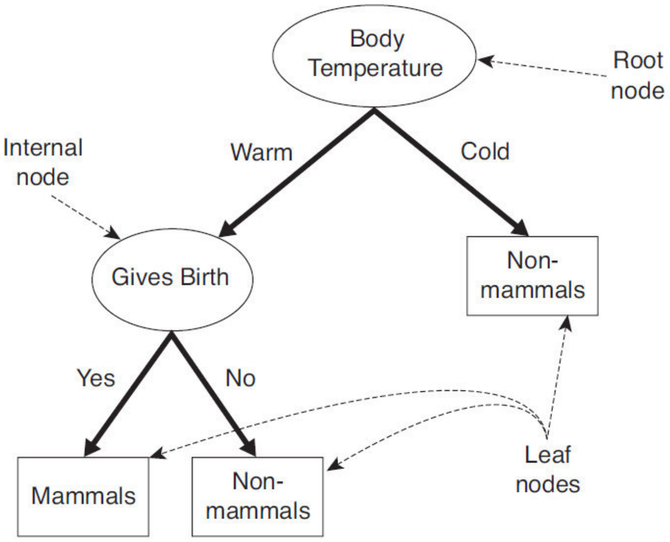


Figure 3.4: Example of a decision tree (Tan et al, 2006, p. 151)

3.2.4 Random Forests

Random Forests, which were first developed by Breiman in 2001, are a bagging method, which consists of a large number of de-correlated trees which are then averaged. The main idea of bagging or bootstrap aggregation is the reduction of the variance of an estimated prediction function. Trees can be especially well used for bagging, since they can explain complex interaction structures and thus they have relatively low bias, if they are grown deep (see Hastie et al., 2011, p. 587). In a Random Forest a large number of trees is grown and if a new object should be classified, each tree gives a classification and the class with most votes wins. In detail, the construction algorithm for a forest works as follows (see Breiman & Cutler, 2015).

In a Random Forest each tree is grown as follows. N is the number of observations in the training set and M the number of input variables.

1. Sample N cases at random, with replacement, from the original data. This will be the training data for the tree.
2. m variables are chosen at random out of the M input variables and the best split on these m is used to split the node. The value of m is constant during the construction of the forest.
3. Each tree is grown to the largest extent possible, without pruning.

With the out-of-bag error an optimal value of m can be found.

Figure 3.5: Algorithm: Random Forest (Source: Breiman & Cutler, 2015)

3.2.4.1 Out-of-bag samples

Since the training set for each tree is drawn by sampling with replacement, some cases are left out of the sample. This oob (out-of-bag) data can be used to get an unbiased estimate of the classification error when trees are added to the forest and it may be also used to get estimates of variable importance (see Breiman & Cutler, 2015).

3.2.4.2 Variable importance

Variable importance of a variable m may be computed by using the oob cases which are put down the forest and the correct number of classifications are counted. Then the values are randomly permuted of variable m in the oob cases and they are again put down the tree. The difference of the correct classifications between the untouched oob cases and the permuted is the raw importance score for variable m (see Breiman & Cutler, 2015).

In case of the Gini importance the Gini impurity criterion is less than the parent node every time a split of a node is made. The sum of the Gini decrease for each individual variable over all trees in the forest gives the Gini variable importance, which is often very consistent with the variable importance measure (see Breiman & Cutler, 2015).

In R Random Forests are created with the package *randomForest* (Liaw & Wiener, 2002). The construction of the forest is based on Breiman and Cutler's original Fortran code (`https://www.stat.berkeley.edu/~breiman/RandomForests/cc_software.htm`).
The forest construction in the R package is implemented in the function *randomForest*, where, for instance, the number of trees in the forest, the number of variables randomly sampled as candidates at each split, the minimum size of terminal nodes or the maximum number of terminal nodes can be adjusted. In addition, the importance of predictors can be assessed and with the function *varImpPlot* it can be plotted. The function plots the Variable Importance and the Gini importance. With the predict method for fitted random forest objects prediction of test data can be applied.

3.2.5 Association rules

Association rules are one of the most common unsupervised learning techniques, which are especially popular for mining commercial databases, such as in market basket analysis. The goal of the association rule analysis is to find frequent item sets: joint values of the variables $X = (X_1, X_2, ..., X_p)$ that appear most frequently in the data base (see Hastie, Tibshirani & Friedman, 2011, p. 487).
Association rules may however also be used in further fields, such as in bioinformatics, medical diagnosis, web mining and scientific data analysis (see Tan et al., 2005, p. 328).

An example of an association rule in the field of the market basket analysis is the statement that *"90% of people that purchase bread and butter also purchase milk"* (see Agrawal et al, 1993, p. 207). The antecedent would be in this case bread and butter, the consequent item is milk. 90% is the confidence of the rule (see Agrawal et al., 1993, p. 2007).

Table 3.1 shows an example market basket data set, represented in a binary format. Each row corresponds to a transaction and each column to an item.

T	Bread	Milk	Butter	Juice
1	1	0	1	0
2	1	1	1	0
3	1	1	1	1
4	1	1	0	0

Table 3.1: Example market basket data

$I = \{i_1, i_2, ...i_d\}$ is the set of all items in the market basket data and $T = \{t_1, t_2, ..., t_N\}$ is the set of all transactions. Each transaction t_i includes a subset of chosen items from I (see Tan et al., 2005, p. 329).

In the example in Table 3.1 the first transaction contains the items $Bread$, $Butter$, but not $Milk$, $Juice$.

An association rule is the expression: $X \rightarrow Y$, such as for example $\{Bread$, $Milk\} \rightarrow \{Butter\}$, which means that "If bread and milk are bought, butter will be bought as well".

Association rules may be described by several properties, which are based on the prevalence of the antecedent and the consequent item in the data set.

The first property is the so called "support" of the rule $T(X \Rightarrow Y)$, which is the fraction of observations in the database of the antecedent and consequent. It can be interpreted as the probability of simultaneously observing both item sets $Pr(X\, and\, Y)$.
The second property is the "confidence", which can be seen as the estimate of $Pr(Y|X)$. The "lift" of the rule is defined as the confidence divided by the expected confidence (see Hastie, Tibshirani & Friedman, 2011, p. 490f.).

The formal definitions are the following (see Hastie, Tibshirani & Friedman, 2011, p. 490f.)

$$Support : S(X \rightarrow Y) = Pr(X \cup Y) \tag{3.7}$$

$$Confidence : C(X \rightarrow Y) = \frac{S(X \cup Y)}{S(X)} \tag{3.8}$$

$$Lift : L(X \rightarrow Y) = \frac{C(X \cup Y)}{S(Y)} \tag{3.9}$$

In the example in Table 3.1 the support of the rule $\{Bread, Milk\} \rightarrow \{Butter\}$ is $2/4 = 0.5$, the confidence is $2/3 = 0.67$, since there are three transactions that contain Bread and Milk.

The Association Rule Mining Problem may be summarized as follows:
"Given a set of transactions T, find all the rules having support > minsup and confidence > minconf, where minsup and minconf are the corresponding

support and confidence tresholds" (Tan et al., 2005, p. 330).

The majority of algorithms beyond the detection of aossociation rules decompose the mining problem into two tasks:

1. Frequent Itemset Generation, where all the frequent items that have a support larger than the minsupport threshold are found
2. Rule Generation, where all the high-confidence rules with a confidence higher than minconf, based on the frequent items are generated (see Tan et al., 2005, p. 331).

In order to find association rules the apriori algorithm can be applied. The main idea of this algorithm is that *"If an itemset is frequent, then all of its subsets must also be frequent* (Tan et al., 2005, p. 333)."
If the itemset $\{c, d, e\}$ is a frequent itemset, then any subset $\{c, d\}, \{c, e\}, \{d, e\}$, $\{c\}, \{d\}$, and $\{e\}$ must be also a frequent itemset (see Tan et al., 2005, p. 333f.).

In detail, the algorithm works as follows: the algorithm first determines the support of each item in the dataset and, for a given support threshold t, all single-item sets with support $> t$ are combined to $L_{1,t}$. Next, all item sets from $L_{1,t}$ are extended with one item and all these item sets of size two with support greater than t define the set of frequent size-two item sets $L_{2,t}$. After $m - 1$ such steps all item sets from $L_{m-1,t}$ are extended with one item and only these size-m item sets with support $> t$ are combined to $L_{m,t}$.
The algorithm continues until all candidate rules from the previous pass have support less than the specified threshold.
The output of the algorithm is the set of item sets with support larger than t: $L_t = \cup_k L_{k,t}$.
Each high-support item set returned by the apriori algorithm is then transformed into a set of association rules. The items $A \cup B = K$ are then generated to the association rule $A \Rightarrow B$ (see Hastie et al., 2011, p. 489f.).

Association rules can be found with the R package *arules* (Hashler et al., 2015). In this package the *Apriori* and *Eclat* algorithms of Borgelt (Borgelt 2003, 2004) are applied. With the functions *apriori* and *eclat* the code is called directly from R. The implementations of *Apriori* and *Eclat* can mine frequent itemsets and *Apriori* can also mine association rules (see Hashler et al., 2015, p. 10).

An extension of the package *arules* is the R package *arulesViz* (Hashler & Chelluboina, 2015), which implements several visualization techniques to explore association rules. In this thesis scatterplots and balloon plots will be applied. The scatterplots use support and confidence on the axes and lift as a color. In the balloon plot the antecedent groups are displayed as columns and consequents as rows (see Hashler & Chelluboina, 2015).

Chapter 4

Imputation

This chapter is the practical part of the thesis, where the aim is an imputation of educational attainment. First, the data sources which further will be used (census and NRN data) are described and analyzed. Next, random forests are grown and association rules are found with this data. For this purpose different versions of forests and association rules (with different numbers of trees, different explanatory variables, support and confidence) were tried. In this chapter some exemplary results of the forests and rules with data of the year 2001 are presented. With the help of these results, a final set-up for the imputation is worked out. Then, the final Random Forests are grown and educational attainment is imputed.
For the analysis the statistical programs Stata (Version 10) and R (Version 3.1.2) are used. In R especially the packages *arules* (Hasler et al., 2015), *arulesViz* (Hashler & Chelluboina, 2015) and *randomForest* (Liaw & Wiener, 2002) were applied.

4.1 Datasets

In this subsection the data which are used for the further analysis is described. For the imputation of educational attainment two different datasets are used: on the one hand, the data of the National Research Network Labor & Welfare State (NRN, 2015), and on the other hand data of the Austrian census of

2001.

4.1.1 NRN data

The first data source for this analysis are the datasets of the National Research Network – Labor & Welfare State (NRN) which are provided by a number of different institutions, such as by

- the Austrian Social Security Institutions (Hauptverband der österreichischen Sozialversicherungsträger);
- the Regional Health Insurance Organisation for Upper Austria and Vorarlberg (Oberösterreichische und Vorarlberger Gebietskrankenkasse);
- the Austrian DRG System (Leistungsorientierte Krankenanstaltenfinanzierung);
- the General Accident Insurance Institution (Allgemeine Unfallversicherungsanstalt);
- the Public Employment Service Austria (Arbeitsmarktservice);
- the Ministry of Finance.

The datasets consist of several years and they contain a multitude of different variables and up to more than 11 million observations. The largest dataset is the data of the Austrian Social Security Institutions, which covers the whole Austrian population. It contains information about the insured person, the employer, the contribution base of the insured, etc. Zweimüller et al. (2009) provide a very detailed description of this data.

The data of the Public Employment Service Austria provides information about all unemployed and is in addition the only reliable source for the data collection of educational attainment, as the Public Employment Service Austria always collects data on the level of education of the unemployed.

4.1.2 Census data

In addition to the datasets of the NRN, the Austrian census is a further data source which provides information about educational attainment.

The classic census which collects demographic and labour market data was carried out every ten years and was last done in 2001. Beside demographic variables, educational attainment, the status of employment, the job and industry, as well as information about the household situation was asked. Results are provided not only for persons, but also for households and families (see Statistik Austria, 2005, p. 4). The census survey is a full sample of all Austrian residents who had the duty to provide information (see Statistik Austria, 2005, p. 3).

The Minnesota Population Center and the University of Minnesota provide with their Integrated Public Use Microdata Series (IPUMS) census microdata for social and economic research. For Austria a 10% sample of the census is available for the years 1971, 1981, 1991 and 2001 (see IPUMS, 2015).

IPUMS structures educational attainment in several different ways. In total, there are four variables that describe educational attainment. To be able to use these variables for the further analysis, they have, however, to match the levels of educational attainment in the other datasets. For this analysis the variable "edattan" was taken. It is structured into eight levels, which have been in order to be able to use them for the analysis, transformed into five levels (Compulsory school (level 1) – Apprenticeship training (level 2) – School without diploma (level 3) – School with diploma (level 4) – College or university (level 5)). Table 4.1 indicates how the different levels of educational attainment have been matched.

4.1.3 Descriptive analysis of the datasets

In this subsection descriptive statistics of the two datasets are carried out.

IPUMS data	**new level**
Compulsory (lower) secondary school	1
Apprenticeship training	2
Intermediate technical and vocational school	3
Higher general secondary	4
Higher technical and vocational secondary school	4
Technical or vocational course	5
(Academic) Intermediate degrees	5
University, college	5

Table 4.1: Transformation of edattan to educ

4.1.3.1 Census 2001

This subsection focuses on the census of 2001, which was hold on the 15th of May 2001.

IPUMS provides a 10% sample of this census, which is a dataset with 803,471 observations. 45 variables that may be interesting for the explanation of educational attainment are included in the data. The variables which may be used for the imputation are listed in Table 4.2. As the R package *randomForest* can only handle complete datasets, missing values in the explanatory variables were replaced with “999”. “999” was chosen, as this is an unrealistic number for the values of the variables, such as for the familysize or the number of born children.

Comparison with full census

As the following results are based on a 10% sample of the complete census, it is interesting to compare the shares of educational attainment with results of the complete census of 2001, published by Statistik Austria, in order to check the quality of the sample.

Table 4.3 shows the comparison of the sample with the full census and it is obvious that the 10% sample represents the census, regarding educational attainment, quite well, as the largest difference of the two samples is 0.09

Variable	Description
nuts2	NUTS2
nuts3	NUTS3
familysize	familysize
nchild	number of children living in family
nchlt5	number of children younger than 5 living in family
eldch	age of the eldest children living in family
yngch	age of the yougest children living in family
birthyear	birthyear
sex	gender
marst	marital status
citizen	citizenship
EU28	member of EU28
educat5	education
eempsta	employment status
class	working class
hrsfull	full or part-time employed
cont	continent
chbornd	number of born children

Table 4.2: Census 2001: variables

percentage points.

	1	**2**	**3**	**4**	**5**
Sample	39.19%	32.14%	10.88%	10.31%	7.48%
Full census	39.23%	32.03%	10.91%	10.30%	7.53%

Table 4.3: Comparison 10% sample with full census

General description

In order to get a first overview about the census data, a brief descriptive analysis of the sample follows.

The sample contains 389,428 men (48.47%) and 414,043 women (51.53%) of all ages. The youngest were born in 2001, the year of the sample, the oldest in 1900. Figure 4.6 shows the detailed distributions of the birthyears, grouped into decades.

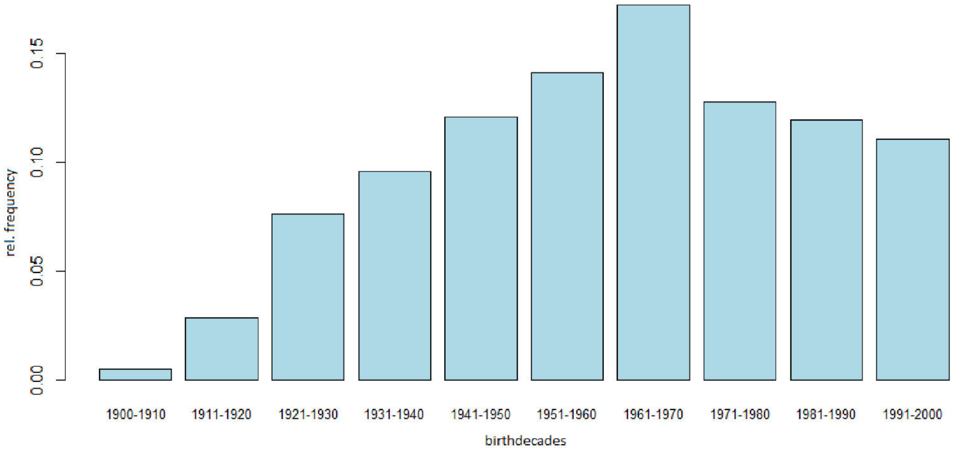


Figure 4.1: Birthdecades

Most of the people in the sample were Austrians (91.06%), there were, however, also many people of other nations that participated in the census. For details see Figure 4.2 which shows the nations where most of the people came from.

Concerning their family status, 42.43% were single and have never been married, 41% are married, 9.47% divorced and 7.10% widowed. Moreover, 28.72% had no children, 21.67% one, 28.03% two, 12.62% three, 4.99% four, 2.06% five and 1.91% more than five children.

In addition, at the date of the census, 46.38% were employed, 3.17% unemployed, 14.58% were students and 22.84% were retired. In detail, 36.69% of the employed were blue-collar workers, compared to 46.39% white collar workers. 9.33% were self-employed, 7% were public officials and 0.58% unpaid family workers.

Moreover, 83.59% of the employed had a full-time and 16.41% a part-time contract.

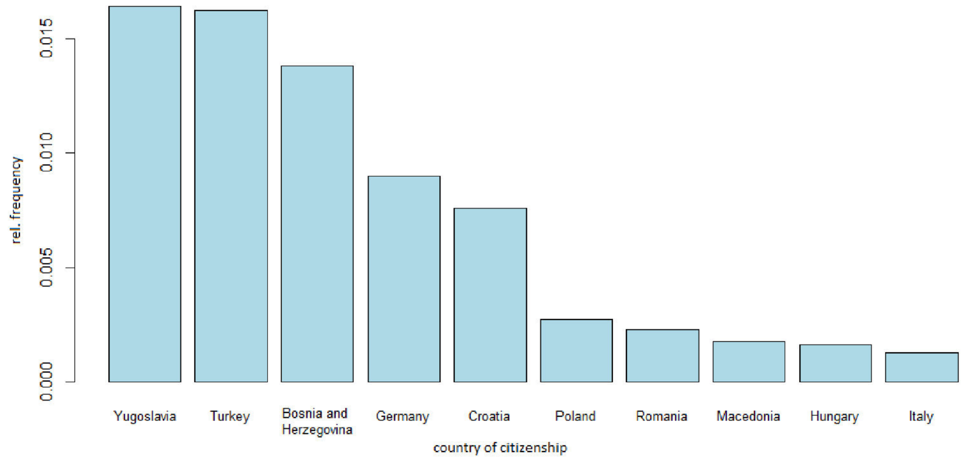


Figure 4.2: Shares of other citizenships

Educational attainment

In the following subsection education attainment is described. In the tables “1” refers to “Compulsory school”, “2” to “Apprenticeship training”, “3” to “School without diploma”, “4” to “School with diploma” and “5” to “College or university”. Regarding educational attainment, 41.44% finished primary school, 11.52% a school without and 9.74% a school with diploma. In addition, 5.68% have a university degree (see Table 4.4).

	1	**2**	**3**	**4**	**5**
educ	41.44%	32.14%	11.52%	9.74 %	5.16%

Table 4.4: Educational attainment Census 01

In order to check if educational attainment differs among citizenships, the relationship between educational attainment and nation will be analyzed in a more detailed way.
Therefore, different countries will be summed up to groups, for example as continents. Afterwards these groups are going to be compared.

At first educational attainment of all people who belonged to a Member State of the EU15 countries will be compared to educational attainment of all others.

	1	**2**	**3**	**4**	**5**
other	63.16%	19.67%	4.06%	8.43%	4.68%
EU15	37.34%	33.10%	11.95%	12.19%	5.42%

Table 4.5: Educational attainment of EU15

Table 4.5 shows that the two groups differ a lot regarding educational attainment. Whereas 63.16% of those who did not belong to a EU15 Member State finished primary school, this share of the EU15 members is 37.34%.

A similar picture shows the comparison of educational attainment of those who are part of the EU28 countries, compared to all other nations. According to Table 4.6 the majority of those who are not a EU28 member had finished primary school and educational attainment of only 6.19% is a school without diploma.

	1	**2**	**3**	**4**	**5**
other	69.41%	17.03%	3.39%	6.19%	3.99%
EU28	37.46%	33.00%	11.84%	12.25%	5.44%

Table 4.6: Educational Attainment of EU28

The last table of this comparison shows a comparison of educational attainment in reference with the continents. Table 4.7 indicates that there are remarkable differences in educational attainment if the people are grouped into continents.

	1	**2**	**3**	**4**	**5**
Africa	60.41%	8.71%	4.32%	13.44%	13.11%
Asia	78.75%	8.53%	2.67%	5.86%	4.20%
Austr./N.Z	43.08%	6.15%	9.23%	12.31%	29.23%
Centr./So. A	55.49%	4.75%	4.75%	18.40%	16.62%
Europe	38.35%	32.70%	11.59%	12.03%	5.33%
North A.	38.48%	5.58%	3.30%	18.03%	34.62%
Oceania	43.75%	6.25%	6.25%	6.25%	3.75%

Table 4.7: Educational attainment continents

4.1.3.2 NRN data 2001

The NRN dataset of 2001 contains 4,611,035 observations, where educational attainment is known. 52.91% are men, 47.09% are women. In addition, 89.44% are Austrians and 10.56% have another citizenship. In reference with the employment status 56.50% were employed, 4.82% unemployed and for the rest the employment status was unknown. Of the employed 42.18% are white collar, 52.89% blue collar workers and 0.84% civil servants. 90.72% have been at least once part-time employed and 37.70% had at least one summer job.

Regarding the number of children, most of the people (68.11%) do not have any child, 11.35% have one child, 14.03% have two, 4.82% three, 1.26% four and 0.43% have five children. When giving birth to the first child the average age of a woman was 25.5 years. 25% of the women were younger than 21 and 75% of them where younger than 29. The average age when giving birth to the second child was 28. The first quartile in this case was 24 years and the third 31. Furthermore, 4.64% of the people in the dataset have already died.

Educational attainment

The descriptive analysis of educational attainment shows that 17.69% have attended compulsory school. Educational attainment of 34.10% is an apprenticeship training and of 12.53% a school without diploma. In addition, 17.03% have completed a school with diploma as the highest educational level of 17.64% is a college or university.

A comparison of educational attainment of Austrians with all other nations indicates that more Austrians than others have completed an apprenticeship training (35.09% compared with 21.12%). However, people from other countries also have a lower probability of having completed a school with or without diploma or a university.

	1	**2**	**3**	**4**	**5**
other	57.48%	21.12%	5.59%	8.14%	7.68%
Austrian	14.67%	35.09%	14.13%	17.71%	18.40%

Table 4.8: Educational attainment of Austrians: NRN data 01

A comparison of educational attainment of those whose nation is a Member State of the EU15 countries with all others shows that there are large differences between the groups. Whereas compulsory school is the highest educational level of 14.78% of those who are part of EU15, this share is 63.15% of those who are not part of EU15.

	1	**2**	**3**	**4**	**5**
other	63.15%	19.27%	5.16%	7.09%	5.33%
EU15	14.78%	35.05%	14.07%	17.67%	18.43%

Table 4.9: Educational attainment of EU15: NRN data 01

The comparison of educational attainment of the Austrians with other nations showed that there are remarkable differences between the groups in both datasets. Therefore, the citizenship is an important explanatory variable which should be included in the models.

4.2 Statistical learning

The final purpose of this thesis is an imputation of educational attainment with the help of two different statistical learning methods: association rules and Random Forests. In the end educational attainment, classified with a level between 1 (Compulsory school) to 5 (College or university), should be available for all Austrians. For this purpose, the following steps will be processed:

1. First of all, a random forest will be grown that shows which levels are easy to predict and which variables are important.
2. Second, association rules with a given minimum support and confidence level will be found.

In order to be able to predict the highest level of education, it is first necessary to find suitable explanatory variables. The variables which may explain educational attainment in the census data were listed in Table 4.2. The list of all 141 explanatory variables in the NRN data is in the Appendix.

4.3 Results

This section will give an overview of some of the results which were obtained with the NRN and census data and finally the set-up for the final imputation will be worked out. The Random Forests are grown with the statistical software R (R Core Team, 2015) and the package *randomForest* (Liaw & Wiener, 2002). In order to find and display association rules the packages *arules* (Hahsler et al., 2015) and *arulesViz* (Hahsler & Chelluboina, 2015) will be applied.

To predict educational attainment different Random Forests (with a different number of trees and number of variables at each split) were tried for Census and NRN data of 1991, 2001 and 2010. In addition, Random Forests with stratified and non stratified samples were calculated, as well as different association rules with different minimum confidence and support.
The following results of association rules and Random Forests are examples of some of the calculations and will focus on data of 2001. In the representation of the results “1” will refer to “Compulsory school”, “2” to “Apprenticeship training” etc.:

1	=	Compulsory school
2	=	Apprenticeship training
3	=	School without diploma
4	=	School with diploma
5	=	College or university

4.3.1 Census data 2001

This subsection is going to present association rules and Random Forests, built with the Census data of 2001.

Random Forests

The Random Forest of the census data which is presented here, was created with 90,000 observations and 18 explanatory variables. This number of observations was chosen due to computation time and the subset was drawn randomly from all observations.

Figure 4.3 shows the development of the out-of-bag (oob) errors over the number of trees which were created. The errors of all classes is decreasing at the beginning. After the creation of around 40 trees, the errors stabilize. As can be seen in Figure 4.3 the out-of-bag (oob) error of this forest is with an average of over 50% quite high. In addition, levels 3 and 4 are with an error of about 90% nearly impossible to predict. The error rate of level 2 is with about 15% the lowest one.

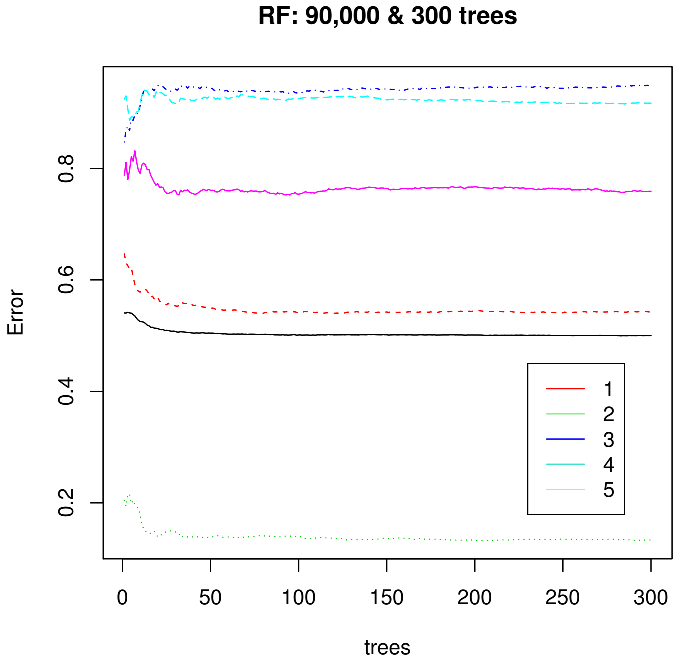


Figure 4.3: RF – Census 2001: error rate

Figure 4.4 plots the Variable importance (left plot) and the Gini importance (right plot). A closer look at the important variables (see Figure 4.4) indicates that the working class (blue or white-collar worker), the birthyear, the gender, the region of residence and size of the family are the most important explanatory variables.

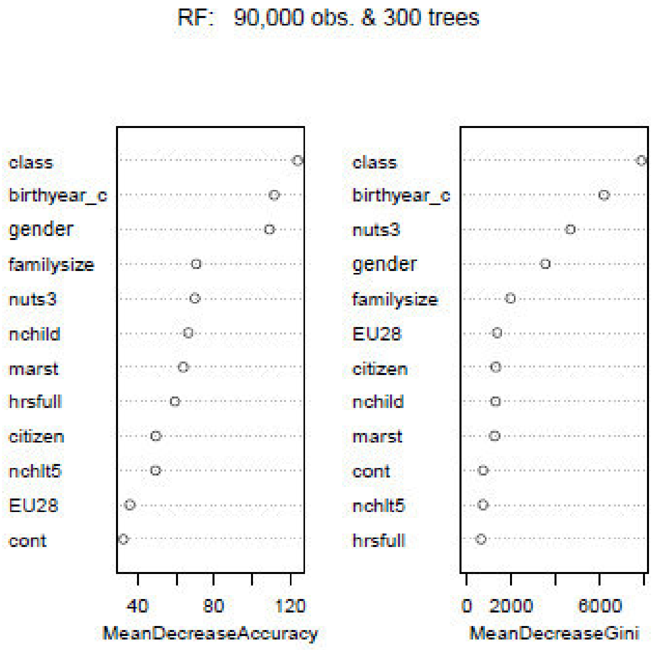


Figure 4.4: RF – Census 2001: most imp. variables

Association rules

With 18 explanatory variables (see Table 4.2), more than 800,000 observations and a minimum support of 0.1% and a confidence of 90% 24,633 rules are found. As a lot of these rules are, however, redundant rules which means that they do not provide further information these redundant rules are deleted so that in the end 698 non redundant rules that may explain educational attainment were left.

Figure 4.5 indicates that some of these 698 rules have a confidence of even 100%, a lot of rules have a confidence around 96% and between 90 and 92%.

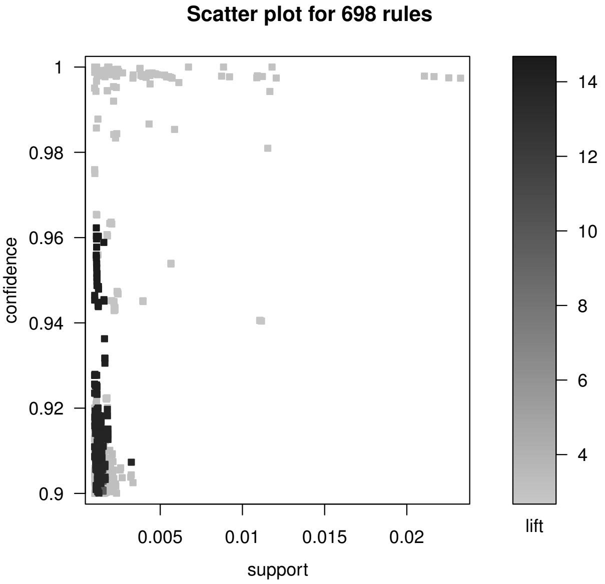


Figure 4.5: Arules – Census 2001: supp. and conf.

The support of nearly all rules lies below 1%.

Figure 4.6 shows the rules in detail. The size of the circles in the figure represents the support, the colour the lift of the rules. "LHS" stands for "Left hand side", which is the antecedent, "RHS" for "Right hand side", the consequent, which is educational attainment. The figure indicates that with these rules levels 1, 2, 4 and 5 may be explained. "School without diploma" can not be predicted with this set of input information. In addition, it is obvious that especially the rules that explain "Compulsory school" have a high support but a low lift.

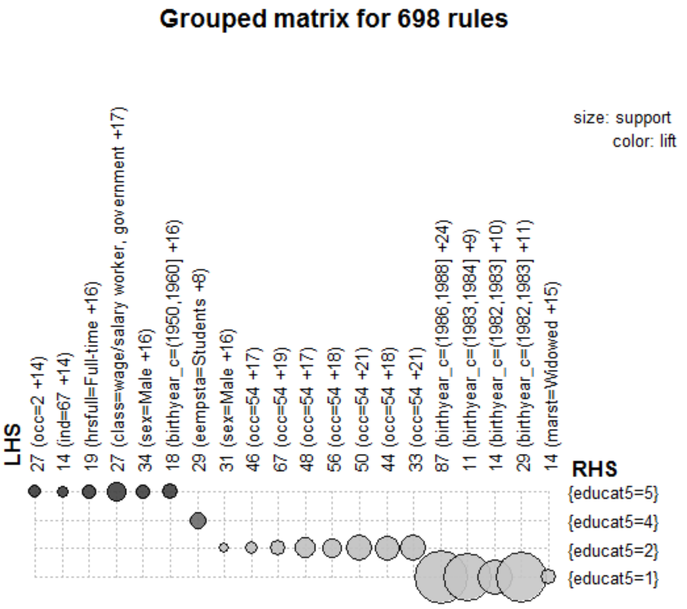


Figure 4.6: Arules – Census 2001: grouped matrix

4.3.2 NRN data 2001

In this subsection association rules and Random Forests will be built with the NRN data of the year 2001.

Random Forests

The example Random Forest of the NRN data which is presented in this subsection was built with 100,000 observations, 111 explanatory variables and 300 trees. Also in this case, due to computation time, the subset was drawn randomly from all observations. The list of the explanatory variables may be found in the Appendix of the thesis.

Figure 4.7 shows that the average oob error is about 23%. Again, level 1 is, with an error below 1% quite easy to predict, levels 5 and 3 are the most difficult ones to impute.

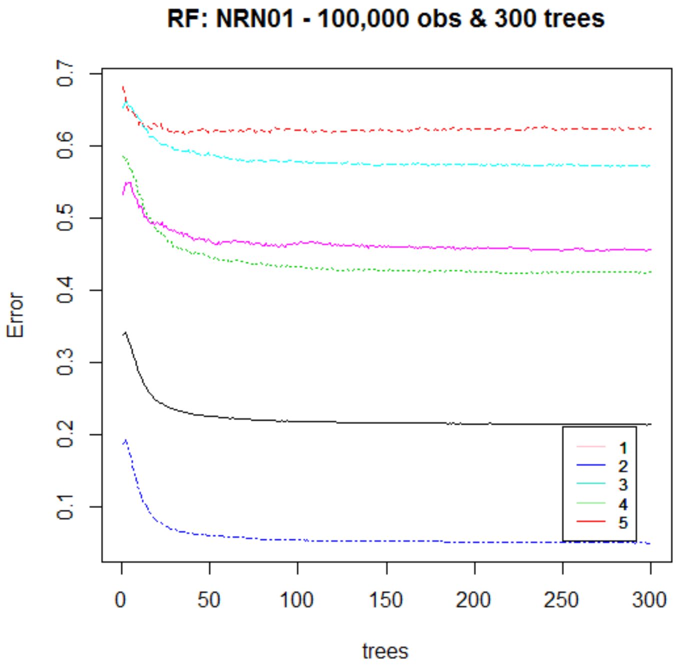


Figure 4.7: RF – NRN 2001: error rate

A look at the most important variables indicates that the age of the entry into the workforce (ej_age_c), the number of workingdays at the age of 20, 25, 30 and 40 (ev_arbeitstage_20/25/30/40), the difference of the dailywage between the age of 26 and the entry into workforce (ev_dif26) and the dailywage at the age of 20 and 25 (ev_dwage_20/25) are the most important explanatory variables (see Figure 4.8).

Association rules

Figure 4.9 shows again a plot of the association rules which were found with 45 variables (in this case the most important variables explaining educational attainment according to the Mean Decrease Accurancy and the Mean Decrease Gini measure in Figure 4.8 were taken), 1 million observations, a minimum support of 1% and a confidence of 90%. With this set of input information 1,775 rules could be found. After the removal of the redundant

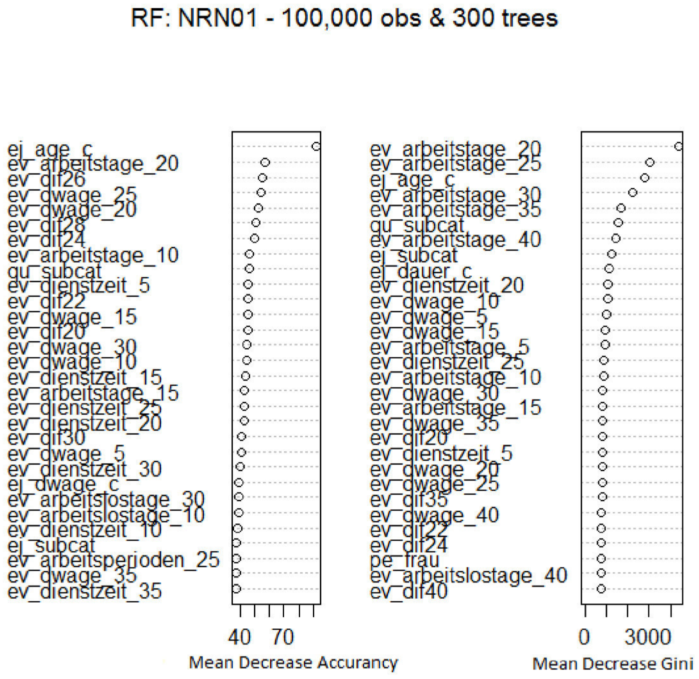


Figure 4.8: RF – NRN 2001: most imp. variables

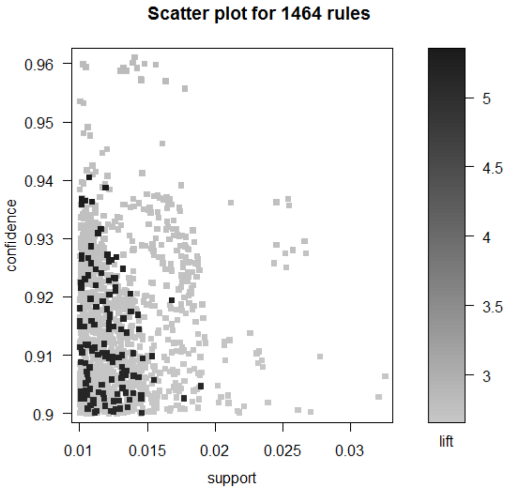


Figure 4.9: Arules – NRN 2001: supp. and conf.

rules 1,464 non redundant rules are left.

A more detailed look at these rules (see Figure 4.10) shows, however, that with this input information only the levels 2 and 5 may be predicted. The other levels can not be imputed with this minimum support and confidence. The figure also indicates that the support of the rules varies a lot.

4.3.3 Summary

For the imputation of educational attainment a lot of different versions of association rules and Random Forests were tried and the results of 2001 were presented in the previous subsections of the thesis.

To sum up all the results up to now, the advantage of the association rules is that they may (if the confidence level is set high) find relationships in the data which have a high probability. The disadvantage is, however, that educational attainment may not be predicted for all Austrians and in a lot of cases some levels of education are not predicted at all. Therefore, only partial imputation would be possible with association rules.

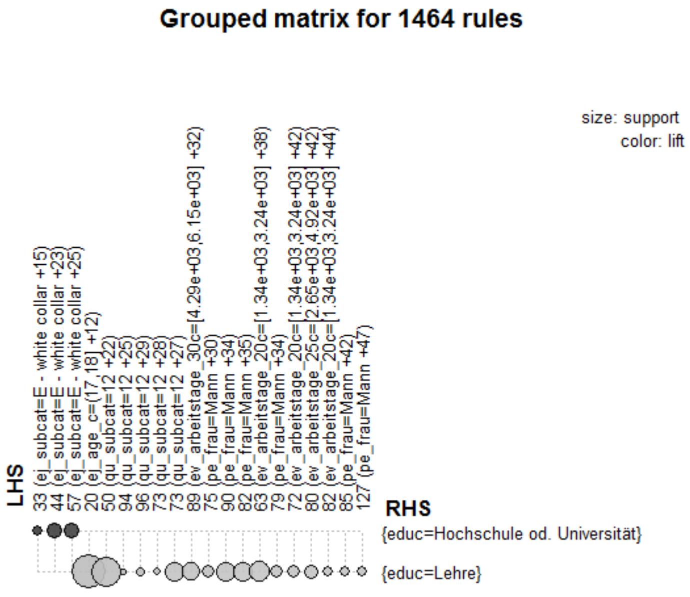


Figure 4.10: Arules – NRN 2001: grouped matrix

With Random Forests educational attainment may be on the other hand predicted for all observations. The average oob error rate is, however, in all cases quite high. The error in the forests which were constructed with the census data were even higher than 50%.

Taking all these information into account a final set up for the final prediction of educational attainment was developed.

4.4 Final imputation set-up

For the final imputation the following considerations were taken into account: The data which will be used will be the NRN data, as it contains much more explaining variables than the census data. In addition, as the Department of Economics in Linz wants to have educational attainment imputed for all Austrians, the final method for imputation will be Random Forests.

Moreover, as there is not much information available for the Austrians born before 1930, the sample for imputation will be restricted to those born between 1930 and 1990.

In addition, in some cases (387,724) there is no information about the working history, the number of children, the qualification, etc. available, but only information about the gender, the birthyear and if the person is a foreigner or not. As this small amount of information can not predict well educational attainment, it will be apriori explained with the distribution published by the Mirco-Census of Statistik Austria, which will be separated by the birth cohorts and gender. The detailed apriori information which was imputed for all cases without any reliable explanatory attributes may be found in the Appendix.

As educational attainment differs between men and women and also between birth cohorts, not only one, but several Random Forests will be grown. In detail, there will be fourteen different forests (7 different birth cohorts – separately for men and women). This approach can be also interpreted as fixed splits at the top of each tree. The first fixed split is the gender, the next fixed split the birthyear.

Moreover, the Random Forests up to now showed that a distinction between levels 3 and 4 (School with and School without diploma) is quite difficult and these levels are therefore hard to predict. For this reason, these two levels will be combined, so that in the end educational attainment will consist of only 4 levels.

- NRN data: a large number of explanatory variables
- Random Forest: imputes educational attainment for all Austrians
- Sample: all people born between 1930 and 1990
- Men and women separately: due to differences in educational attainment
- Birth cohorts seperately: due to the change in educational attainment during time
- Combination of levels 3 & 4 (which were difficult to predict) ⇒ 4 final levels
- If no explanatory attributes: apriori imputation with data from Statistik Austria

Figure 4.11: Summary for the final imputation set-up

4.4.1 Final results

For the final imputation 14 Random Forests (RF), separated by men and women and birth cohorts, were grown. In total, 112 variables were used to explain educational attainment. Missing values in reference to these 112 variables were again replaced by "999".

4.4.1.1 Explanatory variables

Educational attainment is explained with 112 variables. Apart from the birthyear, the marital status and if the person was a foreigner or a member of the EU15 countries was added. Most of the variables concern the working history. Therefore, the job-category and the job-subcategory are included. Also the age at the first job and the duration of the first job are added, as these variables contain important information concerning educational attainment. In addition, the number of marginal or part-time working days, as well as the information, if a person has worked marginal or part-time at a certain age is included. The number of days in unemployment and the number of total working days at a certain age is also considered in the model. Moreover, the information if the person has already died, the ÖNACE classification of the firm, the total number of children, as well as the age when the woman gave birth to her first, second, third, fourth and fifth child is considered. The list of all the 112 variables can be found in the Appendix.

As the imputed variable "educ" will be used for further economic research and as it might be used to explain income with educational attainment, all income related variables will not be used in the forest. A comparison with Random Forests where the income information was included, only showed slightly better results (about 0.5 points lower) regarding the oob errors.

The oob errors differ from forest to forest. In general, it can be said that the imputation works better for younger Austrians.

4.4.1.2 Out-of-bag errors

For all of these forests the oob errors can be estimated and the most important variables may be plotted. For these results “1” represents “Compulsory school”, “2” “School with or without diploma”, “3” “Apprenticeship training” and “4” “College or university”.

Table 4.10 displays all the oob errors for the birth cohorts of all women. It shows the average oob error per cohort, as well as the single class oob errors. As this table indicates, the oob error varies a lot. The average oob error decreases with time from 43.82% for those born between 1940 and 1950 to 8.33% for those born between 1990 and 2000. A look at the detailed class errors shows that this error varies as well. Class 3 (School with or without diploma) can be for instance well predicted for women born between 1970 and 1989, class 4 (University) for those born after 1970.

Women	1930-39	1940-49	1950-59	1960-69	1970-79	1980-89	1990-99
Average	26.59%	43.82%	33.59%	28.17%	28.00%	17.56%	8.33%
1	1.80%	15.01%	25.99%	34.98%	49.46%	43.22%	39.62%
2	96.11%	71.09%	34.08%	22.19%	34.29%	35.00%	1.60%
3	100.00%	89.64%	43.14%	29.34%	21.79%	9.84%	46.27%
4	23.78%	37.75%	31.71%	33.28%	22.67%	19.43%	17.57%

Table 4.10: RF – women: out-of-bag error

Figure 4.12 shows the oob error for the women regarding the birth cohorts graphically. The oob error is on the x-axis, the birth decades may be found on the y-axis. The different colours show the classes of educational attainment. The blue circles represent “Compulsory school”, the pink one “Apprenticeship training”, the green one “School with or without diploma” and the orange one “College or university”.

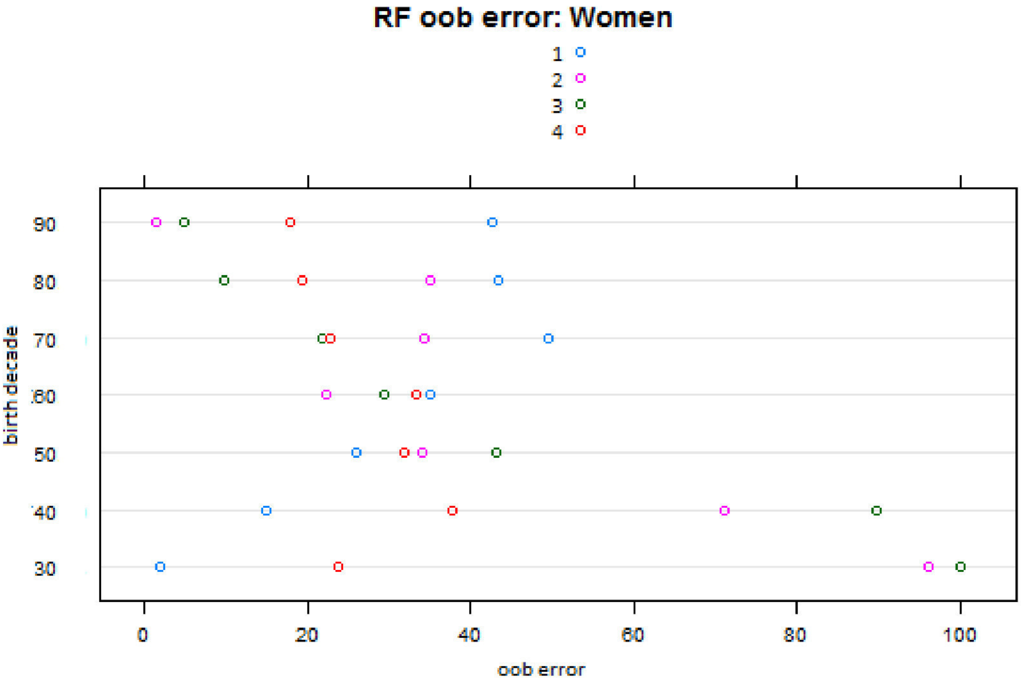


Figure 4.12: RF – women: out-of-bag error

Table 4.11 indicates the oob errors of the separate birth cohorts for all men. Also here the oob errors vary remarkable. The average oob error goes from 44.88% for men born between 1940 and 1949 to 4.94% for men born after 1990 and before 2001. A look at the detailed class shows that in this case class 2 (Apprenticeship) may be predicted quite well for all men born after 1940. Also class 4 (University) may be explained, with an oob error always below 27%, well. The estimated oob error of Class 1 (Compulsory school) lies on the other hand always over 22%. In addition, the oob errors also vary between the birth cohorts. So for men born between 1930 and 1939 it is easy to explain levels 1 and 4, but hard to predict levels 2 and 3.

Men	1930-39	1940-49	1950-59	1960-69	1970-79	1980-89	1990-99
Average	30.42%	41.88%	28.94%	23.46%	24.09%	14.75%	4.94%
1	22.76%	49.59%	46.44%	46.94%	50.61%	44.74%	45.36%
2	56.06%	28.60%	13.23%	9.30%	23.24%	25.67%	0.60%
3	99.84%	82.99%	56.55%	48.67%	22.77%	6.90%	49.83%
4	15.16%	26.18%	26.60%	26.99%	19.52%	23.48%	15.99%

Table 4.11: RF – men: out-of-bag error

Figure 4.13 shows again the oob for men. Also here the birth cohorts can be found on the y-axis and the classes of educational attainment are represented by coloured circles. Figure 4.13 shows well, that the oob error for the first class are constant over all birth cohorts. The oob error of the third class (“School with or without diploma”) varies on the other side a lot.

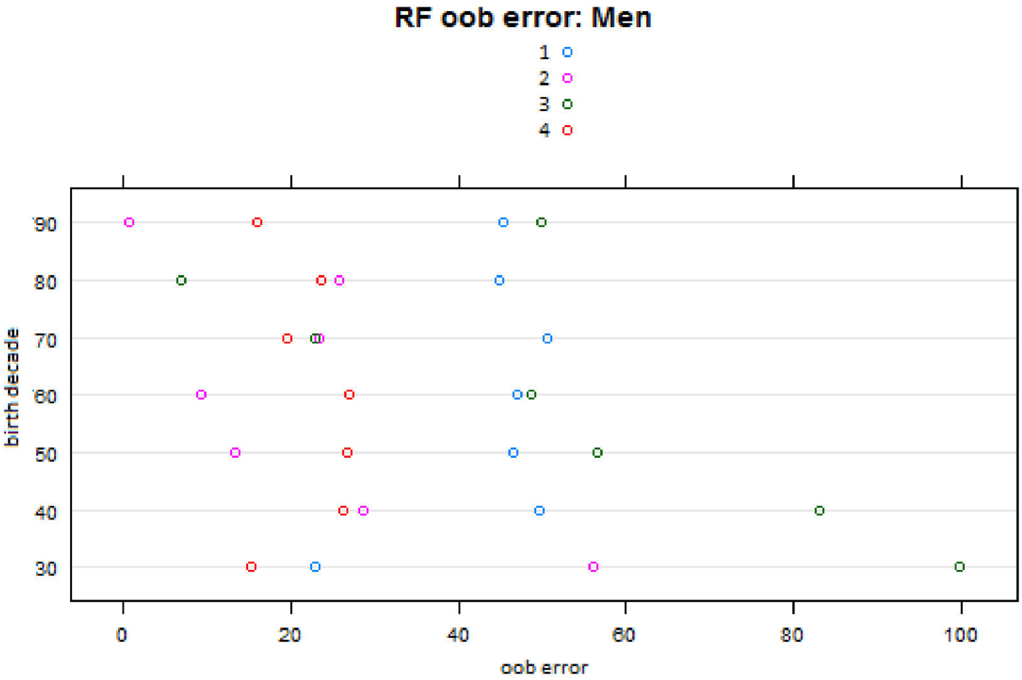


Figure 4.13: RF – men: out-of-bag error

4.4.1.3 Imputation

After the creation of the forests, educational attainment may be predicted with the help of these 14 Random Forests for all persons where information about the highest education completed was missing.

Table 4.12 shows an analysis of educational attainment for all men after the imputation. It contains the men whose educational attainment was known before and also those whose education has been imputed. For 387,724 men no information concerning working history, age at first job, number of children, etc. was available. These men are thus categorized with “NA”. For these people apriori information with data from Statistik Austria was imputed.
An analysis of educational attainment of those whose education is now known shows that most of the men have completed either an apprenticeship training (29.36%) or a university (30.21%).

Education Men	Freq.	Percent	Cum.
1	579,943	12.06	12.06
2	1,411,750	29.36	41.42
3	976,217	20.30	61.72
4	1,452,678	30.21	91.94
NA	387,724	8.06	100.00
Total	4,808,312	100.00	

Table 4.12: Educational attainment of all men

The same analysis for women (see Table 4.13) shows that educational attainment of most of the women is also university (27.85%). 22.44% have completed a school with or without diploma and for about 12.20% there is no information available.

Education Women	Freq.	Percent	Cum.
1	791,619	18.57	18.57
2	807,765	18.95	37.52
3	956,533	22.44	59.95
4	1,187,173	27.85	87.80
NA	520,086	12.20	100.00
Total	4,263,176	100.00	

Table 4.13: Educational attainment of all women

If we take a look at only those whose education was imputed, we see that for men (see Table 4.14) the Random Forest especially predicted the class "University". "Compulsory school" was imputed in very few cases.

Only imputed men	Freq.	Percent	Cum.
1	215,861	9.99	9.99
2	321,085	14.86	24.85
3	231,496	10.71	35.56
4	1,031,056	47.71	83.27
NA	361,435	16.73	100.00
Total	2,160,933	100.00	

Table 4.14: Imputation: men

An analysis of the imputation for women shows that the majority of the predictions was the class "Apprenticeship training", followed by "School with or without diploma" (see Table 4.15).

Apart from the most likely class, Random Forests also predict the probability for all classes, which may be also used for imputation.
Table 4.16 shows an example output of the predicted class probabilities.

Only impued women	Freq.	Percent	Cum.
1	364,082	13.75	13.75
2	1,090,665	41.20	54.95
3	744,721	28.13	83.08
4	421,622	15.93	99.01
NA	26,289	0.99	100.00
Total	2,647,379	100.00	

Table 4.15: Imputation: women

penr	class1pr	class2pr	class3pr	class4pr
1	0.018	0.000	0.000	0.982
2	0.414	0.480	0.076	0.030
3	0.334	0.068	0.036	0.562
4	0.232	0.228	0.176	0.364
5	0.516	0.394	0.074	0.016

Table 4.16: Example output of the imputation

Figure 4.14 shows a histogram of the most class probability for women, Figure 4.15 shows the histogram for men. The two figures indicate that for women, as well as for men, the highest class probability is in a lot of cases around 0.55. The highest class probability of nearly no person lies below 0.3.

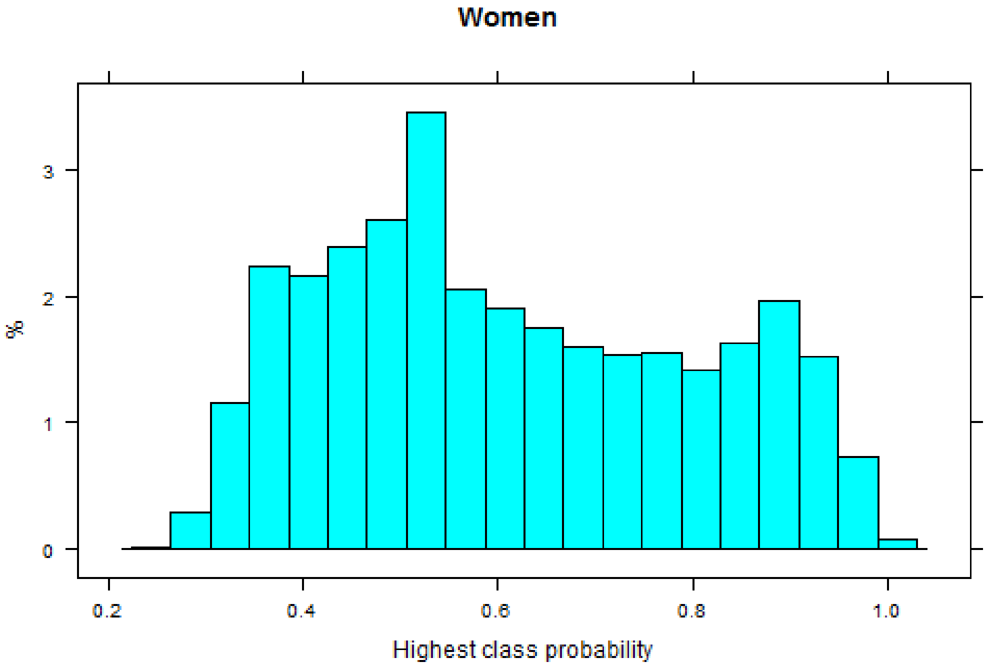


Figure 4.14: Highest class probability: women

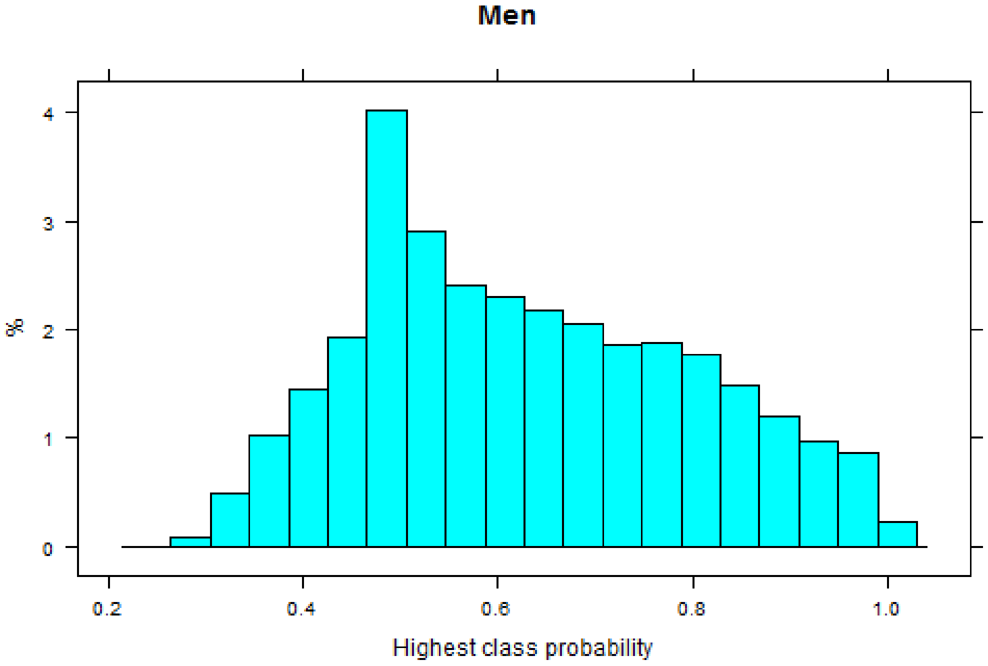


Figure 4.15: Highest class probability: men

4.4.1.4 Comparison with data from Statistik Austria

In order to check the imputation the imputed variable "educ", which contains the persons whose information was known and the persons whose educational attainment was predicted, will be compared to data from Statistik Austria.

As the groups that will be compared have to be based on the same sample composition, the comparison will be restricted to all Austrians at the age 35 to 44 in 2014. Tables 4.17 and 4.18 indicate a comparison with data from the Micro-Census of 2014, published by Statistik Austria. It has to be, however, taken into account that the Micro-Census is based on surveys.

Men	**Imputed Variable**	**Micro-Census**
Compulsory school	4.8%	11.2%
Apprenticeship	52.4%	48.7%
School with or without diploma	22.2%	22.8%
University	20.6%	17.3%

Table 4.17: Comparison imputation with Micro-Census: men

Women	**Imputed Variable**	**Micro-Census**
Compulsory school	10.3%	15.1%
Apprenticeship	35.5%	29.5%
School with or without diploma	33.5%	34.6%
University	20.7%	20.8%

Table 4.18: Comparison imputation with Micro-Census: women

Table 4.17 shows that after the imputation for men level 1 "Compulsory school" is slightly underpredicted and level 4 "University" slightly overestimated.

Regarding women level 1 "Compulsory school" is also slightly underestimated and level 2 "Apprenticeship training" overpredicted, all the other levels correspond, however, quite well to the data published by Statistik Austria.

However, as Random Forests do not only predict the most likely class, but all class probabilities it is also possible to perform multiple imputation and to use this supplementary information when using the imputed variable “educ” for further research.

A comparison of the marginal distribution based on the probabilities with data from Statistik Austria showed, however, similar results than the comparison with the highest likely class. Also here, level 4 “University” was overpredicted and level 1 “Compulsory school” was slightly underpredicted.

Chapter 5

Conclusion

The aim of this Master thesis was the imputation of the variable educational attainment, which is, although it is an important variable for economic research, rarely recorded. Several institutions collect information about the highest level of education of the Austrians, among others for instance the Austrian Social Security Institutions or the Public Employment Service Austria, which is the most reliable source regarding educational attainment. The datasets that were used for the prediction were the NRN data and the Austrian census data of 2001.

For the imputation of a variable a variety of methods exist, but two specific methods, association rules and random forests, were used and different versions (with different data, number of trees, support or confidence, etc.) of these two methods were tried. With association rules educational attainment could be explained for some Austrians with a high probability. It was, however, impossible to predict all levels of educational attainment for all Austrians. Random forests, on the other hand, had quite a high oob error.

The final imputation was then carried out with random forests and the NRN data, separately grown for men and women and for birth cohorts. In addition, levels 3 and 4 (School without and School with diploma) were combined, as these two levels were difficult to distinguish.

The oob error of these forests varied remarkably. Although it was quite high for people born before 1940, it was low for the younger Austrians.

The Random Forests predicted over all level “4” (“College or university”) for men and level “2” (“Apprenticeship training”) for women.
A comparison with data from Statistik Austria showed that for men especially class “4” was slightly overestimated and level “1” (“Compulsory school ”) underestimated. Regarding women level 1 was also underpredicted and level “2” (“Apprenticeship training”) a little bit overestimated.

However, as random forests do not only predict the most likely class, but also class probabilities it might be better to perform multiple imputation and to use this supplementary information when using the imputed variable “educ” for further research.

References

R packages

arules: Hahsler, M., Buchta, C., Grün, B., Hornik, K. & Borgelt, C. (2015). Mining Association Rules and Frequent Itemsets. http://CRAN.R-project.org/package=arules (Download: 22.3.2015)

arulesViz: Hahsler, M. & Chelluboina, S. (2015). Visualizing Association Rules and Frequent Itemsets. http://CRAN.R-project.org/package=arulesViz (Download: 22.3.2015)

randomForest: Liaw, A. & Wiener, M. (2002). Classification and Regression by randomForest. R News 2(3), 18–22. http://CRAN.R-project.org/package=randomForest (Download: 10.6.2015)

Literature

Agrawal, R., Imielinski, T. & Swami, A. (1993). Mining association rules between sets of items in large databases. *SIGMOD '93 Proceedings of the 1993 ACM SIGMOD International Conference on Management of data*, p. 207-216, doi: 10.1145/170035.170072.

Barnard, J. & Meng, X. (1999). Applications of multiple imputation in medical studies: From AIDS to NHANES, *Statistical Methods in Medical Research*, 1999 (8), 17-36, doi: 0962-2802(99)SM173RA .

Borgelt, C. (2003). Efficient Implementations of Apriori and Eclat. In *FIMI'03: Proceedings of the IEEE ICDM Workshop on Frequent Itemset Mining Implementations.*.

Borgelt, C. (2004). Finding Association Rules/Hyperedges with the Apriori Algorithm. Working Group Neural Networks and Fuzzy Systems, Otto-von-Guericke-University of Magdeburg, Universitätsplatz 2, D-39106 Magdeburg, Germany. `http://fuzzy.cs.uni-magdeburg.de/~borgelt/apriori.html`.

Breiman, L. (2001). Random Forest, *Machine Learning*, October 2001, Vol. 45, Issue 1, pp. 5-32.

Breiman, L. & Cutler, A. (2015). Random Forests. `https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm` (Download: 20.10.2015).

Göthlich, S. (2009). Zum Umgang mit fehlenden Daten in großzahligen empirischen Erhebungen, In: Albers et al. (eds.) *Methodik der empirischen Forschung*, Kapitel 2, pp. 119-135, Gabler Verlag.

Hashler, M., Grün, B., Hornik, K. & Buchta, C.(2015). arules - A Computational Environment for Mining Association Rules and Frequent Item Sets. Journal of Statistical Software 14/15. URL: `http://www.jstatsoft.org/v14/i15/` (Download: 10.10.2015).

Hahsler, M. & Chelluboina, S. (2015). Visualizing Association Rules: Introduction to the R-extension Package arulesViz `https://cran.r-project.org/web/packages/arulesViz/vignettes/arulesViz.pdf` (Download: 15.5.2015).

Hastie, T., Tibshirani, R. & Friedman, J. (2011). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Second Edition, Springer.

IBW (2015). Institut für Bildungsforschung der Wirtschaft. The Austrian Education System. `http://www.ibw.at/images/ibw/bbs/bbs_en11.pdf` (Download:

12.4.2015).

IPUMS (2015). Integrated Public Use Microdata Series, International. Minnesota Population Center/University of Minnesota. `https://international.ipums.org/international/index.shtml` (Download: 20.4.2015).

ISCED (2011). International Standard Classification of Education 2011. `http://www.uis.unesco.org/Education/Documents/isced-2011-en.pdf` (Download: 3.3.2015).

James, G., Witten, D., Hastie T., & Tibshirani R. (2013). An Introduction to Statistical Learning with Applications in R, Springer.

Kaiser, L. (2014). Dealing with Missing Values in Data, *Journal of Systems Integration*, 2014(1), pp. 42-51 .

Little, R. & Rubin D. (2002). Statistical Analysis with Missing Data, Second Edition, Hoboken (New Jersey): John Wiley & Sons.

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: `http://www.R-project.org/` (Download: 10.6.2015).

Rubin, D. (1976). Inference and Missing Data (with discussion), *Biometrika 63*, pp. 581-592.

Rokach, L. & Maimon, O. (2008). Data Mining with Decision Trees – Theory and Applications, *Series in Machine Perception and Artifical Intelligence*, Vol. 69, World Scientific Publishing Co

Statistik Austria (2005). Standard-Dokumentation Metainformationen (Definitionen, Erläuterungen, Methoden, Qualität) zur Volkszählung 2001 (Bearbeitungsstand: 15.11.2005) (Download: 20.3.2015).

Statistik Austria (2012a). Arbeitskräfteerhebung – Ergebnisse des Mikrozensus 2011 (Download: 20.3.2015).

Statistik Austria (2012b). Bildung in Zahlen 2012/13 –

Schlüsselindikatoren und Analysen (Download: 25.3.2015)

Statistik Austria (2014a). Standard-Dokumentation Metainformationen (Definitionen, Erläuterungen, Methoden, Qualität) zu Mikrozensus ab 2004: Arbeitskräfte- und Wohnungserhebung (Bearbeitungsstand: 01.08.2014) (Download: 10.3.2015)

Statistik Austria (2014b). Registerzählung. http://www.statistik.at/web_de/statistiken/bevoelkerung/volkszaehlungen_registerzaehlungen_abgestimmte_erwerbsstatistik/index.html (Download: 10.3.2015)

Tan, P., Steinbach, M. & Kumar, V. (2005). Introduction to Data Mining, Pearson.

NRN Labor&Welfare State. (2015). The Austrian Center for Labor Economics and the Analysis of the Welfare State. http://www.labornrn.at/

OECD (2015). OECD iLibrary. OECD Factbook 2013: Economic, Environmental and Social Statistics. Education. Outcomes. Educational attainment. http://www.oecd-ilibrary.org/sites/factbook-2013-en/10/01/06/index.html;jsessionid=117bse0t2ksgs.x-oecd-live-02?contentType=&itemId=/content/chapter/factbook-2013-77-en&_csp_=491cec5152b0cb4856cfa94aec85f33b (Download: 20.3.2015)

Zweimüller, J., Winter-Ebmer, R., Lalive, R. Kuhn, A., Wuellrich J.-P., Ruf, O. & Büchi, S. (2009). Austrian Social Security Database, *NRN Working Paper No. 0903 April*

Appendix A

Tables

Explanatory variables: NRN data (4.3.2)

Variable	Description
penr	ID Number
educ	Educational attainment (1= Compulsory School, 2=Apprenticeship, 3= School with or without diploma, 4=University
year	Year of record
birthyear	Birthyear
pe_ausland	Was born in a foreign country (yes =1, no=0)
pe_familienstand	marital status (3 classes)
pe_familienbeihilfe	familiy allowance
qu_cat	Job-category (2 classes)
qu_subcat	Jobsubcategory (5 classes)
ej_yes	Is there a first job? (yes=1, no=0)
ej_dauer	Duration of first job
ej_age	Age at first job
gf_yes	Has the person ever been marginal employed?
gf_15y	Has person be marginal employed at the age of 15
gf_16y	Has person be marginal employed at the age of 16
gf_17y	Has person be marginal employed at the age of 17

gf_18y	Has person be marginal employed at the age of 18
gf_19y	Has person be marginal employed at the age of 19
gf_20y	Has person be marginal employed at the age of 20
gf_21y	Has person be marginal employed at the age of 21
gf_22y	Has person be marginal employed at the age of 22
gf_23y	Has person be marginal employed at the age of 23
gf_24y	Has person be marginal employed at the age of 24
gf_25y	Has person be marginal employed at the age of 25
gf_26y	Has person be marginal employed at the age of 26
gf_27y	Has person be marginal employed at the age of 27
gf_28y	Has person be marginal employed at the age of 28
gf_15	Number of marginal working days at the age of 15
gf_16	Number of marginal working days at the age of 16
gf_17	Number of marginal working days at the age of 17
gf_18	Number of marginal working days at the age of 18
gf_19	Number of marginal working days at the age of 19
gf_20	Number of marginal working days at the age of 20
gf_21	Number of marginal working days at the age of 21
gf_22	Number of marginal working days at the age of 22
gf_23	Number of marginal working days at the age of 23
gf_24	Number of marginal working days at the age of 24
gf_25	Number of marginal working days at the age of 25
gf_26	Number of marginal working days at the age of 26
gf_27	Number of marginal working days at the age of 27
gf_28	Number of marginal working days at the age of 28
tz_yes	Has person ever worked part-time
sj_yes	Has person ever worked during summer month?
sj_15y	Has person worked during summer month at the age of 15?
sj_16y	Has person worked during summer month at the age of 16?
sj_17y	Has person worked during summer month at the age of 17?

sj_18y	Has person worked during summer month at the age of 18?
sj_19y	Has person worked during summer month at the age of 19?
sj_20y	Has person worked during summer month at the age of 20?
sj_21y	Has person worked during summer month at the age of 21?
sj_22y	Has person worked during summer month at the age of 22?
sj_23y	Has person worked during summer month at the age of 23?
sj_24y	Has person worked during summer month at the age of 24?
sj_25y	Has person worked during summer month at the age of 25?
sj_26y	Has person worked during summer month at the age of 26?
sj_27y	Has person worked during summer month at the age of 16?
sj_28y	Has person worked during summer month at the age of 28?
sj_15	Number of working days during summer month at the age of 15
sj_16	Number of working days during summer month at the age of 16
sj_17	Number of working days during summer month at the age of 17
sj_18	Number of working days during summer month at the age of 18
sj_19	Number of working days during summer month at the age of 19

sj_20	Number of working days during summer month at the age of 20
sj_21	Number of working days during summer month at the age of 21
sj_22	Number of working days during summer month at the age of 22
sj_23	Number of working days during summer month at the age of 23
sj_24	Number of working days during summer month at the age of 24
sj_25	Number of working days during summer month at the age of 25
sj_26	Number of working days during summer month at the age of 26
sj_27	Number of working days during summer month at the age of 27
sj_28	Number of working days during summer month at the age of 28
ki_anzahl	Total number of children
ki_age_1	Age when gave birth to 1. child
ki_age_2	Age when gave birth to 2. child
ki_age_3	Age when gave birth to 3. child
ki_age_4	Age when gave birth to 4. child
ki_age_5	Age when gave birth to 5. child
ev_arbeitstage_5	Number of working days 5 years after entry into workforce
ev_arbeitstage_10	Number of working days 10 years after entry into workforce
ev_arbeitstage_15	Number of working days 15 years after entry into workforce
ev_arbeitstage_20	Number of working days at the age of 20
ev_arbeitstage_25	Number of working days at the age of 25
ev_arbeitstage_30	Number of working days at the age of 30

ev_arbeitstage_35	Number of working days at the age of 35
ev_arbeitstage_40	Number of working days at the age of 40
ev_arbeitslostage_5	Number of days in unemployment 5 years after entry into workforce
ev_arbeitslostage_10	Number of days in unemployment 10 years after entry into workforce
ev_arbeitslostage_15	Number of days in unemployment 15 years after entry into workforce
ev_arbeitslostage_20	Number of days in unemployment at the age of 20
ev_arbeitslostage_25	Number of days in unemployment at the age of 25
ev_arbeitslostage_30	Number of days in unemployment at the age of 30
ev_arbeitslostage_35	Number of days in unemployment at the age of 35
ev_arbeitslostage_40	Number of days in unemployment at the age of 40
ev_arbeitsperioden_5	Number of different employers 5 years after entry into workforce
ev_arbeitsperioden_10	Number of different employers 10 years after entry into workforce
ev_arbeitsperioden_15	Number of different employers 15 years after entry into workforce
ev_arbeitsperioden_20	Number of different employers at the age of 20
ev_arbeitsperioden_25	Number of different employers at the age of 25
ev_arbeitsperioden_30	Number of different employers at the age of 30
ev_arbeitsperioden_35	Number of different employers at the age of 35
ev_arbeitsperioden_40	Number of different employers at the age of 40
ev_dienstzeit_5	Number of working days at current employer 5 years after entry into workforce
ev_dienstzeit_10	Number of working days at current employer 10 years after entry into workforce
ev_dienstzeit_15	Number of working days at current employer 15 years after entry into workforce
ev_dienstzeit_20	Number of working days at current employer at the age of 20

ev_dienstzeit_25	Number of working days at current employer at the age of 25
ev_dienstzeit_30	Number of working days at current employer at the age of 30
ev_dienstzeit_35	Number of working days at current employer at the age of 35
ev_dienstzeit_40	Number of working days at current employer at the age of 40
pe_gest	has person died?
EU15	Is person member of EU15?
ek_yes	Is there a large change in income over time
nace	Önace Classification
ej_dwage	Dailywage of first job
sj_15_dwage	Dailywage of summer job at the age of 15
sj_16_dwage	Dailywage of summer job at the age of 16
sj_17_dwage	Dailywage of summer job at the age of 17
sj_18_dwage	Dailywage of summer job at the age of 18
sj_19_dwage	Dailywage of summer job at the age of 19
sj_20_dwage	Dailywage of summer job at the age of 20
sj_21_dwage	Dailywage of summer job at the age of 21
sj_22_dwage	Dailywage of summer job at the age of 22
sj_23_dwage	Dailywage of summer job at the age of 23
sj_24_dwage	Dailywage of summer job at the age of 24
sj_25_dwage	Dailywage of summer job at the age of 25
sj_26_dwage	Dailywage of summer job at the age of 26
sj_37_dwage	Dailywage of summer job at the age of 27
sj_28_dwage	Dailywage of summer job at the age of 28
ev_dwage_5	Dailywage 5 years after entry into workforce
ev_dwage_10	Dailywage 10 years after entry into workforce
ev_dwage_15	Dailywage 15 years after entry into workforce
ev_dwage_20	Dailywage at the age of 20
ev_dwage_25	Dailywage at the age of 25
ev_dwage_30	Dailywage at the age of 30

ev_dwage_35	Dailywage at the age of 35
ev_dwage_40	Dailywage at the age of 40
ev_dif20	Difference in dailwage at the age of 20 and first job
ev_dif22	Difference in dailwage at the age of 22 and first job
ev_dif22	Difference in dailwage at the age of 22 and first job
ev_dif24	Difference in dailwage at the age of 24 and first job
ev_dif26	Difference in dailwage at the age of 26 and first job
ev_dif28	Difference in dailwage at the age of 28 and first job
ev_dif30	Difference in dailwage at the age of 30 and first job
ev_dif35	Difference in dailwage at the age of 35 and first job
ev_dif40	Difference in dailwage at the age of 40 and first job

Explanatory variables final random forest (4.4.1)

Variable	Description
penr	ID Number
educ	Educational attainment (1= Compulsory School, 2=Apprenticeship, 3= School with or without diploma, 4=University
year	Year of record
birthyear	Birthyear
pe_ausland	Was born in a foreign country (yes =1, no=0)
pe_familienstand	marital status (3 classes)
pe_familienbeihilfe	familiy allowance
qu_cat	Job-category (2 classes)
qu_subcat	Jobsubcategory (5 classes)
ej_yes	Is there a first job? (yes=1, no=0)
ej_dauer	Duration of first job
ej_age	Age at first job
gf_yes	Has the person ever been marginal employed?

gf_15y	Has person be marginal employed at the age of 15
gf_16y	Has person be marginal employed at the age of 16
gf_17y	Has person be marginal employed at the age of 17
gf_18y	Has person be marginal employed at the age of 18
gf_19y	Has person be marginal employed at the age of 19
gf_20y	Has person be marginal employed at the age of 20
gf_21y	Has person be marginal employed at the age of 21
gf_22y	Has person be marginal employed at the age of 22
gf_23y	Has person be marginal employed at the age of 23
gf_24y	Has person be marginal employed at the age of 24
gf_25y	Has person be marginal employed at the age of 25
gf_26y	Has person be marginal employed at the age of 26
gf_27y	Has person be marginal employed at the age of 27
gf_28y	Has person be marginal employed at the age of 28
gf_15	Number of marginal working days at the age of 15
gf_16	Number of marginal working days at the age of 16
gf_17	Number of marginal working days at the age of 17
gf_18	Number of marginal working days at the age of 18
gf_19	Number of marginal working days at the age of 19
gf_20	Number of marginal working days at the age of 20
gf_21	Number of marginal working days at the age of 21
gf_22	Number of marginal working days at the age of 22
gf_23	Number of marginal working days at the age of 23
gf_24	Number of marginal working days at the age of 24
gf_25	Number of marginal working days at the age of 25
gf_26	Number of marginal working days at the age of 26
gf_27	Number of marginal working days at the age of 27
gf_28	Number of marginal working days at the age of 28
tz_yes	Has person ever worked part-time
sj_yes	Has person ever worked during summer month?
sj_15y	Has person worked during summer month at the age of 15?

sj_16y	Has person worked during summer month at the age of 16?
sj_17y	Has person worked during summer month at the age of 17?
sj_18y	Has person worked during summer month at the age of 18?
sj_19y	Has person worked during summer month at the age of 19?
sj_20y	Has person worked during summer month at the age of 20?
sj_21y	Has person worked during summer month at the age of 21?
sj_22y	Has person worked during summer month at the age of 22?
sj_23y	Has person worked during summer month at the age of 23?
sj_24y	Has person worked during summer month at the age of 24?
sj_25y	Has person worked during summer month at the age of 25?
sj_26y	Has person worked during summer month at the age of 26?
sj_27y	Has person worked during summer month at the age of 16?
sj_28y	Has person worked during summer month at the age of 28?
sj_15	Number of working days during summer month at the age of 15
sj_16	Number of working days during summer month at the age of 16
sj_17	Number of working days during summer month at the age of 17

sj_18	Number of working days during summer month at the age of 18
sj_19	Number of working days during summer month at the age of 19
sj_20	Number of working days during summer month at the age of 20
sj_21	Number of working days during summer month at the age of 21
sj_22	Number of working days during summer month at the age of 22
sj_23	Number of working days during summer month at the age of 23
sj_24	Number of working days during summer month at the age of 24
sj_25	Number of working days during summer month at the age of 25
sj_26	Number of working days during summer month at the age of 26
sj_27	Number of working days during summer month at the age of 27
sj_28	Number of working days during summer month at the age of 28
ki_anzahl	Total number of children
ki_age_1	Age when gave birth to 1. child
ki_age_2	Age when gave birth to 2. child
ki_age_3	Age when gave birth to 3. child
ki_age_4	Age when gave birth to 4. child
ki_age_5	Age when gave birth to 5. child
ev_arbeitstage_5	Number of working days 5 years after entry into workforce
ev_arbeitstage_10	Number of working days 10 years after entry into workforce

ev_arbeitstage_15	Number of working days 15 years after entry into workforce
ev_arbeitstage_20	Number of working days at the age of 20
ev_arbeitstage_25	Number of working days at the age of 25
ev_arbeitstage_30	Number of working days at the age of 30
ev_arbeitstage_35	Number of working days at the age of 35
ev_arbeitstage_40	Number of working days at the age of 40
ev_arbeitslostage_5	Number of days in unemployment 5 years after entry into workforce
ev_arbeitslostage_10	Number of days in unemployment 10 years after entry into workforce
ev_arbeitslostage_15	Number of days in unemployment 15 years after entry into workforce
ev_arbeitslostage_20	Number of days in unemployment at the age of 20
ev_arbeitslostage_25	Number of days in unemployment at the age of 25
ev_arbeitslostage_30	Number of days in unemployment at the age of 30
ev_arbeitslostage_35	Number of days in unemployment at the age of 35
ev_arbeitslostage_40	Number of days in unemployment at the age of 40
ev_arbeitsperioden_5	Number of different employers 5 years after entry into workforce
ev_arbeitsperioden_10	Number of different employers 10 years after entry into workforce
ev_arbeitsperioden_15	Number of different employers 15 years after entry into workforce
ev_arbeitsperioden_20	Number of different employers at the age of 20
ev_arbeitsperioden_25	Number of different employers at the age of 25
ev_arbeitsperioden_30	Number of different employers at the age of 30
ev_arbeitsperioden_35	Number of different employers at the age of 35
ev_arbeitsperioden_40	Number of different employers at the age of 40
ev_dienstzeit_5	Number of working days at current employer 5 years after entry into workforce
ev_dienstzeit_10	Number of working days at current employer 10 years after entry into workforce

ev_dienstzeit_15	Number of working days at current employer 15 years after entry into workforce
ev_dienstzeit_20	Number of working days at current employer at the age of 20
ev_dienstzeit_25	Number of working days at current employer at the age of 25
ev_dienstzeit_30	Number of working days at current employer at the age of 30
ev_dienstzeit_35	Number of working days at current employer at the age of 35
ev_dienstzeit_40	Number of working days at current employer at the age of 40
pe_gest	has person died?
EU15	Is person member of EU15?
ek_yes	Is there a large change in income over time
nace	Önace Classification

apriori imputation

Following Tables indicate the apriori information which was imputed when no explanatory covariates were available. The information was seperated between men and women and the birth cohorts.

Men	1930-39	1940-49	1950-59	1960-69	1970-79	1980-1990
1	25.20%	25.20%	16.90%	11.40%	10.60%	10.40%
2	17.30%	17.30%	19.40%	22.70%	23.30%	29.40%
3	47.70%	47.70%	52.20%	51.90%	50.30%	45.60%
4	9.80%	9.80%	11.50%	14.00%	15.80%	14.60%

Table A.3: Apriori information: men

Women	1930-39	1940-49	1950-59	1960-69	1970-79	1980-1990
1	51.80%	51.80%	33.60%	21.90%	15.50%	12.50%
2	19.90%	19.90%	29.80%	29.80%	32.30%	26.80%
3	24.40%	24.40%	27.00%	34.70%	34.90%	39.80%
4	3.90%	3.90%	9.60%	13.60%	17.30%	20.90%

Table A.4: Apriori information: women

Appendix B

Code

For the thesis the two statistical programs Stata (Version 10) and R (Version 3.1.2) were used.

For the descriptive analysis standard functions (Stata: *summarize, table,* etc. & R: *summary, table, barplot,* etc.) were applied.
The Random Forests were created with following R Code, which was grown seperately by the birth cohorts. The Code is an example for the birth decade 1980-1989.

```
#########################
#
#### 80
#load data

d8=read.dta("C:\\Users\\Christina\\Documents\\Christina\\
ImpDatensatz_comp80.dta")
str(d4)
summary(d4)

#only if educ is known
dc=subset(d8, d8$educBek==1)

str(dc)

#males
```

```
dcm=subset(dc, dc$pe_frau=="Mann")
str(dcm)

educ5=NA
educ5[dcm$educ=="keine Pflichtschule"] <- 1
educ5[dcm$educ=="Pflichtschule"] <- 1
educ5[dcm$educ=="Lehre"] <- 2
educ5[dcm$educ=="mittlere Schule (o. Matura)"] <- 3
educ5[dcm$educ=="höhere Schule (m. Matura)"] <- 3
educ5[dcm$educ=="Hochschule od. Universität"] <- 4

educ<-educ5
educ=as.factor(educ)

dcm$nace[is.na(dcm$nace)] <- 0

dcm$ev_arbeitstage_5[is.na(dcm$ev_arbeitstage_5)] <- 0
dcm$ev_arbeitstage_10[is.na(dcm$ev_arbeitstage_10)] <- 0
dcm$ev_arbeitstage_15[is.na(dcm$ev_arbeitstage_15)] <- 0
dcm$ev_arbeitstage_20[is.na(dcm$ev_arbeitstage_20)] <- 0
dcm$ev_arbeitstage_25[is.na(dcm$ev_arbeitstage_25)] <- 0
dcm$ev_arbeitstage_30[is.na(dcm$ev_arbeitstage_30)] <- 0
dcm$ev_arbeitstage_35[is.na(dcm$ev_arbeitstage_35)] <- 0
dcm$ev_arbeitstage_40[is.na(dcm$ev_arbeitstage_40)] <- 0

dcm$ev_arbeitslostage_5[is.na(dcm$ev_arbeitslostage_5)] <- 0
dcm$ev_arbeitslostage_10[is.na(dcm$ev_arbeitslostage_10)] <- 0
dcm$ev_arbeitslostage_15[is.na(dcm$ev_arbeitslostage_15)] <- 0
dcm$ev_arbeitslostage_20[is.na(dcm$ev_arbeitslostage_20)] <- 0
dcm$ev_arbeitslostage_25[is.na(dcm$ev_arbeitslostage_25)] <- 0
dcm$ev_arbeitslostage_30[is.na(dcm$ev_arbeitslostage_30)] <- 0
dcm$ev_arbeitslostage_35[is.na(dcm$ev_arbeitslostage_35)] <- 0
dcm$ev_arbeitslostage_40[is.na(dcm$ev_arbeitslostage_40)] <- 0

dcm$ev_arbeitsperioden_5[is.na(dcm$ev_arbeitsperioden_5)] <- 0
dcm$ev_arbeitsperioden_10[is.na(dcm$ev_arbeitsperioden_10)] <- 0
dcm$ev_arbeitsperioden_15[is.na(dcm$ev_arbeitsperioden_15)] <- 0
dcm$ev_arbeitsperioden_20[is.na(dcm$ev_arbeitsperioden_20)] <- 0
dcm$ev_arbeitsperioden_25[is.na(dcm$ev_arbeitsperioden_25)] <- 0
dcm$ev_arbeitsperioden_30[is.na(dcm$ev_arbeitsperioden_30)] <- 0
dcm$ev_arbeitsperioden_35[is.na(dcm$ev_arbeitsperioden_35)] <- 0
```

```
dcm$ev_arbeitsperioden_40[is.na(dcm$ev_arbeitsperioden_40)] <- 0

attach(dcm)

#EU15
EU15=rep(0,length(pe_nation))
EU15[pe_nation=="D"] <- 1
EU15[pe_nation=="B"] <- 1
EU15[pe_nation=="F"] <- 1
EU15[pe_nation=="I"] <- 1
EU15[pe_nation=="L"] <- 1
EU15[pe_nation=="NL"] <- 1
EU15[pe_nation=="DK"] <- 1
EU15[pe_nation=="IR"] <- 1
EU15[pe_nation=="GB"] <- 1
EU15[pe_nation=="GR"] <- 1
EU15[pe_nation=="P"] <- 1
EU15[pe_nation=="FIN"] <- 1
EU15[pe_nation=="A"] <- 1
EU15[pe_nation=="S"] <- 1
EU15[pe_nation=="E"] <- 1
#EU15[nation=="Sw"] <- 1

dcm$EU15=EU15

gf_15y=NA
gf_15y[gf_15==0] <- 0
gf_15y[gf_15>0] <- 1

gf_16y=NA
gf_16y[gf_16==0] <- 0
gf_16y[gf_16>0] <- 1

gf_17y=NA
gf_17y[gf_17==0] <- 0
gf_17y[gf_17>0] <- 1

gf_18y=NA
gf_18y[gf_18==0] <- 0
gf_18y[gf_18>0] <- 1
```

```
gf_19y=NA
gf_19y[gf_19==0] <- 0
gf_19y[gf_19>0] <- 1

gf_20y=NA
gf_20y[gf_20==0] <- 0
gf_20y[gf_20>0] <- 1

gf_21y=NA
gf_21y[gf_21==0] <- 0
gf_21y[gf_21>0] <- 1

gf_22y=NA
gf_22y[gf_22==0] <- 0
gf_22y[gf_22>0] <- 1

gf_23y=NA
gf_23y[gf_23==0] <- 0
gf_23y[gf_23>0] <- 1

gf_24y=NA
gf_24y[gf_24==0] <- 0
gf_24y[gf_24>0] <- 1

gf_25y=NA
gf_25y[gf_25==0] <- 0
gf_25y[gf_25>0] <- 1

gf_26y=NA
gf_26y[gf_26==0] <- 0
gf_26y[gf_26>0] <- 1

gf_27y=NA
gf_27y[gf_27==0] <- 0
gf_27y[gf_27>0] <- 1


gf_28y=NA
gf_28y[gf_28==0] <- 0
gf_28y[gf_28>0] <- 1
```

```
#sj yes/no
sj_15y=NA
sj_15y[sj_15==0] <- 0
sj_15y[sj_15>0] <- 1

sj_16y=NA
sj_16y[sj_16==0] <- 0
sj_16y[sj_16>0] <- 1

sj_17y=NA
sj_17y[sj_17==0] <- 0
sj_17y[sj_17>0] <- 1

sj_18y=NA
sj_18y[sj_18==0] <- 0
sj_18y[sj_18>0] <- 1

sj_19y=NA
sj_19y[sj_19==0] <- 0
sj_19y[sj_19>0] <- 1

sj_20y=NA
sj_20y[sj_20==0] <- 0
sj_20y[sj_20>0] <- 1

sj_21y=NA
sj_21y[sj_21==0] <- 0
sj_21y[sj_21>0] <- 1

sj_22y=NA
sj_22y[sj_22==0] <- 0
sj_22y[sj_22>0] <- 1

sj_23y=NA
sj_23y[sj_23==0] <- 0
sj_23y[sj_23>0] <- 1

sj_24y=NA
sj_24y[sj_24==0] <- 0
sj_24y[sj_24>0] <- 1
```

```
sj_25y=NA
sj_25y[sj_25==0] <- 0
sj_25y[sj_25>0] <- 1

sj_26y=NA
sj_26y[sj_26==0] <- 0
sj_26y[sj_26>0] <- 1

sj_27y=NA
sj_27y[sj_27==0] <- 0
sj_27y[sj_27>0] <- 1

sj_28y=NA
sj_28y[sj_28==0] <- 0
sj_28y[sj_28>0] <- 1

attach(dcm)
familienstand=99
familienstand[pe_familienstand=="G"]<- "G"
familienstand[pe_familienstand=="L"]<- "L"
familienstand[pe_familienstand=="V"]<- "V"
familienstand[is.na(familienstand)] <- 99

table(familienstand)


dfRF=data.frame(
#penr,
educ,
year, birthyear, pe_ausland,
familienstand, pe_familienbeihilfe, qu_cat, qu_subcat, ej_yes,
ej_dauer, ej_age,  gf_yes, gf_15y,
 gf_16y, gf_17y, gf_18y, gf_19y,  gf_20y, gf_21y, gf_22y,
gf_23y, gf_24y, gf_25y, gf_26y, gf_27y, gf_28y,
gf_15, gf_16, gf_17, gf_18, gf_19, gf_20, gf_21, gf_22, gf_23, gf_24,
gf_25, gf_26, gf_27, gf_28,
tz_yes, sj_yes,
sj_15y, sj_16y, sj_17y, sj_18y, sj_19y, sj_20y, sj_21y,
sj_22y, sj_23y, sj_24y, sj_25y, sj_26y, sj_27y, sj_28y,
sj_15, sj_16, sj_17, sj_18, sj_19, sj_20, sj_21, sj_22, sj_23,
sj_24, sj_25, sj_26, sj_27, sj_28,
ki_anzahl, ki_age_1, ki_age_2, ki_age_3, ki_age_4, ki_age_5,

ev_arbeitstage_5, ev_arbeitstage_10, ev_arbeitstage_15,
ev_arbeitstage_20, ev_arbeitstage_25, ev_arbeitstage_30,
 ev_arbeitstage_35, ev_arbeitstage_40,

ev_arbeitslostage_5, ev_arbeitslostage_10, ev_arbeitslostage_15,
 ev_arbeitslostage_20,  ev_arbeitslostage_25, ev_arbeitslostage_30,
ev_arbeitslostage_35, ev_arbeitslostage_40,

ev_arbeitsperioden_5, ev_arbeitsperioden_10, ev_arbeitsperioden_15,
 ev_arbeitsperioden_20, ev_arbeitsperioden_25, ev_arbeitsperioden_30,
 ev_arbeitsperioden_35, ev_arbeitsperioden_40,

ev_dienstzeit_5, ev_dienstzeit_10, ev_dienstzeit_15,
ev_dienstzeit_20, ev_dienstzeit_25, ev_dienstzeit_30,
 ev_dienstzeit_35, ev_dienstzeit_40,

pe_gest, EU15, ek_yes, nace)

detach(dcm)

summary(dfRF)
head(dfRF)
str(dfRF)

#sample 100,000 observations
s1= dfRF[sample(nrow(dfRF), 100000),]

#forest
fitr11m80oE= randomForest(educ ~ ., data=s1, importance=TRUE, ntree=500)

save(fitr11m80oE, file="80moE.RData")


###################################################################
# females

#only if educ is known
dc=subset(d8, d8$educBek==1)
```

```
str(dc)

dcm=subset(dc, dc$pe_frau=="Frau")
str(dcm)

educ5=NA
educ5[dcm$educ=="keine Pflichtschule"] <- 1
educ5[dcm$educ=="Pflichtschule"] <- 1
educ5[dcm$educ=="Lehre"] <- 2
educ5[dcm$educ=="mittlere Schule (o. Matura)"] <- 3
educ5[dcm$educ=="höhere Schule (m. Matura)"] <- 3
educ5[dcm$educ=="Hochschule od. Universität"] <- 4

educ<-educ5
educ=as.factor(educ)

dcm$nace[is.na(dcm$nace)] <- 0

dcm$ev_arbeitstage_5[is.na(dcm$ev_arbeitstage_5)] <- 0
dcm$ev_arbeitstage_10[is.na(dcm$ev_arbeitstage_10)] <- 0
dcm$ev_arbeitstage_15[is.na(dcm$ev_arbeitstage_15)] <- 0
dcm$ev_arbeitstage_20[is.na(dcm$ev_arbeitstage_20)] <- 0
dcm$ev_arbeitstage_25[is.na(dcm$ev_arbeitstage_25)] <- 0
dcm$ev_arbeitstage_30[is.na(dcm$ev_arbeitstage_30)] <- 0
dcm$ev_arbeitstage_35[is.na(dcm$ev_arbeitstage_35)] <- 0
dcm$ev_arbeitstage_40[is.na(dcm$ev_arbeitstage_40)] <- 0

dcm$ev_arbeitslostage_5[is.na(dcm$ev_arbeitslostage_5)] <- 0
dcm$ev_arbeitslostage_10[is.na(dcm$ev_arbeitslostage_10)] <- 0
dcm$ev_arbeitslostage_15[is.na(dcm$ev_arbeitslostage_15)] <- 0
dcm$ev_arbeitslostage_20[is.na(dcm$ev_arbeitslostage_20)] <- 0
dcm$ev_arbeitslostage_25[is.na(dcm$ev_arbeitslostage_25)] <- 0
dcm$ev_arbeitslostage_30[is.na(dcm$ev_arbeitslostage_30)] <- 0
dcm$ev_arbeitslostage_35[is.na(dcm$ev_arbeitslostage_35)] <- 0
dcm$ev_arbeitslostage_40[is.na(dcm$ev_arbeitslostage_40)] <- 0

dcm$ev_arbeitsperioden_5[is.na(dcm$ev_arbeitsperioden_5)] <- 0
dcm$ev_arbeitsperioden_10[is.na(dcm$ev_arbeitsperioden_10)] <- 0
dcm$ev_arbeitsperioden_15[is.na(dcm$ev_arbeitsperioden_15)] <- 0
dcm$ev_arbeitsperioden_20[is.na(dcm$ev_arbeitsperioden_20)] <- 0
```

```
dcm$ev_arbeitsperioden_25[is.na(dcm$ev_arbeitsperioden_25)] <- 0
dcm$ev_arbeitsperioden_30[is.na(dcm$ev_arbeitsperioden_30)] <- 0
dcm$ev_arbeitsperioden_35[is.na(dcm$ev_arbeitsperioden_35)] <- 0
dcm$ev_arbeitsperioden_40[is.na(dcm$ev_arbeitsperioden_40)] <- 0

attach(dcm)

#EU15
EU15=rep(0,length(pe_nation))
EU15[pe_nation=="D"] <- 1
EU15[pe_nation=="B"] <- 1
EU15[pe_nation=="F"] <- 1
EU15[pe_nation=="I"] <- 1
EU15[pe_nation=="L"] <- 1
EU15[pe_nation=="NL"] <- 1
EU15[pe_nation=="DK"] <- 1
EU15[pe_nation=="IR"] <- 1
EU15[pe_nation=="GB"] <- 1
EU15[pe_nation=="GR"] <- 1
EU15[pe_nation=="P"] <- 1
EU15[pe_nation=="FIN"] <- 1
EU15[pe_nation=="A"] <- 1
EU15[pe_nation=="S"] <- 1
EU15[pe_nation=="E"] <- 1
#EU15[nation=="Sw"] <- 1

dcm$EU15=EU15

gf_15y=NA
gf_15y[gf_15==0] <- 0
gf_15y[gf_15>0] <- 1

gf_16y=NA
gf_16y[gf_16==0] <- 0
gf_16y[gf_16>0] <- 1

gf_17y=NA
gf_17y[gf_17==0] <- 0
gf_17y[gf_17>0] <- 1

gf_18y=NA
```

```
gf_18y[gf_18==0] <- 0
gf_18y[gf_18>0] <- 1

gf_19y=NA
gf_19y[gf_19==0] <- 0
gf_19y[gf_19>0] <- 1

gf_20y=NA
gf_20y[gf_20==0] <- 0
gf_20y[gf_20>0] <- 1

gf_21y=NA
gf_21y[gf_21==0] <- 0
gf_21y[gf_21>0] <- 1

gf_22y=NA
gf_22y[gf_22==0] <- 0
gf_22y[gf_22>0] <- 1

gf_23y=NA
gf_23y[gf_23==0] <- 0
gf_23y[gf_23>0] <- 1

gf_24y=NA
gf_24y[gf_24==0] <- 0
gf_24y[gf_24>0] <- 1

gf_25y=NA
gf_25y[gf_25==0] <- 0
gf_25y[gf_25>0] <- 1

gf_26y=NA
gf_26y[gf_26==0] <- 0
gf_26y[gf_26>0] <- 1

gf_27y=NA
gf_27y[gf_27==0] <- 0
gf_27y[gf_27>0] <- 1

gf_28y=NA
gf_28y[gf_28==0] <- 0
```

```
gf_28y[gf_28>0] <- 1

#sj yes/no
sj_15y=NA
sj_15y[sj_15==0] <- 0
sj_15y[sj_15>0] <- 1

sj_16y=NA
sj_16y[sj_16==0] <- 0
sj_16y[sj_16>0] <- 1

sj_17y=NA
sj_17y[sj_17==0] <- 0
sj_17y[sj_17>0] <- 1

sj_18y=NA
sj_18y[sj_18==0] <- 0
sj_18y[sj_18>0] <- 1

sj_19y=NA
sj_19y[sj_19==0] <- 0
sj_19y[sj_19>0] <- 1

sj_20y=NA
sj_20y[sj_20==0] <- 0
sj_20y[sj_20>0] <- 1

sj_21y=NA
sj_21y[sj_21==0] <- 0
sj_21y[sj_21>0] <- 1

sj_22y=NA
sj_22y[sj_22==0] <- 0
sj_22y[sj_22>0] <- 1

sj_23y=NA
sj_23y[sj_23==0] <- 0
sj_23y[sj_23>0] <- 1

sj_24y=NA
sj_24y[sj_24==0] <- 0
```

```
sj_24y[sj_24>0] <- 1

sj_25y=NA
sj_25y[sj_25==0] <- 0
sj_25y[sj_25>0] <- 1

sj_26y=NA
sj_26y[sj_26==0] <- 0
sj_26y[sj_26>0] <- 1

sj_27y=NA
sj_27y[sj_27==0] <- 0
sj_27y[sj_27>0] <- 1

sj_28y=NA
sj_28y[sj_28==0] <- 0
sj_28y[sj_28>0] <- 1

attach(dcm)
familienstand=99
familienstand[pe_familienstand=="G"]<- "G"
familienstand[pe_familienstand=="L"]<- "L"
familienstand[pe_familienstand=="V"]<- "V"
familienstand[is.na(familienstand)] <- 99

table(familienstand)


dfRF=data.frame(
#penr,
educ,
year, birthyear, pe_ausland, familienstand, pe_familienbeihilfe,
qu_cat, qu_subcat, ej_yes, ej_dauer, ej_age,
gf_yes, gf_15y, gf_16y, gf_17y, gf_18y, gf_19y,
 gf_20y, gf_21y, gf_22y,
gf_23y, gf_24y, gf_25y, gf_26y, gf_27y, gf_28y,
gf_15, gf_16, gf_17, gf_18, gf_19, gf_20, gf_21, gf_22, gf_23, gf_24,
gf_25, gf_26, gf_27, gf_28,
tz_yes, sj_yes,
sj_15y, sj_16y, sj_17y, sj_18y, sj_19y, sj_20y, sj_21y,
sj_22y, sj_23y, sj_24y, sj_25y, sj_26y, sj_27y, sj_28y,
sj_15, sj_16, sj_17, sj_18, sj_19, sj_20, sj_21, sj_22, sj_23,
sj_24, sj_25, sj_26, sj_27, sj_28,

ki_anzahl, ki_age_1, ki_age_2, ki_age_3, ki_age_4, ki_age_5,

ev_arbeitstage_5, ev_arbeitstage_10, ev_arbeitstage_15,
 ev_arbeitstage_20, ev_arbeitstage_25, ev_arbeitstage_30,
 ev_arbeitstage_35, ev_arbeitstage_40,

ev_arbeitslostage_5, ev_arbeitslostage_10, ev_arbeitslostage_15,
ev_arbeitslostage_20, ev_arbeitslostage_25, ev_arbeitslostage_30,
 ev_arbeitslostage_35, ev_arbeitslostage_40,

ev_arbeitsperioden_5, ev_arbeitsperioden_10, ev_arbeitsperioden_15,
 ev_arbeitsperioden_20, ev_arbeitsperioden_25, ev_arbeitsperioden_30,
ev_arbeitsperioden_35, ev_arbeitsperioden_40,

ev_dienstzeit_5, ev_dienstzeit_10, ev_dienstzeit_15, ev_dienstzeit_20,
 ev_dienstzeit_25, ev_dienstzeit_30, ev_dienstzeit_35,
 ev_dienstzeit_40,

pe_gest, EU15, ek_yes, nace)

detach(dcm)

summary(dfRF)
head(dfRF)
str(dfRF)

s1= dfRF[sample(nrow(dfRF), 100000),]

#forest
fitr11f80oE= randomForest(educ ~ ., data=s1, importance=TRUE, ntree=500)

save(fitr11f80oE, file="80foE.RData")

######################################################################
# Imputation
##
##########
#load data
```

```
d8=read.dta("C:\\Users\\Christina\\Documents\\Christina\\
ImpDatensatz_comp80.dta")

str(d5)
summary(d5)

#if educ is not known
dc=subset(d8, d8$educBek==0)

str(dc)

#males
dcm=subset(dc, dc$pe_frau=="Mann")
str(dcm)

educ5=NA
educ5[dcm$educ=="keine Pflichtschule"] <- 1
educ5[dcm$educ=="Pflichtschule"] <- 1
educ5[dcm$educ=="Lehre"] <- 2
educ5[dcm$educ=="mittlere Schule (o. Matura)"] <- 3
educ5[dcm$educ=="höhere Schule (m. Matura)"] <- 3
educ5[dcm$educ=="Hochschule od. Universität"] <- 4

educ<-educ5
educ=as.factor(educ)

dcm$nace[is.na(dcm$nace)] <- 0

dcm$ev_arbeitstage_5[is.na(dcm$ev_arbeitstage_5)] <- 0
dcm$ev_arbeitstage_10[is.na(dcm$ev_arbeitstage_10)] <- 0
dcm$ev_arbeitstage_15[is.na(dcm$ev_arbeitstage_15)] <- 0
dcm$ev_arbeitstage_20[is.na(dcm$ev_arbeitstage_20)] <- 0
dcm$ev_arbeitstage_25[is.na(dcm$ev_arbeitstage_25)] <- 0
dcm$ev_arbeitstage_30[is.na(dcm$ev_arbeitstage_30)] <- 0
dcm$ev_arbeitstage_35[is.na(dcm$ev_arbeitstage_35)] <- 0
dcm$ev_arbeitstage_40[is.na(dcm$ev_arbeitstage_40)] <- 0

dcm$ev_arbeitslostage_5[is.na(dcm$ev_arbeitslostage_5)] <- 0
dcm$ev_arbeitslostage_10[is.na(dcm$ev_arbeitslostage_10)] <- 0
dcm$ev_arbeitslostage_15[is.na(dcm$ev_arbeitslostage_15)] <- 0
dcm$ev_arbeitslostage_20[is.na(dcm$ev_arbeitslostage_20)] <- 0
```

```
dcm$ev_arbeitslostage_25[is.na(dcm$ev_arbeitslostage_25)] <- 0
dcm$ev_arbeitslostage_30[is.na(dcm$ev_arbeitslostage_30)] <- 0
dcm$ev_arbeitslostage_35[is.na(dcm$ev_arbeitslostage_35)] <- 0
dcm$ev_arbeitslostage_40[is.na(dcm$ev_arbeitslostage_40)] <- 0

dcm$ev_arbeitsperioden_5[is.na(dcm$ev_arbeitsperioden_5)] <- 0
dcm$ev_arbeitsperioden_10[is.na(dcm$ev_arbeitsperioden_10)] <- 0
dcm$ev_arbeitsperioden_15[is.na(dcm$ev_arbeitsperioden_15)] <- 0
dcm$ev_arbeitsperioden_20[is.na(dcm$ev_arbeitsperioden_20)] <- 0
dcm$ev_arbeitsperioden_25[is.na(dcm$ev_arbeitsperioden_25)] <- 0
dcm$ev_arbeitsperioden_30[is.na(dcm$ev_arbeitsperioden_30)] <- 0
dcm$ev_arbeitsperioden_35[is.na(dcm$ev_arbeitsperioden_35)] <- 0
dcm$ev_arbeitsperioden_40[is.na(dcm$ev_arbeitsperioden_40)] <- 0

attach(dcm)

#EU15
EU15=rep(0,length(pe_nation))
EU15[pe_nation=="D"] <- 1
EU15[pe_nation=="B"] <- 1
EU15[pe_nation=="F"] <- 1
EU15[pe_nation=="I"] <- 1
EU15[pe_nation=="L"] <- 1
EU15[pe_nation=="NL"] <- 1
EU15[pe_nation=="DK"] <- 1
EU15[pe_nation=="IR"] <- 1
EU15[pe_nation=="GB"] <- 1
EU15[pe_nation=="GR"] <- 1
EU15[pe_nation=="P"] <- 1
EU15[pe_nation=="FIN"] <- 1
EU15[pe_nation=="A"] <- 1
EU15[pe_nation=="S"] <- 1
EU15[pe_nation=="E"] <- 1
#EU15[nation=="Sw"] <- 1

dcm$EU15=EU15

gf_15y=NA
gf_15y[gf_15==0] <- 0
gf_15y[gf_15>0] <- 1
```

```
gf_16y=NA
gf_16y[gf_16==0] <- 0
gf_16y[gf_16>0] <- 1

gf_17y=NA
gf_17y[gf_17==0] <- 0
gf_17y[gf_17>0] <- 1

gf_18y=NA
gf_18y[gf_18==0] <- 0
gf_18y[gf_18>0] <- 1

gf_19y=NA
gf_19y[gf_19==0] <- 0
gf_19y[gf_19>0] <- 1

gf_20y=NA
gf_20y[gf_20==0] <- 0
gf_20y[gf_20>0] <- 1

gf_21y=NA
gf_21y[gf_21==0] <- 0
gf_21y[gf_21>0] <- 1

gf_22y=NA
gf_22y[gf_22==0] <- 0
gf_22y[gf_22>0] <- 1

gf_23y=NA
gf_23y[gf_23==0] <- 0
gf_23y[gf_23>0] <- 1

gf_24y=NA
gf_24y[gf_24==0] <- 0
gf_24y[gf_24>0] <- 1

gf_25y=NA
gf_25y[gf_25==0] <- 0
gf_25y[gf_25>0] <- 1

gf_26y=NA
```

```
gf_26y[gf_26==0] <- 0
gf_26y[gf_26>0] <- 1

gf_27y=NA
gf_27y[gf_27==0] <- 0
gf_27y[gf_27>0] <- 1

gf_28y=NA
gf_28y[gf_28==0] <- 0
gf_28y[gf_28>0] <- 1

#sj yes/no
sj_15y=NA
sj_15y[sj_15==0] <- 0
sj_15y[sj_15>0] <- 1

sj_16y=NA
sj_16y[sj_16==0] <- 0
sj_16y[sj_16>0] <- 1

sj_17y=NA
sj_17y[sj_17==0] <- 0
sj_17y[sj_17>0] <- 1

sj_18y=NA
sj_18y[sj_18==0] <- 0
sj_18y[sj_18>0] <- 1

sj_19y=NA
sj_19y[sj_19==0] <- 0
sj_19y[sj_19>0] <- 1

sj_20y=NA
sj_20y[sj_20==0] <- 0
sj_20y[sj_20>0] <- 1

sj_21y=NA
sj_21y[sj_21==0] <- 0
sj_21y[sj_21>0] <- 1

sj_22y=NA
```

```
sj_22y[sj_22==0] <- 0
sj_22y[sj_22>0] <- 1

sj_23y=NA
sj_23y[sj_23==0] <- 0
sj_23y[sj_23>0] <- 1

sj_24y=NA
sj_24y[sj_24==0] <- 0
sj_24y[sj_24>0] <- 1

sj_25y=NA
sj_25y[sj_25==0] <- 0
sj_25y[sj_25>0] <- 1

sj_26y=NA
sj_26y[sj_26==0] <- 0
sj_26y[sj_26>0] <- 1

sj_27y=NA
sj_27y[sj_27==0] <- 0
sj_27y[sj_27>0] <- 1

sj_28y=NA
sj_28y[sj_28==0] <- 0
sj_28y[sj_28>0] <- 1

familienstand=99
familienstand[pe_familienstand=="G"]<- "G"
familienstand[pe_familienstand=="L"]<- "L"
familienstand[pe_familienstand=="V"]<- "V"
familienstand[is.na(familienstand)] <- 99

table(familienstand)
attach(dcm)

dfRFT=data.frame(
#penr,
#educ,
year, birthyear, pe_ausland, familienstand, pe_familienbeihilfe,
qu_cat, qu_subcat, ej_yes,
ej_dauer, ej_age,
gf_yes, gf_15y, gf_16y, gf_17y, gf_18y, gf_19y, gf_20y, gf_21y,
 gf_22y, gf_23y, gf_24y, gf_25y, gf_26y, gf_27y, gf_28y,
gf_15, gf_16, gf_17, gf_18, gf_19, gf_20, gf_21, gf_22, gf_23, gf_24,
gf_25, gf_26, gf_27, gf_28,

tz_yes, sj_yes,

sj_15y, sj_16y, sj_17y, sj_18y, sj_19y, sj_20y, sj_21y,
sj_22y, sj_23y, sj_24y, sj_25y, sj_26y, sj_27y, sj_28y,
sj_15, sj_16, sj_17, sj_18, sj_19, sj_20, sj_21, sj_22, sj_23,
sj_24, sj_25, sj_26, sj_27, sj_28,

ki_anzahl, ki_age_1, ki_age_2, ki_age_3, ki_age_4, ki_age_5,

ev_arbeitstage_5, ev_arbeitstage_10, ev_arbeitstage_15, ev_arbeitstage_20,
ev_arbeitstage_25, ev_arbeitstage_30, ev_arbeitstage_35, ev_arbeitstage_40,

ev_arbeitslostage_5, ev_arbeitslostage_10, ev_arbeitslostage_15,
 ev_arbeitslostage_20,  ev_arbeitslostage_25, ev_arbeitslostage_30,
  ev_arbeitslostage_35, ev_arbeitslostage_40,

ev_arbeitsperioden_5, ev_arbeitsperioden_10, ev_arbeitsperioden_15,
 ev_arbeitsperioden_20, ev_arbeitsperioden_25, ev_arbeitsperioden_30,
  ev_arbeitsperioden_35, ev_arbeitsperioden_40,

ev_dienstzeit_5, ev_dienstzeit_10, ev_dienstzeit_15, ev_dienstzeit_20,
 ev_dienstzeit_25, ev_dienstzeit_30, ev_dienstzeit_35, ev_dienstzeit_40,

pe_gest, EU15, ek_yes, nace)

detach(dcm)

summary(dfRF)
head(dfRF)
str(dfRF)

#load("80moE.Rdata")

pr=predict(fitr11m80oE, newdata=dfRFT, type="response")
prWS=predict(fitr11m80oE, newdata=dfRFT, type="prob")
```

```
eImp80mpr=data.frame(dcm$penr, pr)
eImp80mprWS=data.frame(dcm$penr, prWS)

save(eImp80mpr, file="eImp80mpr.Rdata")
save(eImp80mprWS, file="eImp80mprWS.Rdata")


############
#females
d8=read.dta("C:\\Users\\Christina\\Documents\\Christina\\
ImpDatensatz_comp80.dta")

str(d5)
summary(d5)

#if educ is not known
dc=subset(d8, d8$educBek==0)

str(dc)

#females
dcm=subset(dc, dc$pe_frau=="Frau")
str(dcm)

educ5=NA
educ5[dcm$educ=="keine Pflichtschule"] <- 1
educ5[dcm$educ=="Pflichtschule"] <- 1
educ5[dcm$educ=="Lehre"] <- 2
educ5[dcm$educ=="mittlere Schule (o. Matura)"] <- 3
educ5[dcm$educ=="höhere Schule (m. Matura)"] <- 3
educ5[dcm$educ=="Hochschule od. Universität"] <- 4

educ<-educ5
educ=as.factor(educ)

dcm$nace[is.na(dcm$nace)] <- 0

dcm$ev_arbeitstage_5[is.na(dcm$ev_arbeitstage_5)] <- 0
dcm$ev_arbeitstage_10[is.na(dcm$ev_arbeitstage_10)] <- 0
dcm$ev_arbeitstage_15[is.na(dcm$ev_arbeitstage_15)] <- 0
```

```
dcm$ev_arbeitstage_20[is.na(dcm$ev_arbeitstage_20)] <- 0
dcm$ev_arbeitstage_25[is.na(dcm$ev_arbeitstage_25)] <- 0
dcm$ev_arbeitstage_30[is.na(dcm$ev_arbeitstage_30)] <- 0
dcm$ev_arbeitstage_35[is.na(dcm$ev_arbeitstage_35)] <- 0
dcm$ev_arbeitstage_40[is.na(dcm$ev_arbeitstage_40)] <- 0

dcm$ev_arbeitslostage_5[is.na(dcm$ev_arbeitslostage_5)] <- 0
dcm$ev_arbeitslostage_10[is.na(dcm$ev_arbeitslostage_10)] <- 0
dcm$ev_arbeitslostage_15[is.na(dcm$ev_arbeitslostage_15)] <- 0
dcm$ev_arbeitslostage_20[is.na(dcm$ev_arbeitslostage_20)] <- 0
dcm$ev_arbeitslostage_25[is.na(dcm$ev_arbeitslostage_25)] <- 0
dcm$ev_arbeitslostage_30[is.na(dcm$ev_arbeitslostage_30)] <- 0
dcm$ev_arbeitslostage_35[is.na(dcm$ev_arbeitslostage_35)] <- 0
dcm$ev_arbeitslostage_40[is.na(dcm$ev_arbeitslostage_40)] <- 0

dcm$ev_arbeitsperioden_5[is.na(dcm$ev_arbeitsperioden_5)] <- 0
dcm$ev_arbeitsperioden_10[is.na(dcm$ev_arbeitsperioden_10)] <- 0
dcm$ev_arbeitsperioden_15[is.na(dcm$ev_arbeitsperioden_15)] <- 0
dcm$ev_arbeitsperioden_20[is.na(dcm$ev_arbeitsperioden_20)] <- 0
dcm$ev_arbeitsperioden_25[is.na(dcm$ev_arbeitsperioden_25)] <- 0
dcm$ev_arbeitsperioden_30[is.na(dcm$ev_arbeitsperioden_30)] <- 0
dcm$ev_arbeitsperioden_35[is.na(dcm$ev_arbeitsperioden_35)] <- 0
dcm$ev_arbeitsperioden_40[is.na(dcm$ev_arbeitsperioden_40)] <- 0

attach(dcm)

#EU15
EU15=rep(0,length(pe_nation))
EU15[pe_nation=="D"] <- 1
EU15[pe_nation=="B"] <- 1
EU15[pe_nation=="F"] <- 1
EU15[pe_nation=="I"] <- 1
EU15[pe_nation=="L"] <- 1
EU15[pe_nation=="NL"] <- 1
EU15[pe_nation=="DK"] <- 1
EU15[pe_nation=="IR"] <- 1
EU15[pe_nation=="GB"] <- 1
EU15[pe_nation=="GR"] <- 1
EU15[pe_nation=="P"] <- 1
EU15[pe_nation=="FIN"] <- 1
EU15[pe_nation=="A"] <- 1
```

```
EU15[pe_nation=="S"] <- 1
EU15[pe_nation=="E"] <- 1
#EU15[nation=="Sw"] <- 1

dcm$EU15=EU15

gf_15y=NA
gf_15y[gf_15==0] <- 0
gf_15y[gf_15>0] <- 1

gf_16y=NA
gf_16y[gf_16==0] <- 0
gf_16y[gf_16>0] <- 1

gf_17y=NA
gf_17y[gf_17==0] <- 0
gf_17y[gf_17>0] <- 1

gf_18y=NA
gf_18y[gf_18==0] <- 0
gf_18y[gf_18>0] <- 1

gf_19y=NA
gf_19y[gf_19==0] <- 0
gf_19y[gf_19>0] <- 1

gf_20y=NA
gf_20y[gf_20==0] <- 0
gf_20y[gf_20>0] <- 1

gf_21y=NA
gf_21y[gf_21==0] <- 0
gf_21y[gf_21>0] <- 1

gf_22y=NA
gf_22y[gf_22==0] <- 0
gf_22y[gf_22>0] <- 1

gf_23y=NA
gf_23y[gf_23==0] <- 0
gf_23y[gf_23>0] <- 1
```

```
gf_24y=NA
gf_24y[gf_24==0] <- 0
gf_24y[gf_24>0] <- 1

gf_25y=NA
gf_25y[gf_25==0] <- 0
gf_25y[gf_25>0] <- 1

gf_26y=NA
gf_26y[gf_26==0] <- 0
gf_26y[gf_26>0] <- 1

gf_27y=NA
gf_27y[gf_27==0] <- 0
gf_27y[gf_27>0] <- 1

gf_28y=NA
gf_28y[gf_28==0] <- 0
gf_28y[gf_28>0] <- 1

#sj yes/no
sj_15y=NA
sj_15y[sj_15==0] <- 0
sj_15y[sj_15>0] <- 1

sj_16y=NA
sj_16y[sj_16==0] <- 0
sj_16y[sj_16>0] <- 1

sj_17y=NA
sj_17y[sj_17==0] <- 0
sj_17y[sj_17>0] <- 1

sj_18y=NA
sj_18y[sj_18==0] <- 0
sj_18y[sj_18>0] <- 1

sj_19y=NA
sj_19y[sj_19==0] <- 0
sj_19y[sj_19>0] <- 1
```

```
sj_20y=NA
sj_20y[sj_20==0] <- 0
sj_20y[sj_20>0] <- 1

sj_21y=NA
sj_21y[sj_21==0] <- 0
sj_21y[sj_21>0] <- 1

sj_22y=NA
sj_22y[sj_22==0] <- 0
sj_22y[sj_22>0] <- 1

sj_23y=NA
sj_23y[sj_23==0] <- 0
sj_23y[sj_23>0] <- 1

sj_24y=NA
sj_24y[sj_24==0] <- 0
sj_24y[sj_24>0] <- 1

sj_25y=NA
sj_25y[sj_25==0] <- 0
sj_25y[sj_25>0] <- 1

sj_26y=NA
sj_26y[sj_26==0] <- 0
sj_26y[sj_26>0] <- 1

sj_27y=NA
sj_27y[sj_27==0] <- 0
sj_27y[sj_27>0] <- 1

sj_28y=NA
sj_28y[sj_28==0] <- 0
sj_28y[sj_28>0] <- 1

familienstand=99
familienstand[pe_familienstand=="G"]<- "G"
familienstand[pe_familienstand=="L"]<- "L"
familienstand[pe_familienstand=="V"]<- "V"
```

```
familienstand[is.na(familienstand)] <- 99

table(familienstand)
attach(dcm)

dfRFT=data.frame(
#penr,
#educ,
year, birthyear, pe_ausland, familienstand, pe_familienbeihilfe,
qu_cat, qu_subcat, ej_yes,
ej_dauer, ej_age,
gf_yes, gf_15y, gf_16y, gf_17y, gf_18y, gf_19y, gf_20y,
gf_21y, gf_22y, gf_23y, gf_24y, gf_25y, gf_26y, gf_27y, gf_28y,
gf_15, gf_16, gf_17, gf_18, gf_19, gf_20, gf_21, gf_22, gf_23, gf_24,
gf_25, gf_26, gf_27, gf_28,
tz_yes, sj_yes,
sj_15y, sj_16y, sj_17y, sj_18y, sj_19y, sj_20y, sj_21y,
sj_22y, sj_23y, sj_24y, sj_25y, sj_26y, sj_27y, sj_28y,
sj_15, sj_16, sj_17, sj_18, sj_19, sj_20, sj_21, sj_22, sj_23,
sj_24, sj_25, sj_26, sj_27, sj_28,

ki_anzahl, ki_age_1, ki_age_2, ki_age_3, ki_age_4, ki_age_5,

ev_arbeitstage_5, ev_arbeitstage_10, ev_arbeitstage_15,
 ev_arbeitstage_20, ev_arbeitstage_25, ev_arbeitstage_30,
  ev_arbeitstage_35, ev_arbeitstage_40,

ev_arbeitslostage_5, ev_arbeitslostage_10, ev_arbeitslostage_15,
 ev_arbeitslostage_20, ev_arbeitslostage_25, ev_arbeitslostage_30,
  ev_arbeitslostage_35, ev_arbeitslostage_40,

ev_arbeitsperioden_5, ev_arbeitsperioden_10, ev_arbeitsperioden_15,
 ev_arbeitsperioden_20, ev_arbeitsperioden_25, ev_arbeitsperioden_30,
  ev_arbeitsperioden_35, ev_arbeitsperioden_40,

ev_dienstzeit_5, ev_dienstzeit_10, ev_dienstzeit_15,
ev_dienstzeit_20,  ev_dienstzeit_25, ev_dienstzeit_30,
 ev_dienstzeit_35, ev_dienstzeit_40,

pe_gest,  EU15, ek_yes, nace)
```

```
detach(dcm)

summary(dfRF)
head(dfRF)
str(dfRF)

load("80foE.Rdata")

pr=predict(fitr11f80oE, newdata=dfRFT, type="response")
prWS=predict(fitr11f80oE, newdata=dfRFT, type="prob")

eImp80fpr=data.frame(dcm$penr, pr)
eImp80fprWS=data.frame(dcm$penr, prWS)

save(eImp80fpr, file="eImp80fpr.Rdata")
save(eImp80fprWS, file="eImp80fprWS.Rdata")
```

The association rules were found and visualized with the following code, which is an example for the rules that were created with the census data of 2001. The rules of the NRN data were created with the same code, but other explanatory variables in the dataframe.

```
#######################################################
#
library(arules)
library(arulesViz)

dVZ_r=data.frame(nuts2,familysize,nchild,nchlt5,eldch_c,yngch_c,
birthyear_c,sex,marst,citizen, EU28,educat5, eempsta,occ,ind,
class, hrsfull, cont, chbornd, nuts3)

dVZ_r=as(dVZvers1,"transactions")


itemFrequencyPlot(dVZ_r, support = 0.3, cex.names=0.8,
main="Microcensus 2001", col="lightblue")
```

```
dVZ_r1 <- apriori(dVZ_r, parameter = list(support = 0.001,
confidence = 0.90))

#dVZ_r1#
summary(dVZ_r1)
rr <- rhs(dVZ_r1) %pin% "educat5="
inspect(dVZ_r1[rr])

#educat==2
rr2 <- rhs(dVZ_r1) %pin% "educat5=2"
inspect(dVZ_r1[rr2])

sink("rul_sup001conf90")
summary(dVZ_r1[rr])
sink()

plot(dVZ_r1[rr],shading="lift")
plot(dVZ_r1[rr],method="grouped")

#write all rules
#write(sort(dVZ_r1[rr], by = "confidence"),"VZ01sup0015conf090.csv")


############
# remove redundant

#inspect(rules)
rules.sorted <- sort(dVZ_r1[rr], by="lift")

subset.matrix <- is.subset(rules.sorted, rules.sorted)
subset.matrix[lower.tri(subset.matrix, diag=T)] <- NA
redundant <- colSums(subset.matrix, na.rm=T) >= 1
which(redundant)

#remove redundant rules

rules.pruned <- rules.sorted[!redundant]

#write redundant
#write(sort(rules.pruned, by = "confidence"),"prunedsup001cond90.csv")
```

```
sink("rules.pruvers1")
summary(rules.pruned)
sink()

plot(rules.pruned, shading="lift")
plot(rules.pruned, method="grouped")
```