Generating synthetic households for microsimulation using machine learning

Master-Arbeit
zur Erlangung des Grades eines
Master of Science

an der Universität Trier
Fachbereich
Wirtschafts- und Sozialstatistik

vorgelegt von

Nicolas Kaiser
Brauerstrasse 22
56743 Mendig
Mtrn: 1448043

Trier, im Dezember 2020

1. Gutachter: Dr. habil. Jan Pablo Burgard
2. Gutachter: Hariolf Merkle

Contents

List of Figures

List of Tables

1 Introduction

A synthetic population is an artificial data set, which preferably represents the characteristics of a real population. There are many reasons for creating a synthetic population: Often it is about investigating the effects of different sample designs on the estimated parameters (Burgard, Muennich and Zimmermann, 2013), checking new estimation methods for their accuracy and efficiency (Javed and Irvan, 2020), evaluating the performance of different methods in the same context (Morris et. al., 2019, 2075) or obtaining a data basis for microsimulation (Lovelace and Dumont, 2016). In short, it is about answering various "what if" questions (Li and O'Donoghue, 2013). In order to do this as well as possible, the synthetic population created must be close to a real population, which means that the individuals in the data set should have realistic characteristics. However, people often do not live exclusively as isolated individuals, but in households with other people. Simply ignoring the household level would make many analyses very unreliable. One particular difficulty here is the creation of such a realistic household structure, i.e the generation of 'correlation structures of persons within a household. If persons within a household are simulated without any social links, then strange results - for example five babies in a five-person household - may appear' (Muennich and Schuerrle, 2003, 4). The present work addresses these difficulties. Our goal will be to create a complete and realistic synthetic population with special emphasis on the household structure. In any case, machine learning methods are used, which seems to be particularly promising for this purpose, as the interrelationships within households are non-linear and interactive (Caiola and Reiter, 2010). We assume that the starting point will always be a household sample from the real population, on the basis of which our strategies will be developed, applied and tested.

The second chapter will deal with basic principles of population generation. We summarize how researchers have dealt with the problem of household generation so far and where their approaches differ from ours. The structure of household data gets briefly described and defined. A special focus is put on the topic statistical modeling for data generation, as this has been an often used strategy which has some similarities to machine learning but also some crucial differences. In addition, possible initial situations of data availability are explained and the sampling methodology is described. Chapter three goes into more detail about the methodological reasons for the use of machine learning methods and then presents the methods used in the context of the topic, namely Random Forests and Neural Nets. Chapter four presents two different strategies for achieving our goals. The first approach picks up loose ideas from the field of partner matching in microsimulation and works by recombining existing elements to obtain a model estimation. The second approach is based on newer ideas from the scientific literature, which works by estimating every single parameter of interest based on the other members in a household. Chapter five explains in detail which criteria we can use to determine the realism of our synthetic population and introduces the AMELIA dataset. We also outline the relevant parameters for a simulation study to evaluate our strategies and methods. The final chapter presents the synthetic populations resulting from the strategies and methods, evaluates their quality and discusses performance and problems. Open questions and remaining gaps are identified for further research. A conclusion summarizes the work in brief.

2 Creating a synthetic household structure

2.1 Literature overview

The advantages of simulating a full population and it's individual characteristics with synthetic data or a census rather than looking at aggregated data were highlighted early on as well as the difficulties to do so (Orcutt, 1957; Kolb, 2013, 1; Tanton and Kimberley, 2013, 3). It would allow in-depth analysis, by avoiding the well-known ecological fallacy and many other benefits would follow (Robinson, 1950; Selvin, 1958). Survey statistics benefits from such data, as its samples are affected by variability through measurement error, sampling error, imputation and editing (Alfons et. al., 2011). The exact effects can be estimated using synthetic data, which is why a close-to-reality approach (Muennich et. al., 2003) is deliberately followed in this area. The generation of microdata plays an equally important role in the broad field of microsimulation. Here the synthetic population is used as a starting point to study it's change over time, given various external influences, such as policys (Lovelace and Dumont, 2016). Using a complete population for analysis, rather than just a sample, could free up unprecedented problem-solving capacity. Unfortunately, such realistic and complete data is hardly available which lead to several suggestions to create one, none of which has been recognized as the best and final solution.

The different approaches are mostly based on general ideas which can be distinguished between structure-giving, structure-acquiring and structure-preserving methods (Kolb, 2013). They often have clearly identifiable advantages and disadvantages, which must be weighed up against their own research objectives.

We start with structure-preserving methods. A particularly simple method of obtaining a synthetic population based on a sample of households would be to simply replicate the elements present in the sample based on appropriate weightings. However, the characteristics of a rarely occurring household would always be copied exactly, which greatly reduces variation. Elements that occur in the population but not in our sample would not occur in the synthetic population - this is called sampling zeros (He et. al, 2014). In addition, this approach often has the disadvantage that correlations between the variables are lost. On the other hand, the distributional characteristics fit quiet well. Such a procedure can be assigned to the so-called structure-preserving procedures, because the structure already found in the sample would be preserved. However, since the main idea is basically reasonable, more complex approaches have been developed, which are summarised under the term 'synthetic reconstruction'. This involves sequentially drawing values from conditional distributions using Monte Carlo sampling (Huang and Williamson, 2001, 1). The necessary conditional distributions are derived from publicly available census tables or other data sets (Stier, 1999, 150). This can be done even without preliminary sample information like Barthelemy and Toint (2013) did it. Another possibility would be to use the alias method (Kronmal and Peterson, 1979) to generate random variables from a discrete distribution, a procedure, which has also been applied by Muennich and Schuerle (2003). There are many further ways to perform synthetic reconstruction, many of them based on iterative proportional fitting (Deming and Stephan, 1940; Fienberg and Rinaldo, 2007). A very famous idea for synthetic population generation was proposed by Rubin (1993): He suggested, to use multiple imputation which was originally developed to deal with missing values in datasets. The approach has been discussed in more Detail by Reiter (2009).

Structure-giving methods may be an option if relatively little preliminary information is available. Then, random numbers can be drawn from previously determined distributions to obtain an initial basis for simulation. In this way, the parameters of a multivariate normal distribution could be established as a basis from which repeated drawings would be made. Of course, many other classes of known distributions are also conceivable. For Sakshaug and Raghunathan (2010) this procedure served as a starting point for their simulation. If one chooses this approach, one obtains at least a population; in the case of questions where the realism of the population is secondary, this may be sufficient. To increase the level of realism, however, additional steps must be taken. For the purposes of this paper, this approach alone would not be sufficient and would at most be complementary.

Structure-capturing approaches are necessarily dependent on preliminary information. If the empirical distributions in classes are known, *acceptance rejection sampling* (Casella et. al, 2004) is a well-used possibility for population generation. However, the use of statistical models is particularly widespread The basic idea is always to predict the parameters of the population using a model, i.e to assume, that the true population was created by an underlying data generating process. If it is able to reconstruct the true data generating process, this should lead to estimates far beyond the preservation of preliminary information. Recently, these approaches have been taken up by Drechsler (2010) and Caiola and Reiter (2010), among others, to generalise them to machine learning methods. It should be mentioned that this is partly about disclosure control, even if the idea presented there can also be used for other purposes. Due to the relevance of this topic for the research done so far and its prominent place within this thesis, we dedicate a separate sub chapter to this topic. Since we assume in this study that empirical information is available and necessary in the form of a sample of households our approach belongs to structure-capturing methods. The present overview could only present the most important approaches and methods, but for the sake of clarity we have limited ourselves to the best known and most frequently used ones.

2.2 Data structure of household data

To understand the next steps in this work and their immediate consequences, it is necessary first of all to look at how we define a household and what it's data structure looks like. We define a household as a row vector h which contains i members with j parameters:

$$\boldsymbol{h} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{ij} \end{bmatrix}$$

so that such a household has a length of $i \times j = m$. Several households can be stored by combining the row vectors into one data set. For instance, a two-dimensional matrix H which

PID	HHS	SEX	AGE	MST
1	2	1	55	2
2	4	2	47	4
3	3	2	28	1
4	1	1	19	1
...	...	...	...	...

Table 1: Person level data: Every row contains an individual

HID	AGE_1	AGE_2	SEX_1	SEX_2	MST_1	MST_2
1	43	12	1	2	5	1
2	25	22	2	2	1	1
3	76	64	2	1	2	2
4	19	3	2	1	1	1
...	...	...	...	...	...	...

Table 2: Household level data: Every row contains a household

covers a number of h households would be of size $h \times m$:

$$\mathbf{H} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{h1} & x_{h2} & \cdots & x_{hm} \end{bmatrix} = (x_{hm}) \in \mathbb{R}^{h \times m}.$$

In contrast to individual-level-data, every row contains a household instead of a person. Tables 1 and 2 give an example. In practice, the assignment of a unique HID allows the data to be converted to the first or second format without loss of information. This means that information on individuals and the composition of households is always available at the same time. In order to ensure that after conversion it is still possible to trace which parameter j belongs to which member i, a unique numbering should be applied. This can be ensured, for example, by sorting the members of a household in descending order of age, starting with the oldest member.

2.3 Data availability and sampling design

We made it clear right at the beginning that the methods to be investigated refer to an initial situation in which preliminary information is available in the form of a sample. However, what exactly does the sampling scheme look like and what other starting points, hereinafter referred to as 'data availability scenarios', could be assumed? Lovelace and Dumont (2016) distinguish between three different data availability scenarios in descending order of detail.

1. Access to a sample of households for which you have information about each member.

2. Access to separate datasets about individuals and households, stored in independent data tables that are not linked by household ID.

3. No access to aggregate data relating to households, but access to some individual level variables related to the household of which they belong.

All procedures to be examined in this thesis are based on scenario 1. Thus, an extremely favorable and optimistic scenario is used, which may not always be encountered in reality. If new methods are tested, those should be tested however first of all always under the best possible conditions and fulfilled assumptions. If they fail already on this level, this can be assumed also for more unfavorable scenarios.

Let us now take a brief look at the mathematical aspects of scenario 1. Technically speaking, this is single stage cluster sampling (Saerndal et al., 2003). Let's consider a finite Population $\mathcal{U}$, consisting of N primary sampling units (PSUs) denoted by the index set

$$\mathcal{U} = \{1, 2, \ldots, N\}.$$

The PSUs are households and the ith psu contains M_i secondary sampling units (SSUs) which are the single individuals. We draw a simple random sample without replacement (SRSWOR) S_i consisting of n PSUs where the selected ith PSU contains m_i SSUs. All elements which belong to a selected PSU are considered. The probability for an individual to be selected by the sample is therefore equal to the probability of it belonging to the PSU to be selected (Lohr, 1999, 134). It is still easy to obtain an unbiased estimate for population means or totals. This is done by concentrating on cluster means and totals instead of considering individual elements, i.e t_i is the total for the elements in PSU i. This could be, for instance, the total of incomes, which is equal to the sum of the individual income observations y_{ij}. The fact that every PSU has the same probability to be selected leads to a self-weighting sample with the weight for each observation unit beeing $w_{ij} = \frac{N}{n}$. The population total t is estimated without bias by

$$\hat{t}_{\text{unb}} = \sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}_i} w_{ij} y_{ij}$$

which is an estimate for the households.

The advantage is that estimates can be made without bias at the individual level as well. For instance, an unbiased estimator could be constructed for the age of a person. Let

$$K = \sum_{i=1}^{N} M_i$$

be the total number of SSUs in the population. Then, the unbiased estimation for the average on person-level $\bar{y}_U$ is

$$\hat{y}_{\text{unb}} = \frac{\hat{t}_{\text{unb}}}{K}.$$

Looking at these properties resulting from the sampling design, it quickly becomes clear why this is the best possible data availability scenario for the generation of synthetic households. On the one hand, the household sample allows us to make a direct inference on the household structure of the population by obtaining the composition of a household. On the other hand, it also allows us to make estimates at the individual level. This leads to the hope of obtaining a complete and realistic household structure while at the same time, reasonable estimates are available for the individuals. Without the information on each individual member of a household, t_i would be unknown and thus no estimation of t would be possible. In case of a worse data availability scenario, t_i would have to be estimated. Lovelace and Dumont (2016) list a number of other methods if scenario 1 does not apply, i.e. no microdata for households are available. These can be found there and represent alternatives to the initial situation described above, but will not be mentioned in the following.

2.4 Modeling as a method of data generation

The methods to be investigated in this thesis rely on modeling, which belongs to the category of structure-capturing methods with regard to population generation. This means that preliminary information already exists, for example in the form of a sample. The information should now be used efficiently to determine the parameters and characteristics of the true population. Examples of such an approach can be found in Chambers and Dunstand (1986) or Kuk (1993). For our overview about modeling methods, we mostly stick to Murphy (2012) and orient ourselves in form and content to the same. Additional information is marked by appropriate citation. Generally speaking, most statistical models aim to estimate a probability function $p\left(y \mid \mathbf{x}, \theta\right)$. Models which follow this principle are also called *parametric models*, because they assume a finite set of parameters θ which allows to make future predictions x, independently from the observed data $\mathcal{D}$, so that

$$p(x \mid \theta, \mathcal{D}) = p(x \mid \theta)$$

holds (Ghahramani, 2013, 8). The basic idea of this can be demonstrated using simple regression models. One of the simplest and most commonly used statistical models is linear regression, which takes the form of

$$y(\mathbf{x}) = \beta^T \mathbf{x} + \epsilon = \sum_{j=1}^{D} \beta_j x_j + \epsilon$$

where $\beta^T \mathbf{x}$ is the inner scalar product between an input vector x and the *beta* coefficient β. The values of a variable y, which we assume to be dependent, are then nothing else but linear combinations of covariates and an error term ε assumed to be independently and normally distributed. The parametric assumptions of the model become more clear, if we write it as a conditional probability density with the form of

$$p(y \mid \mathbf{x}, \boldsymbol{\theta}) = \mathcal{N}\left(y \mid \mu(\mathbf{x}), \sigma^2(\mathbf{x})\right)$$

where μ denotes the mean and σ^2 the variance (Murphy, 2012, 19). Assumptions about distributions can greatly simplify the calculation and interpretation of models, but can also make it much more difficult if they fail. Alternatively, so-called non-parametric models can be estimated, where the conditional density $p(y \mid \mathbf{x})$ is estimated without a fixed set of parameters or distributional assumptions (Ghahramani, 2013, 8). We will get into more details about this later.

In synthetic population generation, the covariates for estimating the variable y are used to estimate the expression of a synthetic variable y_{synth}. Nevertheless, further assumptions of each model used must be taken into account, since these are not insignificant for the predicted synthetic values and their quality. The linear regression assumes, among other things, homoscedasticity, i.e. that the residuals are constantly scattered around the regression line. If we want to simulate heteroskedasticity in the data, the model has to be modified accordingly (Wooldridge, 2013, 849). Basically, these models do not serve in any case for a content-related analysis of the relationships, but exclusively for predicting the values of our synthetic population to be generated. For this reason, the basic principle of thrift, which applies to analytical models (Forster and Sober, 1994), may well be softened.

No matter what model we use, the basic idea that the values of a synthetic variable y_{synth} are predicted by one or more covariates remains basically the same and can be generalized. Let's take a look at another example. In the case of a dependent variable, which has binary categories, a so-called logistic model is estimated. It belongs to the class of *generalized linear models* (Hoffmann, 2003) which generalize the parametric and linear assumptions to situations with different statistical conditions. Applying a linear regression on a binary categorical variable would violate basic assumptions such as normally distributed residuals or homoscedasticity. To facilitate the interpretation of the logistic regression, we must use a function that limits the value range to the interval [0, 1] and makes the respective values interpretable as probabilities. Instead of a normal distribution, we assume a *Bernoulli* distribution and use

$$p(y \mid \mathbf{x}, \beta) = \mathrm{Ber}(y \mid \mu(\mathbf{x}))$$

where $\mu(\mathbf{x}) = \mathbb{E}[y \mid \mathbf{x}] = p(y = 1 \mid \mathbf{x})$. As before, a linear combination of inputs is calculated but this time we pass it through a link function to ensure $0 \leq \mu(\mathbf{x}) \leq 1$. The link function is the sigmoid function

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

leading to a new conditional density

$$p(y \mid \mathbf{x}, \beta) = \mathrm{Ber}\left(y \mid \mathrm{sigm}\left(\beta^T \mathbf{x}\right)\right)$$

and a new model equation

$$y(\mathbf{x}) = \beta_0 + \sum_{i=1}^{D} \beta_i x_i$$

for the logistic regression (Murphy, 2012, 21). For a parametric family of distributions $p(y|\boldsymbol{x};\boldsymbol{\theta})$ the best parameter vector θ has to be found. The model parameters are obtained using *maximum likelihood* instead of ordinary least squares, like in linear regression. The logistic regression then corresponds to the family

$$p(y = 1|\boldsymbol{x}; \boldsymbol{\theta}) = \sigma\left(\boldsymbol{\theta}^\top \boldsymbol{x}\right)$$

which is again a distributional assumption (Goodfellow et. al., 2016, 138).

For dependent variables with multiple categorical characteristics, multinomial models should be estimated (Alfons et, al., 2011). They rely on the assumption of 'independence of irrelevant alternatives' (Ray, 1973) which states that the relative probability of choosing one class over another is independent from the presence or absence of added additional possibilities. This assumption allows us to model the choice of K alternatives as $K - 1$ independent binary choices. Written very compactly, such a model takes the form of

$$p(y = K \mid \mathbf{x}, \beta) = \frac{\exp\left(\beta_k^T \mathbf{x}\right)}{\sum_{\mathbf{k'}=1}^{K} \exp\left(\beta_{k'}^T \mathbf{x}\right)}$$

which will result in several binary models (Murphy, 2012, 252). At this point we refer to further literature, which lists various other statistical models (see, for instance, Kroese and Chan, 2014). The statistical models presented are well suited for many research purposes and their assumptions have been thoroughly researched. Nevertheless, for various reasons we have to question if they are suitable for creating our artificial data set. Linear models in general have the obvious disadvantage that they do not capture interactions between covariates which we need to model the complex non-linear relationships between household members. They are known for not being able to learn the XOR function where $f([0,1], \boldsymbol{\beta}) = 1$ and $f([1,0], \boldsymbol{\beta}) = 1$ but $f([1,1], \boldsymbol{\beta}) = 0$ and $f([0,0], \boldsymbol{\beta}) = 0$ (Goodfellow et. al., 2016, 15). This is because they assume a function $y = f(\boldsymbol{x}; \boldsymbol{\theta})$ and have to adapt to θ. In summary, we can state that there is a need for models or methods that are not based on fixed parameters and assumptions, but adapt to the data set in order to estimate the underlying data generating process. In a very famous article, Leo Breiman (2001b) argued that there are 'two cultures' in statistical modeling. One side relies heavily on the assumption, that the data were generated by a model, the other treats the data mechanism as unknown. The latter property is particularly present in various machine learning methods, which we will now introduce.

3 Machine learning

3.1 Basic ideas of machine learning

Machine learning methods are an essential component of our research. First of all, let us clarify what machine learning actually is and why we hope to obtain good results from these methods for our research question. Goodfellow et. al. (2016, 96) define Machine learning as 'a form of applied statistics with increased emphasis on the use of computers to statistically estimate complicated functions and a decreased emphasis on proving confidence intervals around these functions'. This definition is especially helpful for us, because it is rather general; we often read that machine learning recognizes patterns in example data, which allows us to generalize the gained 'knowledge' to unknown data. The unknown data in our case is the synthetic population we have to generate.

What does the aspect of 'learning' actually refer to, and what does it mean in the context of our research question? A very famous definition by Mitchell (1997) states that 'a computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E'. The Task T to be solved is typically a problem which, due to its complexity, cannot be solved by a computer program written entirely by hand. Instead we give the method an example which is nothing else but a collection of features that have been measured, in our case, the covariates. An example is represented as a vector $\boldsymbol{x} \in \mathbb{R}^n$ with each entry x_i being a feature. In our case, the variables of interest in our sample are the information which the algorithm has to process. In both strategies a classification problem, which is the task T, shall be solved. Generally speaking, the algorithm should correctly assign an input to a category k by producing a function $f : \mathbb{R}^n \to \{1, \ldots, k\}$ or a function f which outputs a probability distribution over classes. Of course, this can be generalized to regression problems, by simply changing the function to $f : \mathbb{R}^n \to \mathbb{R}$ (Goodfellow et. al., 2016, p. 98).

To evaluate the performance of our learning algorithm, we design a performance measure P. In case of our classification problem, this is nothing else but the proportion of examples which were correctly predicted by the algorithm. It should be noted, however, that the sole consideration of accuracy can be problematic. Machine learning methods sometimes suffer from overfitting, which means that they adapt themselves too much to the observed data, so that a generalisation to data not yet seen before fails (Murphy, 2012, 22). This would be the case here if the algorithm only reproduced the exact households from the sample, which would result in the resampling problems mentioned above.

Machine learning algorithms learn by experiencing a dataset. Nevertheless, the researcher can decide what kind of experience E the algorithm can have. All methods used in the thesis belong to the so-called supervised learning, which means that the examples used are linked to a label. For example, as researchers we can specify the number of age categories, or generally decide on the inclusion of covariates. Already by selecting a dependent variable y and explaining it by a covariate x, i.e by estimating the probability function $p\left(y \mid \mathbf{x}, \boldsymbol{\theta}\right)$, we have given the algorithm a specification. In unsupervised learning, which is not considered here, the algorithm would have to derive a meaning from the data set without any guidance.

There, contrary to supervised learning, we would estimate $p\left(\mathbf{x} \mid \boldsymbol{\theta}\right)$ or, in a nonparametric case $p(\mathbf{x})$ which is unconditional density estimation (Murphy, 2012, 9).

The fact that some machine learning methods make no parametric assumptions should not be confused with that they make no assumptions at all or that parametric assumptions are only about distributions. Non-parametric models in general differ from their counterpart because their number of parameters is not fixed but grows with the amount of training data; a fact, which makes them often computationally intractably for large datasets (Murphy, 2012, 16). Limiting the size a decision tree can grow turns it in practice into a parametric algorithm, despite no distributional assumptions were made (Goodfellow et. al., 2016, 146).

The methods, which we will now present, are all part of supervised learning and estimate a density function $p(y \mid \mathbf{x})$. The strategies to be presented in chapter four will avoid unsupervised learning from the outset. What is the reason for this? If, as in our case, a sample is already available, the easiest way to estimate the joint probability distribution would be to use multivariate kernel-density estimation (O'Brien et. al, 2016). Unfortunately, most of these approaches fail in higher-dimensional data sets. Therefore a diversion has to be found.

3.2 Random forest

The term "random forest" was first used in an article by Ho (1995). For many years, the methodological development was repeatedly and sporadically pushed forward (Amit and Geman: 1997; Ho: 1998) in order to finally be systematically summarized and processed by Breimann (1999; 2001a). Our overwiev follows in form and structure mostly Hastie et al. (2017), exceptions are marked via citation. To understand the concept of random forests we first have to understand the idea behind decision tree learning and bootstrap aggregating ("bagging"). Bagging works by accumulating a whole range of noisy and uncorrelated models to average them. If the model to be used is a decision tree, then random forest is nothing more than a large collection of uncorrelated decision trees, with the output generated by the multiplicity of individual decision trees forming the result (Hastie et. al, 2017, 587).

Let us therefore first look at the mathematical details of a decision tree. Depending on whether a discret or continuous outcome should be predicted, a distinction is made between classification trees and regression trees. Often "Classification And Regression Tree" (CART) is used as an umbrella term referring to both methods. Both methods can be described mathematically similar and differ only at a later point. Let us begin with a mathematical description of regression trees which are strongly linked to the description in Hastie et. al. (2017, 305-309).

Decision trees in general divide a feature space into rectangles, where each rectangle contains a simple model. Consider a regression problem with a dependent Variable y called *reponse* and i covariates x_i called *input*. If we divide the feature space, in each part the variable y is modeled by a different constant c_m. The already divided regions can be further divided until a stop rule ends this process. Imagine a data set having N observations while every observation consists of p inputs and a response. Let the response be (x_i, y_i) for $i = 1, 2, \ldots, N$, with inputs $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$. Our algorithm has to split the variables automatically as well as to decide on the split points. It decides to divide our data into M

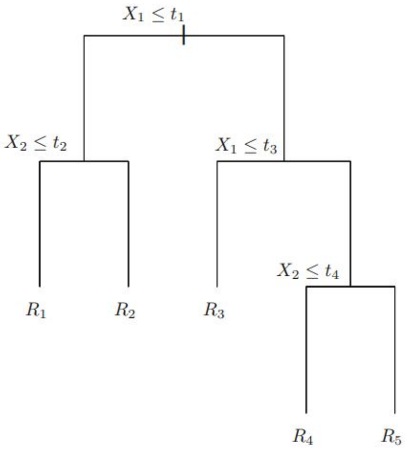


Figure 1: A tree partitioning a feature space (Hastie et. al, 2017, 308)

regions $R_1, R_2, \ldots, R_M$. In each region, the *response* is modelled as a constant c_m as follows

$$f(x) = \sum_{m=1}^{M} c_m I\left(x \in R_m\right)$$

using the sum of squares as criterion for minimization. Then, the best $\hat{c}_m$ is simply the average of y_i in region R_m

$$\hat{c}_m = \operatorname{ave}\left(y_i \mid x_i \in R_m\right).$$

We proceed finding the best binary partion with a computational reasonable effort. Therefore, we introduce a splitting variable j and a split point s which define the pair of half-planes.

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

Then, the equation

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

has to be solved by determining the correct splitting variable j and split point s with the inner minimization solved by

$$\hat{c}_1 = \operatorname{ave}\left(y_i \mid x_i \in R_1(j,s)\right) \text{ and } \hat{c}_2 = \operatorname{ave}\left(y_i \mid x_i \in R_2(j,s)\right).$$

Now the question remains, however, how large the decision tree should grow. As long as the splitting process is not stopped, it will continue to grow. The first step is to define the node size that should stop the growth. In the second step, the tree will be pruned, which can be described mathematically as follows.

Let a tree T_0 grow till it reaches a pre-defined node size as explained. Define a subtree $T \subset T_0$, which is any tree that could be obtained by pruning T_0. Terminal node m represents Region R_m. Let $|T|$ be the number of terminal nodes in T. Assume

$$N_m = \# \{x_i \in R_m\}$$

$$\hat{c}_m = \frac{1}{N_m} \sum_{x_i \in R_m} y_i$$

$$Q_m(T) = \frac{1}{N_m} \sum_{x_i \in R_m} (y_i - \hat{c}_m)^2$$

and define the cost complexity criterion

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

with the intention to find subtree $T_\alpha \subseteq T_0$ for each α to minimize $C_\alpha(T)$. Tree size and prediction accuracy are directly influenced by the tuning parameter $\alpha \geq 0$. Setting $\alpha = 0$ would result in the full tree T_0. We now modify a bit the criteria for splitting nodes and pruning the tree to change the algorithm for classification problems. Details are given in Hastie et. al. (2017, 309-310).

The previously mentioned "bagging" which is essential for the random forest works by drawing bootstrap samples with replacement from the training data to construct each tree and combine them to reduce generalisation error (Breiman, 1994). Technically speaking, each sample constructs a dataset k and applies a model i, each producing a probability distribution $p^{(i)}(y \mid \boldsymbol{x})$. By calculating the arithmetic mean over those distributions as follows

$$\frac{1}{k} \sum_{i=1}^{k} p^{(i)}(y \mid \boldsymbol{x})$$

we obtain the prediction of the ensemble (Goodfellow et al., 2017, 262). However, the resulting reduction of variance can be extended by reducing the correlation between the ensemble of trees; this is the constituting idea of random forest (RF) (Segal and Xiao: 2011, 79). It draws a subset of predictors at random; this additional randomness enables very good prediction accuracy.

Translated into the language of mathematics, a RF is a collection of tree predictors $h(\mathbf{x}; \boldsymbol{\theta}_k), k = 1, \ldots, K$ with x being the observed predictor vector of length p. It is associated with a random vector $\mathbf{X}$ and $\boldsymbol{\theta}_k$ being independently and identically distributed (i.i.d) random vectors. Assume the data to be independently drawn from a joint distribution $(\mathbf{X}, Y)$. Let the unweighted prediction over the collection of trees be

$$\bar{h}(\mathbf{x}) = \frac{1}{K}\sum_{k=1}^{K} h\left(\mathbf{x}; \boldsymbol{\theta}_k\right)$$

with $k \to \infty$ ensuring the Law of large numbers resulting in

$$E_{\mathbf{X},Y}(Y - \bar{h}(\mathbf{X}))^2 \to E_{\mathbf{X},Y}\left(Y - E_\theta h(\mathbf{X}; \boldsymbol{\theta})\right)^2$$

and let the tree be unbiased for all $\boldsymbol{\theta}$ which means nothing else but

$$EY = E_{\mathrm{X}} h(\mathrm{X}; \theta)$$

as shown by Segal and Xiao (2011, 80). Given as pseudo-code, prediction using RF functions as follows:

1. For $b = 1$ to B:
   (a) Start with the training data and draw a boostrap sample $\mathbf{Z}^*$ of size N
   (b) Repeat the following three steps until the minium node size n_{min} is reached. This results in a random forest tree T_b.
      i. Consider p variables and select m variables from it
      ii. Pick best split point amont m
      iii. Node is split into two daughter nodes
2. Result is an ensemble of trees $\{T_b\}_1^B$

Making (regression) predictions works by using B such trees $\{T\left(x; \Theta_b\right)\}_1^B$ and calculating

$$\hat{f}_{\mathrm{rf}}^B(x) = \frac{1}{B}\sum_{b=1}^{B} T\left(x; \Theta_b\right)$$

as it's been shown by Hastie et. al. (2017, 587-590). Random Forest has the advantage of requiring little tuning effort. In most cases, standard hyper parameters lead to similar results.

3.3 Neural Network

Neural networks have enjoyed great popularity in the last years, despite the fact that the field of deep learning is much older than commonly assumed (Goodfellow et al, 2016, 12). This is due to the fact that they easily solve many problems which were considered unsolvable not long ago. They are often successfully used in image recognition or natural language processing. For classification problems, their strengths lie in their ability to deal easily with complex relationships of several variables among each other as well as highly successful tuning procedures to avoid over and underfitting. In the following we would like to give an introduction to the mathematical basics of a neural network, more specifically feed-forward neural networks. Obviously, this is such a large, complex and diversified area that some limitations have to be made. The following introduction will therefore be limited to the description of a very basic neural network in the context of the specific problem to be treated. We mostly follow the descriptions given in Higham and Higham (2018).

A fundamental building block of a neural net are the *neurons*. A neuron takes a proportion i of inputs $x_1, x_2, ..., x_i$ which results in a single output. To calculate the output, every input has to be multiplied with a weight w_i

$$\begin{aligned} &x_1 \rightarrow x_1 * w_1 \\ &x_2 \rightarrow x_2 * w_2 \\ &... \\ &x_i \rightarrow x_i * w_i \end{aligned}$$

which's task is to determine the importance of the related input. Let the neurons output be a binary variable taking the values 0 or 1 and assume an arbitrary *threshold value*. Then the output of the neuron is

$$\text{output } = \begin{cases} 0 & \text{if } \sum_i w_i x_i \leq \text{ threshold} \\ 1 & \text{if } \sum_i w_i x_i > \text{ threshold} \end{cases}$$

At this point we introduce the concept of a *bias* b which measures the likelihood to produce a certain output. We replace it with the aforementioned threshold and obtain a new output which is

$$\text{output } = \begin{cases} 0 & \text{if } w \cdot x + b \leq 0 \\ 1 & \text{if } w \cdot x + b > 0 \end{cases}$$

but we are still not satisfied with it. We would like an output with a predictable form. The elements are now passed trough an activation function $y = f\left(x_1 * w_1 + x_2 * w_2 + b\right)$ which in our case is the very often used sigmoid-function, which is defined by

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

and using it as our activation function changes the neurons output to

$$\frac{1}{1 + \exp\left(-\sum_i w_i x_i - b\right)}$$

as shown in Nielsen (2019). So far, this has only described the basic idea of a neuron but a neural net is way more complex than that. We increase complexity by introducing so called

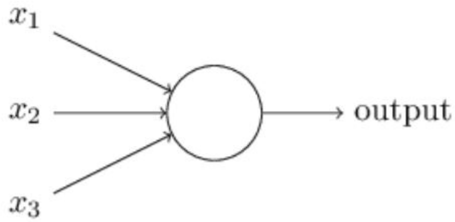


Figure 2: A neuron taking inputs - The basic unit of a neural net (Nielsen, 2019)

layers of neurons (Higham and Higham, 2018). Let $z \in \mathbb{R}^m, \sigma : \mathbb{R}^m \to \mathbb{R}^m$ be the application of the sigmoid function in the described manner that

$$(\sigma(z))_i = \sigma(z_i)$$

holds. Every neuron in each layer gives an output which is handed over to a neuron in the next layer. There, each neuron recalculates the values and applies the sigmoid function. We collect the real numbers produced by a neuron in one layer and store them in a vector a. As a result, the form

$$\sigma(Wa + b)$$

describes the vector of outputs from the next layer with W being a matrix containing the *weights* and b being a vector containing the *biases*. For the ith neuron, the equation above is calculated by

$$\sigma\left(\sum_j w_{ij} a_j + b_i\right).$$

We now generalize these basics into a complete methodical and mathematical explanation for a general framework. Imagine a neural net having L layers and assume that layer l, for $l = 1, 2, 3, \ldots, L$ consists of n_l neurons, therefore, the network maps from $\mathbb{R}^{n_1}$ to $\mathbb{R}^{n_L}$. Moreover, let $W^{[l]} \in \mathbb{R}^{n_l \times n_{l-1}}$ be the matrix of weights at layer l with $w_{jk}^{[l]}$ beeing the weight which neuron j at layer l uses to evaluate the output from neuron k at layer $l - 1$.

Based on these preliminary considerations, we now find a final description of the way the network handles an input and produces an output. Assume an input $x \in \mathbb{R}^{n_1}$ with $a_j^{[l]}$ beeing the output from neuron j at layer l. Then, it is true that

$$a^{[1]} = x \in \mathbb{R}^{n_1}$$

and

$$a^{[l]} = \sigma\left(W^{[l]} a^{[l-1]} + b^{[l]}\right) \in \mathbb{R}^{n_l}, \quad \text{for } l = 2, 3, \ldots, L$$

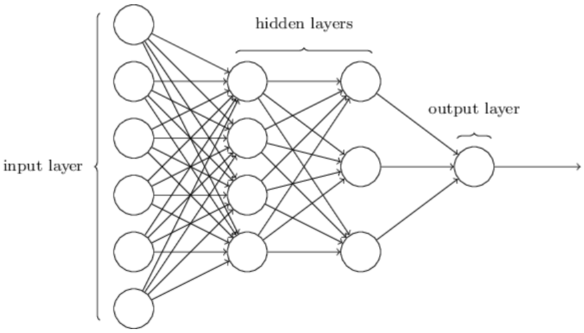


Figure 3: A four layer network with two hidden layers (Nielsen, 2019)

produce an output $\boldsymbol{a}^{[L]} \in \mathbb{R}^{n_L}$. The generalized cost function the network has to minimize is

$$\text{cost } = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{2}\left\|y\left(x^{\{i\}}\right)-a^{[L]}\left(x^{\{i\}}\right)\right\|_2^2.$$

In simple terms, weights and bias should be arranged in the network in such a way that an optimal solution is created (Nielsen, 2019). A solution which is able to correctly assign outputs for the training data will result in the cost function $Cost \approx 0$. However, this does not yet explain how the network "learns".

The algorithm that makes this possible is called *gradient descent* with the idea to compute iteratively a sequence of vectors (Higham and Higham, 2018). Let the current vector be p with perturbation Δp. Applying a taylor series expansion on the cost function we obtain

$$\operatorname{cost}(p+\Delta p) \approx \operatorname{cost}(p)+\sum_{r=1}^{s}\frac{\partial \operatorname{cost}(p)}{\partial p_r}\Delta p_r$$

with $\partial \operatorname{cost}(p)/\partial p_r$ referring to the partial derivative with respect to the rth parameter. Let the vector of partial derivatives be written as $\nabla \operatorname{cost}(p) \in \mathbb{R}^s$. How can we find the best Δp for letting the cost function be as close to zero as possible? With the cauchy-schwarz inequality we can show that Δp should lie in the direction $-\nabla \operatorname{cost}(p)$. As a result,

$$p \rightarrow p-\eta\nabla \operatorname{cost}(p)$$

holds with η being the stepzise, now called *learning rate*. Gradient descent can be understood as a step-by-step approach to reduce the cost function. Knowing that our partial derivative $\nabla \operatorname{cost}(p)$ is equal to the sum over the individual partial derivatives in the training data one can show that

$$\nabla \operatorname{cost}(p)=\frac{1}{N}\sum_{i=1}^{N}\nabla C_{x^{(i)}}(p)$$

is true. Since this method is computationally expensive, the stochastic gradient descent can be used as an alternative; more detailed explanations can be found in (Goodfellow et al., 2017, 151-155).

However, the explanations given so far do not allow for an exact calculation of the gradient decent of the cost function. Therefore we finally consider the algorithm "Back Propagation", which does this. It considers the *weights* ad *biases* to compute $\partial C/\partial w$ and $\partial C/\partial b$ for the cost function. Let assume that the cost function can be written as $C = \frac{1}{n}\sum_x C_x$ which is the average over cost functions C_x. We see that, for a single training example,

$$C = \frac{1}{2}\left\|y - a^{[L]}\right\|_2^2$$

holds. We choose this approach because Back Propagation allows to solve partial equations for single training examples, which is $\partial C/\partial w^l_{jk}$ and $\partial C/\partial b^l_j$. The solution of those one is our final goal. Lets introduce some notation to make this easier. Let

$$z^{[l]} = W^{[l]}a^{[l-1]} + b^{[l]} \in \mathbb{R}^{n_l}, \qquad \text{for } l = 2, 3, \ldots, L$$

where the weighted input for neuron j at layer l is indexed by $z^{[l]}_j$. We now introduce the concept of *error*. Let $\delta^{[l]} \in \mathbb{R}^{n_l}$ be defined by

$$\delta^{[l]}_j = \frac{\partial C}{\partial z^{[l]}_j}, \qquad \text{for } 1 \leq j \leq n_l \quad \text{and} \quad 2 \leq l \leq L$$

which is the *error* for the jth neuron in layer l. The final equation is derived by a highly complex application of the hadamard product, as well as transformations in a number of steps. However, this is beyond the scope of this subchapter, so we will only show the final formula here. Details are given in Higham and Higham (2018). The formula is

$$\frac{\partial C}{\partial w^{[l]}_{jk}} = \sum_{s=1}^{n_l} \frac{\partial C}{\partial z^{[l]}_s}\frac{\partial z^{[l]}_s}{\partial w^{[l]}_{jk}} = \frac{\partial C}{\partial z^{[l]}_j}\frac{\partial z^{[l]}_j}{\partial w^{[l]}_{jk}} = \frac{\partial C}{\partial z^{[l]}_j}a^{[l-1]}_k = \delta^{[l]}_j a^{[l-1]}_k.$$

In practice, it has often proved to be sufficient to use a one layer. Many problems can already be solved and the computational complexity as well as the tuning effort is limited. For this reason, many standard statistics packages, like *nnet* (Venables and Ripley, 2002) in R limit the number of hidden layers to one from the outset.

4 Two strategies for data generation

In the previous chapter the difficulty was worked out to estimate the true underlying data generating density function. The two strategies that we will now present serve to provide possible solutions to this problem. Let us assume in the following that the starting point of the two strategies is the sampling design described in chapter 2.3. It should be made explicitly clear that it is only a matter of enabling the creation and application of models in principle, not of the concrete design or performance of these models. This will be the subject of chapter 5. Of course, the approaches to be presented now could also be implemented with the help of statistical models instead of machine learning methods.

4.1 The combination approach

We have called the first strategy to be presented now ”combination approach”. This is loosely inspired by ideas from the field of microsimulation, as they occur in couple matching (Zinn, 2012), but also in household allocation.

The namegiving idea of the approach works by reallocating existing individuals in the sample into random households. Written as pseudo-code, we proceed as follows:

1. Create a dichotomous variable [0, 1] and assign ”1” to every household in the sample.
2. Randomly reallocate the individuals into c new households and code them with ”0”.
3. Bind the two datasets together and obtain a dataset with the number of rows equal to n + c = Z.
4. Apply a model on dataset Z which uses the dichotomous variable as dependent variable and predict it by using the $i \times j = m$ individual characteristics as covariates.

A model to be applied shall learn what distinguishes a ”true” (i.e an observed) household from an unrealistic (i.e randomly allocated) household. This also means, that for every household size a seperate model must be estimated where the number of covariates is equal to the number of members multiplied with the variables of interest. For instance, we have six covariates for a two-person Household with three variables of interest (AGE, SEX, MST) which would be AGE1, AGE2, SEX1,...,MST2.

The decision to create such a model with a dependent dichotomous variable has far-reaching consequences for the further procedure, more precisely, for the application of the model. To be able to decide if a household should be part of the synthetic population, it must first be presented with a selection of possible households from which to choose as well as a scheme to decide which one to pick. In order to avoid having to rely on sources of information other than the sample when applying the strategy, we develop a data generation mechanism that can be produced exclusively using sample information.

Continue the began pseudo-code from above and proceed as follows:

1. For each covariate, draw a value from it’s empirically observed probability density function $f(x)$ in the sample.

2. By binding the covariates, obtain a column vector which number of columns M is equal to the number of covariates.

3. Repeat the two steps above N-times and store the randomly created households in a matrix of size N * M.

4. Predict the probability P, that a household i from the matrix is a realistic one, by using the models described in the first part of the pseudo-code and calculate P_i for everyone.

5. Apply acceptance-rejecting-sampling: Draw random-number r from a univariate distribution in the interval of [0, 1]. If $r_i < P_i$ the household becomes part of the synthetic dataset.

6. If the resulting number of synthetic households is not big enough, repeat steps 1-5.

Through this random generation of households we hope to overcome the problem of sampling zeros. We will take a closer look at the advantages and disadvantages of this approach later.

4.2 The Modeling approach

The "modeling approach" opened in the following is based on concrete, previous work by various authors. Of particular note are the works of Caiola and Reiter (2010) and Drechsler (2010). Basically, the idea of sequential modelling is taken up, which was previously carried out using parametric statistical models. This approach can be generalized to machine learning.

We take our sample and use a variable y_2 of interest as a dependent variable to impute a variable y_{synth} for the synthetic data set. y_2 is regressed on by y_1 and y_3 by y_1, y_2 and so on. With regard to the modeling of households, this means that the characteristics of one member are considered dependent on the characteristics of the other members. In this way we hope to capture the complex dependencies within a household, although marginal distributions are estimated instead of the joint distribution. This results in numerous models, exactly one for each parameter of interest within a household. In contrast to the combination approach, the household size serves as an additional prediction variable in the model.

The model is applied to a data set of N rows, corresponding to the desired size of the synthetic population. At least one variable (y_1) is needed as a starting point. Otherwise, we would lack of a needed covariable and couldn't start the synthesizing. One way of generating this is to use the sample information to calculate the distribution of age among the oldest people in a household, for example, and then sample it N times. This can be done for all parameters of the oldest person or only for a specific one; in either case, it is then possible to apply the models. The synthetic data set now grows in each further step by one more parameter, that of the next person in a household, and so on. This is done by assigning a probability to each possible expression of the parameter of a person within the household. On this basis, a weighted random drawing then decides on the expression of the parameter.

There is no mathematical rule for the order in which the variables are created. Even if there is at least an order by age, the choice of which parameter (AGE, SEX, MST) is simulated first must be made arbitrarily. It can be assumed that different sorting will produce different results, but these should not differ too much.

4.3 Discussion of both approaches

Irrespective of the performance of the methods still to be discussed, decisive advantages and disadvantages can already be discussed at this point. The included data generating process of randomly drawing possible households and using acceptance-rejecting sampling evaluates the complete household structure as a whole. Additionally, the models could also be used to classify realistic households in a foreign dataset. For example, it would be conceivable to have the synthetic households of the model approach evaluated by the combination approach. Moreover, the model works regardless of whether we use categorical variables or continuous variables. Using the combination approach, any composition of a household, no matter where the data comes from, can be checked for realism. This is however at the same time also the largest disadvantage; such a selection of households to be evaluated must first be created somehow. Although this creation process can be carried out at will, it also influences the result, because only those households that are available for selection at all can be included in the synthetic population. At the same time this offers a chance for early logical editing, because values which should not be sampled can easily be excluded by setting their probability to 0. Although we have already proposed a useful mechanism, which also overcomes the problem of sample zeros, it is highly computationally intensive. First of all, possible households have to be generated (the first additional step compared to the model approach) and these have to be evaluated (the second step). If, as in our case, accepance-rejecting sampling is also used, the generation of the synthetic population can take even longer. This problem increases the better the model fit. A good model will sort out potential households more rigorously, thus prolonging the process. The more parameters of interest we want to simulate, the more random household structures will be possible, which might raise the dimensionality of the data to an unreasonable level. Basically, the combination approach is more computationally intensive than the modelling approach, which is a particular problem with large populations.

At this point the difference to the modeling approach becomes clear: A pre-selection of households to be selected is not necessary for the generation of synthetic households. After the algorithm has done its work, a complete synthetic population with desired N households is available in any case. A disadvantage is that the method was created for categorical data and the change to continuous variables requires some adjustments. In addition, a large number of separate models with different numbers of covariates have to be programmed and estimated, whereas the number of covariates remains the same in the combination approach and separate models for different household sizes have to be estimated.

5 Simulation setup

In this thesis we use a simulation study (see, for instance, Morris et al, 2019) to analyze how promising the discussed strategies and methods would probably be in a realistic situation. For this purpose two basic approaches can be chosen: Either one assumes a finite and superordinate true population, whose true parameters are represented as accurately as possible by the estimated parameters, or one draws from so-called superpopulation models, which often represent a class of distributions. The first one is called *design based*-simulation, the latter *model based*-simulation (Vink, 2016, 7). Our starting point is a single-stage cluster sample of households drawn from a finite population (Chapter 2.3), which is why we have chosen to draw from an already known synthetic dataset, which will mimic a true population. Based on this sample, the respective method must reconstruct the true population as best as possible. The success of this will determine the evaluation of it's performance.

Admittedly, it can be objected to this approach that it is only possible to draw conclusions about the performance of the data set used and that other true populations would possibly lead to completely different results. This problem, however, applies in principle to all simulation studies, including model-based approaches. An unambiguous answer could at most be given by a study whose actual population would be a collection of all possible populations. Since this is not possible, we limit ourselves to data sets that most closely resemble the typical populations that are actually used. It should also be noted that the 'no free lunch theorem' (Wolpert and Mcready, 1996) applies to machine learning. This theorem states that no method is fundamentally superior to any other method across all possible data sets. So it always depends on the concrete situation and simulation studies like this can only provide a reliable indication for a specific situation.

Statistical modelling and machine learning are basically only as good as the data used for it. The comparison between the different methods should show which methods can process the available information most efficiently. It cannot be expected that poor input data will lead to excellent results, so the comparison between the methods should always be considered relative. We will therefore investigate which methods, under the same conditions, can use existing data as efficiently as possible, i.e. as close as possible to the true distribution.

5.1 AMELIA

According to the previous remarks, a true population must already exist in order for the synthetic population to be compared. Therefore, we use the AMELIA dataset (Burgard et al., 2017). This data set fulfills some properties that make it especially useful for our purposes. On the one hand, the data set is freely available for reproductive research, thus ensuring the scientific theoretical requirement of intersubjective traceability and reproducibility (Popper, 1934, 18) of the results. Furthermore AMELIA is based on the EU-SILC data which leads to a dataset highly comparable to a realistic population.

AMELIA contains of far more variables than needed. In the context of the research question, we focus on three specific parameters that typically provide a high degree of interdependence between household members and are essential for a realistic population due to their relevance. These are age (AGE), gender (SEX) and marital status (MST). Since the

Variable	Categories/Range
AGE	0-80
AGE_Cat	1: 0-17 2: 18-28 3: 29-39 4: 40-50 5: 51-61 6: 62-72 7: 73-77 8: 77+x
SEX	1: Men 2: Woman
MST	1: Never married 2: married 3: seperated 4: widowed 5: divorced

Table 3: Variables used

Modelling approach can only process categorical variables without modification, a categorical variable with 8 values is created based on AGE. The first category includes persons between 0-17 years of age to allow a distinction from persons defined as adults. Afterwards, all further categories are completed in steps of 10 and 5 for last two categories (18-28, 29-39, ..., 73-77, 77+x). AGE will still be used by the combination approach. Table 1 gives an overview about the variables used.

Herewith we receive at first only individual data. Due to the additional household ID (HID) marking contained in the data set, the data set can be easily transferred to the wide format, which also allows the true distribution of households to be seen. In order to limit the number of models and the effort of the simulation, households with size >= 6 are combined into one category. This seems appropriate, as households with more than 6 members hardly exist in AMELIA in any significant number.

The question now is how to measure the structure of households. As we have seen in Chapter 2, the totals t_i can be calculated for a household i. For a continuous variable such as AGE, in our case, we can simply calculate the average (or higher moments) of the average age within a household. In the case of categorical variables, combining them results in all possible combinations, whose relative frequencies we measure. For example, in a two-person household there may be three combinations of genders (both male, both female, mixed). It should be noted that this reduces dimensionality, as it is not possible anymore to conclude who are the exact individuals which are men or woman. As the size of the household increases, so does the complexity of the internal household structure. Combining these combined variables again, now with other combinations of characteristics (e.g. the gender ratio in a household with the marital status of those involved) produces higher-dimensional distributions, which should also be compared with the true population. It should be said in advance that the correct modelling of these higher-dimensional distributions is particularly difficult. The reason for this can be easily illustrated. Already the possible number of one-dimensional manifestations of the variable MSTcomb in a 6-person household is quite high because it is relying on higher dimensional data (combining MST1, MST2, ..., MST6). We calculate the possible combinations C with repetition by $C'_k(n) = \left(\begin{array}{c} n+k-1 \\ k \end{array} \right)$ with k = 6 (household size) and n = 5 (number of categories MST) the result is 210 for MSTcomb. To calculate the proportion of possible combinations over multiple attribute's, we multiply 210 with 1716 possible combinations of age (AGEcomb) leading to 360.360 possible household outcomes for size six based only on these two variables. However, due to its relatively small size, our sample will not be able to cover all possible breakdowns with an adequate number of observations. In a cross tabulation, this would mean that many fields would have no observations at all. Nevertheless, the function estimated by the ML algorithm should in the end provide a realistic estimate for these cases as well. The problem is well known in mathematics as "Course of 'dimensionality' (Verleysen and Francois, 2005).

5.2 Requirements of a synthetic population

At best, the synthetic population behaves in all analyses as the real population would. This requirement can be specified more precisely in a number of respects. The following requirements (Kolb, 2013) should be made:

1. Distributional characteristics
2. Correlation structures
3. Clustering and stratification
4. Hierarchical congruence
5. Structural consistency
6. Disclosure control

We will essentially limit ourselves to the first point and will only touch on the second and fifth at most. The reason why only a fraction of the desirable properties are checked is that correct distribution properties are often a basic requirement for the other properties to be fulfilled as well. If the methods already fail at this crucial point, this would be a clear argument against their use. Moreover, population generation based on models is known to be able to preserve correlation structures in most cases - but often at the expense of distribution. This is the reverse problem of resampling, where the distributions are well mapped but the correlations are lost. However, this should not be misinterpreted as if the distribution properties were the only relevant parameters or that correlation structures shouldn't be checked.

The requirement of correct distributional characteristics can be relaxed a little, so that it is sufficient if the distributional characteristics are approximately the same:

$$F_{\text{Synth}}(\mathbf{x}) \sim \mathbf{F}_{\text{TP}}(\mathbf{x})$$

Measures of the central tendency are the arithmetic mean, the weighted mean, the median and the mode. Other moments between true and synthetic populations should also be compared. For categorical variables, those measures can not be applied which is why we switch to the Devation V_D to compare relative freuquencies of categories in variables. That means that

$$V_D\left(x_{\text{Synth },l}\right) \approx V_D\left(x_{\text{TP},l}\right)$$

with l being a number of variables $l = 1 \ldots$ vars for which the condition has to be fulfilled simultaneously.

Correlation structures are usually investigated by comparing the covariance-matrix of true and synthetic variables

$$\operatorname{cov}\left(X_{1,\text{TP}}, \ldots, X_{k,\text{TP}}\right) = \operatorname{cov}\left(X_{1, \text{ synth }}, \ldots, X_{k, \text{ synth }}\right)$$

which works fine for continuous data. For categorical data, some other criteria have to be used.

We would also like to pay special attention to the issue of structural consistency (Kolb, 2013, 52-53). As already mentioned in chapter two, when using a sample as a starting point, there is a risk that elements which do not occur in the sample will not occur in the synthetic population - called sample zeros. However, there is also the risk that properties are modelled in the synthetic population that are not present in the sample or in the true population. We call this structural zeros. One way to address this problem is to use editing rules. A structural zero, such as the five babies in a five-person household mentioned above, could be prevented by logical rules defined in advance. In a particularly qualitative data set, editing rules would not be necessary at all. The higher the proportion of structural zeros and the editing effort, the worse the performance of the synthetic population is evaluated.

5.3 Evaluation criteria

Measuring the difference between household distributions in the AMELIA dataset, called Q, and those in the synthetic population, called P, is easier said than done. What we actually want is to represent both data sets as a whole with a suitable metric, which is then compared. Unfortunately, there is still no unique solution for such a problem. However, there are many approaches to compare different distributions and we have to choose one. We first decide to examine households separately according to the number of their members. We do this because the different number of members leads to more or less variables at the individual level. In other words, large households are at a different dimensional level than small ones.

In the first step, we check to what extent variables resulting from the combination of household members meet the defined requirements. Here, we are initially concerned with one-dimensional aggregate data (AGEcomb, SEXdiff, MSTgroup) and their deviation. Let both a synthetic and a true variable x have k categories with quantity of all possible outcomes being $\mathcal{T} = \{x_1, x_2, \ldots, x_k, \ldots\}$. The deviation is then nothing else but the difference between the relative frequency's in $\mathcal{T}$ for both aggregate variables. The value range varies between 0 (no deviation) and 100 (maximum deviation). More difficult is the verification of higher-dimensional distributions. Although it is certainly possible to define criteria for a household, such as ”Mixed sexes, one person divorced, one person younger than 18 years of age”, it would be possible to define considerably more definitions of a household. So we have to decide either to find a set of definitions of a household that are as reasonable as possible but ultimately arbitrary, or to opt for a complex mathematical comparison between two multivariate distributions. However, a comprehensive comparison between two multivariate distributions is difficult to make at this stage of research. There are possibilities for comparison, such as the Kullback-Leibler divergence, but this is no longer reliable with increasing dimensionality. One possible solution could be to use kernel embedding of distributions, where a probability distribution is represented as an element of a reproducing kernel Hilbert space (Smola et. al, 2007). However, this is by no means trivial and would go beyond the scope of this paper, but should be considered for future comparative research.

The verification of correlations will be done by means of paired correlation checks according to pearson's r (Pearson, 1895)

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

where *cov* is the covariance, with σ_X the standard deviation for variable X and σ_Y for variable Y. For categorical variables like MST and SEX we will use Cramer's V

$$V = \sqrt{\frac{\chi^2}{n \cdot (\min[J,K] - 1)}}$$

with J being the number of categories in variable x and K the ones in variable y (Cramer, 1946, 282).

Finally, we define some simple editing rules and check how many households had to be corrected using this rule. The following rules apply to all synthetic households with at least two people:

1. A household must consist of at least one person older than 17 years.
2. Persons under 18 years of age always have the marital status ”single”.

5.4 Settings

Let us first assume that the complete implementation of the described strategies and methods with their respective resulting synthetic populations represents a simulation run r. This in-

Package	Function	Weigthing possible
randomForest	randomForest	No
nnet	nnet	Yes
sampling	srswor	Yes
data.table	dcast	No

Table 4: Overview of used packages in R

cludes drawing a single stage cluster sample of size n from the synthetic population AMELIA with N households and M individuals, applying one of the two strategies and using exclusively one of the described machine learning algorithms. The resulting synthetic population contains synthetic variables of interest whcih can be viewed as an estimate $\hat{\theta}$ for the true parameter θ. However, multiple simulation runs are extremely computationally intensive. Therefore, we limit ourselves to one single synthetic population.

The conditions we choose are:

- Draw a 1 percent single stage cluster sample out of AMELIA (n = 37813)
- Use Two strategies (Combination/Model approach)
- Apply different machine learning algorithms (RF and NN) and logit/multionimial statistical models for comparison

leading to $1 \times 2 \times 3 = 6$ different Simulation scenarios with 20 results, considering the subgroups of household sizes.

A number of packages are used to execute the methods. Table 4 gives an overview of these and whether it is possible to define weightings. This is particularly relevant in cases where complex surveys are used.

The data generating process developed depends on how many random household combinations c are defined based on the sample. If c is increased, the model fit converges towards it's optimum. For our Simulation, we have decided to always generate a proportion of random combinations so that the total data set Z consists of exactly 300,000 households. This determination is arbitrary, but leads to a consistently high model fit.

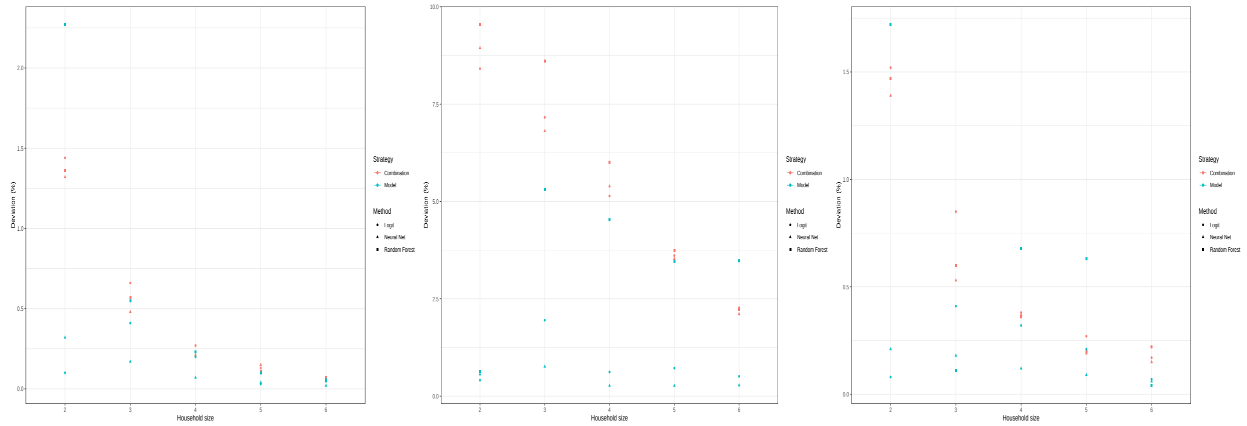
Furthermore, some specifications have to be made regarding the hyper parameters of the models. The algorithm of Breiman was executed via the R package *random Forest* (Liaw and Wiener, 2002). We choose standard tuning parameters, like 500 trees to constitute a random forest or using sampling with replacement for bootstrapping. All other possible hyperparameters are defined within the package. For the neural net, *nnet* only allows for single-hidden-layer neural networks. Nevertheless, this makes the neural net way less complicated to tune and leads to several advantages over a multiple-hidden-layer network (Nakama, 2011). Standard hyperparameters as they are described in *nnet* were used with one major exception: The number of units in the hidden layer was set to four. This number was chosen after the function *nnettrain* from the package *caret* (Kuhn, 2008) was used to select the best hyperparameter between a size from 1 to 10. Not all possible models could be checked for this, but the number 4 proved to be the best value in most cases to achieve the highest possible model fit.

6 Simulation results

6.1 Results

Let us first look at how much the synthetic populations resemble the AMELIA population considering the aggregated variables (Figure 4). The Y-axis shows the deviation in percent, while the X-axis differentiates the results by household size. Different strategies are indicated by the colours red and blue, while the shape of the symbols distinguishes the methods used (Logit/Multinomial Model, Random Forest, Neural Net). The Y-axis must be considered individually in each case, since the deviation was different in size depending on the variable under consideration.

Overall, the estimation of all three previously defined aggregate variables shows a similar picture, even if the quality of the estimate differs at one point. The first finding is that the model approach produced the more accurate results at this level overall. This is particularly true for the neural network: in almost all cases, this led to more accurate estimates than those of the random forest. Interestingly, the neural network also almost always produced better results in the combination approach, even if these do not come close to the quality of the neural network in the model approach. In fact, sometimes the combination approach even performed better than the random forest estimate in the model approach. Thus, regardless of the strategy used, the random forest in most cases led to significantly worse results than the use of a neural network. Another observation was that the results of the combination approach differed only very slightly and no clearly superior method could be identified for this strategy. We will examine this uniformity of results in more detail later. A positive surprise was the multinomial model in the model approach. It performed significantly better than the random forest and often came very close to the results of the neural net. The highest deviation was observed in the gender composition of a household, especially in the combination approach. Overall, for all aggregate variables, the deviation decreased with increasing household size. This result will also be the subject of further analysis.



(a) Age (b) Sex (c) Marital Status

Figure 4: Deviation in aggregate variables for different Methods

To better understand these initial results, we create a variable that measures the standard deviation of the age of household members from the respective average age in the household (Figure 5). The values were calculated separately for the different household sizes. The five resulting graphs show density plots, with the three estimates from the combination approach in addition to the true distribution. Basically, the Combination approach must work with the data we have previously generated using the proprietary data generation process. This uniformity applies to all household sizes. Except for HHS6, all distributions also deviate significantly from the true distribution. This may be due to the nature of the distributions of the combination approach: If we look at them more closely, they tend to resemble a normal distribution. However, if the true distribution is not normal, the Combination approach is unable to reflect this correctly. At this point, doubts must be expressed as to whether the individual methods, based on the same data-generating process, even lead to interesting differences. If not, this would be a reason to look at this data generating process in particular again. While our methods in the combination approach are able to estimate whether a household is based on the true distribution, they cannot accurately estimate the frequency of a realistic household. A household composition that occurs 1000 times is just as realistic as one that exists only once. Because our data-generating process, based on the marginal distributions of individuals, randomly throws dice, the Central Limit Theorem (see, for instance, Johnson, 2004) takes effect. However, if the actual distributions are not normal, the estimate is strongly distorted. The true distribution of age deviations in bigger households, in turn, tends to be normally distributed, which could be why the synthetic bias decreases significantly in larger households.

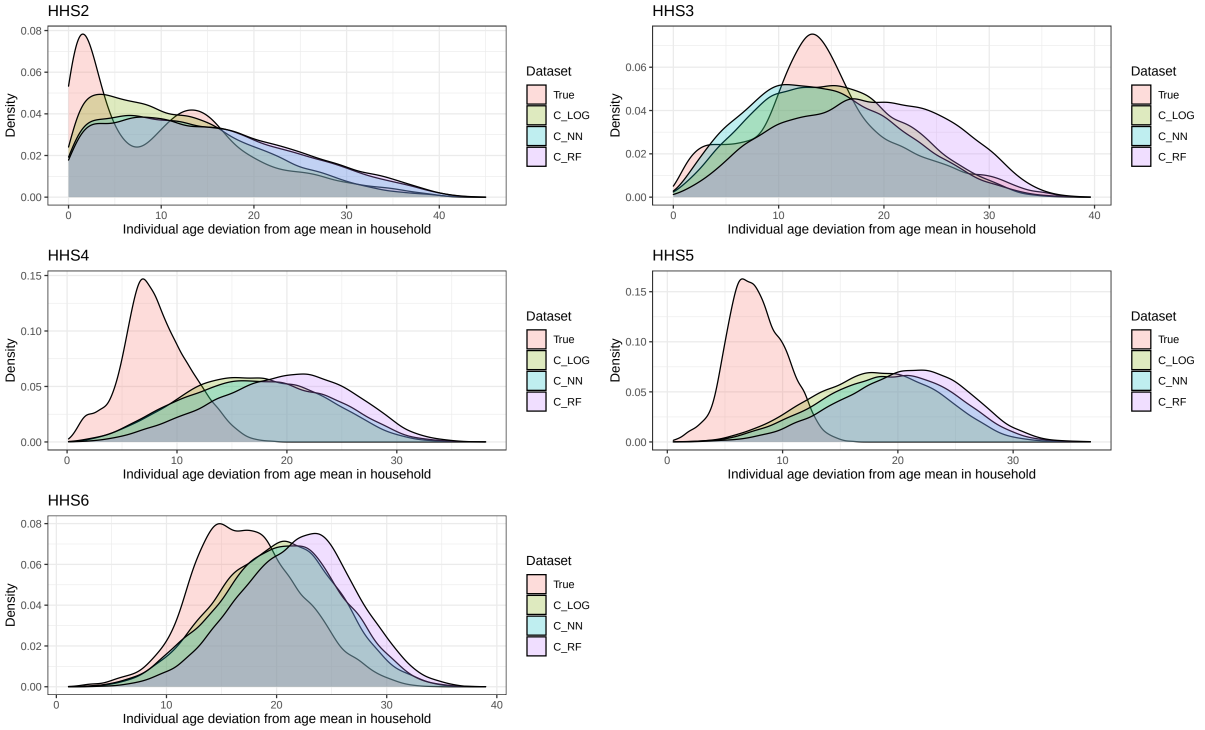


Figure 5: Density of age deviation from mean age in households

The above arguments are supported by further empirical results. First, we would like to look at a number of bivariate distributions. For this purpose we consider the paired frequencies between individual attributes of individual members, such as the age between the first, second, etc. person of a household. Such comparisons can of course also be made between different parameters, for example between the marital status of person 1 and the gender of person 2. The larger the household, the more combinations are possible. Unfortunately, due to the large number of possibilities we have to limit ourselves to an arbitrary selection. But for HHS2, it is possible to show at least all pairings between members with the same parameter of interest. Two-dimensional frequency tables can be depicted with the help of mosaic plots. The size of the areas illustrates the relative proportion within the table. The advantage is that large deviations can be recognised at a glance.

So let us first look at the pairwise comparisons for size two households in Figure 6. Especially for the distribution of age between the members there are considerable differences as well as interesting patterns. First of all, it can be said that the even distribution of relative frequencies among the individual categories, as we have seen so far with the combination approach, continues for those strategies. Although C-LOG differs slightly more than C-NN from C-RF, the basic pattern is pointing in the same direction. For M-LOG and M-NN, both distributions strongly resemble the true distribution. However, this cannot apply to M-RF, which produces a significantly different synthetic distribution. This first result raises the question whether a simple multinomial model might be as suitable as a neural network, or whether these are only positive outliers. In the following we will consider whether these results continue. The paired martial status is again only correctly represented by M-LOG and M-NN. Surprisingly, the results of the C-series are all closer to the true result than M-RF. This confirms the suspicion that there are true distributions which ”accommodate” the data generating process in the combination approach. At the same time, the result again proves that M-RF is obviously not suitable for estimating the true distribution. Only in the sex variable the whole M-series showed a very good result. Here it is particularly evident that the results of the C-series not only hardly differ, but that the relative frequencies are very evenly distributed.

Are these results tenable for larger households? Let’s look at Figure 7, which this time is limited to the analysis of age, as this is where the greatest difficulties have been encountered. Also, there are more pairs to compare. For the C-series, the arguments already mentioned apply again, which is why we will not consider them for further analyses of this kind. For the M-series, similar observations can be made as before, but there is one exception: For the comparison between the first and third person, M-LOG shows minor inaccuracies while these do not exist for M-NN. This is the first indication of possible weaknesses of M-LOG.

Let us now make a separate analysis for marital status for the same household size. This time we will look only at the two most promising methods, M-LOG and M-NN. Here, in Figure 8 we find strong differences between the true distribution and M-LOG, especially between persons 2 and 3 and 4 on the other side. Apparently some categories of the table have not been filled in at all, making others clearly too large. M-NN, on the other hand, again consistently shows very good results.

So far, it seems that the problems discussed are mainly related to the age of the persons. In size four households (Figure 9), however, the Marital status is also sometimes very incorrectly reflected and similar patterns are observed as in size two households. M-RF is again the furthest from the true distribution. M-LOG had problems in estimating the smaller relative frequencies that occur, which is particularly evident when comparing persons 1 and 3. The

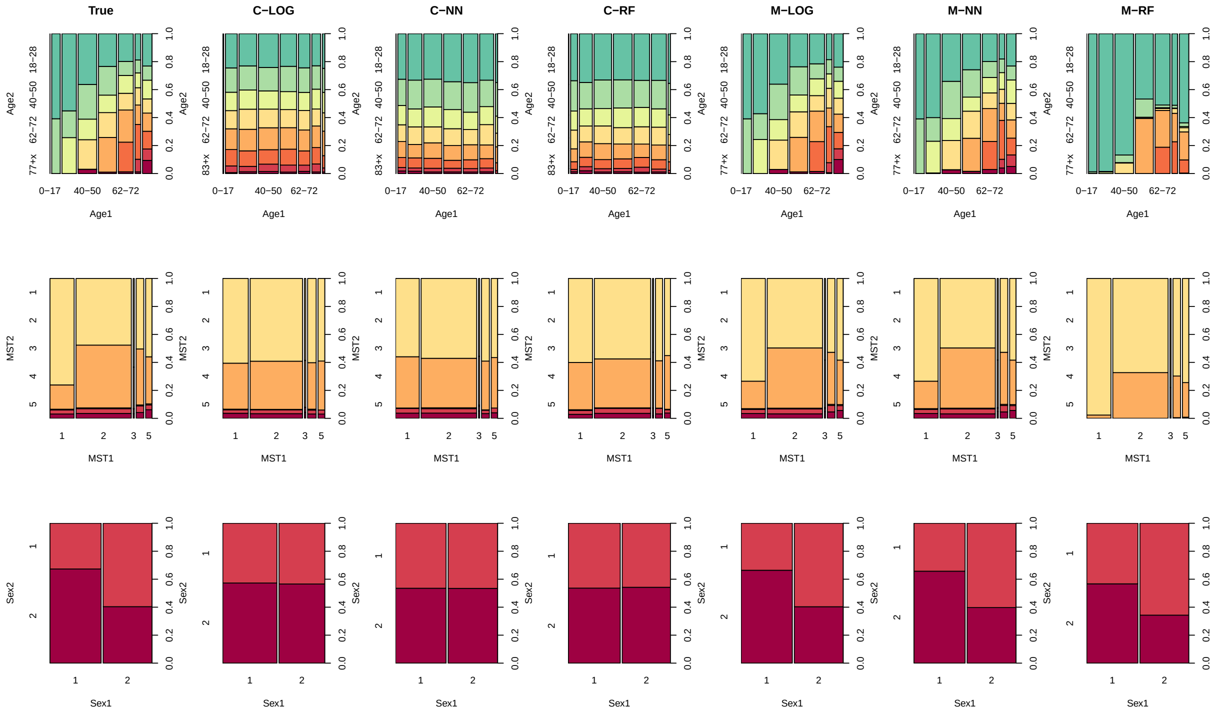


Figure 6: Distribution of parameters between household members (HHS2)

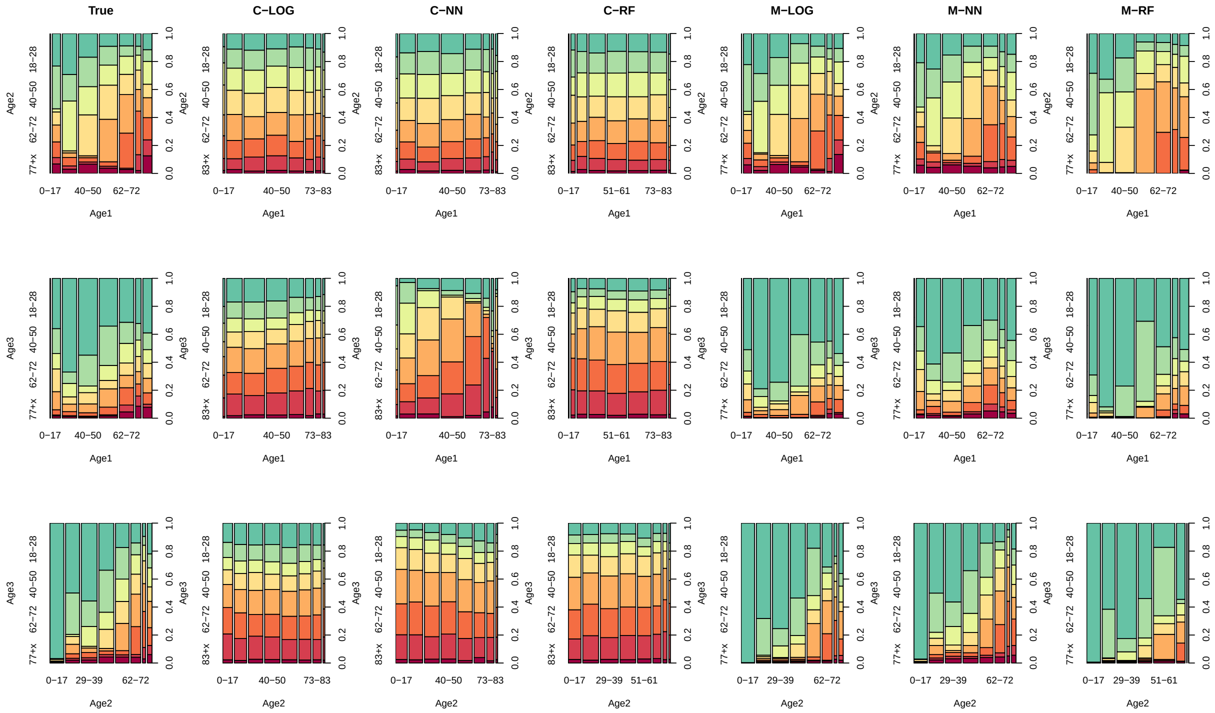


Figure 7: Distribution of categorical age between household members (HHS3)

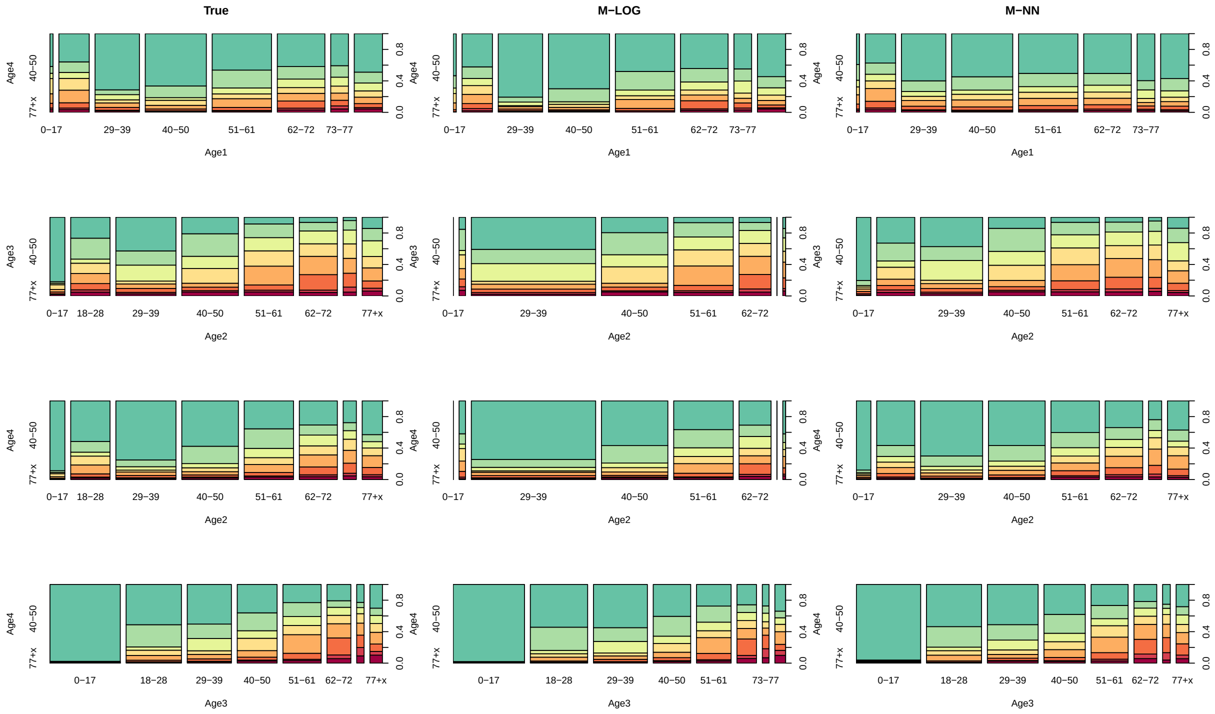


Figure 8: Distribution of categorical age between household members (HHS4)

C methods come closer to the true distribution here, but again an exact match could not be observed. Despite minimal inconsistencies, M-NN again proved to be the best method.

The last bivariate distribution we look at concerns the martial status of size 6 households, for which we have chosen four possible comparison pairs this time (Figure 10). These showed that all M-methods reproduced the true distribution well. This is in line with previous results, which showed a decrease in bias with increasing household size. The distributions of the C-methods are also closer to the true distribution than before, but show greater deviations in detail.

In the following, we will now look at the number of unique households in the populations. From this, conclusions can be drawn as to how many households were actually ”re-modelled”. Since we used only categorical variables for the purpose of dimensional reduction, these can be easily determined without minimal deviations making a comparison difficult. Table 5 first tells us how many elements per household are in the real population (N) and how many in the sample (n). Then three values were recorded for each method used: ’U’ indicates how many unique households were generated, ’IP’ shows how many of them are in the true population and ’IS’ tells us how many were in the sample. The relative share is expressed as a percentage, where ’U’ and ’IP’ refer to the true number of unique households, and ’IS’ refers to the share relative to the sample. In the best case, all three values would be 100 percent. However, there may be rigid differences between ’U’ and ’IP’ and IS’, which must be interpreted precisely. If ’U’ exceeds the true number of unique households, more unique households have been created than actually exist. However, this does not tell us anything about whether these households are not only unique but also consistent with the true population. For example, there may be twice as many unique households as in the real population, but none of them match the

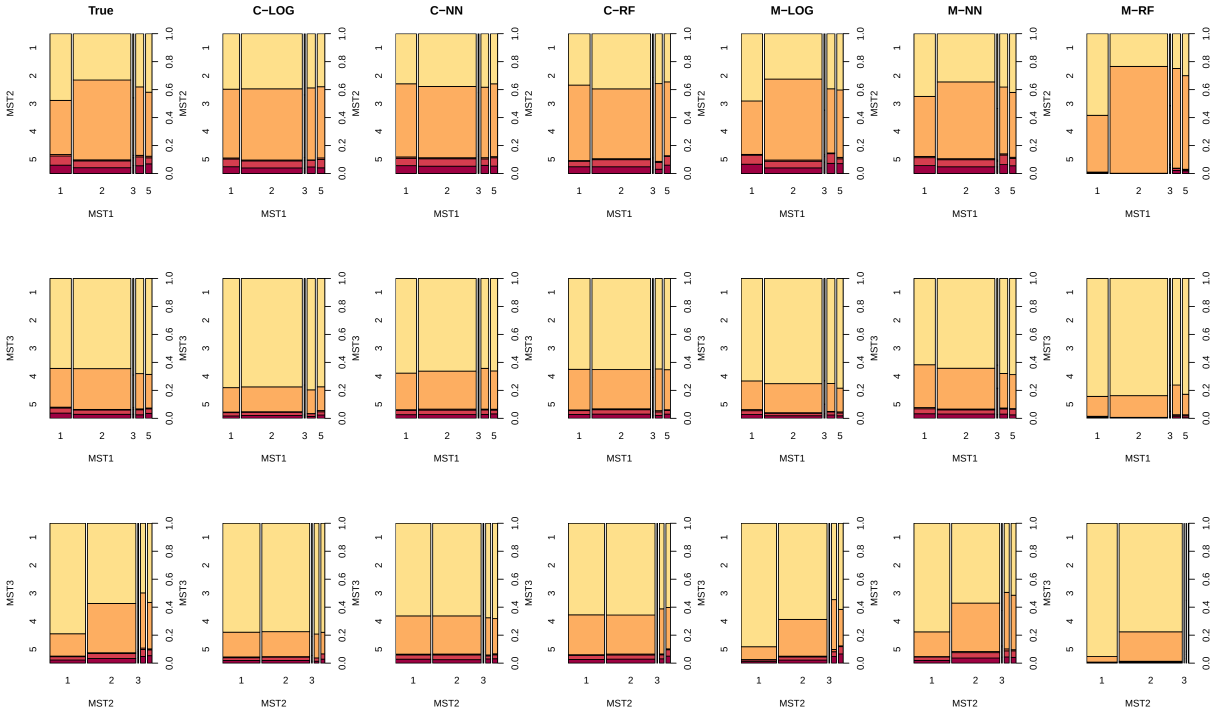


Figure 9: Distribution of marital status between household members (HHS4)

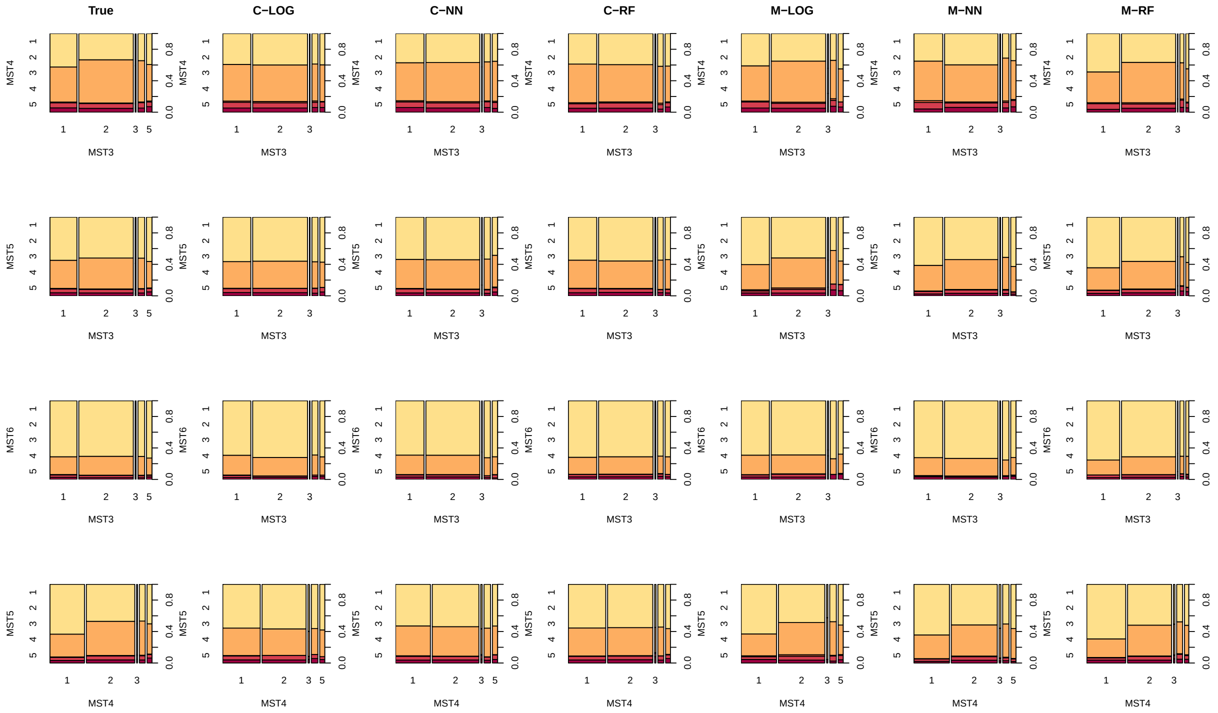


Figure 10: Distribution of marital status between household members (HHS6)

real population. Nor do we receive information about how much a non-unique household differs from a true unique household. Especially in the case of large households with many dimensions, even slight deviations can prevent an exact match. In the case of M-NN, the results so far indicate that the distributions have been reproduced very well and that the additional number of uniqueness of households is so rare or so slightly different from the true households that it is not relevant to the overall results. In short, surplus unique households should not be interpreted as problematic structural zeros. To this end, we will carry out a separate analysis below to determine the proportion of households that have breached the rules we have defined.

The exact number of unique households was not correctly estimated by any strategy or method, but once again a clear pattern is emerging. The synthetic data sets generated using the combination approach consistently overestimate the true number of unique households. Also, their results do not differ that much, which supports the idea, that the data-generating process itself is the problem and that the methods are not able to distinguish an often occurring household from a seldom one. Common to all methods is that the relative proportion of true unique households decreases sharply as the size of the household increases. The combination approach methods start with a share of 87 percent of size two households, but tend towards zero from size 5 onwards. A similar trend can be observed in the model approach, even if there are differences in detail between the methods, which still need to be discussed. In principle, these partial results are not surprising. On the one hand, as the size of the household increases, the number of elements in the sample, and thus the information content, decreases, while at the same time dimensionality increases exponentially. In the case of large households, it hardly seems possible to reproduce the exact composition of the household. Here the question must be asked as to the extent to which this is a problem and how strong the individual deviations are in detail.

Looking at 'IP', M-NN showed the best results. Although the previously discussed decrease in relative share is also progressing, the method consistently shows the highest share of matching households. Depending on the type of household, between 20 and 50 per cent of unique household members are created too much. M-LOG shows similar results, but differs in detail in that fewer surplus unique households are produced, while at the same time 'IP' is worse than the neural network. The results of the random forest must be considered very unreliable. For households of sizes two to four, the worst results for 'IP' by far are available. At the same time, this was the only method that, with one exception (HHS6), produced too few unique households. It is therefore suspected that the random forest reproduced elements from the sample too frequently instead of re-modeling an appropriate number of households.

It is quite difficult to interpret the importance of the proportion of households in the synthetic population that match those in the sample ('IS'). For logical reasons alone, this value must be 100 percent if 'IP' also takes on the value 100, because the elements of the sample are always a subset of the elements of the population. Nevertheless, a high value here does not necessarily have to be something positive; it is conceivable that the elements of the sample were reused, but that no modelling beyond this took place. The example M-RF also shows that an inverse relationship cannot be observed either: For households of size two, M-RF underestimated both the share of unique households and 'IP'. The 'IS' had a low value, which suggests that the random forest took only half of the sample into account and then reproduced it inappropriately often. To better understand these results, let us look at Table

6. This table shows how often the most frequently repeating households appear in the real population, the sample and the synthetic populations. Households in the C-series repeat far too seldom, and between the sizes 2 to 4, they often correspond to only between a third and a quarter of the true frequency of repetition. Although it has so far appeared that the synthetic populations of the C-series only differ to a limited extent, this table shows a somewhat more differentiated picture. The random forest tended to repeat households more often instead of modelling new ones. Since, as mentioned above, the C-range generally creates too many new households, Random Forest actually performs best in this area. However, this becomes a problem in the M-series. Here individual households are repeated ten times too often. This fits in well with the results in Table 5, where Random Forest modelled less unique households than there actually are. These results confirm the fears of the previous analyses. The best results were again achieved by M-LOG and M-NN, with M-NN being slightly closer to the true values overall. Only with HHS4 did M-LOG perform better.

However, the question still remains as to how many of the generated households are to be assessed as structural zeros according to our rules defined in chapter 5.3. Table 7, which shows the share of households that violate the first rule (FR) or the second rule (SR), excludes this question. In general, it can be seen that FR was violated only very rarely. C-LOG seems to have been the only method that produced structural zeros by violating FR in any significant way. Most striking, however, is that the C-series violates SR extremely often, while the M-series produces far fewer such violations. The share of structural zeros in the total number of households varies in the C series between 17 and 41 percent, with C-RF showing by far the worst values. The M-series values range from 0 to 5.45 percent, a massive difference. Again, the order is similar: most structural zeros are produced by M-RF, M-LOG scores significantly better and M-NN is the best. Thus, with the methods of the M-series, significantly fewer households would have to be logically edited afterwards.

Next, we consider how well higher-dimensional distributions were estimated. Figure 11 shows the relative frequency of arbitrarily selected multivariate household types, seperately given for different household size. This shows that the C-Series, as expected after the analysis of aggregate data, performs largely worse than the M-Series. The results of C-LOG, C-NN and C-RF are similarly biased and it is difficult to identify a clearly superior method. As expected, estimating higher dimensional distributions becomes easier for large households, probably due to increasingly multivariate normally distributed data, as well as more covariates for estimating complex dependency structures. Basically, the results of NN and LOG are similar, but NN consistently shows the better results, regardless of the strategy used. In the model approach, the random forest almost always fails to provide a correct estimate. The neural network, on the other hand, hits all arbitrarily selected higher-dimensional distributions almost exactly. Thus, the use of the model approach together with a neural net has proven to be a clearly superior approach for our simulation. The classical statistical multinomial model showed the second best results and performed significantly better than the random forest. This shows that machine learning methods are not superior in principle - the basic parametric assumptions of many statistical models can be a serious obstacle, but they can also be an easy path to excellent results.

Table 5: Unique households in true, synthetic and sample population

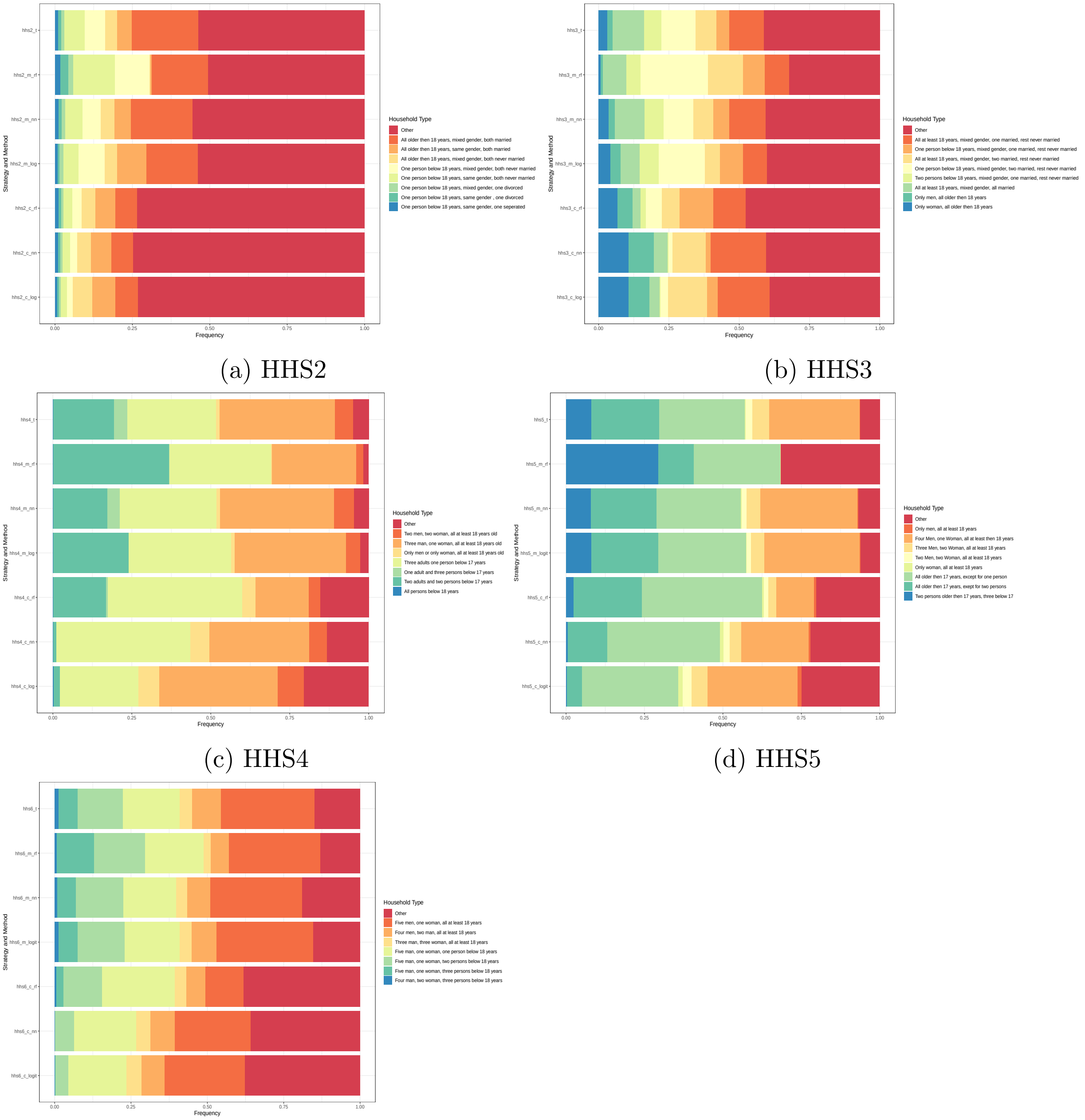
Size	Descriptive		Unique Households							
HHS	N	n	True	Sample	C-LOG	C-NN	C-RF	M-LOG	M-NN	M-RF
2	1178764	11731	2502	1391	U: 5260 (210 %) IP: 2165 (87 %) IS: 918 (66 %)	U: 5189 (207 %) IP: 2180 (87 %) IS: 921 (66%)	U: 5360 (214 %) IP: 2179 (87%) IS: 920 (66 %)	U: 3042 (122 %) IP: 2455 (98 %) IS: 1391 (100 %)	U: 3210 (128 %) IP: 2468 (99 %) IS: 1390 99 %)	U: 1065 (43 %) IP: 867 (35 %) IS: 654 (47 %)
3	703453	7062	35700	3833	U: 71432 (200%) IP: 12549 (35%) IS: 2450 (63%)	U: 70148 (196%) IP: 13091 (36%) IS: 2632 (68%)	U: 72100 (202%) IP: 12765 (35%) IS: 2577 (67%)	U: 54403 (152%) IP: 19511 (54%) IS: 3367 (88%)	U: 67369 (188%) IP: 23909 (66%) IS: 3505 (91%)	U: 18507 (52%) IP: 7911 (22%) IS: 2172 (56%)
4	678141	6766	151701	5565	U: 235643 (155%) IP: 19976 (13%) IS: 1987 (35%)	U: 237560 (156%) IP: 20456 (13%) IS: 1956 (35%)	U: 234502 (155%) IP: 20012 (13%) IS: 2004 (36%)	U: 216960 (143%) IP: 36109 (23%) IS: 2876 (51%)	U: 286524 (188%) IP: 45387 (30%) IS: 3522 (63%)	U: 70341 (46%) IP: 16610 (11%) IS: 2723 (49%)
5	249307	2494	128731	2394	U: 184564 (143%) IP: 1766 (1%) IS: 234 (10%)	U: 183112 (142%) IP: 2345 (2%) IS: 315 (11%)	U: 179856 (140%) IP: 1965 (1%) IS: 326 (11%)	U: 199528 (154%) IP: 9273 (7%) IS: 665 (27%)	U: 186193 (144%) IP: 11025 (9%) IS: 775 (32%)	U: 101882 (79%) IP: 7771 (6%) IS: 1453 (60%)
6	110497	1134	86898	1124	U: 109764 (126 %) IP: 8 (0 %) IS: 2 (0%)	U: 110463 (127%) IP: 34 (0%) IS: 1 (0%)	U: 110392 (127%) IP: 93 (0%) IS: 8 (0%)	U: 107559 (124%) IP: 1096 (1%) IS: 56 (5%)	U: 103620 (119%) IP: 891 (1%) IS: 111 (10%)	U: 88976 (102%) IP: 1808 (2%) IS: 684 (60%)

HHS	True	Sample	C-LOG	C-NN	C-RF	M-LOG	M-NN	M-RF
2	1: 22457 2: 20842 3: 20787	1: 221 2: 216 3: 215	1: 7043 2: 6985 3: 6453	1: 7175 2: 7094 3: 6704	1: 8050 2: 7816 3: 7812	1: 19733 2: 17974 3: 16940	1: 21827 2: 20845 3: 19794	1: 55787 2: 54353 3: 35127
3	1: 5822 2: 5735 3: 4336	1: 58 2: 55 3: 47	1: 1678 2: 1455 3: 1399	1: 1786 2: 1601 3: 1520	1: 2245 2: 2109 3: 2009	1: 4277 2: 4171 3: 3911	1: 4790 2: 4294 3: 3975	1: 21104 2: 18142 3: 15459
4	1: 2519 2: 2511 3: 2470	1: 26 2: 23 3: 22	1: 923 2: 654 3: 511	1: 1123 2: 1023 3: 967	1: 1554 2: 1334 3: 1299	1: 3523 2: 2688 3: 2553	1: 1196 2: 1196 3: 1039	1: 16349 2: 13764 3: 13213
5	1: 348 2: 329 3: 327	1: 6 2: 4 3: 3	1: 25 2: 23 3: 17	1: 45 2: 39 3: 32	1: 120 2: 108 3: 99	1: 185 2: 128 3: 127	1: 235 2: 216 3: 215	1: 2249 2: 1494 3: 1414
6	1: 45 2: 41 3: 32	1: 3 2: 2 3: 2	1: 2 2: 2 3: 2	1: 2 2: 2 3: 2	1: 2 2: 2 3: 2	1: 18 2: 14 3: 13	1: 32 2: 26 3: 23	1: 487 2: 415 3: 358

Table 6: Three most often repeated households

Finally, we take a brief look at the correlation structures within the data sets. As mentioned, an undirected correlation can be determined for categorical variables by means of Cramer's V. For metric variables, the standard Pearson correlation coefficient is used. Again, it would be possible to list a large number of correlations, but for reasons of economy we will limit ourselves to a small selection. Therefore, we look at the strength of the correlation between the categorical variables SEX and MST according to Cramer's V for households of size two. Table 8 reveals that the correlation structures between the categorial variables have been almost completely lost. Of course, this only leads to a good result if there are no correlations in the true population either. This result speaks for the expressed hypothesis that the data-generating process of the combination approach triggers the Central Limit Theorem, whereby the variables no longer show any correlations among themselves. On the other hand, this is only a selective sample and further analyses will have to be awaited. The results of the model approach, on the other hand, show that it can differentiate between strong and weak correlations. M-LOG and M-NN show excellent results, but M-RF often deviates significantly from the true value. The analysis of the correlation structures also fits into the interpretation of the previous overall context in terms of its content.

Does the combination approach always cause the correlations between the variables to disappear? A look at the pearsons'r correlation for the age of the persons in HHS4 (Table 9) shows that this is not always the case. For example, C-NN even shows a medium-strong correlation of 0.30 for persons 1 and 3. Unfortunately, these values are again far from the true values. The lack of precision has not changed, but it can no longer be assumed that the results of the C-series always make the correlations between the variables disappear completely. M-LOG and M-NN again proved to be the best methods, although M-NN produced less precise



(a) HHS2

(b) HHS3

(c) HHS4

(d) HHS5

(e) HHS6

Figure 11: Relative frequencies of multivariate household types

HHS	C-LOG	C-NN	C-RF	M-LOG	M-NN	M-RF
2	FR: 0,08 SR: 14,32	FR: 0,00 SR: 13,95	FR: 0,04 SR: 14,62	FR: 0,00 SR: 0,00	FR: 0,00 SR: 0,01	FR: 0,00 SR: 0,51
3	FR: 0,10 SR: 28,32	FR: 0,00 SR: 17,64	FR: 0,00 SR: 34,95	FR: 0,08 SR: 0,01	FR: 0,07 SR: 0,00	FR: 0,10 SR: 0,66
4	FR: 0,14 SR: 38,42	FR: 0,00 SR: 29,58	FR: 0,00 SR: 44,31	FR: 0,00 SR: 0,70	FR: 0,02 SR: 0,08	FR: 0,00 SR: 0,78
5	FR: 0,18 SR: 33,44	FR: 0,00 SR: 25,43	FR: 0,00 SR: 36,01	FR: 0,02 SR: 1,78	FR: 0,01 SR: 0,82	FR: 0,00 SR: 2,36
6	FR: 0,21 SR: 37,45	FR: 0,00 SR: 28,43	FR: 0,00 SR: 41,06	FR: 0,00 SR: 2,45	FR: 0,01 SR: 0,07	FR: 0,00 SR: 5,45

Table 7: Percentage of structural zeros in synthetic populations

results than M-LOG in some cases. It is difficult to say whether these deviations are due to chance or have a systematic reason.

6.2 Discussion

The way the simulation was performed had some limitations, which we will discuss below. Compared to the combination approach, the model approach had the initial advantage that the values of the oldest person in a household did not have to be estimated, but were already available. This artificially increased the performance of the model approach, at least in part. However, it can be stated that this advantage was not the reason for the better performance. Despite its initial advantage in the model approach, the random forest sometimes still performed worse than some methods in the combination approach. If at all, these results can be interpreted in such a way that the random forest again proved to be clearly inferior to the neural network and the multinomial model.

A further limitation is the lack of variance estimation (Muennich, 2008) in the simulation. Randomness exists in the use of our strategies at several levels: When drawing the sample, varying the ML model and modelling households by drawing from the estimated probabilities. In the combination approach, another random element exists in the data-generating process. Even if we assume on the basis of our own experiments that the results available do not differ significantly if repeated several times, there is still no form of variance estimation. Further studies should try to arrive at this estimate, but this will require an enormous computational effort.

It should also be noted that our simulation is not only based on the best data availability scenario, but also requires a perfect sampling process. The various additional errors in sampling that arise here (Campanelli: 2009), as well as the problem of validly and reliably measuring an object by means of indicators, which is particularly common in the social sciences, should in reality lead to problems (Hox: 2009).

As already indicated, a clear answer to the question which ML method is fundamentally

	SEX1	SEX2	MST1	MST2
SEX1	True: 1.00 C-LOG: 1.00 C-NN: 1.00 C-RF: 1.00 M-LOG: 1.00 M-NN: 1.00 M-RF: 1.00	True: 0.27 C-LOG: 0.01 C-NN: 0.00 C-RF: 0.00 M-LOG: 0.26 M-NN: 0.26 M-RF: 0.22	True: 0.02 C-LOG: 0.02 C-NN: 0.01 C-RF: 0.03 M-LOG: 0.02 M-NN: 0.02 M-RF: 0.02	True: 0.12 C-LOG: 0.01 C-NN: 0.03 C-RF: 0.01 M-LOG: 0.11 M-NN: 0.11 M-RF: 0.16
SEX2	True: 0.27 C-LOG: 0.01 C-NN: 0.00 C-RF: 0.01 M-LOG: 0.26 M-NN: 0.26 M-RF: 0.22	True: 1.00 C-LOG: 1.00 C-NN: 1.00 C-RF: 1.00 M-LOG: 1.00 M-NN: 1.00 M-RF: 1.00	True: 0.03 C-LOG: 0.01 C-NN: 0.01 C-RF: 0.01 M-LOG: 0.04 M-NN: 0.03 M-RF: 0.05	True: 0.13 C-LOG: 0.01 C-NN: 0.01 C-RF. 0.02 M-LOG: 0.12 M-NN: 0.13 M-RF: 0.26
MST1	True: 0.02 C-LOG: 0.02 C-NN: 0.01 C-RF: 0.01 M-LOG: 0.02 M-NN: 0.02 M-RF: 0.02	True: 0.03 C-LOG: 0.01 C-NN: 0.02 C-RF: 0.01 M-LOG: 0.03 M-NN: 0.03 M-RF: 0.06	True: 1.00 C-LOG: 1.00 C-NN: 1.00 C-RF: 1.00 M-LOG: 1.00 M-NN: 1.00 M-RF: 1.00	True: 0.12 C-LOG: 0.02 C-NN: 0.01 C-RF: 0.02 M-LOG: 0.13 M-NN: 0.11 M-RF: 0.16
MST2	True: 0.12 C-LOG: 0.01 C-NN: 0.03 C-RF: 0.01 M-LOG: 0.11 M-NN: 0.11 M-RF: 0.16	True: 0.13 C-LOG: 0.01 C-NN: 0.03 C-RF: 0.01 M-LOG: 0.12 M-NN: 0.13 M-RF: 0.26	True: 0.12 C-LOG: 0.02 C-NN: 0.01 C-RF: 0.02 M-LOG: 0.13 M-NN: 0.11 M-RF: 0.16	True: 1.00 C-LOG: 1.00 C-NN: 1.00 C-RF: 1.00 M-LOG: 1.00 M-NN: 1.00 M-RF: 1.00

Table 8: Cramer's V for paired categorical variables (HHS2)

	Age_cat_1	Age_cat_2	Age_cat_3	Age_cat_4
Age_cat_1	True: 1.00	True: 0.30	True: 0.09	True: 0.06
	C-LOG: 1.00	C-LOG: 0.01	C-LOG: 0.09	C-LOG: 0.11
	C-NN: 1.00	C-NN: -0.01	C-NN: 0.31	C-NN: 0.03
	C-RF: 1.00	C-RF: 0.04	C-RF: 0.18	C-RF: 0.04
	M-LOG: 1.00	M-LOG: 0.26	M-LOG: 0.11	M-LOG: 0.09
	M-NN: 1.00	M-NN: 0.31	M-NN: -0.04	M-NN: -0.06
	M-RF: 1.00	M-RF: 0.44	M-RF: 0.26	M-RF: 0.13
Age_cat_2	True: 0.30	True: 1.00	True: 0.30	True: 0.25
	C-LOG: 0.01	C-LOG: 1.00	C-LOG: -0.02	C-LOG: -0.02
	C-NN: -0.01	C-NN: 1.00	C-NN: 0.01	C-NN: 0.01
	C-RF: 0.04	C-RF: 1.00	C-RF: 0.05	C-RF. 0.05
	M-LOG: 0.26	M-LOG: 1.00	M-LOG: 0.32	M-LOG: 0.26
	M-NN: 0.31	M-NN: 1.00	M-NN: 0.30	M-NN: 0.26
	M-RF: 0.44	M-RF: 1.00	M-RF: 0.58	M-RF: 0.37
Age_cat_3	True: 0.09	True: 0.30	True: 1.00	True: 0.50
	C-LOG: 0.09	C-LOG: -0.02	C-LOG: 1.00	C-LOG: -0.10
	C-NN: 0.31	C-NN: 0.01	C-NN: 1.00	C-NN: -0.04
	C-RF: 0.18	C-RF: 0.05	C-RF: 1.00	C-RF: 0.03
	M-LOG: 0.11	M-LOG: 0.32	M-LOG: 1.00	M-LOG: 0.48
	M-NN: -0.04	M-NN: 0.30	M-NN: 1.00	M-NN: 0.47
	M-RF: 0.26	M-RF: 0.58	M-RF: 1.00	M-RF: 0.54
Age_cat_4	True: 0.06	True: 0.25	True: 0.50	True: 1.00
	C-LOG: 0.11	C-LOG: -0.02	C-LOG: -0.10	C-LOG: 1.00
	C-NN: 0.03	C-NN: 0.01	C-NN: -0.04	C-NN: 1.00
	C-RF: 0.04	C-RF. 0.05	C-RF: 0.03	C-RF: 1.00
	M-LOG: 0.09	M-LOG: 0.26	M-LOG: 0.48	M-LOG: 1.00
	M-NN: -0.06	M-NN: 0.26	M-NN: 0.47	M-NN: 1.00
	M-RF: 0.13	M-RF: 0.37	M-RF: 0.54	M-RF: 1.00

Table 9: Perason's r for paired categorical age (HHS4)

better is difficult to answer because of the 'no free lunch theorem'. However, the available results, even if they only refer to a known finite population, indicate a fundamental superiority of the model approach in combination with a neural network to such an extent that this may be assumed at first. A more precise answer can only be obtained through additional experience and further research.

It was mentioned several times at the beginning that a synthetic population can serve as a basis for microsimulation. Therefore, in a further step it would be exciting to investigate how the populations created here behave within a microsimulation. Small differences, such as between M-LOG and M-NN, could possibly lead to major effects. Further investigations should therefore not only create a synthetic population, but also explore its effects on the simulation process of a microsimulation.

Overall, as mentioned more often, a selection of results had to be made. It would have been beyond the scope to list and compare every possible distribution, every possible statistical parameter or every possible correlation structure. In statistics, there is currently no possibility to summarise higher-dimensional data sets and their parameters in a short time. As long as this is not possible, the researcher has to justify why he chose the selected part of the results and not another. The results selected in this paper were chosen because they are similar in their basic message to the rest of the results not shown. However, it cannot be ruled out that interesting results were simply overlooked. For the presentation of multivariate household types, a selection had to be made from several possibilities. It is possible that other definition criteria would yield different insights. It would also have been possible to depict three-dimensional distributions and observe them from different angles.

Overall, we used a fairly large sample in the simulation (1 percent sample). However, such large samples are rather rare in reality. Therefore, it can be assumed that the results for smaller samples and especially smaller households would be even worse. In the future, it should also be researched how large a sample should be in order to obtain good results using the methods discussed here.

Finally, the question also remains whether the number of parameters to be simulated within a synthetic population has a positive or negative effect on the quality of the estimate. In our case, we have limited ourselves to three variables (age, gender, degree of relationship), as these are strongly interrelated and are essential for a qualitative database. But what would happen if other variables, such as income and education, or spatially disaggregated information were added? The literature not only speaks of a 'Curse of dimensionality' but also of a 'Blessing of dimensionality' (Gorban and Tyukin, 2018). It is conceivable that many additional variables could even increase the quality of the simulation. However, this would first have to be further investigated for this specific question.

7 Conclusion

The aim of the work was to create a complete synthetic population based on a single-stage cluster sample using machine learning methods. For this purpose two different strategies were investigated: The combination approach and the model approach. Two ML methods were used, namely random forest and a feedforward neural net. To enable a comparison with 'classical' statistical methods, a logit model was included in the analysis for the combination approach and a multinomial model for the model approach. The strategies were tested using the AMELIA data set: The aim was that a certain approach, based on the sample drawn, would successfully reproduce the AMELIA household structure.

Our simulation showed that the model approach performed better overall than the combination approach. It was best to use it in combination with a Neural Net. The Random Forest almost always showed worse results than the use of a Neural Net with the same strategy. In some cases the performance of the random forest in the model approach was even the worst of all four approaches. Moreover, the neural network in the model approach was the only way to correctly map higher-dimensional distributions; all other approaches failed in this task, which could also be easily understood by looking at the distributions between household members. The second best was the Model apporach with a multionimal model. This performed significantly better than the random forest and only showed weaknesses in comparison to the neural network in a detailed analysis.

The combination approach only led to relatively good results for large household sizes. The problems of this strategy became apparent in the data generation process designed for its purposes. By randomly drawing the individual parameters based on their observed empirical density in the sample, this process led to normally distributed data that did not match the distribution in AMELIA. The combination approach was thus able to detect whether a household might be from the real population, but not how often it actually occurs. It may be that the combination approach is better suited to check and cleanse an existing data set for unrealistic households than to perform a realistic data-generating process itself. In addition, the Combination Approach was hardly able to estimate the correlation structures between the variables correctly; often these were simply lost or deviated massively from the true values. Once again, only the two best methods of the Model approach delivered good results; the Random Forest also misestimated the correlation structures. Future research should try to vary the present results with other data sets and further methods as well as hyperparameters. Especially useful here would be an additional meaningful variance estimation, the use of the created synthetic populations in a realistic microsimulation as well as further statistical comparison methods. Work such as this can only offer a small and very limited insight. However, the strategies and comparative methods developed should pave the way to enable further structured and comparable research.

8 Bibliography

Alfons, Andreas; Kraft, Stefan; Templ, Matthias; Filzmoser, Peter (2011): *Simulation of close-to-reality population data for household surveys with application to EU-SILC. In: Stat Methods Appl 20 (3), p. 383 - 407.*

Amit, Y. ; Geman, D. (1997): *Shape quantization and recognition with randomized trees. Neural Computation, 9, p. 1545 - 1588.*

Barthelemy, Johan; Toint, Philippe L. (2013): *Synthetic Population Generation Without a Sample. In: Transportation Science 47 (2), p. 266 - 279.*

Breiman, Leo (1994): *Heuristics of instability in model selection, Technical Report, Statistics Department, Univer- sity of California at Berkeley (to appear, Annals of Statistics).*

Breiman, Leo (1999): *Using adaptive bagging to debias regressions, Technical Report 547, Statistics Dept. UCB.*

Breiman, Leo (2001a): *Random Forests. In: Machine Learning 45 (1), p. 5 - 32.*

Breiman, Leo (2001b): *Statistical modeling: The two cultures. Quality Engineering, 48, 81-82.*

Burgard, Jan-Pablo; Muennich, Ralf; Zimmermann, Thomas (2013): *The impact of Sampling Designs on Small Area Estimates for Business Data, Journal of Official Statistics 30, 4, p. 749 - 771 .*

Burgard, Jan Pablo; Kolb, Jan-Philipp; Merkle, Hariolf; Muennich, Ralf (2017): *Synthetic data for open and reproducible methodological research in social sciences and official statistics. In: AStA Wirtsch Sozialstat Arch 11 (3-4), p. 233 - 244.*

Caiola, Gregory; J. Reiter (2010): *Random Forests for Generating Partially Synthetic, Categorical Data. Trans. Data Priv. 3, p. 27 - 42.*

Casella, George, Christian P. Robert, and Martin T. Wells (2004): *Generalized Accept-Reject Sampling Schemes. Lecture Notes-Monograph Series, vol. 45, p. 342 - 347.*

Campanelli, Pamela (2009): *Testing Survey Questions . In: Leeuw, Edith, Desiree de; Hox, Joop J.; Dillman, Don A. (Hg.): International handbook of survey methodology. Repr. New York, NY: Psychology Press. Wiley Series in Probability and Statistics, p. 176 - 200.*

Chambers, R. L.; Dunstan, R. (1986): *Estimating distribution functions from survey data. In: Biometrika, 73 (3), p. 597 - 604.*

Cramer, Harald (1946): *Mathematical Methods of Statistics. Princeton: Princeton University Press.*

Deming, W. Edward; Frederick F. Stephan (1940): *On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals Are Known. The Annals of Mathematical Statistics, 11 (4), p. 427 - 444.*

Drechsler, Joerg (2010): *Using Support Vector Machines for Generating Synthetic Datasets. In: Josep Domingo-Ferrer und Emmanouil Magkos (Hg.): Privacy in statistical databases. UNESCO Chair in Data Privacy International Conference, PSD 2010, Corfu, Greece, September 22 - 24, 2010 ; proceedings, Bd. 6344. Berlin: Springer (Lecture Notes in Computer Science, 6344), p. 148 - 161.*

Fienberg, Stephen; Rinaldo, Alessandro (2007): *Three centuries of categorical data analysis: Log-linear models and maximum likelihood estimation. Journal of Statistical Planning and Inference, 137, p. 3430 - 3445.*

Forster, M; Sober, E (1994): *How to Tell When Simpler, More Unified, or Less Ad Hoc Theories Will Provide More Accurate Predictions. In: British Journal for the Philosophy of Science 45, p. 1 - 35.*

Ghahramani, Zoubin (2013): *Bayesian non-parametrics and the probabilistic approach to modelling. In: Philosophical Transactions of The Royal Society A Mathematical Physical and Engineering Sciences 371 (1984), p. 211 - 553.*

Goodfellow, Ian; Bengio, Yoshua; Courville, Aaron (2016): *Deep learning. Cambridge, Massachusetts, London, England: MIT Press.*

Gorban, A. N.; Tyukin, I. Y. (2018): *Blessing of dimensionality: mathematical foundations of the statistical physics of data. In: Philosophical transactions. Series A, Mathematical, physical, and engineering sciences 376 (2118).*

Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome H. (2017): *The elements of statistical learning. Data mining, inference, and prediction. Second edition, corrected at 12th printing 2017. New York, NY: Springer (Springer series in statistics).*

Hoffmann, J. (2003): *Generalized linear models: An applied approach. Boston, MA: Allyn and Bacon.*

Hox, Joop J. (2009): *Accommodating measurement errors In: Leeuw, Edith Desiree, de; Hox, Joop J.; Dillman, Don A. (Hg.): International handbook of survey methodology. Repr. New York, NY: Psychology Press. Wiley Series in Probability and Statistics, p. 387 - 402.*

Huang, Z.; Williamson, P. (2001): *A comparison of synthetic reconstruction and combinatorial optimisation approaches to the creation of small-area microdata. Population Microdata Unit, Department of Geography, University of Liverpool.*

Higham, Catherine F.; Higham, Desmond J. (2018): *Deep Learning: An Introduction for Applied Mathematicians.*

He, Hua; Tang, Wan; Wang, Wenjuan; Crits-Christoph, Paul. (2014): *Structural zeroes and zero-inflated models, Shanghai archives of psychiatry. 26, p. 236 - 242.*

Ho, Tin Kam (1995): *Random decision forests. In: Proceedings of the third International Conference on Document Analysis and Recognition, Calif: IEEE Computer Society Press, p. 278 - 282.*

Ho, Tin Kam (1998): *The random subspace method for constructing decision forests. IEEE Trans. Pattern Analysis and MachineIntelligence, 20, p. 832 - 844.*

Javed, Maria; Irfan, Muhammad (2020): *A simulation study: new optimal estimators for population mean by using dual auxiliary information in stratified random sampling. In: Journal of Taibah University for Science 14 (1), p. 557 - 568.*

Johnson, Oliver (2004): *Information theory and the central limit theorem. London : Imperial College Press.*

Kroese, D. P.; Chan, J. C. C. (2014): *Statistical Modeling and Computation, Springer.*

Kuk, A. (1993): *A kernel method for estimating finite population distribution functions using auxiliary information. In: Biomeirika, 80, p. 385 - 392.*

Kuhn, M. (2008): *Building Predictive Models in R Using the caret Package. Journal of Statistical Software, 28(5), 1 - 26.*

Liaw, A.; Wiener, M. (2002): *Classification and Regression by randomForest. R News 2(3), 18–22.*

Lovelace, Robin; Dumont, Morgane (2016): *Spatial microsimulation with R. Boca Raton, London, New York: CRC Press Taylor and Francis Group a Chapman and Hall Book.*

Mitchell, Tom M. (1997): *Machine learning. International ed., New York, NY: McGraw-Hill.*

Morris, Tim P.; White, Ian R.; Crowther, Michael J. (2019): *Using simulation studies to evaluate statistical methods. In: Statistics in Medicine 38 (11), p. 2074 - 2102.*

Muennich, R.; Schuerle, J. (2003): *On the Simulation of Complex Universes in the Case of Applying the German Microcensus, DACSEIS Research Paper Series.*

Muennich, R. ;Bihler, W. ;Bjornstad, J. ;Zhang, Li-Chun ;Davison, A. ;Sardy, S. ;Haslinger, A. ;Knottnerus, P. ;Laaksonen, S. ;Ohly, D. ;Schuerle, J. ;Wiegert,R. ;Oetliker, U. ;Renfer, Jean-Pierre ;A., Quatember ;Skinner, C. ;Berger, Y. (2003): *Data Quality in Complex Surveys, IST-2000-26057 DACEIS, 2003. Research Project Report.*

Muennich, Ralf (2008): *Varianzschaetzung in komplexen Erhebungen. Austrian Journal of Statistics, 37 (4), 319 - 334.*

Murphy, Kevin P. (2012): *Machine learning. A probabilistic perspective. Cambridge, Mass.: MIT Press (Adaptive computation and machine learning series).*

Nakama T. (2011): *Comparisons of Single- and Multiple-Hidden-Layer Neural Networks. In: Liu D., Zhang H., Polycarpou M., Alippi C., He H. (eds) Advances in Neural Networks. Lecture Notes in Computer Science, vol 6675. Springer, Berlin, Heidelberg.*

Nielsen, Michael (2019): *Neural Networks and Deep Learning, neuralnetworksanddeeplearning.com*

O'Brien, Travis A.; Kashinath, Karthik; Cavanaugh, Nicholas R.; Collins, William D.; O'Brien, John P. (2016): *A fast and objective multidimensional kernel density estimation method: fastKDE (PDF). Computational Statistics and Data Analysis, 101, p. 148 - 160.*

Kronmal, R. A. and Peterson Jr., A. V. (1979): *On the alias method for generating random variables from a discrete distribution. The American Statistician, 33, p. 214 - 218.*

Kolb, Jan-Phillip (2013): *Methoden zur Erzeugung synthetischer Simulationsgesamtheiten, Dissertation.*

Li, J.; O'Donoghue, C. (2013): *A survey of dynamic microsimulation models: uses, model structure and methodology, International Journal of Microsimulation 6(2), pp. 3-55.*

Lohr, S.L. (1999): *Sampling: Design and Analysis. Duxbury Press, Pacific Grove, CA.*

Orcutt, G. H. (1957): *A New Type of Socio-Economic System. Review of Economics and Statistics, 39(2), p. 116-123.*

Pearson, Karl (1895): *Notes on regression and inheritance in the case of two parents. Proceedings of the Royal Society of London. 58, p. 240â€"242.*

Ray, Paramesh (1973): *Independence of Irrelevant Alternatives. Econometrica, 41 (5), p. 987-991.*

Robinson, W.S. (1950): *Ecological Correlations and the Behavior of Individuals. American Sociological Review. 15 (3), p. 351 - 357.*

Reiter, J. (2009): *Using multiple imputation to integrate and disseminate confidential microdata. In: International Statistical Review, 77 (2), p. 179 - 195.*

Rubin, Don, B. (1993): *Satisfying Confidentiality Constraints Through Use of Synthetic Multiply-imputed Microdata. Journal of Of ocial Statistics, 9, 461-468.*

Saerndal, Carl-Erik; Swensson, Bengt; Wretman, Jan Hakan; Saerndal-Swensson-Wretman (2003): *Model assisted survey sampling. 1. softcover print. New York, NY: Springer*

Selvin, Hanan C. (1958): *Durkheim's Suicide and Problems of Empirical Research. American Journal of Sociology. 63 (6), p. 607 - 619.*

Segal, Mark; Xiao, Yuanyuan (2011): *Multivariate random forests. In: WIREs Data Mining Knowl Discov 1 (1), p. 80 - 87.*

Smola A.; Gretton A.; Song L.; Schoelkopf B. (2007): *A Hilbert Space Embedding for Distributions. In: Hutter M., Servedio R.A., Takimoto E. (eds) Algorithmic Learning Theory. ALT 2007. Lecture Notes in Computer Science, vol 4754. Springer, Berlin, Heidelberg.*

Stier, W. (1999): *Empirische Forschungsmethoden. Springer.*

Sakshaug, J. W. und Raghunathan, T. E. (2010): *Synthetic data for small area estimation. In: Proceedings of the 2010 international conference on Privacy in statistical databases.*

Tanton, Robert; Edwards, Kimberley L. (Hg.) (2013): *Spatial microsimulation. A reference guide for users. Dordrecht: Springer.*

Vink, G. (2016): *Towards a standardized evaluation of multiple imputation routines.*

Venables WN, Ripley BD (2002): *Modern Applied Statistics with S, Fourth edition. Springer, New York.*

Verleysen, Michel; Francois, Damien (2005): *The Curse of Dimensionality in Data Mining and Time Series Prediction. In: Computational Intelligence and Bioinspired Systems, 8th International Work-Conference on Artificial Neural Networks, IWANN 2005, Vilanova i la GeltrÃº, Barcelona, Spain, June 8-10, 2005, Proceedings 3512, p. 758 - 770.*

Wolpert, David; Macready, William (1996): *No Free Lunch Theorems for Search.*

Wooldridge, Jeffrey (2013): *Introductory Econometrics. A Modern Approach. South-Western, Cengage Learning, Mason, Ohio 2013*

Zinn, Sabine (2012): *A Mate-Matching Algorithm for Continuous-Time Microsimulation Models, International Journal of Microsimulation, 5, issue 1, p. 31-51.*

ERKLÄRUNG ZUR BACHELORARBEIT / MASTERARBEIT

Hiermit erkläre ich, dass ich die Bachelorarbeit / Masterarbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt und die aus fremden Quellen direkt oder indirekt übernommenen Gedanken als solche kenntlich gemacht habe.

Die Arbeit habe ich bisher keinem anderen Prüfungsamt in gleicher oder vergleichbarer Form vorgelegt. Sie wurde bisher nicht veröffentlicht.

Datum

Unterschrift