EUROPEAN COMMISSION
Community Research

SEVENTH FRAMEWORK PROGRAMME

AMELI
Advanced Methodology for European Laeken Indicators

## Deliverable 6.2

# Synthetic Data Generation of SILC Data

Version: 2011

Andreas Alfons, Peter Filzmoser, Beat Hulliger, Jan-Philipp Kolb, Stefan Kraft, Ralf Münnich and Matthias Templ

The project **FP7–SSH–2007–217322 AMELI** is supported by European Commission funding from the Seventh Framework Programme for Research.

http://ameli.surveystatistics.net/

# Contributors to deliverable 6.2

**Chapter 1:** Andreas Alfons and Matthias Templ, Vienna University of Technology Jan-Philipp Kolb and Ralf Münnich, University of Trier.

**Chapter 2:** Andreas Alfons and Matthias Templ, Vienna University of Technology Jan-Philipp Kolb and Ralf Münnich, University of Trier.

**Chapter 3:** Andreas Alfons and Matthias Templ, Vienna University of Technology.

**Chapter 4:** Jan-Philipp Kolb and Ralf Münnich, University of Trier.

# Main Responsibility

Matthias Templ, Vienna University of Technology; Ralf Münnich, University of Trier

# Evaluators

**Internal expert:** Risto Lehtonen, University of Helsinki.

# Aim and objectives of deliverable 6.2

The objective of this deliverable is to give an overview of the state of the art of data generation mechanisms. The generated populations which serve as the simulation basis are presented in this deliverable. Two main synthetic universes have been generated: on the one hand the AAT-SILC population and on the other hand the AMELIA data set. The University of Vienna is responsible for the generation of the AAT-SILC population whereas the University of Trier is responsible for the generation of the AMELIA population.

# Contents

# Chapter 1

# Introduction

The set of targets of the AMELI project comprehends the testing of different estimation methods and variance estimation for the Laeken Indicators. In general, micro-simulation is often used to control for the interplay between data structure, sampling scheme, and properties of estimators. This can be done with a design-based micro-simulation approach. To apply such an approach a synthetic population is necessary, which is bigger than the scientific use file (SUF) delivered by Eurostat. The requirements for such a population are manifold. In this work an overview of the methodology for the reproduction of the most important population characteristics and for the minimization of the disclosure risk is presented.

The benefit obtained from the stand-alone Laeken Indicators is small, whereas the indicators give added value when a comparison is possible. Comparisons can be made over time and across different regional entities. The big echo in the press after the publication of the *atlas of poverty* by the German Charity Group (Deutscher Paritätischer Wohlfahrtsverband) showed that there is a big need and interest in regional analysis. Further, it is interesting to evaluate the reaction of the indicators to extreme data situations such as the presence of outliers. These reactions can be tested in micro-simulation approaches.

Generally, it is always the optimum to conduct micro-simulations on real data, like data from the census or complete statistics like the basic file of the *European Union Statistics on Income and Living Conditions* (EU-SILC). However, due to different reasons the real data are often not available. Possible reasons are disclosure problems but also lacking recent survey material on the topic of interest.

Nevertheless, it is possible to apply scenario analysis to synthetic data if the real data is not available. To catch the structure of the original data is then the most important issue, with the disadvantage that statements concerning the content are no longer possible but also not intended.

Unfortunately, information from censuses or comparable data for most of the countries in the European Union is not available. The absence of data in an appropriate manner necessitates the generation of synthetic universes. One important example for the requirement of a suitable data set is the situation where the performance of point or variance estimators have to be checked in a Monte Carlo experiment. The synthetic universe can

then be helpful to test how different sampling schemes influence the inference. It is remarkable that reasons for generating a synthetic population are different. Furthermore, the requirements for such a population are manifold and differ between research tasks. Moreover, the underlying data basis is of different quality and often various data sets are available. This also implies that the methods for generating a synthetic population differ from each other. Therefore, it is not possible to recommend one proceeding which may be regarded as the best one overall.

For the AMELI project different data sets were applied. The Austrian SILC data set, delivered by the Austrian national statistical office, and the scientific use file of the EU-SILC data set, delivered by Eurostat, are the basis for synthetic populations. This data should be treated confidentially, it becomes subject to non disclosure. Therefore, it is necessary to generate a synthetic population for which it is impossible to link entries with real live persons. The importance and difficulty of this task depend very much on the basic data set. However, when applying synthetisation techniques, one should keep in mind that the structure of the original data should be identifiable in the synthetic universe.

Concluding, it is a difficult task to take all exigencies into consideration while generating a synthetic data set. At the same time it is a chance because the combination of these structures in one freely available data set is seldom. Our target in the present case is to produce a public dataset which is free available and which can be used to compare different estimation methods.

In chapter 2, synthetic data generation is discussed in general. Requirements for synthetic universes and the state of the art in the generating synthetic universes are presented. Further, the data framework for the simulations within AMELI is introduced within section 2.3. Chapter 3 covers the generation of the synthetic Austrian population data AAT-SILC. The synthetic population AMELIA, which is dedicated to small area estimation, is presented in Chapter 4. In both chapters a rough overview about the proceeding for the generation of the universes and selected results is presented. Finally, a summary is given in Chapter 5.

# Chapter 2

# Synthetic data generation

This chapter provides a general discussion on synthetic data generation. Section 2.1 addresses requirements for synthetic populations. A short review of common methods for data simulation is given in section 2.2. Lastly, section 2.3 is focused on EU-SILC data.

## 2.1 Requirements for synthetic universes

It was mentioned in the introduction that requirements may result from specific research tasks. One example concerns regional arrangements which have to be available if it is the task to compare survey designs. Other requirements result from the original structure of the basic data set which should be as far as possible preserved. It is, for example, important to have the real correlation between variables as well as a realistic heterogeneity between the populations in different stratas. At the same time the household structure of the real data should be maintained. That is for example of special interest for the indicators of social exclusion because some indicators are implemented on household level. The indicator *Persons living in jobless households* (SIP5) is one example. Further, it is assumed that strong interactions exist between the household members, which can be important for estimations in following simulations. It is, for example, possible that the current education activity (PE010 in EU-SILC) of children has to be estimated but is not known. However, maybe the ISCED level (PE040 in EU-SILC) attained by their parents is known and can be taken as auxiliary information.

If a strong relationship between the educational backgrounds of different household members is observed in the basic data set. The so-called household structure should be preserved in the synthetic population. The household structures are of prior interest for some Laeken Indicators and have to be respected. Because it is quite difficult to generate a completely synthesized household structure, the original structure contained in the data set should be preserved. These structures can vary between communities of different size. Beside the interest to preserve micro structures, there is also an interest in preserving macro structures. This point concerns the heterogeneities which are assumed to have a great impact on the estimates. Especially in the case of the DACSEIS data set this was a disadvantage. The question if heterogeneities are sufficiently mapped in the data set is closely linked with the question whether a close to reality spatial structure is realized.

This circumstance is one aspect of the exigency to create consistent structures for different regional levels. Thus, not only the micro structures (household and maybe address membership) but also macro structures (communal, regional compilations) have to be coherent.

Concerning the structure of the data set, the distributions for categorical and discrete variables as well as conditional frequencies and interactions between variables should be correct.

A complete new task is the longitudinal structure. For some variables it is easy to forecast a value into the future (e.g. the age of a person), while for others it is not. One easy example here is the variable age. Knowing the recent age of a person, it is easy to predict the age for one year later.

Another requirement affects the interdependencies between variables and geographic entities which have to be taken into account. The comparison of the within-group variance with the between-group variance should lead towards the same results for the synthetic and the original data set.

Moreover, the data basis needs to be of sufficient size. A bigger data set than the EU-SILC scientific use file is necessary because this data set should be treated as universe. Afterwards, different samples will be drawn to control for the interplay between data structure, sampling scheme, and properties of estimators. This can conduce to the elaboration of methods for measuring the adequacy of synthetically generated data with respect to accuracy, distributional aspects and confidentiality.

The proposed solutions presented in this report fulfill these requirements. Since a synthetic universe cannot be perfect, the aim of the presented solutions is to generate synthetic universes which are as realistic as possible.

## 2.2 State of the art in data generation

The simulation of population micro data is closely related to the field of microsimulation, which is a well-established methodology within the social sciences. Microsimulation models attempt to reproduce the behavior of individual units such as persons, households, vehicles or firms. Therefore, most microsimulation studies involve the creation of an adequate, large-scale micro data set as a first step.

The main purpose of microsimulation models is to allow for policy analysis at the micro-level. By contrast, within the AMELI project synthetic populations are generated solely as a basis for extensive simulation studies. Hence, there are some differences in the requirements for data generation. However, some of the methods used within microsimulation have been integrated in the development of the simulation scheme.

There are several main approaches for the generation of synthetic micro data. Two very important approaches are synthetic reconstruction and combinatorial optimization. Synthetic reconstruction normally involves sampling from conditional distributions derived from published contingency tabulations (Huang and Williamson, 2001). In contrast, combinatorial optimization uses reweighting of existing, publicly available micro data

sets, as released by many countries. Both approaches share the advantage that data from different sources can easily be linked through methods like iterative proportional fitting (IPF).

A detailed comparison of these methods and their application to the Sample of Anonymised Records from the 1991 Census in Britain is provided by HUANG and WILLIAMSON (2001). NORMAN (1999) gives a practical introduction to IPF together with a comprehensive overview of related literature.

Two examples of microsimulation models simulating entire populations are the SimBritain model (BALLAS et al., 2005) and the SVERIGE model (HOLM et al., 2006). Both models are dynamic spatial microsimulation models, which means that they simulate the population over many years at the small area level. SYNTHESIS (cf. e.g. BIRKIN and CLARKE 1988) is another example in this context.

An alternative approach for the generation of synthetic data sets is discussed by RUBIN (1993). He addresses the confidentiality problem connected with the release of publicly available micro data and proposes the generation of fully synthetic micro data sets using multiple imputation. RAGHUNATHAN et al. (2003); DRECHSLER et al. (2008a); REITER (2009) discuss this approach in more detail. DRECHSLER et al. (2008a) compare the regression coefficients of the imputed data with the original ones. However, it is impossible to generate categories that are not represented in the (original) sample with their approach. In addition, they do not consider outliers and missing values, or the possible generation of structural zeros in combinations of variables.

The generation of population micro data as a basis for a Monte Carlo simulation study is described by MÜNNICH and SCHÜRLE (2003) and MÜNNICH et al. (2003). Their work also acts as a starting point for the development of a simulation scheme for EU-SILC populations.

As stated above, different methods exist to construct a synthetic data set. The idea of a synthetic data set goes back to RUBIN (1976). This approach to create synthetic data is embedded in the multiple imputation framework. DRECHSLER et al. (2008b) implemented this approach in a German example to take care for disclosure problems. Their target was to minimize the disclosure risk while maximizing the data utility.

The way of creating a synthetic population depends on the purpose of the study, it is in fact a multidisciplinary research interest. Synthetic populations are, for example, necessary in disease research but also for socio-demographic research topics. In general, many different approaches compete on the task of population generation.

It is important to analyze the process of data collection, especially if extreme sampling weights exist. At least it is possible to create several public datasets with the same underlying data like ABOWD and LANE (2004) mentioned it. Then it can be possible to serve different requirements from different user groups without running in disclosure problems. It is then possible to avoid this dilemma because sometimes only the combination of information is critical.

The area of application for synthetic micro data expanded greatly. The so-called synthetic baseline population is used in travel demand models for example by BECKMAN et al. (1996). BALLAS and CLARKE (2000) applied a microsimulation for local labour market

analysis. Another field of application of synthetic micro data exists in the context of spatial microsimulation models, applied for example by Chin and Harding (2006). Hanaokaa and Clarke (2007) combined the examination of these spatial microsimulation models with the content analysis of retail markets.

The use of microsimulation models for the analysis of tax systems is widespread, Chin et al. (2005) processed such an analysis. Other examples for microsimulation models with synthetic data exist for the examination of firm behaviors which was analysed by Kumar and Kockelman (2007). Such microsimulation approaches are also applied in fields which are very close to the examined issues in the AMELI context. Harding et al. (2004) performed a spatial microsimulation approach for the assessment of poverty and inequality.

Often the aim is to enable the usage of multiple sources which can be micro or macro data. Also in the present case information from different sources has to be processed. However, less publications have been published concerning this topic. Kohnen and Reiter (2009) is one example for the combination of data from two agencies which should be treated confidentially.

One aim of the AMELI project was to investigate robust estimation of the Laeken Indicators. For this purpose, Alfons et al. (2011b) developed a data generation framework, which is implemented in the R package `simPopulation` (Alfons and Kraft, 2010). Based on Austrian EU-SILC sample data, the synthetic population AAT-SILC was generated with this framework (see Chaper 3). AAT-SILC was designed to resemble a representative country. A further objective was that the population data should not contain any large outliers, as these are included in the samples during the simulations for full control over the amount of outliers (see Alfons et al., 2011c).

Obtaining estimates for small regional areas and domains was another target of the simulation study in the AMELI project. The investigation of regional breakdowns to relevant sub-populations is of great importance in the context of the Laeken Indicators. In this context the AMELIA population (see Chapter 4) was generated to complement the AAT-SILC population. Moreover, the AMELIA population was generated based on the ideas of Voas and Williamson (2000).

## 2.3 Simulation of SILC populations

In section 2.3.1 a brief description of the basic EU-SILC data provided by Eurostat is given. Some requirements for synthetic EU-SILC data as basis for simulation studies in robust statistics and small area estimation are then discussed in sections 2.3.2 and 2.3.3.

### 2.3.1 The basic EU-SILC data set

The basis for the quantification of poverty and social exclusion is EU-SILC. This data set, which exists since 2004, is also the basic data set for the data generation within the AMELI project. The data delivered for the AMELI project by Eurostat contains four subsequent years from 2004 to 2007. Thus, cross-sectional data are available as well as

longitudinal data sets. For one year four different files are available, that is the household register, the personal register, the household data and the personal data (see table 2.1). The personal register is the most extensive file, for 2004 it contains 307 666 entries and 536 993 entries for 2006.

| **Dataset** | **Variables** | **2004** | **2005** | **2006** |
|---|---|---|---|---|
| household register (D) | 14 | 116 743 | 197 657 | 202 975 |
| personal register (R) | 34 | 307 666 | 527 189 | 536 993 |
| household data (H) | 65 | 116 743 | 197 657 | 202 978 |
| personal data (P) | 87 | 241 796 | 422 400 | 435 169 |
| Countries | - | 15 | 26 | 26 |

Table 2.1: Size of different EU-SILC data sets.

EU-SILC is a rotating panel (cf. HAUSER 2007, p. 2) which is collected differently in the member states of the EU-SILC survey. Following the data production process in the countries, an ex-post output harmonization is implemented.

In November 2006 Eurostat organized a conference on requirements concerning the EU-SILC data set (cf. HAUSER 2007, p. 8) in Helsinki. The three criteria *accuracy*, *reliability* and *international comparability* were of special interest in this conference. The conference gave important hints on the problems which stem from different data collection strategies across the European countries.

In Germany, for example, the Microcensus is the base for the EU-SILC data. Attendants for the access panel are recruited from the discarded quarter of the Microcensus. The access panel is a pool of households with the willingness to attend further interviews. This proceeding gives cause for the discussion about the question whether the German EU-SILC sample can be seen as correct random selection.

The unknown distortions resulting from different selection procedures can affect the computation of sampling errors. Thus, it can be helpful to analyze this process, especially when extreme sampling weights exist. The composition of the German panel for EU-SILC does not allow for the calculation of methodologically correct sampling errors and confidence intervals. Therefore, the proposed strategy for the simulation within the AMELI project was to reconstruct the different sampling designs applied in Europe to control for the effects.

Another problem is the extrapolation of income variables. As an example we take again the German case where income variables, which are based on the German Microcensus, are supposed to give rise to problems. The income reference period defined in EU-SILC is the whole precedent year, whereas the net incomes for the Microcensus are captured as income classes and monthly. Thus, a high non-response has to be ascertained for questions in relation to the income. Therefore, a bias especially for the low income classes has to be expected.

Additionally, data is not available for every country in all three years, e.g., data for Germany lacks in the case of 2004. Especially to get the time structure it is preferable to have information about every year.

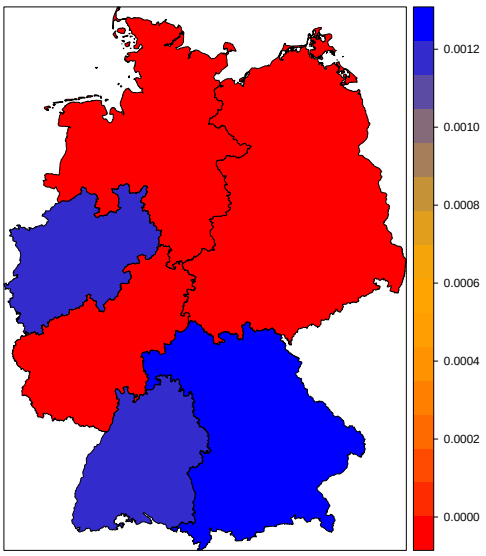


Figure 2.1: Sampling fractions of the German EU-SILC data for the six available regions.

Furthermore, the data set contains no information about regional arrangements apart from the information about the region of the entry. Germany, for example, is divided into six regions (cf. figure 2.1) which is quite crude). Regional disaggregated analysis is difficult due to lacking regional indicators. The colouring of figure 2.1 shows the approximated sampling rate for Germany. It is visible that the red color which represents a small approximated sampling rate predominates the graphic.

### 2.3.2 Synthetic SILC populations and robust estimation

One aim of the AMELI project is, to evaluate advanced estimation techniques for the Laeken Indicators under common data problems. It is certainly of interest to see how such data problems affect the estimation of the indicators on national level. For this purpose, it is necessary to generate synthetic population data for a representative country.

One data problem frequently occurring in practice is the presence of non-representative outliers, i.e., observations that are either incorrect or can be considered unique in the population. In simulation studies, whose purpose is to investigate how robust the developed estimation methods are against such deviating observations, it is crucial to have full control over the amount of non-representative outliers in the samples. Thus, the underlying population data should not contain any non-representative outliers, instead they should be included in the samples (see Alfons et al., 2011c).

Since the Austrian EU-SILC sample from 2006 does not contain any large non-representative outliers in the income variables, it is perfectly suited as a basis for generating synthetic population data to be used in simulations focusing on robustness issues. The generation of the resulting synthetic population data AAT-SILC is discussed in detail in Chapter 3.

### 2.3.3 Synthetic SILC populations and small area estimation

The AAT-SILC population[1] was created to have a data set which is close to the real EU-SILC data of one country, in this case Austria. The Austrian EU-SILC survey sample from 2006, published by Statistics Austria, was the basic sample for this synthetic population.

The EU-SILC data set as a whole is a conflation of different surveys of many European countries, it can be supposed that this data set is very heterogeneous. The data comes from different surveys which are independent from each other. Afterwards, an ex-post output harmonization is processed. Nevertheless, the data structure can be very different for the countries. This fact contributes to a situation which can cause problems for the estimation of results for smaller areas.

One part of the AMELI simulation study deals with the question how to provide reliable estimates for small areas. The focus of these parts of the simulation study lies more on small area investigation and the different survey designs. For the realization of the complex survey designs it is necessary to have information about administrative boundaries. The survey designs are in a comparable way realized for the AAT-SILC population. Some aspects, which are very important for the simulation with the AAT-SILC data set, are of less importance for simulations targeting at the evaluation of small area effects.

One difference concerns the question of outlyingness. For an adequate synthetic population, outlying areas are more interesting than single outliers. Here it is not so important to have the maximum control over the amount of contaminated observations. It is more interesting to have awareness about the relation between the different nested and disjoint areas. BALLAS et al. (1999) motivated the need for spatially disaggregated micro-data as basis for microsimulation approaches.

The Austrian SILC data set can be seen as a more homogeneous one than the data of the public-use-file for EU-SILC delivered by Eurostat. Therefore, a requirement is moving into the focus which can be neglected for the Austrian data set. That is the requirement that the statistics on poverty and social exclusion have a close to reality level for every regional and contentual subarea or subdomain.

It is of course to be welcomed to have one simulation environment which is based on one population. Unfortunately, the requirements and starting points are extreme different. Therefore, it seemed sensible to the AMELI project team to produce two different data sets. It was the target to dispose a population which is on the one hand heterogeneous but on the other hand also synthetic.

[1] This synthetic population is described in Chapter 3.1.

# Chapter 3

# The AAT-SILC data set

The generation of the synthetic data set AAT-SILC with emphasis on the included variables is described in section 3.1, whereas section 3.2 presents selected results for the simulated data.

## 3.1 Generation of the synthetic data AAT-SILC

In this section, the generation of the synthetic population data AAT-SILC is described. It was generated in the statistical environment R (R Development Core Team, 2011) using the data simulation framework developed by Kraft (2009) and Alfons et al. (2011b), which is implemented in the add-on package `simPopulation` (Alfons and Kraft, 2010). Note that this section is focused on describing the variables that are included in AAT-SILC. For a detailed mathematical description of the models involved in the data simulation process the reader is referred to Alfons et al. (2011b).

The data basis for the synthetic population data AAT-SILC is the Austrian EU-SILC survey sample from 2006, which was provided by Statistics Austria. Consequently, the abbreviation AAT-SILC stands for *Artificial Austrian Statistics on Income and Living Conditions*. The motivation for using this particular sample to generate a synthetic universe is twofold. First, it was desired to generate synthetic population data that resemble a representative country as close to reality as possible, so that the simulation studies performed on these data give meaningful results with respect to the performance of the indicators on the national level. Second, the Austrian sample from 2006 did not contain any non-representative outliers, i.e., large incomes that are either incorrectly recorded or can be considered unique in the population. This is important for simulation studies focused on the evaluation of robust methods, where it is crucial that the amount of outliers in the samples can be controlled precisely. Thus, the underlying synthetic population should not contain any non-representative outliers, instead they should be included in the samples (see Alfons et al., 2011c).

Section 3.1.1 gives a detailed description of the variables available in AAT-SILC, including their possible outcomes. Afterwards, section 3.1.2 summarizes the simulation models and parameter settings used to generate the variables.

### 3.1.1 Description of the variables

In table 3.1 the basic variables of the synthetic population data AAT-SILC and their possible outcomes are listed. While `eqIncome` (*equivalized disposable income*) is of course of main interest for the simulation studies, most of the basic variables are categorical. Note that some categories of `pl030` (*self-defined current economic status*) and `pb220a` (*citizenship*), respectively, have been combined due to low frequencies of occurrence in the underlying survey sample. Such combined categories are marked with an asterisk (*) in table 3.1. It should also be noted that these two variables are only conducted in the survey for persons aged 16 or above. In order to avoid missing values in the synthetic population data for persons below age 16, a new category (*Not applicable*) has been added. This added category is marked with two asterisks (**) in table 3.1. Furthermore, the variables `hsize` (*household size*), `age`, `eqSS` (*equivalized household size*), `eqIncome` (*eqivalized disposable income*) and `main` (*main income holder*) are not included in the standardized format of EU-SILC data and have been derived from other variables for convenience. For a complete description of the variables included in EU-SILC and their possible outcomes, the reader is referred to EUROSTAT (2004).

In addition to the basic variables, most income components conducted in EU-SILC are available in the synthetic population data AAT-SILC. Nevertheless, some components were excluded from the data simulation process because they contain too few non-zero values in the underlying survey sample, e.g., the components `py020` (*non-cash employee income*) and `hy120n` (*regular taxes on wealth*) did not contain any non-zero values. Including those components would only cause an unnecessary increase in the file size of AAT-SILC. It is further important to note that the personal income components are only recorded in the survey for persons aged 16 or above. The values of persons below age 16 have thus been set to zero to avoid missing values in the synthetic population. This strategy is reasonable since the income of persons below age 16 is recorded in the household income component `hy110n`. In any case, tables 3.2 and 3.3 list the personal income components and household income components, respectively, which are included in AAT-SILC.

However, using all 16 available income components to evaluate complex multivariate procedures in simulation studies with a large number of samples would be computationally extremely expensive. Hence, ALFONS et al. (2011a) suggested to limit the multivariate setting for the simulation studies to four aggregated components. The aggregated income components available in AAT-SILC are listed in table 3.4.

### 3.1.2 Models used to generate the variables

A detailed mathematical description of the models is given in (ALFONS et al., 2011b). How to use the R package `simPopulation` (ALFONS and KRAFT, 2010) is illustrated in the package vignette `simPopulation-eusilc` (ALFONS et al., 2010). If `simPopulation` is installed, the following command can be used to view the vignette from within R:

```
R> vignette("simPopulation-eusilc")
```

Table 3.1: Basic variables in the synthetic population data AAT-SILC.

| **Variable** | **Name** | **Possible outcomes** | |
|---|---|---|---|
| Household ID | `db030` | | Unique integer identifier of household |
| Household size | `hsize` | | Number of persons in household |
| Region | `db040` | 1 | Burgenland |
| | | 2 | Lower Austria |
| | | 3 | Vienna |
| | | 4 | Carinthia |
| | | 5 | Styria |
| | | 6 | Upper Austria |
| | | 7 | Salzburg |
| | | 8 | Tyrol |
| | | 9 | Vorarlberg |
| Degree of urbanisation | `db100` | 1 | Densely populated area |
| | | 2 | Intermediate area |
| | | 3 | Thinly populated area |
| Age | `age` | | Age (for the previous year) in years |
| Gender | `rb090` | 1 | Male |
| | | 2 | Female |
| Main activity status during the income reference period | `rb170` | 1 | At work |
| | | 2 | Unemployed |
| | | 3 | In retirement or in early retirement |
| | | 4 | Other inactive person |
| Self-defined current economic status | `pl030` | 1 | Working full-time |
| | | 2 | Working part-time |
| | | 3 | Unemployed |
| | | 4 | Pupil, student, further training or unpaid work experience or in compulsory military or community service* |
| | | 5 | In retirement or in early retirement or has given up business |
| | | 6 | Permanently disabled or/and unfit to work or other inactive person* |
| | | 7 | Fulfilling domestic tasks and care responsibilities |
| | | 8 | Not applicable** |
| Citizenship | `pb220a` | 1 | Austria |
| | | 2 | EU* |
| | | 3 | Other* |
| | | 4 | Not applicable** |
| Equivalized houshold size | `eqSS` | | Household size according to modified OECD scale |
| Equivalized disposable income | `eqIncome` | 0 | No income |
| | | $> 0$ | Income |
| Main income holder | `main` | `TRUE` | Person holds largest income in household |
| | | `FALSE` | Otherwise |

* combined categories
** added category to avoid `NA`s

Table 3.2: Personal income components in the synthetic population data AAT-SILC.

| Variable | Name | Possible outcomes | |
|---|---|---|---|
| Employee cash or near cash income | py010n | 0 | No income |
| | | $> 0$ | Income |
| Cash benefits or losses from self-employment | py050n | $< 0$ | Losses |
| | | 0 | No income |
| | | $> 0$ | Benefits |
| Unemployment benefits | py090n | 0 | No income |
| | | $> 0$ | Income |
| Old-age benefits | py100n | 0 | No income |
| | | $> 0$ | Income |
| Survivor's benefits | py110n | 0 | No income |
| | | $> 0$ | Income |
| Sickness benefits | py120n | 0 | No income |
| | | $> 0$ | Income |
| Disability benefits | py130n | 0 | No income |
| | | $> 0$ | Income |
| Education-related allowances | py140n | 0 | No income |
| | | $> 0$ | Income |

Table 3.3: Household income components in the synthetic population data AAT-SILC.

| Variable | Name | Possible outcomes | |
|---|---|---|---|
| Income from rental of a property or land | hy040n | $< 0$ | Losses |
| | | 0 | No income |
| | | $> 0$ | Income |
| Family/children related allowances | hy050n | 0 | No income |
| | | $> 0$ | Income |
| Housing allowances | hy070n | 0 | No income |
| | | $> 0$ | Income |
| Regular inter-household cash transfer received | hy080n | 0 | No transfer |
| | | $> 0$ | Transfer |
| Interest, dividends, profit from capital investments in unincorporated business | hy090n | 0 | No income |
| | | $> 0$ | Income |
| Income received by people aged under 16 | hy110n | 0 | No income |
| | | $> 0$ | Income |
| Regular inter-household cash transfer paid | hy130n | 0 | No transfer |
| | | $> 0$ | Transfer |
| Repayments/receipts for tax adjustment | hy145n | $< 0$ | Receipts |
| | | 0 | No income |
| | | $> 0$ | Repayments |

Table 3.4: Aggregated income components in the synthetic population data AAT-SILC.

| Variable | Name | Possible outcomes | |
|---|---|---|---|
| Personal income from employment | `pye` | $< 0$ | Losses |
| | | 0 | No income |
| | | $> 0$ | Income |
| Personal income from transfers | `pye` | 0 | No income |
| | | $> 0$ | Income |
| Household income from capital | `pye` | $< 0$ | Losses |
| | | 0 | No income |
| | | $> 0$ | Income |
| Household income from employment and transfers | `pye` | $< 0$ | Losses |
| | | 0 | No income |
| | | $> 0$ | Income |

Note that observations with negative personal net income or household net income of less than $-10\,000$ € (i.e., too large losses on the household level) were disregarded for the generation of AAT-SILC, since this lead to a considerable number of households with negative equivalized disposable income and poor fit in the lower tail of the distribution. However, while only very few observations of the original sample were removed because of this removal criterion, the resulting improvement in the fit of the equivalized disposable income is substantial.

### Household structure

The household structure is simulated by resampling households from the survey data conditional on the variables `db040` (*region*) and `hsize` (*household size*). First, the number of households in the population for each combination of region and household size is determined by the Horvitz-Thompson estimator (Horvitz and Thompson, 1952), i.e., by the sum of the sample weights of the corresponding observations. Second, households are resampled separately for each combination of region and household size. The probability of each sample household to be chosen is thereby determined by its sample weight. For each household in the population, the values of all household members for certain basic variables are adopted from the respective sample household. Note that the variables `db040` and `hsize` are immediately available in the synthetic population data as a by-product. Resampling the variables `age` and `rb090` (*gender*) ensures sensible correlation structures within the households. The variable `db030` (*household ID*) is then simply generated by assigning the simulated households consecutive integer numbers.

### Additional categorical variables

For the synthetic data set AAT-SILC, the main aim for the simulation of additional categorical variables is to generate good predictors for the income variables.

Table 3.5: Categorized variables created for use as predictors during the simulation of the synthetic population data AAT-SILC.

| **Variable** | **Categories** |
|---|---|
| Age category | $\leq 15$, $(15, 20]$, $(20, 25]$, $(25, 30]$, $(30, 35]$, $(35, 40]$, $(40, 45]$, $(45, 50]$, $(50, 55]$, $(55, 60]$, $(60, 65]$, $(65, 70]$, $(70, 75]$, $(75, 80]$, $> 80$ |
| Personal net income category (for multinomial model) | $0$, $(0, 800]$, $(800, 2800]$, $(2800, 5012]$, $(5012, 8431.59]$, $(8431.59, 11200]$, $(11200, 13664]$, $(13664, 15428.26]$, $(15428.26, 17675]$,$(17675, 20066.67]$, $(20066.67, 23520]$, $(23520, 29085.30]$,$(29085.30, 36000]$,$(36000, 56548.35]$, $> 56548.35$ |
| Personal net income category (for components) | $0$, $(0, 800]$, $(800, 2800]$, $(2800, 5012]$, $(5012, 8431.59]$, $(8431.59, 13664]$, $(13664, 17675]$, $(17675, 23520]$, $(23520, 29085.30]$, $(29085.30, 36000]$, $(36000, 56548.35]$, $> 56548.35$ |
| Equivalized personal net income category | $0$, $(0, 2088]$, $(2088, 6500]$, $(6500, 8610.59]$, $(8610.59, 10666.67]$, $(10666.67, 12798.88]$, $(12798.88, 14826.67]$, $(14826.67, 16800.61]$, $(16800.61, 18823.07]$, $(18823.07, 21480.17]$, $(21480.17, 24693.18]$, $(24693.18, 30000]$, $(30000, 36000]$, $(36000, 53519.11]$, $> 53519.11$ |
| Household net income category | $[-10000, -5000)$, $[-5000, -2500)$, $[-2500, 0)$, $0$, $(0, 431]$, $(431, 1342]$, $(1342, 2471.60]$, $(2471.60, 4293.20]$, $(4293.20, 5676.80]$,$(5676.80, 7161.80]$, $(7161.80, 9011]$, $(9011, 11705.04]$, $(11705.04, 14994.20]$, $(14994.20, 21790]$, $> 21790$ |

For the simulation of additional variables on the personal level, age categories are built in order to reduce the computational effort. Table 3.5 lists the age categories thereby used. This categorization is retained throughout the rest of the data generation, so whenever age is mentioned in this section from now on, it actually refers to age categories rather than the precise age.

The additional categorical variables are each generated with the following procedure, which is performed separately for each region (given by variable `db040`).

1. Fit a multinomial logistic regression model with suitable predictors to the sample data taking the sample weights into account.
2. Predict the probabilities for each outcome of the response conditional on the outcomes of the predictor variables.
3. Draw the realization for each observation in the synthetic population from the respective conditional probability distribution.

**Degree of urbanization** The variable `db100` (*degree of urbanization*) is generated on the household level, i.e., households are used as observations rather than persons. It

Table 3.6: Generation of `rb170` (*main activity status during the income reference period*) from `pl030` (*self-defined current economic status*).

| `rb170` | | `pl030` | |
|---|---|---|---|
| 1 | At work | 1 | Working full-time |
| | | 2 | Working part-time |
| 2 | Unemployed | 3 | Unemployed |
| 3 | In retirement or early retirement | 5 | In retirement or in early retirement or has given up business (if `age` $> 45$) |
| 4 | Other inactive person | 4 | Pupil, student, further training or unpaid work experience or in compulsory military or community service |
| | | 5 | In retirement or in early retirement or has given up business (if `age` $\leq 45$) |
| | | 6 | Permanently disabled or/and unfit to work or other inactive person |
| | | 7 | Fulfilling domestic tasks and care responsibilities |
| | | 8 | Not applicable |

should be noted that the region *Vienna* is treated as a special case. Since the whole region is densely populated, the value of `db100` is set to one for all observations. For each other region, `db100` is simulated by a weighted multinomial model with predictor `hsize` (*household size*) as described above.

**Main activity status and economic status** First of all, it is important to note that the variable `rb170` (*main activity status during the income reference period*) is not available in the original EU-SILC sample provided by Statistics Austria. Therefore, the variable `pl030` (*self-defined current economic status*) is simulated beforehand. Then, variable `rb170` is constructed by combining categories from `pl030`. The conversion of categories is shown in table 3.6. In any case, `pl030` is simulated by weighted multinomial models with predictors *age category*, `rb090` (*gender*), `hsize` (*household size*) and `db100` (*degree of urbanization*).

**Citizenship** The variable `pb220a` (*citizenship*) is simulated by weighted multinomial models with predictors *age category*, `rb090` (*gender*), `hsize` (*household size*), `db100` (*degree of urbanization*) and `pl030` (*self-defined current economic status*).

### Income variables

Concerning the income variables, `eqIncome` (*equivalized disposable income*) is of main interest. It is generated from two parts which are simulated separately: the *personal net income* and the *household net income*. Each of these parts is then further split into components for the evaluation of multivariate procedures in simulation studies.

The generation of personal net income and household net income is based on the procedure for categorical variables described above:

1. Discretize the continuous income variable in the sample data.
2. Simulate the income categories for the synthetic population with the procedure based on multinomial logistic regression models.
3. Draw values of the observations in the synthetic population from uniform distributions within the assigned income category, except for the largest category. There the values are drawn from a truncated generalized Pareto distribution (GPD; e.g. KLEIBER and KOTZ, 2003) which is fitted to the sample data.

KRAFT (2009) and ALFONS et al. (2011b) also proposed a procedure based on two-step regression models, but the results they present clearly favor the approach based on multinomial logistic regression models.

In addition, the income components for each of these two variables are generated based on conditional resampling of fractions. Only very few highly influential categorical variables should thereby be used as conditioning variables. The procedure for simulating income components is summarized by the following two steps:

1. According to the value of the conditioning variables, draw the fractions of the components from the respective subset in the sample data. The probability of selection for each observation in the sample is thereby proportional to its sample weight.
2. Multiply the simulated fractions by the total income of the corresponding observation in the synthetic population in order to obtain absolute values.

This simplified procedure based on resampling is chosen for two reasons. First, the dependencies between the components are too complex to consider all of them. Second, the income components in the survey sample are very sparse, i.e., they contain a large amount of zeros. Figure 3.1 shows the percentage of zeros in the income components. Note that the sample weights are considered in the computation of the percentages and the household income components are obtained using the households as observations rather than the persons.

However, before the simulation of the income variables is further discussed, the generation of the variable `eqSS` (*equivalized household size*) needs to be described. Not only is `eqSS` necessary for the computation of `eqIncome` (*equivalized disposable income*), but it is also used for the simulation of the household net income.

**Equivalized household size** The variable `eqSS` (*equivalized household size*) is computed according to the modified OECD scale: for each household, a weight of 1.0 is given to the first adult, 0.5 to other household members aged 14 or over, and 0.3 to household members aged less than 14 (EUROSTAT, 2004, 2009).

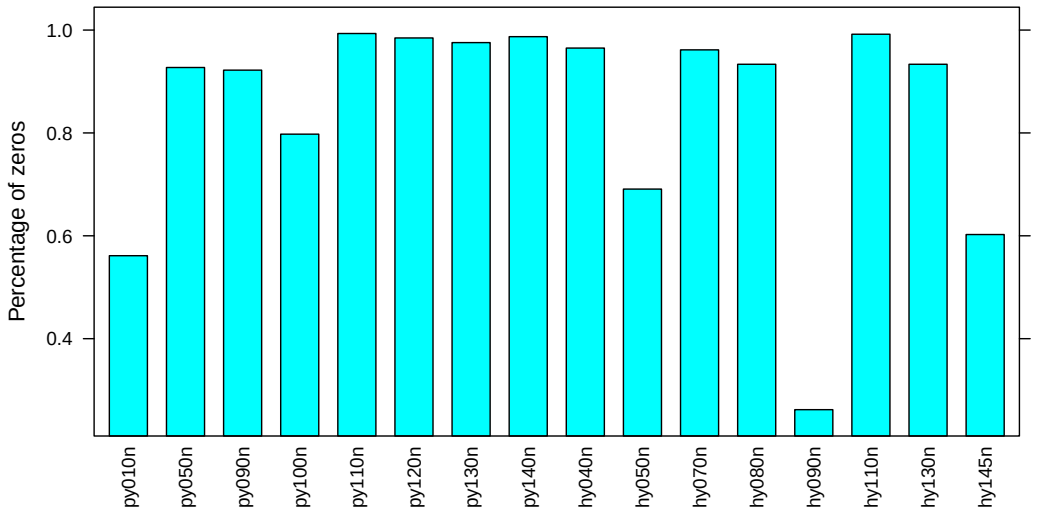


Figure 3.1: Percentage of zeros in the income components. Percentages for the household components are computed on the household level rather than the personal level.

**Personal net income** Since personal net income is a semi-continuous variable, zero is a category of its own in the categorization of the variable. The other breakpoints are given by the weighted 1%, 5%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 95% and 99% quantiles of the positive values. The resulting categories are listed in table 3.5. The personal net income is then simulated with the procedure based on multinomial logistic regression models as described above, thereby using the predictors *age category*, `rb090` (*gender*), `hsize` (*household size*), `db100` (*degree of urbanization*), `pl030` (*self-defined current economic status*) and `pb220a` (*citizenship*). It should be noted that the personal net income is not included in the final AAT-SILC data set to keep the file size reasonable, as it can easily be reconstructed from the personal income components.

**Personal net income components** For the generation of the personal income components listed in table 3.2, it is quite natural to use the the categorized personal net income as one of the conditioning variables. However, Kraft (2009) suggests to use fewer income categories than for the multinomial models in the simulation of personal net income. Thus, the breakpoints for the categorization are limited to the weighted 1%, 5%, 10%, 20%, 40%, 60%, 80%, 90%, 95% and 99% quantiles of the positive values. Table 3.5 lists the resulting categories. Then, the personal income components are simulated by resampling fractions conditional on those broader personal net income categories and `pl030` (*self-defined current economic status*).

**Main income holder** As the name suggests, the variable `main` (*main income holder*) is simply given by assigning `TRUE` to the persons with the highest personal net income in the respective households, and `FALSE` to all other persons.

**Household net income** First of all, the household net income is generated on the household level, using households as observations rather than persons. Moreover, household net income is somewhat more complicated to simulate than personal net income. It contains a considerable amount of negative values, and its distribution is more right-skewed, as there are many zeros and small positive values, but also some very high values. Consequently, the categorization for the multinomial models is more complex. For the negative values, the breakpoints $-10\,000$, $-5\,000$ and $-2\,500$ are used. Since the household net income is semi-continuous, 0 is a category of its own. In addition, the breakpoints for the positive values given by their 40%, 50%, 60%, 70%, 80%, 85%, 90%, 95%, 97.5% and 99% quantiles. The resulting categories are listed in table 3.5. For the simulation procedure based on weighted multinomial logistic regression models, the following predictors are used: *age category*, `rb090` (*gender*), `hsize` (*household size*), `db100` (*degree of urbanization*), `pl030` (*self-defined current economic status*), `pb220a` (*citizenship*), *number of persons below age 16*, and *equivalized personal net income category*. Values for *age category*, `rb090`, `pl030` and `pb220a` thereby refer to the values of the main income holder. As the name suggests, *number of persons below age 16* for each household simply counts the number of persons aged under 16. However, the construction of the predictor *equivalized personal net income category* is more complex. It is generated by computing the sum of the personal net income of all persons in a household and dividing this sum by the equivalized household size. Afterwards, it is categorized using the weighted 1%, 5%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 95% and 99% quantiles, while zero is a category of its own (see table 3.5 for the resulting categories).

**Household net income components** Of course, also the household net income components are generated using households as observations rather than persons. Using the number of persons below age 16 as one of the conditioning variables seems reasonable for the simulation of the household income components, since many of those components are related to families or children, e.g., `hy050n` (*family/children related allowances*), `hy110n` (*income received by people aged under 16*), `hy080n`/`hy130n` (*inter-household cash transfer received/paid*). In short, the household income components are simulated by resampling fractions conditional on the number of persons below age 16 and the household net income category. Note that unlike for the personal net income components, all household net income categories from the multinomial models in the simulation of household net incomes are used for conditioning.

**Equivalized disposable income** For each household, the value of `eqIncome` (*equivalized disposable income*) is obtained by first computing the sum of the personal net income of all persons in a household plus the household net income, and then dividing this sum by the equivalized household size (for details, see EUROSTAT, 2004, 2009).

**Aggregated income components** In order to simplify the multivariate settings for the simulation studies within the project, the four aggregated income components are computed from the available 16 components in the following manner (using R syntax, see also ALFONS et al., 2011a):

- Personal income from employment: `pye <- py010n + py050n`

- Personal income from transfers:
  `pyt <- py090n + py100n + py110n + py120n + py130n + py140n`
- Household income from capital: `hyc <- hy040n + hy090n`
- Household income from employment and transfers:
  `hyet <- hy050n + hy070n + hy080n + hy110n - hy130n - hy145n`

## 3.2 Selected results

In this section the quality of the synthetic population data AAT-SILC is evaluated by diagnostic plots and by comparing certain quantities of interest to their counterparts from the underlying survey sample.

Figure 3.2 shows mosaic plots that display the expected and realized frequencies of gender, region and household size (*top*), as well as gender, economic status and citizenship (*bottom*). In both pairs of plots, very similar structures in the sample and population data are visible. Note that in the bottom plots, the added category *not applicable* for economic status and citizenship (see table 3.1) is disregarded. Since this category affects the same observations in both variables, disregarding it actually makes the plots more readable. Furthermore, these two specific combinations of variables have been selected representatively. The number of possible combinations of categorical variables is simply too large to show all of them. Nevertheless, the interactions between all categorical variables are in general very well reflected in AAT-SILC, which is further documented later on by the contingency coefficients in table 3.7.

However, while the two mosaic plots at the top of figure 3.2 are nearly identical, small differences can be seen in the two plots at the bottom. The differences in the latter two plots are due to the use of multinomial logistic regression models. On the one hand, the expected frequencies of the different combinations are determined by their Horvitz-Thompson estimates, i.e., by the sum of the corresponding sample weights. On the other hand, the multinomial models in the data generation procedure allow to simulate combinations that do not occur in the sample, but are likely to occur in the real population. As a consequence, a possible interpretation of the small differences is that they are corrections of the expected frequencies.

In addition to the diagnostic mosaic plots, the interactions between the categorical variables in AAT-SILC are evaluated by contingency coefficients. Pearson's coefficient of contingency is a measure of association for categorical data. It is defined as

$$P = \sqrt{\frac{\chi^2}{n+\chi^2}}, \tag{3.1}$$

where $\chi^2$ is the test statistic of the $\chi^2$ test of independence and $n$ is the number of observations (see, e.g., Kendall and Stuart, 1967).

Table 3.7 compares contingency coefficients obtained from the survey sample to those obtained from the synthetic population AAT-SILC. The values for the sample data are thereby based on weighted distributions. The only significant difference occurs for the

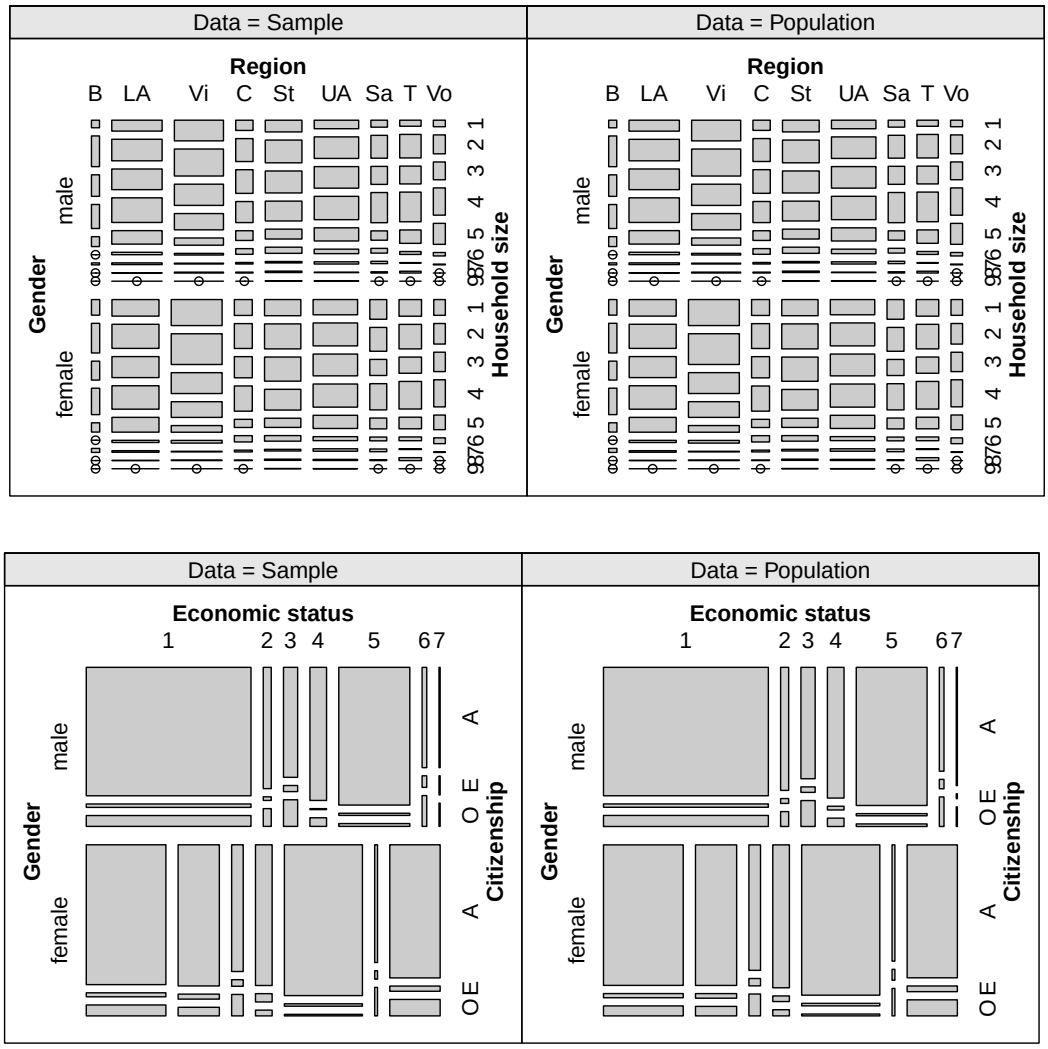


Figure 3.2: *Top*: Mosaic plots of gender, region and household size. *Bottom*: Mosaic plots of gender, economic status and citizenship.

contingency coefficient of `db100` (*degree of urbanization*) and `age`. This is a result of using simplified multinomial models to simulate `db100` on the household level which do not consider the age of the household members. All in all, the correlation structure of the underlying sample is very well reflected in AAT-SILC.

In figure 3.3 (*left*), the cumulative distribution functions (CDF) of the equivalized disposable income in AAT-SILC is compared to the empirical CDF obtained from the underlying survey sample. For the latter, sample weights are taken into account by adjusting the step height. Note that the plot shows only the main parts of the data (from 0 to the weighted 99% quantile of the positive values in the sample) for better visibility of the differences. Even though there are some deviations, the CDFs indicate an excellent fit. These differences are due to the complex structure of the household income. ALFONS et al. (2011b) showed that the proposed data simulation methodology leads to almost indistinguishable CDFs for the personal net income, although it should be noted that the models have been slightly adjusted for the final version of AAT-SILC.

Table 3.7: Pairwise contingency coefficients of the categorical variables for the survey sample and synthetic population data AAT-SILC.

| | | db040 | db100 | age | rb090 | pl030 | pb220a |
|---|---|---|---|---|---|---|---|
| Sample | hsize | 0.216 | 0.208 | 0.546 | 0.083 | 0.479 | 0.371 |
| | db040 | | 0.587 | 0.262 | 0.020 | 0.137 | 0.156 |
| | db100 | | | 0.150 | 0.015 | 0.097 | 0.150 |
| | age | | | | 0.118 | 0.823 | 0.714 |
| | rb090 | | | | | 0.356 | 0.034 |
| | pl030 | | | | | | 0.712 |
| Population | hsize | 0.216 | 0.208 | 0.546 | 0.082 | 0.479 | 0.371 |
| | db040 | | 0.587 | 0.262 | 0.020 | 0.137 | 0.156 |
| | db100 | | | 0.100 | 0.015 | 0.098 | 0.149 |
| | age | | | | 0.118 | 0.820 | 0.711 |
| | rb090 | | | | | 0.356 | 0.033 |
| | pl030 | | | | | | 0.712 |
| Relative difference (in %) | hsize | 0.042 | 0.111 | 0.012 | −0.408 | 0.051 | −0.189 |
| | db040 | | −0.004 | −0.099 | 2.425 | 0.136 | −0.139 |
| | db100 | | | −33.532 | −1.353 | 0.171 | −1.092 |
| | age | | | | −0.135 | −0.415 | −0.378 |
| | rb090 | | | | | 0.115 | −1.256 |
| | pl030 | | | | | | −0.028 |

Figure 3.3 (*right*) compares the distributions using box plots. In order to better visualize semi-continuous variables, the box plots are adapted in the following way. While box and whiskers are computed only for the non-zero part of the data, the box widths are proportional to the ratio of non-zero observations to the total number of observed values. For the survey sample, the box plot statistics and the box widths are computed such that sample weights are taken into account. However, points outside the extremes of the whiskers are not plotted due to the large number of observations in the synthetic population. Clearly, the box plots suggest that the proportion of individuals with zero income and the distribution of non-zero income for the main part of the data are well reflected in AAT-SILC.

Table 3.8 evaluates the simulated equivalized disposable income based on various quantities of interest: the percentage of zeros, 5% quantile, median, mean, 95% quantile and standard deviation. For the survey sample, the sample weights are of course considered for the computation of the quantities of interest. Even though the relative difference for the percentages of zeros seems quite large, this is not a big problem as the absolute values are so small that this does not have a considerable impact on the quality of the synthetic population data. Otherwise, there is a noteworthy deviation in the 5% quantile. Considering the complex structure of the income data in EU-SILC, the relative differences indicate that the fit is quite good.

In order to investigate whether heterogeneities are well reflected in AAT-SILC, figure 3.4 contains box plots of the conditional distributions of the equivalized disposable income

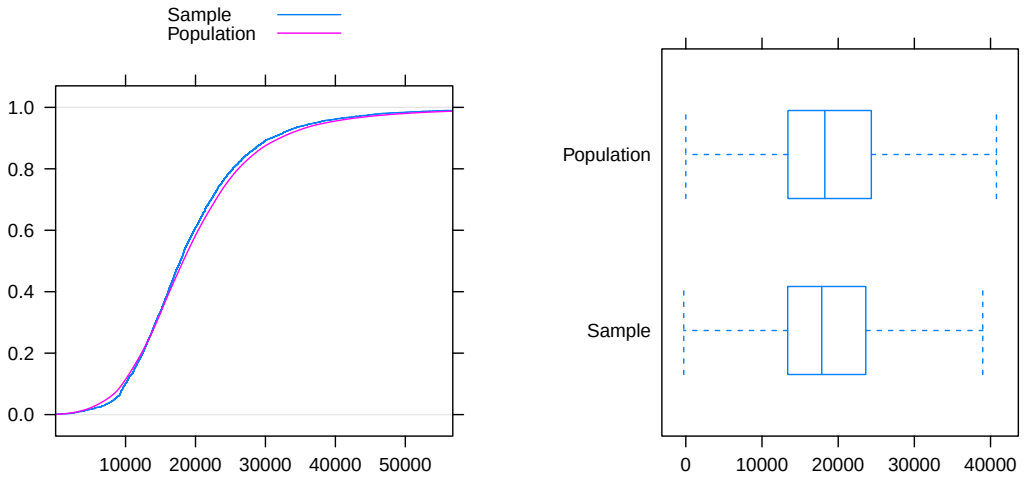


Figure 3.3: *Left*: Cumulative distribution functions of the equivalized disposable income. For better visibility, the plot shows only the main parts of the data. *Right:* Box plots of the equivalized disposable income. Points outside the extremes of the whiskers are not plotted.

Table 3.8: Evaluation of the equivalized disposable income based on the percentage of zeros, 5% quantile, median, mean, 95% quantile and standard deviation.

| | **%Zeros** | **5%** | **Median** | **Mean** | **95%** | **SD** |
|---|---|---|---|---|---|---|
| Sample | 0.02 | 8401.15 | 17845.33 | 19732.64 | 37225.56 | 10287.62 |
| Population | 0.02 | 7329.59 | 18247.41 | 20129.66 | 38703.37 | 10917.81 |
| Rel. difference (in %) | 35.89 | −12.75 | 2.25 | 2.01 | 3.97 | 6.13 |

with respect to gender (*top left*), citizenship (*top right*), region (*bottom left*) and economic status (*bottom right*). Only some of the smaller subgroups show significant deviations in the equivalized disposable income, in general the realized distributions are an excellent fit. While the heterogeneities in the underlying survey sample are not very distinct, the data simulation procedure succeeds in reflecting them in the synthetic population.

Last but not least, the aggregated income components are evaluated via box plots in figure 3.5. Since the components are semi-continuous variables i.e., contain a large number of zeros, the box plots are adapted in the following way. While box and whiskers are computed only for the non-zero part of the data, the box widths are proportional to the ratio of non-zero observations to the total number of observed values. For the survey sample, the box plot statistics and the box widths are computed such that sample weights are taken into account. Furthermore, the box plots for the household income components are computed using the households as observations rather than the persons. The box plots suggest that the synthetic population data performs well regarding the proportion of individuals with zero income and the distribution of non-zero income for the main part of the data. There are some significant differences for the component `hyc` (*household*

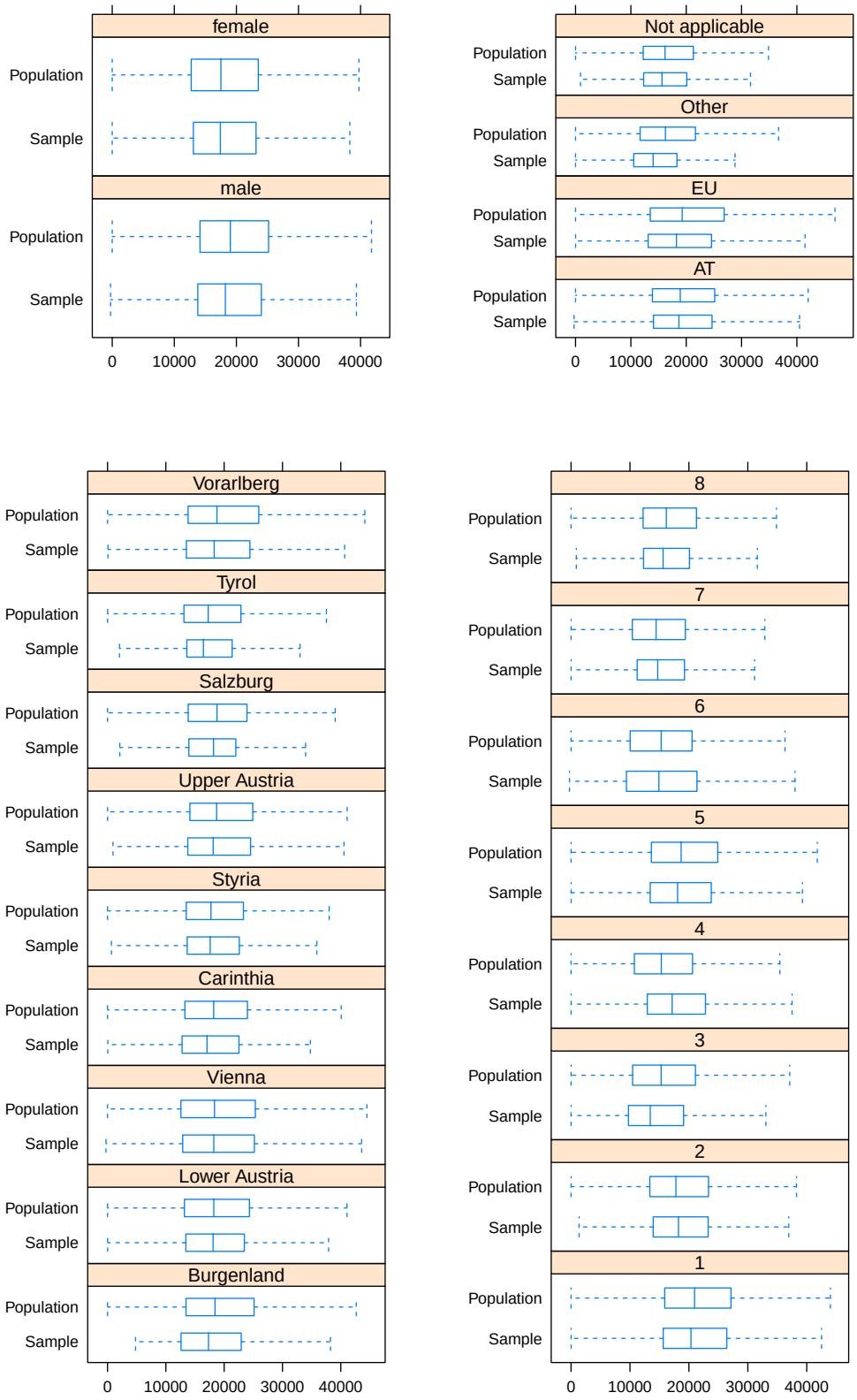


Figure 3.4: Box plots of the equivalized disposable income split by gender (*top left*), citizenship (*top right*), region (*bottom left*) and economic status (*bottom right*). Points outside the extremes of the whiskers are not plotted.

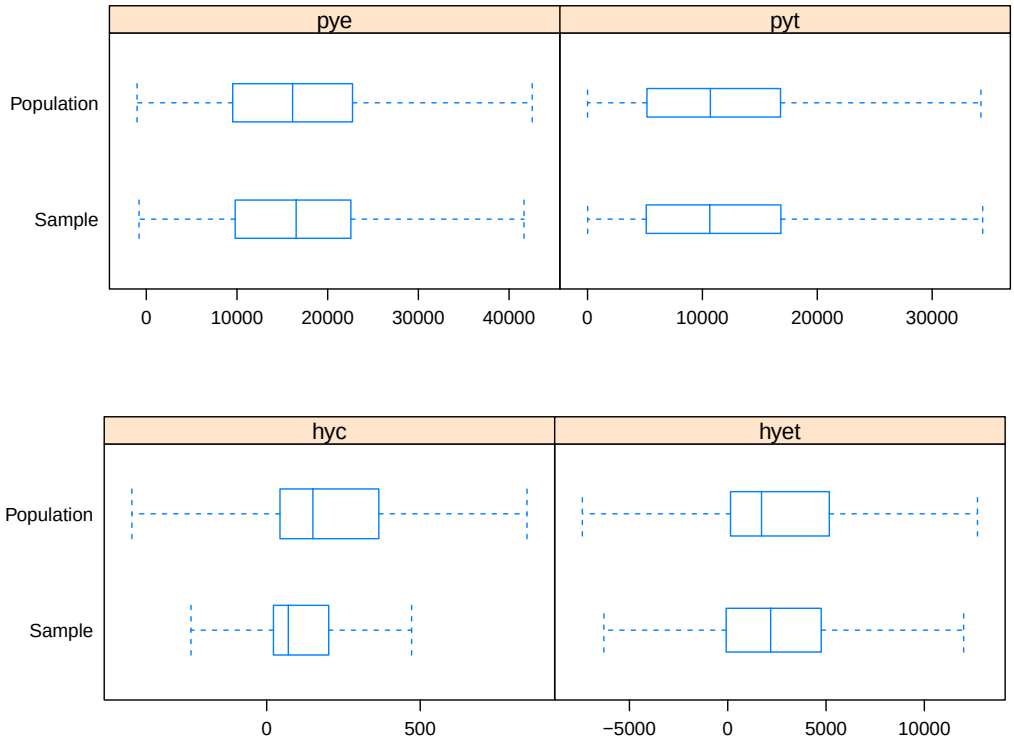


Figure 3.5: Box plots of the aggregated income components. Points outside the extremes of the whiskers are not plotted.

*income from capital*), but, due to the mostly small values compared to the other income components, these differences are negligible.

Additional results for the simulation of population data based on the Austrian EU-SILC sample from 2006 can be found in KRAFT (2009) and ALFONS et al. (2011b). The results in KRAFT (2009) correspond to a preliminary version of AAT-SILC and also include $\chi^2$ goodness of fit tests for categorical variables. ALFONS et al. (2011b) focus on the evaluation of the data simulation procedure itself. They present average results from multiple simulations based on the real survey sample, but also from multiple samples that have in turn been drawn from one specific synthetic population. In the latter case, even different sampling designs and sample sizes are investigated.

# Chapter 4

# The AMELIA data set

Beside the AAT-SILC, a second synthetic universe was generated within the AMELI project. As described in the introduction the reasons and the requirements for generating a synthetic universe as well as the underlying data basis can be different. This is the case for the situation at hand. One of the main differences between the above depicted AAT-SILC data set and the AMELIA data set, which will be described in the following, is the differing underlying data base. The Austrian part of the EU-SILC, the AT-SILC is the basis for the AAT synthetic population, while the scientific-use-file of the complete EU-SILC population delivered by Eurostat is the basis for the AMELIA population. Although this scientific-use-file of the EU-SILC data is not as critical under confidentiality reasons, disclosure control should arouse interest. In either case, it is useful to provide a free accessible data set. This is reasonable to have the possibility to reproduce, compare and understand results of simulation studies based on these data sets. The big advantage of a synthetic data set is that it can be published. Therefore, applied methods can be tested and evaluated by others on the same data. For these datasets it has to be impossible to link entries with real live persons. Like AAT-SILC, the AMELIA data set was also generated with the statistical environment R (R Development Core Team, 2011).

Another argument for the generation of a synthetic data set, beside a solution for the disclosure problem and the chance to get micro and macro structures, is the size of a potential size of the synthetic data frame. The personal EU-SILC data set for 2004 includes 241 796 entries for the EU-SILC countries. Some countries are not included in the data set at this stage. Compared to the whole population the sampling fraction is quite small. However, it is also interesting to simulate the process of data collection, whereas it is complicated to realize these simulations with the present EU-SILC data set.

Within the AMELI project, it was the target to investigate questions like the impact of influential units on the estimation results for small regional or conceptual units. Therefore, the supply of synthetic administrative structures beyond the regional level is necessary. The AMELIA synthetic universe is designed to allow for small area estimation. These administrative structures are of great importance to consider auxiliary information from other areas or spatial correlations. This is one potential advantage of synthetic micro data because regional information often lacks for real data sets due to disclosure reasons. Additionally, it is desirable to visualize regional indicators in maps, as this is an intuitive portrayal. The AMELIA data set is a synthetic data set and it is important to prevent

potential misleadings. Therefore, it is essential to make clear that it is not possible to draw conclusions from the contentual results. No statements with respect to content are intended. It is, for example, not intended to make the statement that the ARPR is at a certain level in Eastern Germany. The synthetic data set was generated to test methods. For making statements about content-related issues the real EU-SILC data should be utilized. For the AMELIA data set a synthetic map was designed alongside the generation of the synthetic universe to enable the realization of this visualization technique. In the following, the steps to create the synthetic data set AMELIA will be presented shortly.

## 4.1 Steps for the generation of AMELIA

To generate the synthetic data set, the following steps are proposed. The first step is to analyze the delivered EU-SILC data. It is necessary to structure and to classify the data set. General parameters have to be determined, one example here is the number of regions and the size of the synthetic population. The planned synthetic data set called *AMELIA* will comprehend about $N = 10$ millions entries. To allow for complex sampling schemes, it is crucial to generate administrative areas. Five levels of administrative areas are generated, these are depicted in table 4.1. For the AMELIA data set it was decided to create four regions. The idea behind this was to catch differences between Europe's regions. Further, an analysis of the incomes for the EU-SILC 2005 data set showed that outliers occur for some regions, whereas values in this magnitude do not occur for other countries. The regions have been created to level these differences between the countries.

The distribution of the EU-SILC countries into the four regions is shown in figure 4.1. Four regions are chosen to guarantee enough observations in every region. Additionally, it was the target to have the same number of observations in every region. Great Britain, for example, is contained in the northern region to have enough observations in this region.

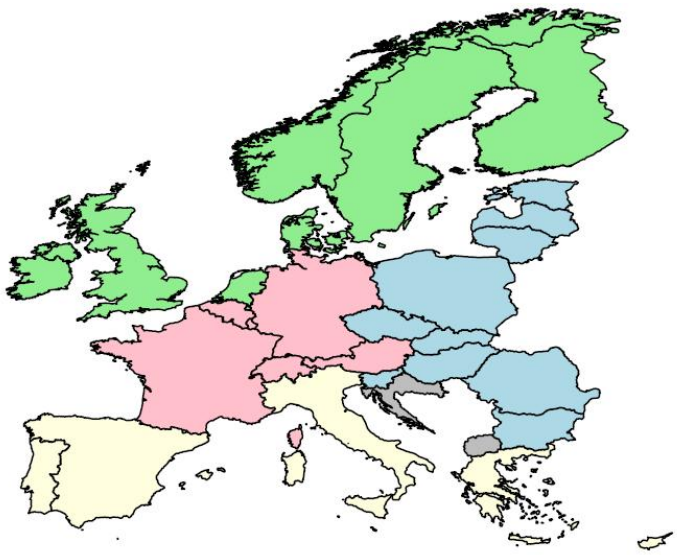

Figure 4.1: The classification of the EU-SILC countries into four regions.

As indicated above, it will be also interesting to control for spatial effects. Therefore, also a synthetic map was created.

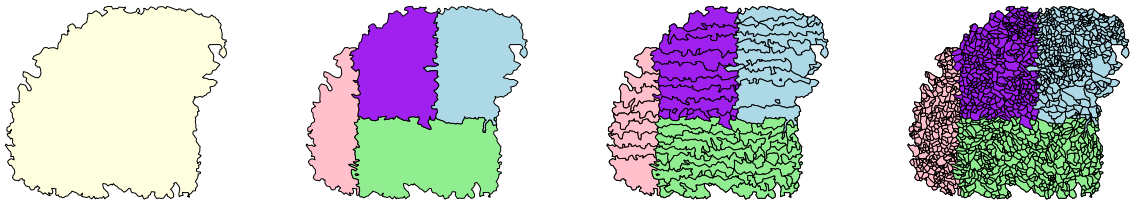

Figure 4.2: The administrative structure of AMELIA

As can be seen in figure 4.2, the four regions are further divided into counties (third picture), districts (fourth picture) and communities. The equivalent NUTS level can be seen in table 4.1 as well as the mean size, the minimum and maximum size of these levels is displayed in table 4.2.

| **NUTS level** | **Variable** | **Description** | **Number** |
|---|---|---|---|
| NUTS0 | AML | AMELIA complete | 1 |
| NUTS1 | REG | region | 4 |
| NUTS2 | NUTS2 | county | 11 |
| NUTS3 | DIS | district | 40 |
| LAU1 | CIT | community | 1592 |

Table 4.1: Administrative levels in AMELIA.

| **Variable** | (mean) **population** | Minimum | Maximum |
|---|---|---|---|
| AML | 10012600 | - | - |
| REG | 2503150 | 2230382 | 2673002 |
| NUTS2 | 910236.4 | 535313 | 1382586 |
| DIS | 250315 | 173485 | 369379 |
| CIT | 6289.322 | 2 | 268689 |

Table 4.2: Size of the administrative entities in AMELIA.

The German data set *Statistik lokal*[1] was the basis for the distribution of community sizes in the synthetic population. A peculiarity of this empirical distribution is the occurrence of very small communities, in the example at hand the smallest community has only two

[1]Information on this data base can be found at http://www.destatis.de/jetspeed/portal/cms/Sites/destatis/Internet/DE/Navigation/Publikationen/Querschnittsveroeffentlichungen/StatistikLokal,templateId=renderPrint.psml__nnn=true.

inhabitants. This can be problematic for the realization of the sampling scheme since too heterogeneous sizes of the entities are assumed to cause problems for the variance estimation. Therefore, an alternative indicator was introduced at this level.

As a next step the generation sequence (building blocks) has to be defined, therefore, it is necessary to know the structure of the basic data. The found structures should be preserved and new requirements for the synthetic data set result.

The analysis of the EU-SILC data set showed that this data set comprises different blocks of variables. Here it is possible to detect seven blocks of variables. These can be designated as income, education, economic profile, health, household structure, and structure variables in general.

The first generated block of variables affects household variables and variables which contain information about the biographical situations of persons.

As mentioned in the literature review in Chapter 2.2, the easiest advancement to create a synthetic data set would be to implement a kind of sampling with replacement. A reweighting method is often chosen to produce a data set. On the one hand, this proceeding is easy to implement but, on the other hand, the proceeding has the disadvantage that the rate of replication is quite high. In addition, disclosure is not made more complicated. The next possible method is the one proposed by BALLAS and CLARKE (2000). Here, one observation is generated by drawing from a theoretic distribution and by the subsequent generation of a variable outcome. The disadvantage here is the long duration of the iterative proceeding. Normally, IPF is used to generate tables by adjusting to known aggregated values. In the present case, this approach is not usable because the aggregated values for the synthetic country are not known. Another method implies the usage of linear or logit models, which are used for the generation of the AAT-SILC population. The reason why these methods are not applicable for the generation of the AMELIA population is the lack of high correlated covariates for the EU-SILC data set. Thus, a composition of these methods is used in the case at hand.

Sampling with replacement is necessary to create a bigger data set with more households than the original data set. Households are sampled with some key variables like sex, age, and marital status of the household members. Moreover, belonging to a different cluster is part of the first block of variables , e.g. whether one person belongs to an education cluster. Afterwards, one outcome is sampled out of the possible variable combinations in this cluster.

Implementing this approach the *household cross-sectional weight* (DB090) could serve as information to create inclusion probabilities $\pi_j$. However, the households are taken as sampling units to preserve the household structures. The task to create realistic synthetic household structures is difficult, therefore households rather than persons are taken as sampling units in the first stage. The problem with this approach is the high risk of replication within the data set. The average household size in the EU-SILC data set for 2004 is 2.63 persons per household, this implies about $M = 3.7$ million households for the synthetic population. High weights imply that the risk of replication is quite high for some households. Furthermore, uniqueness should be avoided, which means that characteristics in a table are so seldom that they allow the intruder to match them with other data bases which include the same variables (FIENBERG 2003). The approach is slightly modified in order to avoid disclosure problems.

Another problem, which has to be kept in mind, is that the data are output of stochastic processes. Therefore, a sampling error can occur, especially for spatial subgroups. If a sampling error exists for the basic data, it will be reproduced for the synthetic population (Huang and Williamson 2001, p.8).

To handle these problems, a first alternation to the basic sampling with replacement strategy is done. The variable set $f_i = (x_{i,1}, \ldots, x_{i,k}, \ldots, x_{i,K})$ has to be generated for every individual. $\text{Per}_i$ is spread into several blocks, whereas the blocks should be as far as possible independent from each other. The following symbols are proposed as a nomenclature for the generation of the AMELIA population.[2]

| | | | | |
|---|---|---|---|---|
| Strat | stratum | $1, \ldots, h, \ldots, H$ | H | number of strata |
| HH | household | $1, \ldots, j, \ldots, M$ | M | number of households |
| Per | person | $1, \ldots, i, \ldots, N$ | N | number of persons |
| X | variable | $1, \ldots, k, \ldots, K$ | K | number of variables |
| f | variable set | $1, \ldots, cl, \ldots, B$ | B | number of variable sets |

A first block $f^{1*} = (x_1^*, \ldots, x_K^*)$ then comprehends demographic variables like age, sex and marital status of the household members. The household ID is also one of the variables which belongs to the first block of variables to be generated. This is necessary to preserve the household structure. The real ID is directly signed over with a synthetic ID.

Key figures can be computed from this first block, like the *size of the household* (HX040 EU-SILC), the number of people per sex and the age group of the people in the household. These numbers are in parts relevant for the computation of equivalent household income.

To construct these first variables the proceeding proposed by Lehtonen et al. (2008) is used. This means that the weights $a_{jSILC} = 1/\pi_{jSILC}$ are used to replicate the households in the EU-SILC data set. Thus, a sample S is drawn with replacement $S = \text{HH}_1^*, \ldots \text{HH}_M{}^*$, where $\text{HH}_1^*, \ldots, \text{HH}_M^*$ are the households in the synthetic population. Together with the household. we do a sample a variable set of the first key variables $x^{1*}$ like it is described above. Further information about the membership of the variable combinations of individuals to latent classes is drawn within the same step: $f_{ij}^1 = f_{ij,1}, \ldots, f_{ij,2}, \ldots, f_{ij,K}$ for every individual $i$ in household $j$.

As described above, the original EU-SILC data set is divided into four regions. To introduce regional effects at this stage, it is possible to create one regional indicator per artificial region and to draw then only households with that region indicator. For region 1 for example, only households from northern countries can be drawn.

An alternative would be to allocate the households to regions according to socio-demographic criteria. Arbitrative for the allocation may be the average age of the first person in the household and the average *lowest monthly income to make ends meet* (HS130 EU-SILC). This allocation could be conducted with a cluster analysis. However, in this case the population would be very homogeneous within and very heterogeneous between the regions. The target of this proceeding is to account for heterogeneity between regions.

[2]Parameters with a asterisk at the top belong to the synthetic population and parameters without a star rely on the original population.

The number of households to be drawn per region $n^*_{hh,r}$ is then calculated with:

$$n^*_{hh,r} = \frac{n_{p,r}}{\sum n_{p,r} / \sum n_{hh,r}} \tag{4.1}$$

Where $n_{p,r}$ is the number of persons and $n_{hh,r}$ is the number of households in that region respectively. The population size per region, defined in the step before, is thus only a proxy, the resulting number of people in the artificial population can differ from this number.[3] It is important to mention that not every variable available in the EU-SILC data set is replicated with this kind of drawing. Only the first block of variables is generated with this proceeding.

The EU-SILC data set comprehends very specific variables with lots of outcomes. If the households would be replicated with the whole variable combination, it would be easy to identify households coming from the original data set. Therefore, only the set of variables $f^1_{ij}$ associated to persons in the drawn households are accounted.

In the following, recursive modelling is applied. A multivariate distribution $F(x_1, x_2, \ldots, x_n)$ has to be determined. The starting point are the variables $x_1, \ldots, x_k$ which are generated from the EU-SILC distributions, further variables can be calculated with:

$$F(x_{k+1}, \ldots, x_n | x_1, \ldots, x_k) = \frac{F(x_1,\ldots,x_n)}{F(x_1,\ldots,x_k)} \tag{4.2}$$

In general, the following is given:

$$F(x_1, \ldots, x_n) = F(x_1) \cdot F(x_2|x_1) \cdot \ldots \cdot F(x_n|x_1, \ldots, x_{n-1}) \tag{4.3}$$

Thus, the constituent (blocks of) variables can be calculated recursively. For every person in the data set an outcome for every variable is produced with draws from this multivariate distribution. In the following, a latent class analysis is used to generate further blocks for variables (see LINZER and LEWIS 2007 for further details).

### Modelling the income

The categorical variables resulting from this drawing procedure are used to create continuous variables like the *total household gross income* (HY010). This is only one of the variety of variables which treat the income aspect. Various income components are included in the EU-SILC data on personal and household level. The importance of these components is visible in figure 4.3. On the left-hand side, the personal income components are visible, whereas the significance of the various household income components is visible on the right. The graphics show the average income for each component in one income reference period. The most important income component on the personal level is the *Employee cash or near cash income* (PY010). The most important income on household level is, of course, the *total household gross income* (HY010).

[3] $n_{hh,r}$ has to be an integer, therefore the result has to be rounded.

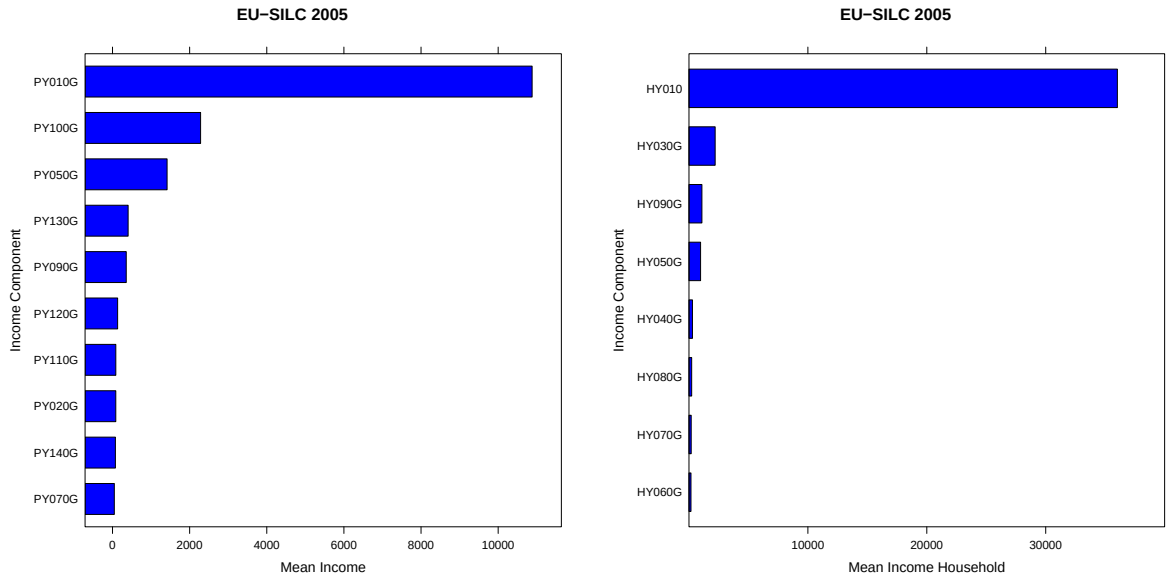


Figure 4.3: The importance of the income components.

The income components are generated step by step. Here, the interdependencies between the income components have to be considered. These interdependencies are visible in figure 4.4. In this graphic, the Bravais-Pearson correlation coefficient is plotted for every combination of income components. Red fields occur for combinations with very high correlation, blue fields display a weak correlation for this variable combination. Therefore, every field in the diagonal has to be red, because the correlation is one. The graphic shows that there are not as many income components with high correlations. One reason might be the high share of zero incomes for some income components.

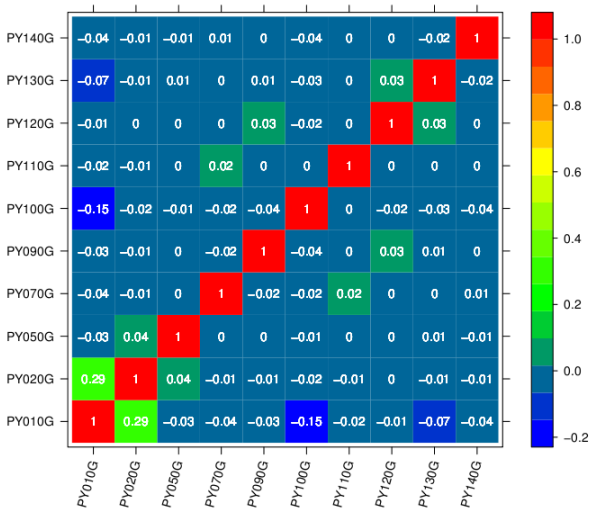


Figure 4.4: The interdependence between the income components

The basis for the computation of most indicators in the context of social exclusion and poverty is based on the equivalized household income. This income has to be computed

from different income components. Moreover, to receive a consistent picture, the income components are generated separately. The equivalized household income is computed in the following way:

$$\begin{aligned} HY010 = &\sum PY010G + \sum PY020G + \sum PY050G + \sum PY090G + \sum PY100G \\ &+ \sum PY110G + \sum PY120G + \sum PY130G + \sum PY140G + HY040G \\ &+ HY050G + HY060G + HY070G + HY080G + HY090G + HY110G \end{aligned} \tag{4.4}$$

The equivalized household income is computed with the total income of the household divided by the equivalent size of the household. The equivalent size is computed with the modified OECD scale. Following this scale, the first adult has a weight of 1.0. The weight 0.5 is assigned to every further person, provided that this person is older than 14 years. All other persons have the weight of 0.3.

After the computation of first indicators on poverty and social exclusion, it became apparent that adjustments to the income components are necessary because the values for Laeken Indicators computed from the synthetic population should be close to the real values for these indicators. After the basic data set is generated, slight changes will be implemented to this data set , i.e. the data set will be restructured. These changes are subject of scenario analysis, which implies that the basic data set stays unmodified. As a result of this, several entities will belong to another regional community after the restructuring.

Another possibility is to introduce extreme values by the recoding of the overall income. The most important insight is that the aggregate values do not change, in other words, the comparability of the data sets is preserved. Changes are only visible in some scenario variables.

The first scenario analysis treats the heterogeneities between and homogeneities within entities. These structures are assumed to have a big impact on estimation results. Therefore, a scenario with more heterogeneity was generated.

The last step is to check the above formulated requirements and to check for logic inconsistencies. Furthermore, it should be checked that disclosure is guaranteed.

## 4.2 Selected results

The AMELIA data set consits of 10 012 600 entries. These persons are distributed to 3 781 289 households. The data set has more than 40 variables including many income components. Table 4.3 gives an overview of some selected variables.

Table 4.3: Basic variables in the synthetic population data AAT-SILC.

| Variable | Name | Possible outcomes | |
|---|---|---|---|
| Household ID | HID | | Unique integer identifier of household |
| Household size | HHG | | Number of persons in household |
| Region | REG | 1 | Middle region |
| | | 2 | Northern region |
| | | 3 | Southern region |
| | | 4 | Eastern region |
| Degree of urbanization | DOU | 1 | Densely populated area |
| | | 2 | Intermediate area |
| | | 3 | Thinly populated area |
| Age | AGE | | Age (for the previous year) in years |
| Gender | SEX | 1 | Male |
| | | 2 | Female |
| Basic activity status | RB210 | 1 | At work |
| | | 2 | Unemployed |
| | | 3 | In retirement or in early retirement |
| | | 4 | Other inactive person |

The main focus of this population was to consider heterogeneities. Therefore, mainly these population characteristics are described in the following. The share of unemployed persons is shown in figure 4.5. It is visible that this share differs widely between the regions in the basic EU-SILC data set and the AMELIA population. On the left-hand side, the share of unemployed people per region in the EU-SILC data set of 2005 is presented. Every bar represents the share of this population group in one region. The share differs between about 12 % and 2 %. It is also visible that this information is not available for every region. In the figure on the right, the shares for the artificial AMELIA population are depicted. Here it is visible that more regions exist, also the information about unemployed people is available for every regional entity.

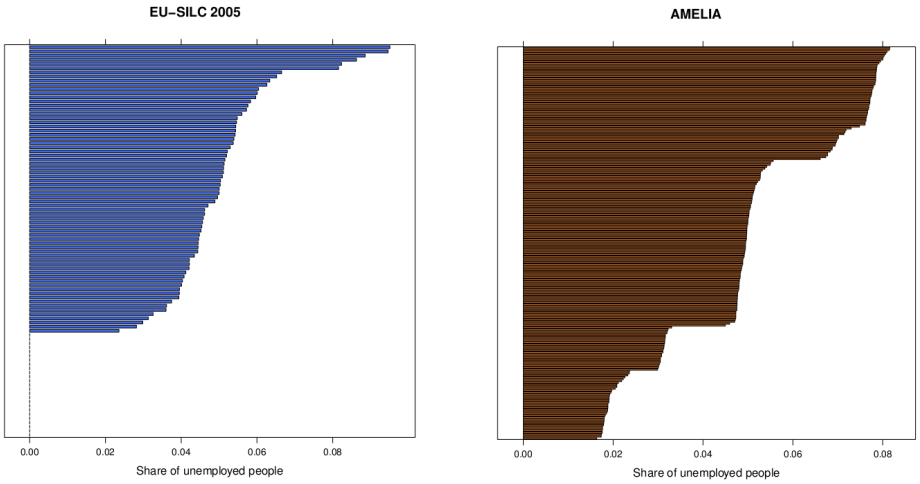


Figure 4.5: Share of unemployed persons per area.

As mentioned in the introduction, it is interesting and intended by the European Commission to compare results for poverty and social exclusion statistics between different thematic subgroups. The synthetic data set AMELIA is obviously based on the EU-SILC data files. However, there is more information which should be taken into consideration to generate the synthetic data set. The target of the AMELI project is to evaluate estimation functions for the indicators on poverty and social exclusion. Values for these statistics should be in an interval, in which they are in reality as well.

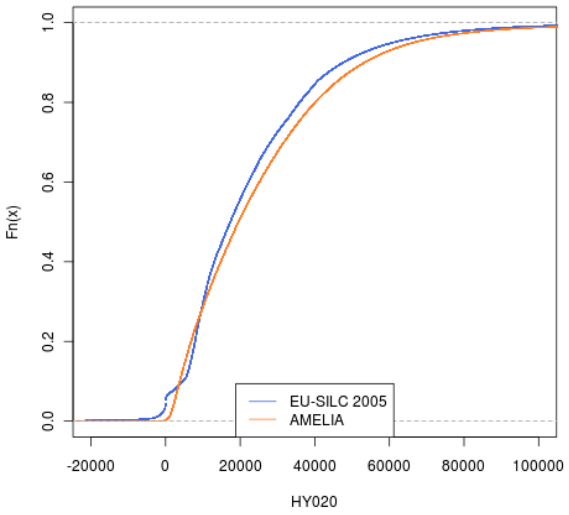


Figure 4.6: Cumulative distribution functions of the household income HY020.

Figure 4.6 shows the CDF of the household income HY020 for the basic EU-SILC population and the artificial AMELIA population. The CDF for the AMELIA distribution lies slightly below the CDF of the EU-SILC population. This is due to the adjustments which were necessary to reach the realistic range of poverty and social inclusion indicators.

It is possible in the case of the EU-SILC population, that distortions from this values do occur if the statistics are computed on the scientific-use-file. Some of these problems occur, because special observations are treated differently. Negative incomes and extremely high incomes are designated as special observations in this context.

In figure 4.7 the mean income per area is plotted. The green dots are representing mean incomes in the different cities of the synthetic universe, whereas the blue points are representing the mean incomes per country in the regions of the EU-SILC 2005 population. The areas are sorted in descending order by average income. It is, first of all, visible that the number of cities in the synthetic universe is much higher than the number of regions in the EU-SILC basic population. In both cases the smallest average income is about 5000 Euro in one income reference period. The highest average income for one region in the EU-SILC data set is much higher than the highest average income in the case of AMELIA. This implies that the synthetic population might be much more homogeneous than the basic EU-SILC population.

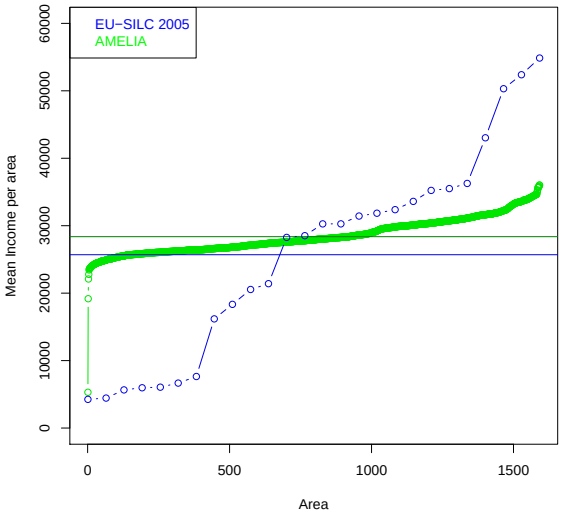


Figure 4.7: Average incomes for areas in the EU-SILC 2005 and the AMELIA population.

This impression was the point of departure for the introduction of the scenario analysis. It was the objective to get a scenario where the observations of the data set remain unchanged but the allocation of the persons to cities within one region are changed. With this reallocation more heterogeneity was reached in general (see figure 4.8). In addition the income for the personal income component PY010 is allocated with more spatial heterogeneity, which is visible in the right panel of figure 4.8.

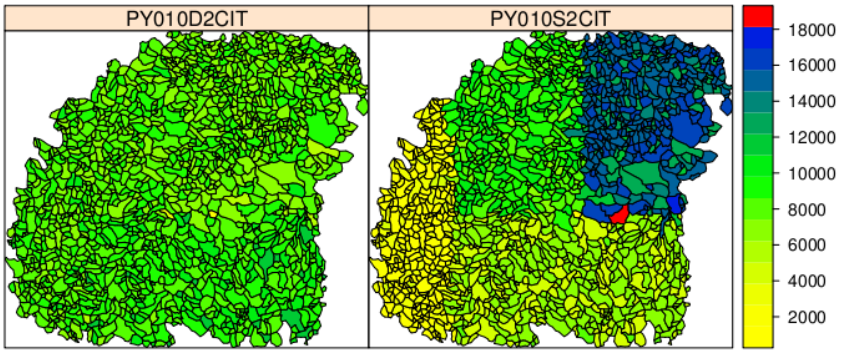


Figure 4.8: Two scenarios for the spatial distribution of the income component PY010

# Chapter 5

# Summary

To run simulations and to investigate the methodology developed in WP 2, 3, and 4 in a close-to-reality environment, realistic data sets have to be simulated. This allows for elaborating recommendations on methods and its applications for estimating the social inclusion indicators.

The target of this work package was to generate a data set which is an appropriate basis for the simulations within the AMELI project. For the generation of a synthetic data set different problems had to be addressed. A fine administrative structuring was necessary to realize the complex sampling designs. Furthermore, many income components and their correlation structure had to be reproduced for the synthetic data set.

It is important to keep in mind that many methods are conceivable to produce a synthetic data set and the best method depends always on the starting point and the requirements coming from the simulations. The description of work scheduled a lot of simulations and different data sets have been available. Due to these reasons it was decided to generate two different synthetic data sets, which are the AAT-SILC data set and the AMELIA data set. This two methods to simulate population microdata have been discussed in this contribution.

The first approach discusses how the Austrian EU-SILC household population data can be generated from a given sample by sequentially simulating variables using a model-based approach. While in this contribution mainly the application to the Austrian data is discussed, the approach and the corresponding software (Alfons and Kraft, 2010) can be used to simulate population data in a quite general manner (Alfons et al., 2011b). In Templ and Alfons (2010) it is shown that such synthetic data sets fulfills the requirements of confidentiality.

In the second approach, regional effects are considered when generating the universe. The difference to the first approach is that the user can vary such differences, while in the first approach the regional differences are modelled from the underlying information of the sample only. This allows for investigations in small area applications.

Both data sets can be obtained from the AMELI website, and the software used to simulate the Austrian EU-SILC data is freely available at the comprehensive R archive network (http://cran.r-project.org/package=simPopulation).

As mentioned above, with the synthetic data sets generated, all the simulations can be carried out in a realistic manner. The analysis of these simulations to support in policy making is undertaken in WP 7, 9, and 10.

One has to bear in mind that only those effects can be measured in microsimulation approaches which have been included into the data set. There is always the risk to get in a garbage in garbage out process. Estimation techniques can work perfectly well in a microsimulation model which do not work in reality. The risk of doing so can never be completely excluded. To catch some of these uncertainties a scenario analysis was applied.

# Bibliography


**Abowd, J. M.** and **Lane, J. I.** (**2004**): *New Approaches to Confidentiality Protection Synthetic Data, Remote Access and Research Data Centers*. Technical report, U.S. Census Bureau, LEHD Program.
URL http://lehd.did.census.gov/led/library/techpapers/tp-2004-03.pdf

**Alfons, A.**, **Bruch, C.**, **Filzmoser, P.**, **Graf, M.**, **Graf, E.**, **Kolb, J.-P.**, **Lehtonen, R.**, **Hulliger, B.**, **Lussmann, D.**, **Meraner, A.**, **Münnich, R.**, **Nedyalkova, D.**, **Schoch, T.**, **Templ, M.**, **Valaste, M.**, **Veijanen, A.** and **Zins, S.** (**2011**a): *Report on the Analysis of the Simulation Results*. Deliverable 7.1, AMELI project.
URL http://ameli.surveystatistics.net

**Alfons, A.** and **Kraft, S.** (**2010**): `simPopulation`: Simulation of synthetic populations for surveys based on sample data. R package version 0.2.1.
URL http://CRAN.R-project.org/package=simPopulation

**Alfons, A.**, **Kraft, S.**, **Templ, M.** and **Filzmoser, P.** (**2011**b): *Simulation of close-to-reality population data for household surveys with application to EU-SILC*. Statistical Methods & Applications, accepted for publication.

**Alfons, A.**, **Templ, M.** and **Filzmoser, P.** (**2010**): *Simulation of EU-SILC population data: Using the R package* ***simPopulation***. Research Report CS-2010-5, Department of Statistics and Probability Theory, Vienna University of Technology.
URL http://www.statistik.tuwien.ac.at/forschung/CS/CS-2010-5complete.pdf

**Alfons, A.**, **Templ, M.**, **Filzmoser, P.**, **Kraft, S.**, **Hulliger, B.**, **Kolb, J.-P.** and **Münnich, R.** (**2011**c): *The AMELI simulation study*. Deliverable 6.1, AMELI project.
URL http://ameli.surveystatistics.net

**Ballas, D.** and **Clarke, G.** (**2000**): *GIS and microsimulation for local labour market analysis*. Computers, Environment and Urban Systems, 24, pp. 305–330.

**Ballas, D.**, **Clarke, G.**, **Dorling, D.**, **Eyre, H.**, **Thomas, B.** and **Rossiter, D.** (**2005**): *SimBritain: A Spatial Microsimulation Approach to Population Dynamics*. Population, Space and Place, 11 (1), pp. 13–34.

**Ballas, D.**, **Clarke, G.** and **Turton, I.** (**1999**): *Exploring Microsimulation Methodologies for the Estimation of Household Attributes*. Conference on GeoComputation.
URL http://www.geog.leeds.ac.uk/papers/99-11/99-11.pdf

**Beckman, R. J.**, **Baggerly, K. A.** and **McKay, M. D.** (**1996**): *Creating synthetic baseline populations*. Transportation Research Part A: Policy and Practice, 30 (6), pp. 415–429.

**Birkin, M.** and **Clarke, M.** (**1988**): *SYNTHESIS – a synthetic spatial information system for urban and regional analysis: methods and examples*. Environment and Planning A, 20 (12), pp. 1645–1671.

**Chin, S.-F.** and **Harding, A.** (**2006**): *Regional Dimensions: Creating Synthetic Small-area Microdata and Spatial Microsimulation Models*. Technical report, National Centre for Social and Economic Modelling University of Canberra.

**Chin, S.-F.**, **Harding, A.**, **Lloyd, R.**, **Namara, J. M.**, **Phillips, B.** and **Vu, Q. N.** (**2005**): *Spatial microsimulation using synthetic small-area estimates of income, tax and social security benefits*. Australasian Journal of Regional Studies, Vol. 11, No. 3.

**Drechsler, J.**, **Bender, S.** and **Rässler, S.** (**2008**a): *Comparing fully and partially synthetic datasets for statistical disclosure control in the German IAB Establishment Panel*. Transactions on Data Privacy, 1 (3), pp. 105–130.

**Drechsler, J.**, **Dundler, A.**, **Bender, S.**, **Rässler, S.** and **Zwick, T.** (**2008**b): *A new approach for disclosure control in the IAB establishment panel - multiple imputation for a better data access*. AStA Advances in Statistical Analysis, 92 (4), pp. 439–458.

**Eurostat** (**2004**): *Common cross-sectional EU indicators based on EU-SILC; the gender pay gap*. EU-SILC 131-rev/04, Unit D-2: Living conditions and social protection, Directorate D: Single Market, Employment and Social statistics, Eurostat, Luxembourg.

**Eurostat** (**2004**): *Description of target variables: Cross-sectional and longitudinal*. EU-SILC 065/04, Eurostat, Luxembourg.

**Eurostat** (**2009**): *Algorithms to compute social inclusion indicators based on EU-SILC and adopted under the Open Method of Coordination (OMC)*. Doc. LC-ILC/39/09/EN-rev.1, Unit F-3: Living conditions and social protection, Directorate F: Social and information society statistics, Eurostat, Luxembourg.

**Fienberg, S.** (**2003**): *Allowing Access to Confidential Data: Some Recent Experiences and Statistical Approaches*. Presented at Statistics Sweden.
URL lib.stat.cmu.edu/~fienberg/Sweden/swedenworkshop-8-03final.ppt

**Hanaokaa, K.** and **Clarke, G. P.** (**2007**): *Spatial microsimulation modelling for retail market analysis at the small-area level*. Computers, Environment and Urban Systems, 31, pp. 162–187.

**Harding, A.**, **Lloyd, R.**, **Bill, A.** and **King, A.** (**2004**): *Assessing Poverty and Inequality at a Detailed Regional Level: New Advances in Spatial Microsimulation*. Working Papers UNU-WIDER Research Paper, World Institute for Development Economic Research (UNU-WIDER).

**Hauser, R.** (**2007**): *Probleme des deutschen Beitrags zu EU-SILC aus der Sicht der Wissenschaft – Ein Vergleich von EU-SILC, Mikrozensus und SOEP*. SOEPpapers on Multidisciplinary Panel Data Research.

**Holm, E.**, **Lindgren, U.**, **Lundevaller, E.** and **Strömgren, M.** (**2006**): *The SVERIGE spatial microsimulation model*. Paper presented at the 8th Nordic Seminar on Microsimulation Models, Oslo.

**Horvitz, D.** and **Thompson, D.** (**1952**): *A generalization of sampling without replacement from a finite universe*. Journal of the American Statistical Association, 47 (260), pp. 663–685.

**Huang, Z.** and **Williamson, P.** (**2001**): *A comparison of synthetic reconstruction and combinatorial optimisation approaches to the creation of small-area microdata*. Working Paper 2001/2, Department of Geography, University of Liverpool.

**Kendall, M.** and **Stuart, A.** (**1967**): The Advanced Theory of Statistics, vol. 2. London: Charles Griffin & Co. Ltd., 2nd ed.

**Kleiber, C.** and **Kotz, S.** (**2003**): Statistical Size Distributions in Economics and Actuarial Sciences. Hoboken: John Wiley & Sons, ISBN 0-471-15064-9.

**Kohnen, C. N.** and **Reiter, J. P.** (**2009**): *Multiple imputation for combining confidential data owned by two agencies*. Journal of the Royal Statistical Society Series A, 172 (2), pp. 511–528.

**Kraft, S.** (**2009**): Simulation of a population for the European Income and Living Conditions survey. Diploma thesis, Department of Statistics and Probability Theory, Vienna University of Technology, Vienna, Austria.

**Kumar, S.** and **Kockelman, K. M.** (**2007**): *Microsimulation of Household and Firm Behaviors: Coupled Models of Land Use and Travel Demand in Austin, Texas*. Technical report, Center for Transportation Research University of Texas at Austin.
URL http://swutc.tamu.edu/publications/technicalreports/167262-1.pdf

**Lehtonen, R.**, **Valaste, M.** and **Veijanen, A.** (**2008**): *Reconstructing artificial population from a sample*. Unpublished Paper.

**Linzer, D.** and **Lewis, J.** (**2007**): *poLCA: Polytomous Variable Latent Class Analysis*. R package version 1.1.
URL http://userwww.service.emory.edu/~dlinzer/poLCA

**Münnich, R.** and **Schürle, J.** (**2003**): *On the simulation of complex universes in the case of applying the German Microcensus*. DACSEIS research paper series No. 4, University of Tübingen.
URL http://w210.ub.uni-tuebingen.de/volltexte/2003/979/

**Münnich, R.**, **Schürle, J.**, **Bihler, W.**, **Boonstra, H.-J.**, **Knotterus, P.**, **Nieuwenbroek, N.**, **Haslinger, A.**, **Laaksonen, S.**, **Eckmair, D.**, **Quatember, A.**, **Wagner, H.**, **Renfer, J.-P.**, **Oetliker, U.** and **Wiegert, R.** (**2003**): *Monte Carlo simulation study of European surveys*. DACSEIS Deliverables D3.1 and D3.2, University of Tübingen.
URL http://www.dacseis.de

**Norman, P.** (**1999**): *Putting iterative proportional fitting on the researcher's desk*. Working Paper 99/03, School of Geography, University of Leeds.

**Raghunathan, T. E.**, **Reiter, J. P.** and **Rubin., D. B.** (**2003**): *Multiple imputation for statistical disclosure limitation.* Journal of Official Statistics, pp. 19:1–16.

**R Development Core Team** (**2011**): R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.
URL http://www.R-project.org

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

**Rubin, D.** (**1976**): *Inference and missing data.* Biometrika, 63 (3), pp. 581–592.

**Rubin, D.** (**1993**): *Discussion: Statistical disclosure limitation.* Journal of Official Statistics, 9 (2), pp. 461–468.

**Templ, M.** and **Alfons, A.** (**2010**): *Disclosure risk of synthetic population data with application in the case of EU-SILC.* **Domingo-Ferrer, J.** and **Magkos, E.** (editors) Privacy in Statistical Databases, *Lecture Notes in Computer Science*, vol. 6344, pp. 174–186, Heidelberg: Springer, ISBN 978-3-642-15837-7.

**Voas, D.** and **Williamson, P.** (**2000**): *An Evaluation of the Combinatorial Optimisation Approach to the Creation of Synthetic Microdata.* International Journal of Population Geography, 6, pp. 349–366.