DACSEIS

IST-2000-26057

Workpackage 3

Monte-Carlo Simulation Study of European Surveys

Deliverables 3.1 and 3.2

2003

List of contributors:

Wolf Bihler, Statistisches Bundesamt
Harm-Jan Boonstra, Paul Knottnerus, Nico Nieuwenbroek, Statistics Netherlands
Alois Haslinger, Statistics Austria
Seppo Laaksonen, Statistics Finland
Doris Eckmair, Andreas Quatember, Helga Wagner, JKU/IFAS Linz
Jean-Pierre Renfer, Ueli Oetliker, OFS-Office fédéral de la Statistique Suisse
Ralf Münnich, Josef Schürle, Rolf Wiegert, University of Tübingen

Main responsibility:

Ralf Münnich and Josef Schürle, Tübingen

IST–2000–26057–DACSEIS

The DACSEIS research project is financially supported within the IST programme of the European Commission. Research activities take place in close collaboration with Eurostat.

`http://europa.eu.int/comm/eurostat/research/`

`http://www.cordis.lu/ist/`

`http://www.dacseis.de/`

Preface

The DACSEIS recommended practice study mainly depends on two items:

- a thorough study of estimators and variance estimators in the given field and
- well-constructed universes as the basis of the Monte-Carlo simulation study to test the estimators and variance estimators in practical complex environments.

The universes to be considered consist of household and individual data and in two cases also include labour force and related surveys as well as household budget and consumption surveys. Details of the given surveys from Austria, Finland, Germany, Switzerland, and The Netherlands can be drawn from the report on workpackage 2 of the DASCEIS project (Quatember et al. (2002): DACSEIS WP2 report).

The development of this workpackage was highly influenced by the teams that provided the datasets and that helped to find an adequate solution when problems with the practical implementation occurred. Thanks go to Paul Knottnerus and Harm-Jan Boonstra (Dutch Labour Force Survey), Seppo Laaksonen (Finnish Labour Force Survey), Jean-Pierre Renfer and Ueli Oetliker (Swiss Household and Budget Survey), Alois Haslinger (Austrian Microcensus), Wolf Bihler (German Microcensus and Sample Survey of Income and Expenditure), to the external and internal evaluators, Clæs Andersson and Siegfried Gabler, as well as to the whole DACSEIS team for many discussions and valuable comments and suggestions.

Helga Wagner and Doris Eckmair were responsible for the implementation of the Austrian Microcensus. The implementation and description of all other surveys was mainly done by Josef Schürle. Ralf Münnich was responsible for integrating the Monte-Carlo study within the project aims as well as coordinating the general procedure and communication between the various activities.

Ralf Münnich and Josef Schürle — Tübingen, April 2003

Contents

List of figures **VII**

List of tables **IX**

1 Introduction **1**

2 The Model - simulation setup **3**

2.1 Description of the Model 3

2.2 Generation of Random Variables for the Pseudo Universe 5

3 Labour Force Surveys (LFS) **9**

3.1 The Dutch LFS Pseudo Universe 9

3.1.1 Description of the Data and Application of the General Model . . . 9

3.1.2 Selected Results of the Simulation 11

3.1.3 Description of the Implemented Sampling Procedure 15

3.2 The Finnish LFS Pseudo Universe 17

3.2.1 Description of the Data and Application of the General Model . . . 17

3.2.2 Selected Results of the Simulation 17

3.2.3 Description of the Implemented Sampling Procedure 21

4 Microcensus Surveys (MC) **23**

4.1 The German Microcensus Pseudo Universe (GMC) 23

4.1.1 Description of the Data and Application of the General Model . . . 23

4.1.2 Selected Results of the Simulation 26

4.1.3 Description of the Implemented Sampling Procedure 32

4.2 The Austrian Microcensus Pseudo Universe (AMC) 33

4.2.1 Description of the Data and Application of the General Model . . . 33
4.2.2 Selected Results of the Simulation 37
4.2.3 Description of the Implemented Sampling Procedure 40

5 Household and Budget Surveys (HBS) 45

5.1 The German Sample Survey of Income and Expenditure Pseudo Universe (EVS) 45
5.1.1 Description of the Data and Application of the General Model . . . 45
5.1.2 Selected Results of the Simulation 47
5.1.3 Description of the Implemented Sampling Procedure 49
5.2 The Swiss HBS Pseudo Universe 51
5.2.1 Description of the Data and Application of the General Model . . . 51
5.2.2 Selected Results of the Simulation 54
5.2.3 Description of the Implemented Sampling Procedure 57

6 Summary 61

References 63

List of Figures

1.1 Outline of the simulation study 2

2.1 Main simulation principle. 3

3.1 Realized marginal frequency distributions within the Dutch pseudo universe and expected frequency distributions on the basis of the Dutch survey data distributions. 13

3.2 Marginal expected and realized frequency distributions within Zuid - Holland and Flevoland. 15

3.3 Marginal frequency distributions within the Finnish February 2000 pseudo universe 19

4.1 Marginal frequency distributions within the German pseudo universe. . . . 27

4.2 Marginal frequency distributions within the German federal states BAW and MVP 28

4.3 Realized marginal age frequency distributions within the German house size classes 1 to 5. 31

4.4 Differences between the expected and the realized marginal age frequency distributions within the German house size classes 1 to 5. 32

4.5 Marginal gender and employment frequency distributions within the German house size classes 4 and 5. 33

4.6 Marginal frequency distributions within the AMC pseudo universe and the AMC data 38

4.7 Marginal frequency distributions within federal states TIR and WIE of the AMC pseudo universe 40

4.8 Marginal frequency distributions within Parts A and B of the AMC pseudo universe 42

5.1 Marginal frequency distributions within the quarter 1 and quarter 2 EVS pseudo universes. 48

5.2 Marginal frequency distributions within selected Swiss *a lot of classes* pseudo universes. 55

5.3 Marginal frequency distributions within the Swiss *a lot of classes* pseudo universe. 56

List of Tables

3.1 Variables included in the Dutch pseudo universe. 10

3.2 Number of households (nohh) and persons within the regions and the four big cities of the Dutch pseudo universe and the total number of persons in reality according to information of the CBS (reality). 12

3.3 Realized contingency coefficients within the Dutch pseudo universe and expected contingency coefficients on the basis of the Dutch survey data distributions. 14

3.4 Contingency coefficients within Zuid-Holland and Flevoland. 16

3.5 Variables included in the Finnish pseudo universes. 18

3.6 p-values when applying a χ^2 goodness of fit test to the Finnish LFS realized and expected marginal frequency distributions. 19

3.7 Contingency coefficients within the Finnish pseudo universes. 20

3.8 p-values when applying a χ^2 goodness of fit test to the expected and realized Finnish LFS multivariate distributions. 21

4.1 Variables included in the German pseudo universe. 24

4.2 Partition of the federal states into regional classes and number of classes (noc) within each federal state. 24

4.3 Number of households and persons within the federal states of the German pseudo universe. 25

4.4 p-values when applying a χ^2 goodness of fit test to the expected and realized German MC marginal distributions. 26

4.5 Contingency coefficients within the German pseudo universe. 29

4.6 p-values when applying a χ^2 goodness of fit test to the expected and realized marginal frequency distributions within BAW and MVP. 29

4.7 Contingency coefficients within the German federal states BAW and MVP. 30

4.8 Number of households and persons within the five house size classes of the German pseudo universe. 31

4.9 Contingency coefficients within the German house size classes 4 and 5. . . 34
4.10 Variables included in the AMC pseudo universe 35
4.11 Partition of federal states of the AMC pseudo universe into Parts and strata 35
4.12 Number of households and persons within federal states and Parts of the AMC pseudo universe 37
4.13 Contingency coefficients within the AMC data and the Austrian pseudo universe 39
4.14 Contingency coefficients within the Austrian federal states TIR and WIE . 41
4.15 Contingency coefficients within Parts A and B of the AMC pseudo universe 43

5.1 Variables included in the German EVS pseudo universe. 46
5.2 Number of households included in the German EVS pseudo universes. . . . 47
5.3 p-values when applying a χ^2 goodness of fit test to the expected and realized marginal frequency distributions within quarter 1 to quarter 4 EVS pseudo universes. 47
5.4 Contingency coefficients within the quarter 1 to quarter 4 EVS pseudo universes. 49
5.5 Pearson correlation coefficient values for income and expenditure within the quarter 1 to quarter 4 EVS survey data and pseudo universes. 49
5.6 Statistics for income and expenditure within the quarter 1 to quarter 4 survey data and the pseudo universes. 50
5.7 Target number of households to be surveyed per federal state in the implemented German EVS drawing procedure. 51
5.8 Variables included in the Swiss pseudo universe. 52
5.9 Number of households included in the Swiss pseudo universes. 52
5.10 Classification of expenditure and income. 53
5.11 p-values when applying a χ^2 goodness of fit test to the expected and realized marginal distributions within the Swiss *a lot of classes* pseudo universe. . . 55
5.12 Contingency coefficients within the Swiss *a lot of classes* pseudo universe. . 56
5.13 p-values when applying a χ^2 goodness of fit test to the expected and realized marginal distributions within the Swiss *a lot of classes* pseudo universe regions Plateau central, Zurich and Tessin. 57
5.14 Contingency Coefficients within the Swiss *a lot of classes* pseudo universe regions Plateau central, Zurich and Tessin. 58
5.15 Pearson correlation coefficient values for income and expenditure within the Swiss survey data and the Swiss pseudo universes. 59

5.16 Statistics for income and expenditure within the pseudo universes and the survey data. 59

5.17 Number of households surveyed per stratum in the implemented Swiss survey procedure. 59

Chapter 1

Introduction

The aim of workpackage 3 is to yield the basis for the Monte-Carlo simulation study which is placed at workpackage 1 interacting with the methodology from the workpackages 5 to 11. The basis for the simulation study consists of the generation of adequate universes and suitable interfaces for the national sampling schemes, the estimators and variance estimators, as well as the possible inclusion of different peculiarities of the national surveys. The universes of interest are the Dutch and Finish labour force survey (LFS), the Austrian and German Microcensus, as well as the Swiss household and budget survey. Since for these surveys no full universes, e.g. census data, are available, the universes had to be created from samples.

The simulation study itself will follow a general flow scheme which can be drawn from figure 1.1. Within workpackage 3, adequate *true* universes had to be generated. Since all universes do not consist of census data, they had to be generated from samples as best as possible. However, one major conflict had to be solved:

1. The best possible mechanism to create the universes should allow to rebuild marginal distributions as well as interaction between variables;
2. This mechanism should also allow for heterogeneities between subgroups, especially for regional aspects;
3. Pure replication of units, which seems to best suit the first two items, should be avoided because this generally leads to an extremely small variability of units within smaller subgroups;
4. The use of many variables on the microlevel must not end up in nondisclosure difficulties.

This seems to be an unsolvable conflict. However, one has to bear in mind the target of the simulation study. The aim is to find *best* recommendations for the practical use of variance estimators in a most appropriate practical environment. Therefore, several assumptions have to be made which will be presented in connection with the corresponding surveys. However, these assumptions will not negatively influence the generation process. Once suspicious results occur from the simulation study, one has to perform an additional study which may result in a sensitivity analysis.

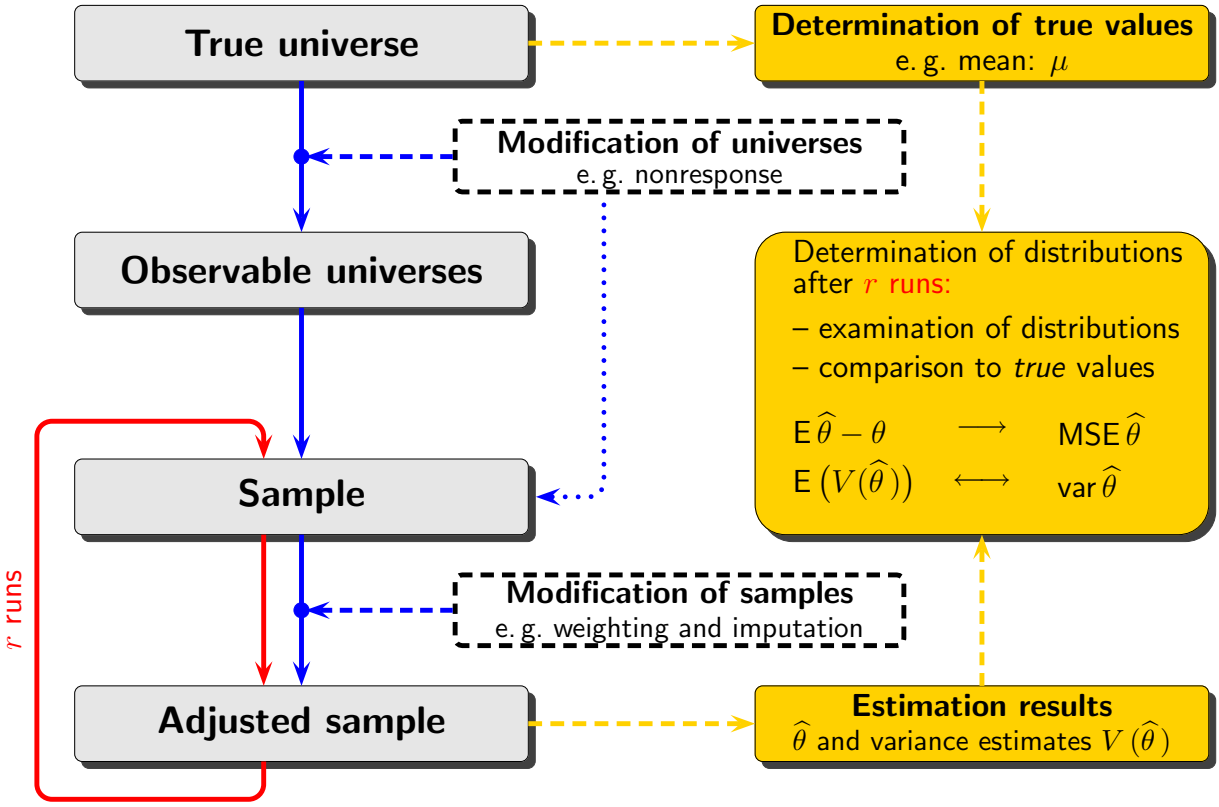

Figure 1.1: Outline of the simulation study

The possible difference between *true* and *observable* universe will be part of workpackage 1 where different processes of interest have to be presented systematically. The interface for integrating different type of processes is included in the programs of this workpackage. An example for this type of process may be a nonresponse variable, if simulated within the universe, or changes of variables between two or more points of time to investigate methods for variance estimation for change. The latter will be applied to the Finish Labour Force Survey. The presentation and implementation of nonresponse models will also take place within workpackage 1.

A further aspect of workpackage 3 is to implement the national sampling schemes that are used for the surveys. Since these sampling schemes are too specialized to allow for sampling schemes that are considered under the workpackages 5 to 11, further sampling schemes will be considered in the simulation study in workpackage 1.

The next chapter describes the general model of creating the universes for household and individual surveys. A very important topic is to take into consideration the national figures like size, regions, and population characteristics. The survey specific implementation will be shown in the subsequent chapters. These will include the variable lists and some important figures of the data.

Chapter 2

The Model - simulation setup

2.1 Description of the Model

Within this section, the main principle for the DACSEIS pseudo universe simulations is described. The simulation processes of all country universes work according to this main principle which is outlined in Figure 2.1. The concrete realization depends on the respective survey process and on the data available.

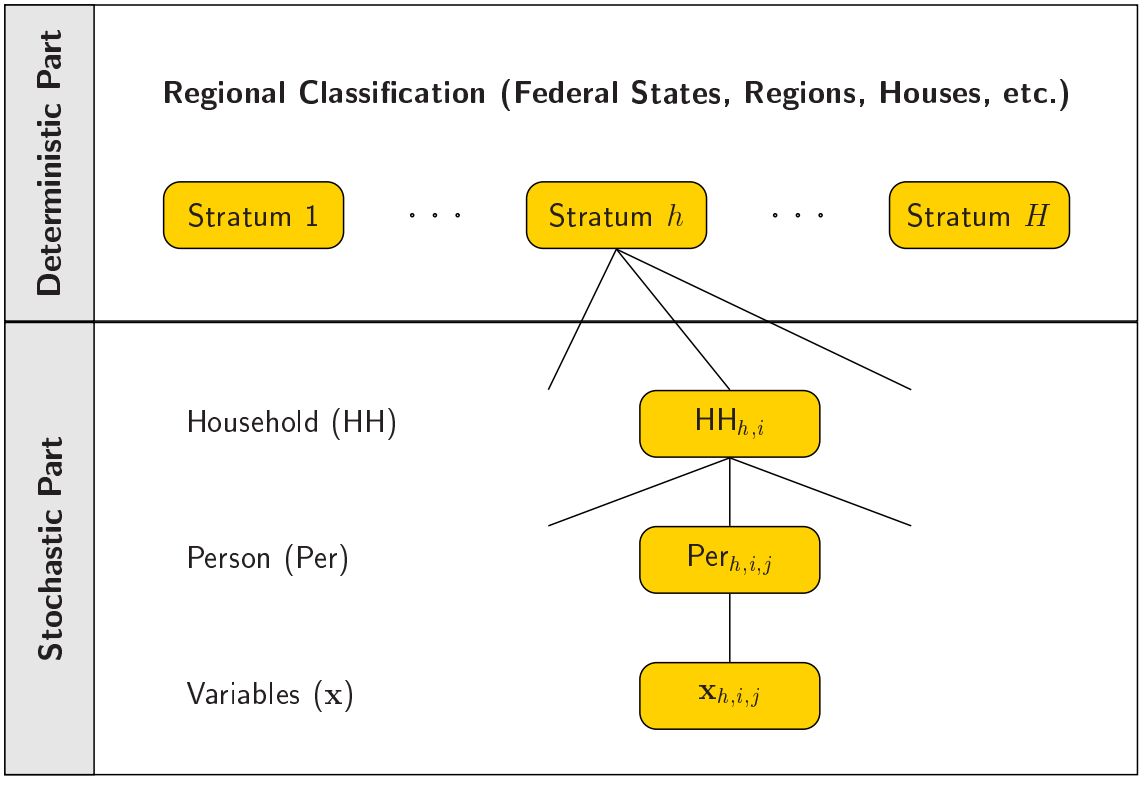

Figure 2.1: Main simulation principle.

Before a survey is performed, the respective universe generally is subdivided into several strata. Usually this subdivision is done on the basis of geographical aspects, e.g. on

the basis of the place of residence. Often the strata are built according to already existing regions like for example federal states or municipalities. The subdivision is done independently of the concrete realization before surveying and on the basis of well know attributes. Therefore it is a deterministic process. For each survey, the real subdivision is assigned to the pseudo universe by defining strata according to it. This stratification leads to separate units which are simulated independently of each other. To do this, individual distributions are taken for each stratum to simulate household and individual data at random. This proceeding should account for the homogeneity or heterogeneity

- within a stratum reflected by the respective stratum-distributions
- between the strata reflected by using different distributions.

The distributions needed for the simulation depend on the procedure how households and individuals are simulated. This proceeding is described subsequently.

A main problem of the simulation of universes is the creation of households, i.e. to generate correlation structures between persons within a household. If persons within a household are simulated without any social links, then strange results - for example five babies in a five-person household - may appear. As a result of this, households within the pseudo universes would be only a technical pooling of units with no further meaning. To avoid such inconsistencies, correlation structures have to be considered. This would be complex if correlations between all variables would be taken into account. For example if k variables are considered, then the realizations for the k variables of a given person in a 5-person household have to be conditioned by the realization of the 4 other persons. The exact realization of the correlation structures has main disadvantages. First of all, it is difficult to get data which satisfies the needs of the approach. And second data protection problems appear. Therefore a simplification is needed. It is an empirical fact that the age and gender of a person has a great influence on other variables of interest like for example employment or income. As a consequence of this, within a household only the correlations between age and gender are considered. The other correlations are only implicitly generated by the influence of age and gender. There are effects which are not considered but the method seems to be a reasonable compromise between efforts to realize the simulation and the quality of the results.

In detail the simulations will be performed as follows. First of all, strata are defined according to the survey process. Subsequently the number of households in each stratum is determined. This is either done by a simulation on the basis of a given distribution or by pre-determination. The number of persons in each household is simulated at random on the basis of a given distribution. Next the variables of interest are simulated. As described above, correlations are only explicitly considered with respect to the variables age and gender. Therefore a realistic age and gender structure is taken for each household. For example, if a household with k persons is considered, then the age and gender of the k persons are taken by drawing a realistic k-person household from a data set and assigning each pseudo unit the age and gender of a realistic unit. By doing so, unrealistic combinations are precluded. At last, the remaining variables of interest are simulated. Therefore, a given multivariate distribution for the variables of interest is taken and for each unit within a stratum the conditional distribution - conditional to the age and gender

of the unit - is calculated. After that, the remaining variables of interest are simulated by using the conditional distributions.

In this subsection the distributions used for simulating are said to be given. In fact, they are unknown and therefore they have to be estimated. This is done on the basis of the data available for the respective universes. Descriptions of how the distributions are obtained in the particular cases are given in Section 3.1 to 5.

2.2 Generation of Random Variables for the Pseudo Universe

Within this section, the methods used for generating random variables are briefly described. Those methods are widely known standard techniques. Therefore only a rough overview is given here. A more detailed and precise description can be found for example in Devroye (1986), Johnson (1987), Press *et al.* (1992) and Kronmal and Peterson Jr. (1979) as well as in the references therein.

For generating univariate discrete distributions, two methods are in use. The first is the so called inversion method with sequential search (cf. Devroye, 1986, p. 85f). The random variable to be generated is named X and is non-negative integer valued with distribution function $F(x) := \sum_{i<x} P(X = i)$. First a random variable U is generated which is uniformly distributed on $[0, 1]$. Sequential search means that for a given realization u of U - which is created using a standard random number generator (cf. Press *et al*, 1992) - it is sequentially tested, which of the values $x = 0, 1, 2, \ldots$ solve the equation

$$F(x-1) \leq u < F(x). \tag{2.1}$$

Then

$$P(X = i) = \begin{cases} F(i) - F(i-1) & \text{if } i \in \mathbb{N}_0 \\ 0 & \text{else.} \end{cases}$$

The sequential search method is very slow in general but has the advantage that no setup is needed. Therefore the method is used when only a very small number of random variables from a given distribution is needed.

The second method used is the so called alias method (cf. Kronmal and Peterson Jr., 1979). The integer valued random variable X with probability function $f(x)$ has a finite number of outcomes. Define

$$\mathcal{S} := \{x \in \mathbb{N}_0 \mid f(x) > 0\} \qquad \text{and} \qquad m := \mid \mathcal{S} \mid .$$

Then m two-point distributions are calculated in a specific way from $f(x)$[1]. For each of the m distributions, the two outcomes as well as the probabilities for the two outcomes are determined. This is called the setup phase. After the setup is complete, random variables

[1] A detailed description of how the two-point distributions are calculated is given in Kronmal and Peterson Jr. (1979).

could be simulated. Therefore, a random variable U which is uniformly distributed on $[1, m+1)$ is created. The number

$$\lfloor u \rfloor := \max\{x \in \mathbb{N}_0 \mid x \leq u\}$$

is taken to select one out of the m two-point distributions with equal probability. The first outcome of the selected two-point distribution $\lfloor u \rfloor$ is j and k is the second. The probabilities for those outcomes are $p_{\lfloor u \rfloor}(j)$ and $p_{\lfloor u \rfloor}(k) = 1 - p_{\lfloor u \rfloor}(j)$. Then

$$x = \begin{cases} j & \text{if } u - \lfloor u \rfloor < p_{\lfloor u \rfloor}(j) \\ k & \text{else.} \end{cases}$$

In KRONMAL and PETERSON JR. (1979) it is shown that when applying the alias method, X has the desired distribution with probability function $f(x)$. The disadvantage of the method is the setup phase needed. But on the other hand, when the setup phase is finished the alias method is very fast - independently of n. This is, because only one $[0; 1)$ random variable has to be created. Hence, the alias method is very fast, if a big quantity of random variables from the same distribution is needed.

As already mentioned, the inversion method is relatively fast when only a small quantity of random numbers is needed. But on the other hand the average time needed for generating one random number is constant while the total number of random variables generated increases. In opposition to this, the alias method needs a setup phase and therefore is relatively slow when only a small quantity of random numbers is needed. On the other hand it gains efficiency with the number of random numbers generated from the same distribution. To account for this the inversion method is used within the simulation when only one random numbers is generated from a given distribution. This is the case when specific conditional distributions for each unit are calculated. In all other cases, the alias method is applied.

Within the simulations not only random variables but also random vectors have to be created. Because the random vectors needed are mainly discrete, this is done by transforming the multivariate into a univariate distribution. The desired random vector $X = (X_1, \ldots, X_n)$ is composed out of the non-negative integer valued random variables X_j which are bounded above. Define

$$\max X_j := \max(x \in \mathbb{N}_0 \mid P(X_j = x) > 0).$$

The random variable Y is defined as

$$Y := X_1 + \sum_{j=2}^{n} \left(X_j \cdot \prod_{i=1}^{j-1} (\max X_j + 1) \right).$$

what is generally known as coding function (cf. DEVROYE, 1986, p. 559). Hence

$$P\left(X = (x_1, \ldots, x_n)\right) = P\left(Y = x_1 + \sum_{j=2}^{n} (x_j \cdot \prod_{i=1}^{j-1} (\max X_j + 1)) \right).$$

The random variable Y is non-negative integer valued and bounded above. Therefore the inversion and the alias method could be applied to create random numbers which are

P_Y distributed. After a univariate random number is created it has to be re-transformed again. This is done by using the relationship

$$x_j = \begin{cases} \left\lfloor \dfrac{y}{\prod_{i=1}^{n-1}(\max X_i + 1)} \right\rfloor & \text{if } j = n \\ \left\lfloor \dfrac{y - \sum_{i=j+1}^{n}(x_i \cdot \prod_{l=1}^{i-1}(\max X_l + 1))}{\prod_{i=1}^{j-1}(\max X_i + 1)} \right\rfloor & \text{if j=1,2,...,n-1,} \end{cases}$$

which is referred to as decoding function (cf. Devroye, 1986, p. 559). As a result, by creating random numbers with the distribution P_Y and transforming them by using a decoding function, the resulting random vectors have the desired distribution P_X.

For the simulation of the Swiss pseudo universe, multivariate normal distributed random variables are needed. They are generated in the following way. The mean vector $\mu = (m_1, \ldots, m_d)^T$ and the $d \times d$ covariance matrix Σ are given. A $d \times d$ lower triangular matrix C is calculated for which

$$CC^T = \Sigma$$

is true. This is done by applying the Cholesky decomposition. For the generation of one multivariate normal distributed random vector X, d independent standard normal distributed random numbers Z_i, $i = 1, \ldots, d$, are generated. Therefore the ratio of uniforms method is used (cf. Devroye, 1986, p. 194 et seqq.). Next the vector $Z = (Z_1, \ldots, Z_d)^T$ is transformed by

$$X = CZ + \mu.$$

The resulting random vector X has the desired multinormal distribution with parameters μ and Σ (cf. Johnson, 1987, p. 50).

Chapter 3

Labour Force Surveys (LFS)

3.1 The Dutch LFS Pseudo Universe

3.1.1 Description of the Data and Application of the General Model

For the simulation of the Dutch pseudo universe, several distributions gained from Dutch survey data were available. Those distributions were determined by the Centraal Bureau voor de Statistiek (CBS). Included are the number of persons per household distributions for 12 different regions and the multivariate distributions for the variables displayed in Table 3.1 distinguished between 12 different regions and 4 big cities. Also, the number of persons and households within each region is given. The data covers a wide range of what is needed, but not everything. For example, the regional number of persons per household distributions include only the classes 1, 2, 3, 4 and 5^+ where 5^+ means 5 and more persons. The composition of the 5^+ classes is not known. Also, within the data there is no information about the persons within the households. Therefore the German Microcensus data is used to fill existing gaps. The procedure is described subsequently.

The strata are built along 12 geographical regions which are distinguished in the survey process. Within those regions an additional partition along the municipalities was made. For simplification, an internal reordering of the municipalities was done. Each municipality was given an internal number within the respective geographical region. The data in use contains the number of households within the municipalities. Therefore this number is given and deterministic. Also included in the data are the distributions of the number of persons within a household, one distribution for each municipality. A main problem is that the distributions only contain information about 1 to 4 and 5 and more (5^+) person households. The distribution within the 5^+ class is not known. To handle this shortcoming, the partition of 5^+ from the German survey data is used as substitution. Therefore the number of households with 5, 6, 7 ... persons in the German survey data was counted. The absolute number was divided by the total number of households with 5 or more persons in it. This leads to the German 5^+ distribution. This distribution is adopted by multiplying its elements with the 5^+ probability of the respective municipalities. On the basis of the respective distributions, the number of persons in the households

Table 3.1: Variables included in the Dutch pseudo universe.

Variables	possible outcomes
age	0 15 - 24 1 25 - 34 2 35 - 44 3 45 - 54 4 55 - 64 5 65+
gender	0 male 1 female
marital status	0 married 1 divorced or widowed 2 unmarried
ethnicity	0 Dutch 1 other European 2 non European
employment	0 employed labour force 1 unemployed labour force 2 non labour force

are simulated. Therefore each household is created independently of the others by using the alias method.

Next the age and gender structures within the households have to be created. Within the Dutch data there is no information about people living in the same household. The distributions of the variables of interest only include individual but no household information. Instead of the missing Dutch household structures German data is used. All households in the data with the same number of persons in it are pooled. Then for each household in the Dutch pseudo universe, a household with the corresponding number of people in it is drawn at random from the German data. The age and gender structure of the real household is adopted for the artificial household. Because the Dutch data does only distinguish between 6 classes of age, the age of the persons resulting from the German data are transformed into those six classes. Additionally a seventh class for people which are younger than 15 is defined. This proceeding has an influence on the resulting distributions within the pseudo universe. If the age and gender structure within the German households differ from the ones in the Dutch data, then the age and gender distributions in the Dutch pseudo universe differ from the distributions within the Dutch survey data. Because the distributions of marital status, ethnicity and employment are generated conditional on age and gender realizations, the marginal distributions of those three variables are also biased. Hence, the solution is not optimal in a global way, but a reasonable approach on the basis of the given data.

Finally the remaining 3 variables are created. This is done similar to the German pseudo universe simulation. The starting point is the multivariate frequency distribution of the variables region, age, gender, marital status, ethnicity and employment. This distribution

contains estimates for the number of person in the real Dutch universe with the respective multivariate realizations of the 6 variables. First of all the distribution is split into 12 distributions by using the region variable. Hence, for each region an own multivariate frequency distribution is calculated. The four big cities Amsterdam, Den Haag, Utrecht and Rotterdam are treated separately by using individual distributions. For each person in the pseudo universe the respective frequency distribution is taken and all combinations are considered which share the age and gender of that person. If the age is under 15, then no simulation is done. This is, because the multivariate frequency distribution only considers persons with 15 years and older. If the person is 15 years or older, the resulting conditional multivariate frequency distribution is divided by the absolute number of cases remaining. The resulting relative frequency distribution is taken as probability distribution. The three variables are simulated by applying these distributions in conjunction with a coding function, the inversion method and a decoding function.

The partly use of the German survey data will lead to a bias away from the Dutch survey data distributions. Therefore, the estimates gained from the Dutch pseudo universe should of course not be interpreted as true Dutch universe values. But even if there is a bias in the data, the global structure of the universe is widely kept. Hence, standard estimates will be interpretable. If strange results appear when seldom events or events within small regions are estimated then checkups are necessary.

3.1.2 Selected Results of the Simulation

The differences between the distributions in the Dutch survey data and in the Dutch pseudo universe are of main interest. They are due to the fact that a mix of the Dutch and the German survey data is used for the simulation. The age and gender structures within the households are generated on the basis of the German data, while the variables marital status, ethnicity and employment are based on the conditional distributions from the Dutch data. The distributions are conditioned to the age and gender of each person. Thus, if the two-dimensional age and gender distributions within the German and the Dutch data are not equivalent, then the 5-dimensional distribution within the Dutch pseudo universe will differ from the one in the Dutch survey data. The dimension of this effect is shown subsequently.

The number of persons in the households is partly created on the basis of the German data, because the composition of the 5^+-class within the Dutch data is not known. In Table 3.2 the realized numbers of persons within the 12 regions and the four big cities are displayed. Also the number of persons which should be realized according to the Dutch data is displayed. Within the table, there is a distinction between people aged 0 to 14 and people older than 14. This results because the multivariate distributions for the variables of interest are only for people aged 15 years and older. On the other hand, the person per household distributions in the Dutch survey data do not distinguish between these two categories of age. The existing gap is filled by the German data. For the variables marital status, ethnicity and employment, values are only simulated if the respective person is 15 years or older. Thus, for investigations concerning one or more of those three variables, only units aged 15 and older could be included. The numbers in Table 3.2 show that by using the German person per household distribution, the number of persons is a bit lower

than it should be. Hence, the distribution of the Dutch 5^+ is not completely met. But the numbers are at least very close to what they should be.

Table 3.2: Number of households (nohh) and persons within the regions and the four big cities of the Dutch pseudo universe and the total number of persons in reality according to information of the CBS (reality).

		number of persons			
	nohh	aged 15^+	aged ≤ 14	total	reality
Regions					
Groningen	255,297	466,450	86,532	552,982	558,017
Friesland	255,262	503,281	110,995	614,276	618,115
Drenthe	186,225	377,958	82,564	460,522	464,672
Overijssel	424,198	854,202	201,700	1,055,902	1,063,527
Gelderland	768,072	1,530,367	344,704	1,875,071	1,895,656
Utrecht	466,488	887,944	187,496	1,075,440	1,088,621
Noord-Holland	1,146,675	2,077,434	385,178	2,462,612	2,486,105
Zuid-Holland	1,478,285	2,768,492	555,146	3,323,638	3,359,047
Zeeland	152,604	301,422	62,830	364,252	369,949
Noord-Brabant	936,621	1,884,593	420,517	2,305,110	2,319,262
Limburg	473,368	932,661	191,980	1,124,641	1,137,935
Flevoland	112,796	234,542	57,689	292,231	293,286
Total	6,655,891	12,819,346	2,687,331	15,506,677	15,654,912
Cities					
Amsterdam	399,696	626,352	81,077	707,429	718,151
DenHaag	226,669	376,336	58,258	434,594	442,799
Rotterdam	295,583	503,666	80,195	583,861	590,478
Utrecht	123,570	199,166	29,141	228,307	232,744

The marginal distributions of age, gender, marital status, ethnicity and employment are displayed in Figure 3.1. Within this figure, the realized frequencies as well as the ones expected on the basis of the Dutch survey data distributions - frequencies are displayed. At first glance, differences are obvious. The use of the German data leads to a lower number of persons from 15 to 54, while the number of units aged 55 and elder is higher. Because age has an effect on the other variables as well, the marginal distributions change. An elder population tends to have a larger women fraction than a younger population. This is, because men die earlier on average. If there are more elder people, then there is also a tendency to a higher number of widowed persons. And of course, if there are more elder people in the universe, then the number of people in the non labour force category is larger. The differences between the expected and realized distributions are approved by a χ^2 goodness of fit test[1] which shows a p-value of 0 for all variables.

[1]The test statistic $\chi^2 := \sum_{j=0}^{m} \frac{(n_j - n_j^*)^2}{n_j^*}$ was used. Thereby, n_j is the realized number of observations with outcome j and n_j^* is the respective number which is expected on the basis of the survey sample distribution.

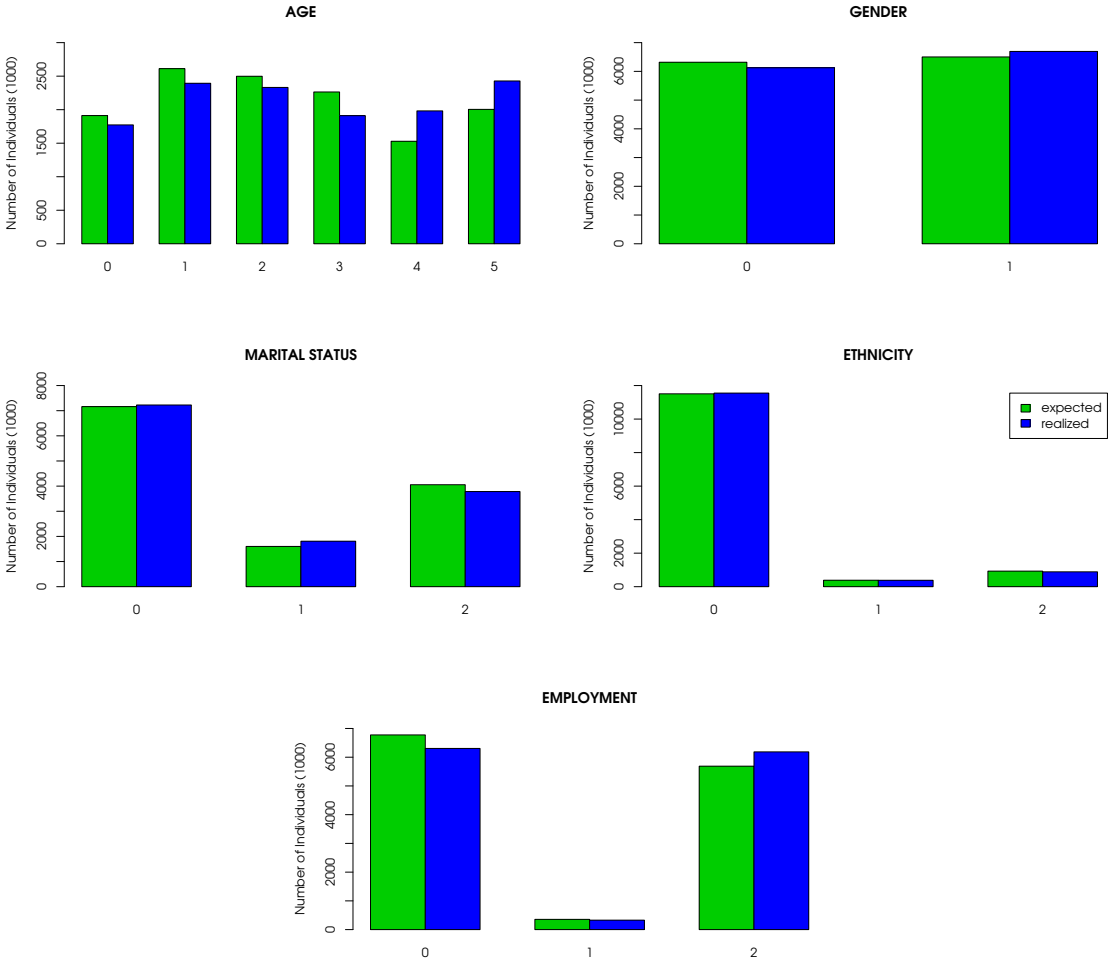

Figure 3.1: Realized marginal frequency distributions within the Dutch pseudo universe and expected frequency distributions on the basis of the Dutch survey data distributions.

The desired two-dimensional correlations do not differ from the realized as much as the marginal distributions. The contingency coefficients[2] are displayed in Table 3.3. Obviously, there are differences in the age-gender correlation. This is because the age and gender structure is completely taken from the German data and therefore the whole difference between the German and the Dutch data appears. On the other hand, the differences in expected and realized contingency coefficients are smaller for other variable combinations. Beside age and gender, the biggest absolute differences appear in the marital status/employment, gender/marital status and age/employment combinations. The other absolute differences are very small and therefore can not be attributed to a bias in the pseudo universe distributions. The results indicate that the desired correlation structures are

[2]The definition $c := \sqrt{\frac{\chi_*^2}{n+\chi_*^2}}$ with $\chi_*^2 := \sum_{j=0}^{m}\sum_{k=0}^{l}\frac{(n_{jk}-n_{jk}^*)^2}{n_{jk}^*}$ was used. Here, n is the total number of observations, n_{jk} is the number of observations with the realization j regarding the first and k regarding the second variable and n_{jk}^* is the respective number of observations, which is expected on the basis of the survey sample distribution.

approximately kept, although German data is partly used for the simulation.

Table 3.3: Realized contingency coefficients within the Dutch pseudo universe and expected contingency coefficients on the basis of the Dutch survey data distributions.

data source		gender	mar. stat.	ethnicity	employment
survey data	age	0.0628	0.5911	0.0783	0.5079
	gender	-	0.1411	0.0148	0.2488
	mar. stat.	-	-	0.0266	0.2043
	ethnicity	-	-	-	0.0891
pseudo universe	age	0.1023	0.5958	0.0729	0.5215
	gender	-	0.1644	0.0152	0.2576
	mar. stat.	-	-	0.0239	0.2332
	ethnicity	-	-	-	0.0827
relative differences	age	62.9%	0.8%	-6.9%	2.7%
	gender	-	16.5%	2.7%	3.6%
	mar. stat.	-	-	-10.2%	14.1%
	ethnicity	-	-	-	-7.2%

Because the age and gender structures of the households within the pseudo universe are all drawn from the same data set, the universe loses heterogeneity. The differences in the marginal distributions between two regions are in this case only caused by different number of persons per household distributions. This effect is shown for Zuid-Holland and Flevoland in Figure 3.2. Flevoland is the region with the smallest number of persons and conversely Zuid-Holland the one with the biggest number. Zuid-Holland includes the cities DenHaag and and Rotterdam which represent about 30% of the population within the region. The graphics show that the realized differences between the regions in the age distributions are smaller than they should be on the basis of the survey data. The altered age and gender structure has also an influence on other variables. In Figure 3.2, additionally the employment distributions are presented which approve this. On the other hand, the marginal distributions of ethnicity are hardly changed in comparison to the expected ones. Hence, not all variables are influenced. In Table 3.4, the contingency coefficients are presented. They show that there are differences between the two-dimensional correlations within the survey data and the pseudo universe. Although the relative differences are sometimes large, the absolute ones are in general very small. Hence, the correlation structure from the survey data is widely kept.

The analyses show, that the proceeding for the creation of the Dutch pseudo universe leads to a bias in the marginal distributions. This is, because the age and gender structure from the German survey data differs from the one implicitly given in the Dutch data. Also, the usage of the German data for the whole Dutch pseudo universe reduces the heterogeneity between the regions. On the other hand, the correlations between the variables are widely kept.

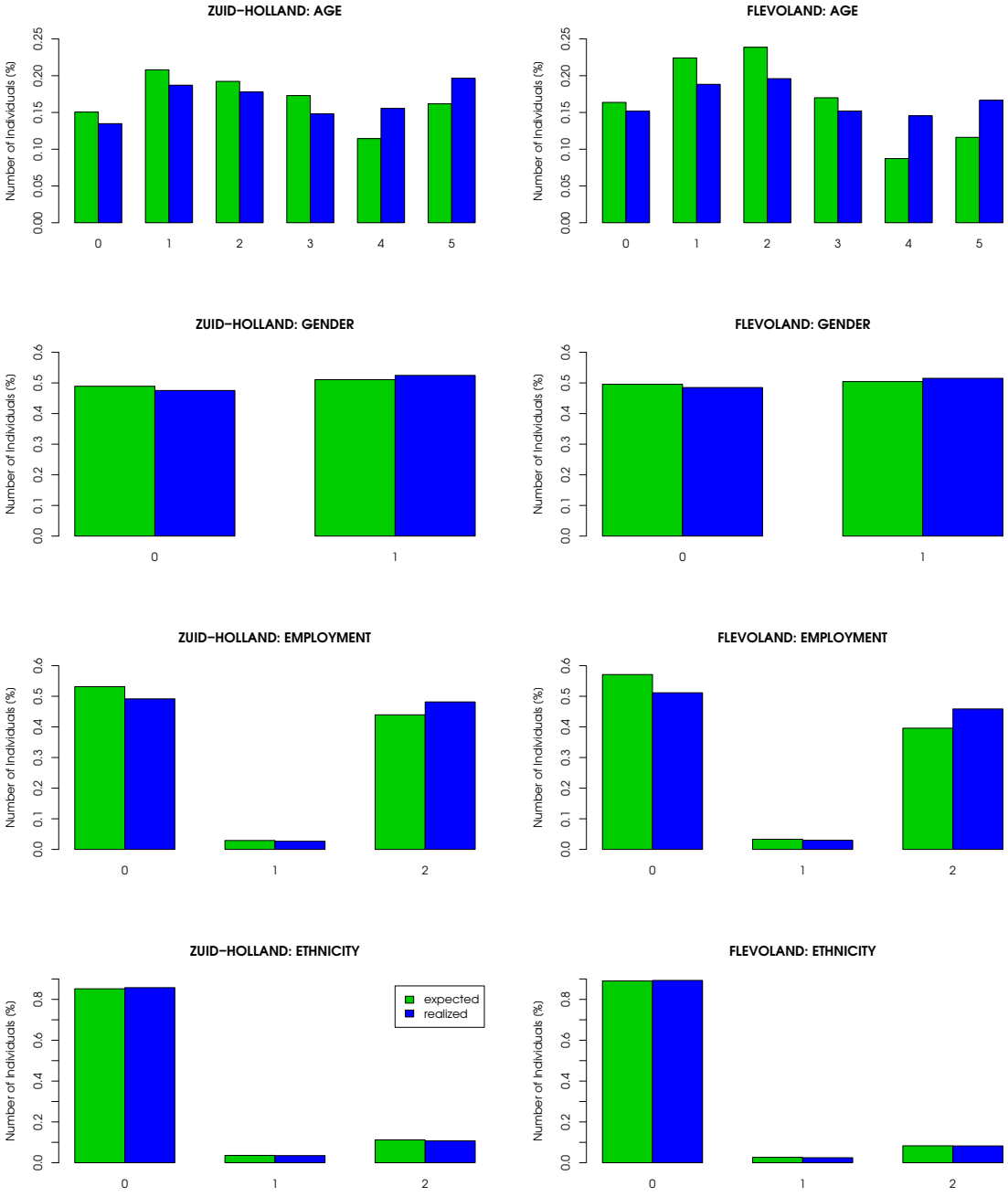

Figure 3.2: Marginal expected and realized frequency distributions within Zuid - Holland and Flevoland.

3.1.3 Description of the Implemented Sampling Procedure

In reality, the Dutch LFS consists of twelve subsamples, one for each month in a year. Because there is only one Dutch pseudo universe - without including time effects - the samples are aggregated and all addresses are sampled in one go. The number of addresses to be sampled is approximately $132,000$. Beside the fact that no monthly subsamples are considered, the simulated drawing procedure is kept close to the real one.

Table 3.4: Contingency coefficients within Zuid-Holland and Flevoland.

data source		gender	mar. stat.	ethnicity	employment
Zuid-Holland survey data	age	0.0700	0.5812	0.0954	0.5130
	gender	-	0.1528	0.0076	0.2425
	mar. stat.	-	-	0.0413	0.2130
	ethnicity	-	-	-	0.1243
Zuid-Holland pseudo universe	age	0.1116	0.5842	0.0952	0.5273
	gender	-	0.1768	0.0063	0.2534
	mar. stat.	-	-	0.0363	0.2416
	ethnicity	-	-	-	0.1171
Zuid-Holland relative differences	age	59.4%	0.5%	-0.2%	2.8%
	gender	-	15.7%	-17.1%	4.5%
	mar. stat.	-	-	-12.1%	13.4%
	ethnicity	-	-	-	-5.8%
Flevoland survey data	age	0.0435	0.6004	0.0876	0.4935
	gender	-	0.1179	0.0515	0.2646
	mar. stat.	-	-	0.0487	0.1711
	ethnicity	-	-	-	0.1137
Flevoland pseudo universe	age	0.0859	0.6149	0.0932	0.5169
	gender	-	0.1578	0.0464	0.2690
	mar. stat.	-	-	0.0487	0.2089
	ethnicity	-	-	-	0.1102
Flevoland relative differences	age	97.8%	2.4%	6.4%	4.7%
	gender	-	33.8%	-9.9%	1.7%
	mar. stat.	-	-	0%	22.1%
	ethnicity	-	-	-	-3.1%

The sampling procedure implemented - analogous the real one - is divided into two stages. Within the first stage, municipalities are selected out of the 12 regions by using a systematic sampling procedure. All municipalities for which the number of addresses is higher than the sampling interval are pre-selected because of an inclusion probability of 1. Those municipalities are referred to as self-representing municipalities. Because no addresses are available within the pseudo universe, households are used instead. In the second stage, households are sampled out of the selected municipalities. Therefore, a random sample without replacement is drawn within each selected municipality. The sampling intervals and sampling fractions are calculated according to the real Dutch LFS procedure.

3.2 The Finnish LFS Pseudo Universe

3.2.1 Description of the Data and Application of the General Model

For the simulation of the Finnish pseudo universe, LFS panel data including 25,021 units was used. Available was data from three different interview months:

- February 2000
- May 2000
- February 2001.

There was partly an overlap in the data, i.e. some units were interviewed in two of the three or even in all three months. To get data sets referring to one point in time, the LFS survey data was split into three different data sets. This was done on the basis of the interview date. All units interviewed in the same month were grouped and taken as separate data. Because of the overlap, some units are included in two or three data sets, but with their specific statements in the respective interview month. The results are three different data sets containing 12,391, 12,260 and 12,624 units respectively. For the simulation, a pre-selection of variables included in the data was made. The variables chosen and the respective realizations are presented in Table 3.5.

The Finnish Labor Force Survey is a systematic sampling procedure from the Central Population Register. This means that the individuals from the register are the sampling units. Hence, the sampling procedure of the Finnish LFS collects no household information. Therefore, no household structures are needed in the artificial universe, either. The simulation process aims at the creation of a pseudo population register, which is used as Finnish Universe.

The number of individuals in the pseudo universe is taken to be deterministic and set as 3,900,000. On the basis of the LFS data sets, multivariate frequency distributions including the variables described in Table 3.5 are calculated. For this, the calibration weights included in the data sets are used. The frequency distributions are taken as multidimensional probability distributions for the units in the pseudo population register. By applying a coding function, the alias method and a decoding function, the outcomes of the 5 variables are simulated for all units in the universe. Because there are three different data sets representing different points in time, three different universes are created.

3.2.2 Selected Results of the Simulation

The investigations presented in this subsection focus mainly on the marginal distributions of the variables and the correlations between the variables. Also conclusions on the differences between the data - which depend on the fact that the respective data sets were surveyed at different time - are possible.

Table 3.5: Variables included in the Finnish pseudo universes.

Variables	possible outcomes
region	0 Uusimaa 1 Southern Finland 2 Eastern Finland 3 Mid-Finland 4 Northern Finland 5 Aland
age	0 15 - 19 1 20 - 24 2 25 - 29 3 30 - 34 4 35 - 39 5 40 - 44 6 44 - 49 7 50 - 54 8 55 - 59 9 60 - 64 10 65 - 69 11 70 - 74
gender	0 male 1 female
labor force characteristics (lfstat)	0 employed 1 unemployed 2 conscripts 3 students 4 disabled 5 pensioners 6 persons performing domestic work 7 others 8 nonresponse 9 overcoverage
highest level of education (isced)	0 no answer 1 upper secondary education 2 post secondary non-tertiary education 3 5B-programmes 4 5A-programmes 5 second stage of tertiary education 6 level unknown

The marginal distributions for the five variables within the February 2000 universe are presented in Figure 3.3. Within this figure, the realized as well as the expected frequencies are presented. It can be seen that the marginal distributions in the pseudo universe almost equal the distributions in the survey data. Because all units in the pseudo universe

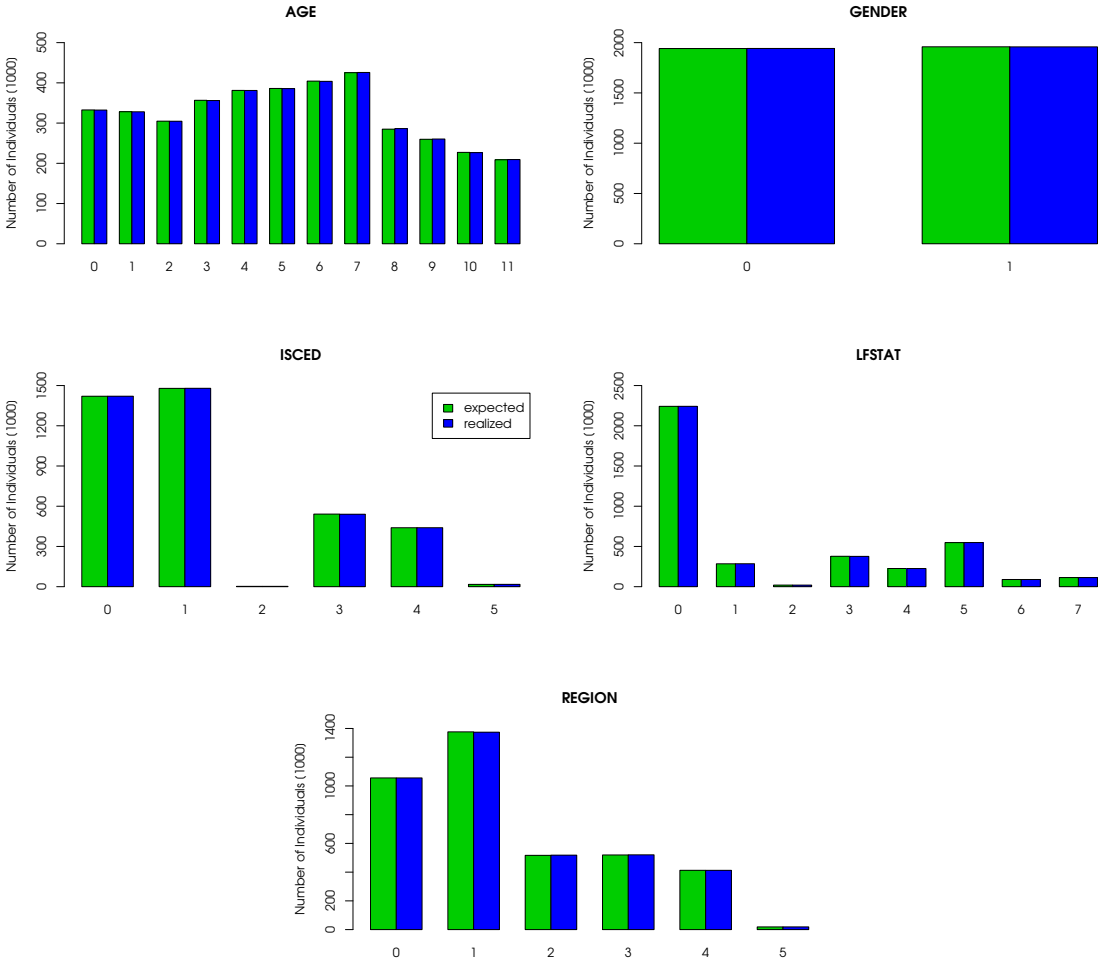

Figure 3.3: Marginal frequency distributions within the Finnish February 2000 pseudo universe

are simulated on the basis of the same distribution, the differences are almost negligible. Similar results appear when the other two universes are considered. The p-values when applying a χ^2 goodness of fit test to the expected and realized marginal frequency distributions within the three universes are presented in Table 3.6.

Table 3.6: p-values when applying a χ^2 goodness of fit test to the Finnish LFS realized and expected marginal frequency distributions.

Universe	region	age	gender	lfstat	isced
Feb 2000	0.0701	0.2664	0.4271	0.2983	0.4083
May 2000	0.4928	0.0145	0.2352	0.8120	0.5494
Feb 2001	0.9533	0.9821	0.1720	0.5731	0.8169

Next, the multivariate distributions within the three universes are considered. The contingency coefficient values are displayed in Table 3.7.

Table 3.7: Contingency coefficients within the Finnish pseudo universes.

data source		age	gender	lfstat	isced
Feb 2000 survey data	region	0.1016	0.01889	0.1485	0.1356
	age	-	0.0497	0.7632	0.4763
	gender	-	-	0.1693	0.0969
	lfstat	-	-	-	0.3768
Feb 2000 pseudo universe	region	0.1018	0.0188	0.1476	0.1367
	age	-	0.05077	0.7633	0.4765
	gender	-	-	0.1694	0.0960
	lfstat	-	-	-	0.3772
Feb 2000 relative differences	region	0.2%	-0.5%	-0.6%	-0.8%
	age	-	2.2%	0.01%	0.04%
	gender	-	-	0.06%	-0.9%
	lfstat	-	-	-	0.1%
May 2000 survey data	region	0.1028	0.0182	0.1317	0.1442
	age	-	0.0528	0.7531	0.4829
	gender	-	-	0.1642	0.0906
	lfstat	-	-	-	0.3729
May 2000 pseudo universe	region	0.1031	0.0184	0.1321	0.1439
	age	-	0.0533	0.7532	0.4830
	gender	-	-	0.1641	0.0898
	lfstat	-	-	-	0.3733
May 2000 relative differences	region	0.3%	1.1%	0.3%	-0.2%
	age	-	0.9%	0.01%	0.02%
	gender	-	-	-0.06%	-0.9%
	lfstat	-	-	-	0.1%
Feb 2001 survey data	region	0.1135	0.0185	0.1387	0.1504
	age	-	0.0516	0.7648	0.4835
	gender	-	-	0.1663	0.0754
	lfstat	-	-	-	0.3811
Feb 2001 pseudo universe	region	0.1128	0.0188	0.1384	0.1504
	age	-	0.0519	0.7648	0.4835
	gender	-	-	0.1667	0.0749
	lfstat	-	-	-	0.3814
Feb 2001 relative differences	region	-0.6%	1.6%	-0.2%	0
	age	-	0.6%	0	0
	gender	-	-	0.02%	-0.7%
	lfstat	-	-	-	0.08%

It is obvious that for the three universes the two dimensional dependencies are almost the same as in the data. Also a χ^2 goodness of fit test was applied to the expected and realized 5-dimensional distributions in the three universes. This was done by transforming them into univariate distributions by using a coding function. The classes were built with a minimum number of 5 elements in it. The resulting number of classes and p-values are

displayed in Table 3.8.

Table 3.8: p-values when applying a χ^2 goodness of fit test to the expected and realized Finnish LFS multivariate distributions.

Universe	number of classes	p-value
Feb 2000	1467	0,3334
May 2000	1372	0,3757
Feb 2001	1436	0.3595

The results indicate, that the distributions within the universes are the same as the expected distributions. The lowest p-value is around 33% with a very high number of 1372 classes. Hence, the hypothesis that the expected and realized distributions are the same cannot be rejected. It can be concluded, that the desired and the realized multivariate distributions are identical and particulary the desired correlation structures are reproduced.

The investigations show that the method used for the simulation leads to the desired distributions. Also the results indicate that the differences between the three distributions are only small. The marginal distributions and the contingency coefficients show only small differences. Of course this result is not surprising because the time lag is not very large.

3.2.3 Description of the Implemented Sampling Procedure

The sampling procedure of the Finnish LFS is a systematic random sampling with implicit geographical stratification (cf. Statistics Finland, 2001, p. 5). In Statistics Finland (2001) it is emphasised that there are not indications of a selection bias due to the sampling procedure. Therefore the selection process can be approximated by simple random sampling without replacement (cf. Statistics Finland, 2001, p. 5).

Beside the simplification that a simple random sample is implemented within the simulated sampling procedure, all sampling units are drawn at one go. This means that no separate waves are drawn. This is because of the temporary character of the pseudo universe. Hence, for one survey, $12,000$ units are sampled from the pseudo universe by simple random sampling without replacement.

Chapter 4

Microcensus Surveys (MC)

4.1 The German Microcensus Pseudo Universe (GMC)

4.1.1 Description of the Data and Application of the General Model

For the simulation the German pseudo universe, real Microcensus survey data from 1996 was taken. The raw data contains information of more than $700,000$ individuals. The variables covered and the respective possible outcomes are presented in Table 4.1. The data contains information about the place of residence, about the individual status as well as information about the labor force status of the persons.

As described in Chapter 2, a partition of the universe has to be done. The partition of the German universe is done by using the variables regional class and house size class, i.e. each regional class and house size class combination represents one stratum. Hence, $5 \times 214 = 1,070$ strata are build. Within each stratum there are a number of selection areas. This number is taken to be deterministic. The respective frequencies are obtained by taking 100 times the number of selection areas within the appropriate regional class and house size class combination in the survey data. This is done because approximately one-hundredth of the selection areas are selected during the survey process. In Table 4.2, the partition of the federal states into regional classes is displayed. The deterministic setup phase is followed by the creation of households and individuals. Therefore, the information within the data is used to get the distributions needed for the simulation.

Within each surveyed selection area, complete households are sampled. Therefore household structures are needed within the universe. To create them, first of all the number of households within each selection area is created at random. The distributions for the number of households per selection area are extracted from the data in the following way. All persons with the same regional class and house size class outcome are grouped. Within each group, the individuals are pooled according to the selection area and the household number. Next the number of households within each selection area is counted for each regional class and house size class combination. The absolute numbers are divided by

Table 4.1: Variables included in the German pseudo universe.

Variables	possible outcomes	
age	0 - 94	age in years
	95	95 or more years
gender	0	male
	1	female
ethnicity	0	German
	1	EU foreigner
	2	non EU foreigner
duration of job-seeking (dojs)	0	missing or non-seeking
	1	up to 6 months
	2	6 to 12 months
	3	more than 12 months
employment	0	employed labour force
	1	unemployed labour force
	2	non labour force
registered at the employment center (rec)	0	employed
	1	unemployed

Table 4.2: Partition of the federal states into regional classes and number of classes (noc) within each federal state.

	federal state	regional classes	noc
1	Schleswig-Holstein (SWH)	1 - 7	7
2	Hamburg (HAM)	8 - 9	2
3	Niedersachsen (NIE)	10 - 30	21
4	Bremen (BRE)	31 - 32	2
5	Nordrhein-Westfalen (NRW)	33 - 76	44
6	Hessen (HES)	77 - 93	17
7	Rheinland-Pfalz (RLP)	94 - 106	13
8	Baden-Württemberg (BAW)	107 - 132	26
9	Bayern (BAY)	133 - 166	34
10	Saarland (SAL)	167 - 169	3
11	Berlin (BER)	170 - 174	5
12	Brandenburg (BRA)	175 - 179	5
13	Mecklenburg-Vorpommern (MVP)	180 - 185	6
14	Sachsen (SAC)	186 - 199	14
15	Sachsen-Anhalt (SAA)	200 - 205	6
16	Thüringen (THN)	206 - 214	9

the number of selection areas within the respective regional class and house size class combination. The resulting $1,070$ relative frequency distributions are taken as probabil-

ity distributions for the number of households per selection area in the 1,070 strata. On the basis of this distributions, the number of households in each selection area within the 1,070 classes are simulated at random by using the alias method.

The next step is the simulation of the number of persons in the households. Therefore the number of persons per household in the survey data is counted and the resulting frequency distribution is taken as probability distribution. This is done separately for each stratum. On the basis of the resulting 1,070 distributions, for each household the number of persons in it is generated. The alias method is the selected method for the simulation. After this operation the individual variables have to be created. This is done in two steps.

As described in Chapter 2, the correlation structure within a household is created by using the variables age and gender. To get realistic age and gender structures for the households, all households in the data with the same number of persons in it are grouped. This is done for each stratum individually. For each household size a data file is built in which the age and gender of all persons within the households of the respective size are included. For each household of the pseudo universe a household with the same size is drawn at random from all real households in the same stratum with the respective person number. The age and gender of all household members are taken from the drawn household and are assigned to the persons in the simulated households. This leads to a realistic age and gender structure.

Table 4.3: Number of households and persons within the federal states of the German pseudo universe.

federal state	Number of households	Number of persons
SWH	1,343,312	2,924,508
HAM	952,755	1,794,844
NIE	3,282,729	7,363,264
BRE	347,640	687,042
NRW	7,858,812	17,357,578
HES	2,748,344	6,128,783
RLP	1,824,484	4,155,488
BAW	4,774,624	10,590,105
BAY	5,612,198	12,719,037
SAL	514,464	1,087,930
BER	1,870,256	3,580,162
BRA	1,116,202	2,647,942
MVP	747,752	1,793,669
SAC	2,078,257	4,655,381
SAA	1,209,769	2,799,846
THN	1,128,283	2,629,173
TOTAL	37,409,881	82,914,752

To design the missing variables conditional distributions are calculated. First of all, the multivariate realizations of the variables age, gender, nationality, duration of job

seeking, labor force status and registration at the employment center are considered. The absolute number of occurrence of each multivariate realization in the survey data is counted. This is done for each stratum separately. The frequency distributions are taken as a basis for the simulation of the remaining variables. Each person in the pseudo universe is treated independently of the others. For a given person, the multivariate frequency distribution of the respective stratum is taken and all cases for which the age and gender of that person is appropriate are separated into a new multivariate distribution with only four variables, excluding age and gender. The number of all cases in the new distribution are counted and the absolute frequencies are divided by this number. The result is a four dimensional conditional relative frequency distribution, which is taken as multivariate conditional probability distribution for the four remaining variables. By using the inversion method in conjunction with a coding and decoding function, the variables are generated.

4.1.2 Selected Results of the Simulation

As described above, regional and house size class combinations within the German pseudo universe are simulated independently of each other and then are aggregated to the whole universe. Therefore it is of main interest if the realized global structure equals the structure in the data used for simulation. Also, differences between the small areas and the different house size classes are of interest. The heterogeneity of the universe should reflect the heterogeneity within the real universe indicated by the survey data.

First of all, the global figures are treated. In Table 4.3, the number of households and person within the pseudo universe are displayed. In total, $82,914,752$ people are created. Within Figure 4.1, the expected and realized marginal distributions of the 6 variables of interest are presented. It can be seen that the numbers expected on the basis of the survey data almost equal the numbers realized in the pseudo universe. Nevertheless, there are differences which are a result of the fact that the small areas are aggregated. Differences included therein are cumulated and result in the global differences. They are also reflected by a χ^2 goodness of fit test. The resulting p-values are presented in Table 4.4.

Table 4.4: p-values when applying a χ^2 goodness of fit test to the expected and realized German MC marginal distributions.

age	gender	ethnicity	dojs	employment	rec
≈ 0	0.0494	0.0294	0.0021	0.0024	0.0007

The two-dimensional correlations within the pseudo universe equal the respective ones in the survey data. This is shown by the contingency coefficients displayed in Table 4.5. The differences are negligible. It can be concluded that the global structure of the German pseudo universe is close to the global structure of the German survey data used for simulation.

The reason for simulating small regions is to keep the differences between those regions within the pseudo universe. Therefore it is of importance, whether they are included and

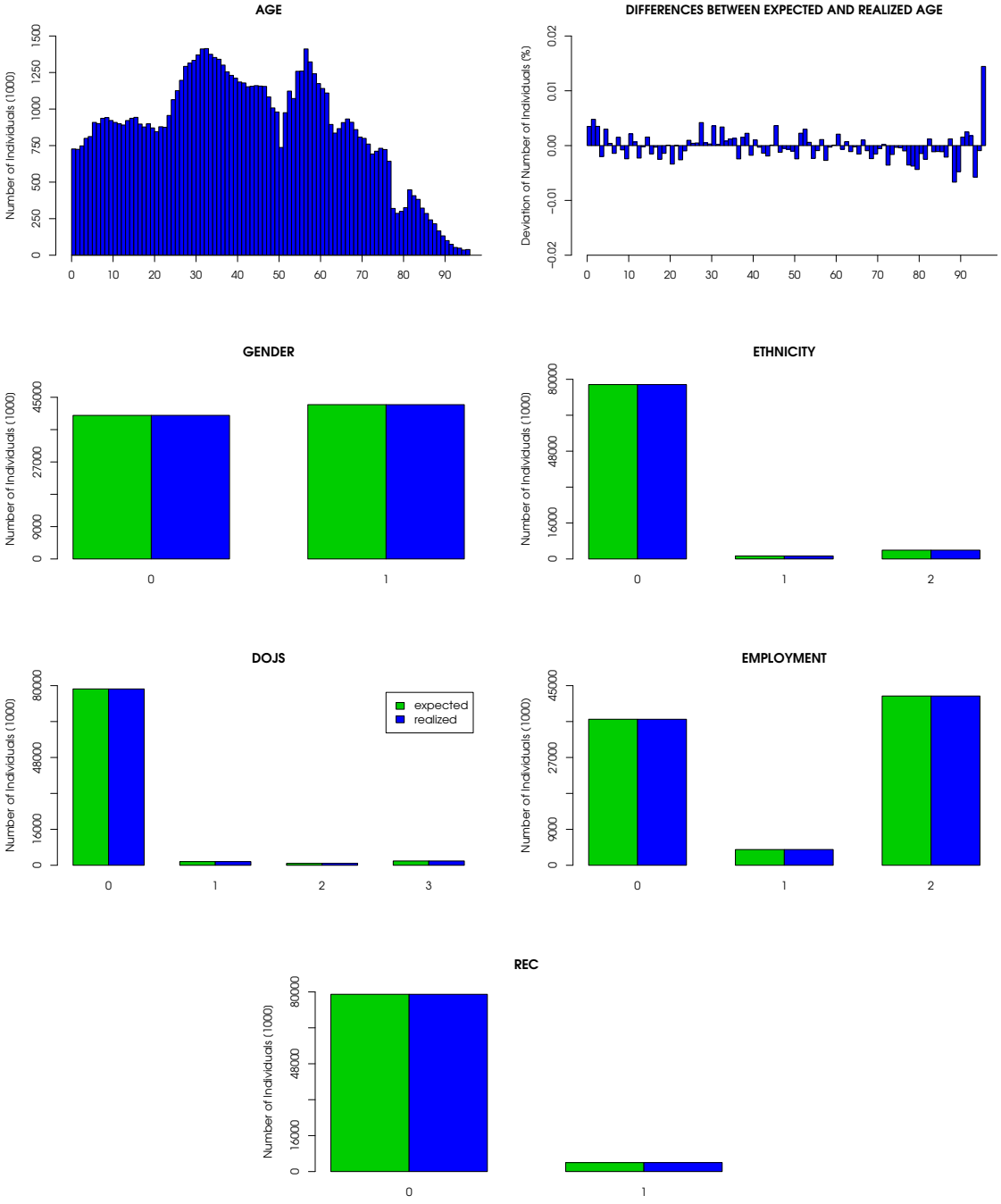

Figure 4.1: Marginal frequency distributions within the German pseudo universe.

which effects they have. In Figure 4.2 the marginal distributions within Baden - Württemberg and Mecklenburg - Vorpommern as two federal states that differ significantly are displayed. To be able to compare the distributions, the relative frequencies for the variables age, employment and rec are displayed. Differences between the two federal states are obvious. There is a strong difference in the age structure. Also the unemployment rate within Mecklenburg-Vorpommern is much higher than it is in Baden-Württemberg. Hence, regional structures are reflected within the pseudo universe. That the distributions

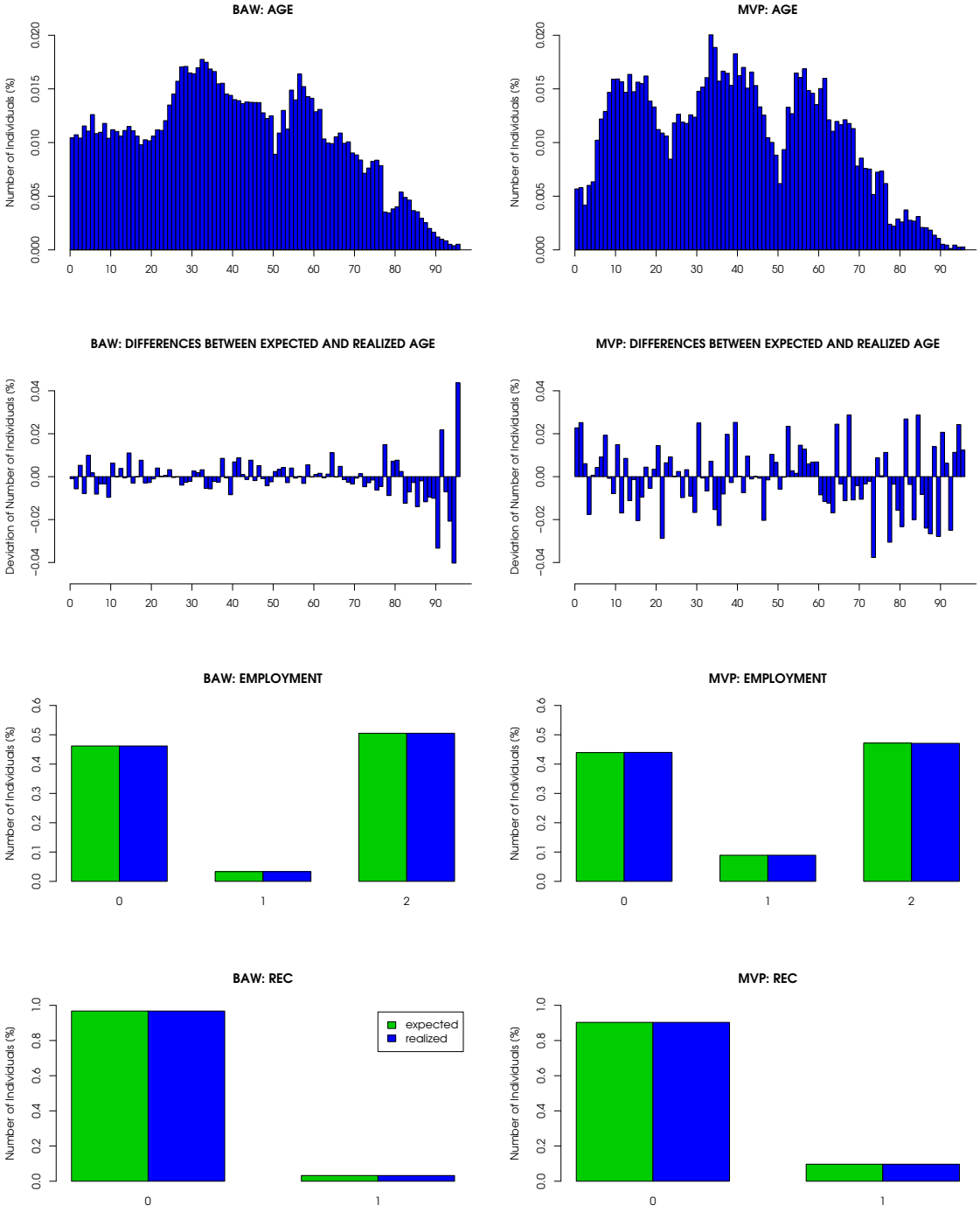

Figure 4.2: Marginal frequency distributions within the German federal states BAW and MVP

resulting are similar to them in the survey data is shown by the p-values when the expected and realized marginal distributions are tested with a χ^2 goodness of fit test (Table 4.6). Except the variable age the p-values are ok. The small p-values for age come from the fact that a lot of classes are used. Because of this the χ^2-test reacts very sensitive on small differences.

Table 4.5: Contingency coefficients within the German pseudo universe.

source		gender	ethn.	dojs	empl.	rec
survey data	age	0.1130	0.1361	0.1983	0.6052	0.1830
	gender	-	0.0231	0.0209	0.1589	0.0162
	ethn.	-	-	0.0430	0.0619	0.0430
	dojs	-	-	-	0.6339	0.6412
	empl.	-	-	-	-	0.6764
pseudo universe	age	0.1129	0.1362	0.1985	0.6051	0.1831
	gender	-	0.0234	0.0209	0.1590	0.0162
	ethn.	-	-	0.0428	0.0618	0.0429
	dojs	-	-	-	0.6339	0.6411
	empl.	-	-	-	-	0.6764
relative differences	age	-0.09%	0.07%	0.01%	-0.02%	0.05%
	gender	-	1.3%	0	0.06%	0
	ethn.	-	-	-0.5%	-0.01%	-0.2%
	dojs	-	-	-	0	-0.02%
	empl.	-	-	-	-	0

Table 4.6: p-values when applying a χ^2 goodness of fit test to the expected and realized marginal frequency distributions within BAW and MVP.

federal state	age	gender	ethnicity	dojs	employment	rec
BAW	≈ 0	0.2761	0.4041	0.9774	0.2753	0.3404
MVP	≈ 0	0.2409	0.4183	0.0888	0.0938	0.0588

The contingency coefficients also show that not only the marginal distributions but also the correlation structure differs between the federal states. The numbers displayed in Table 4.7 also show that the correlations reflect the values within the survey data. Hence, the heterogeneity within the German pseudo universe is a result of the heterogeneity within the survey data.

As shown above, there are significant differences between the federal states in the German pseudo universe. The next question is if there are major differences between the distributions within the house size classes. There should be differences because the classes reflect different character of living. The number of households and person within the five classes are presented in Table 4.8. In Figure 4.3 and 4.4, the marginal age distributions within the house size classes 1 to 5 and the differences between the expected and the realised age distributions are presented. Indeed, there are major differences between the distributions. Especially within class 4 and 5, there are completely different age structures in comparison to the global one. In class 4 there are a lot of old people while there are disproportionate young people living in house size class 5. Of course, the different age structures have also influence on the distributions of the other variables. The marginal distributions of gender and employment within class 4 and 5 are displayed in Figure 4.5. The completely different age structures lead to significantly different gender and employment realizations. Within

Table 4.7: Contingency coefficients within the German federal states BAW and MVP.

data source		gender	ethnicity	dojs	empl.	rec
BAW survey data	age	0.1102	0.1693	0.1768	0.6006	0.1517
	gender	-	0.0270	0.0296	0.1619	0.0386
	ethnicity	-	-	0.0712	0.0857	0.0722
	dojs	-	-	-	0.6194	0.6209
	empl.	-	-	-	-	0.6689
BAW pseudo universe	age	0.1106	0.1693	0.1771	0.6006	0.1522
	gender	-	0.0275	0.0294	0.1622	0.0386
	ethnicity	-	-	0.0713	0.0857	0.0723
	dojs	-	-	-	0.6189	0.6208
BAW relative differences	age	0.3%	0	0.2%	0	0.3%
	gender	-	1.9%	-0.7%	0.2%	0
	ethnicity	-	-	0.1%	0	0.01%
	dojs	-	-	-	-0.08%	-0.02%
	empl.	-	-	-	-	-0.01%
MVP survey data	age	0.1328	0.0994	0.3050	0.6574	0.2814
	gender	-	0.0213	0.0662	0.1179	0.0625
	ethnicity	-	-	0.0228	0.0313	0.0189
	dojs	-	-	-	0.6349	0.6536
	empl.	-	-	-	-	0.6787
MVP pseudo universe	age	0.1329	0.0987	0.3059	0.6577	0.2816
	gender	-	0.0208	0.0677	0.1192	0.0634
	ethnicity	-	-	0.0237	0.0310	0.0180
	dojs	-	-	-	0.6342	0.6538
	empl.	-	-	-	-	0.6781
MVP relative differences	age	0.08%	-0.7%	0.3%	0.05%	0.07%
	gender	-	-2.3%	2.3%	1.1%	1.4%
	ethnicity	-	-	3.9%	-1%	-4.8%
	dojs	-	-	-	-0.1%	0.03%
	empl.	-	-	-	-	0

class 4, the proportion of women is much higher than in class 5. Also, the non labour force proportion within class 4 is higher than 80% while it is lower than 50% in class 5. The differences between the two house size classes are also confirmed by the contingency coefficients presented in Table 4.9. Especially the two-dimensional correlations between age and the other 5 variables differ significantly.

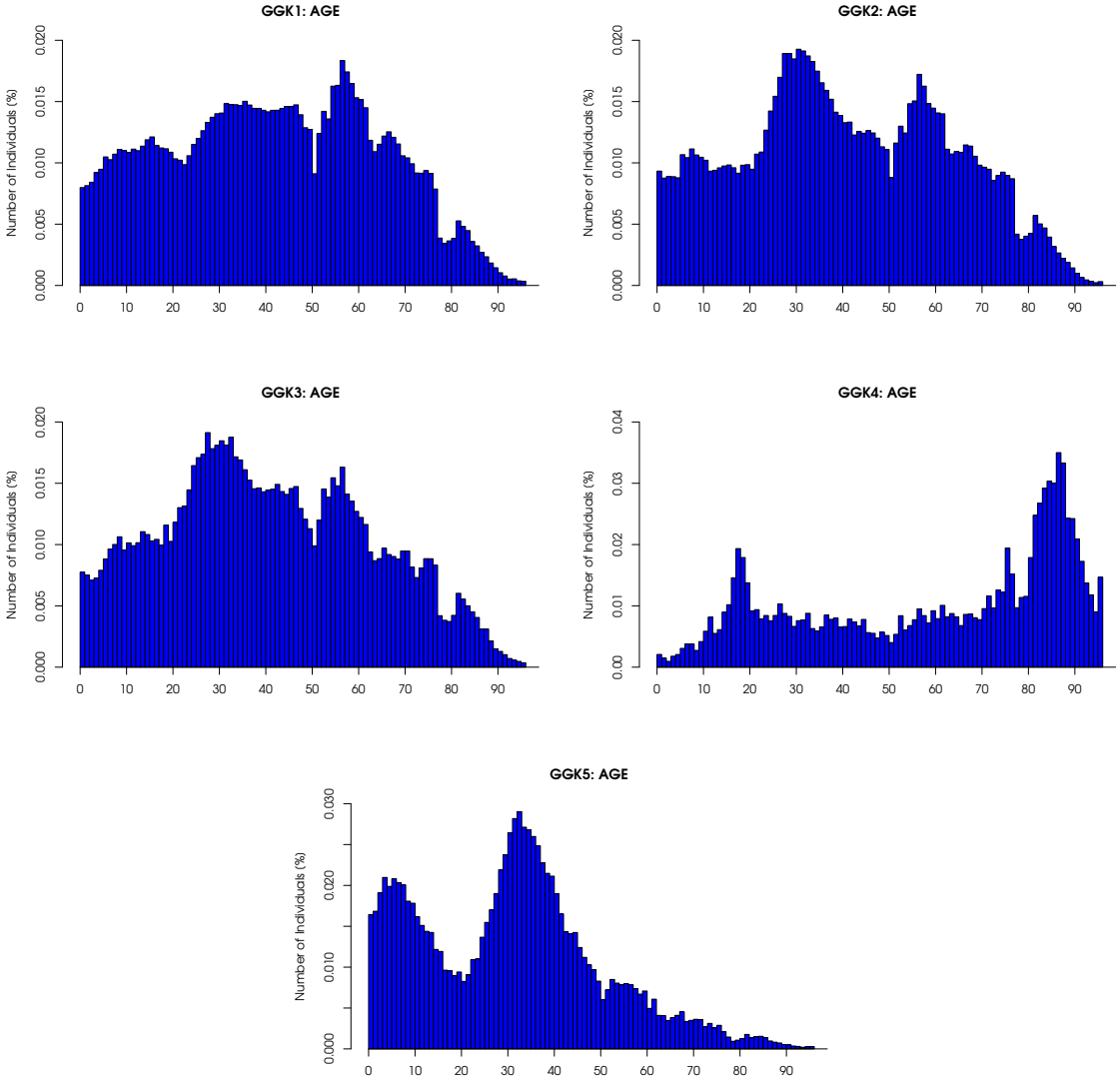

Figure 4.3: Realized marginal age frequency distributions within the German house size classes 1 to 5.

Table 4.8: Number of households and persons within the five house size classes of the German pseudo universe.

GGK	Number of households	Number of persons
1	20,501,492	48,357,739
2	8,580,881	16,957,105
3	5,949,803	10,995,498
4	120,493	763,473
5	2,257,212	5,840,937
TOTAL	37,409,881	82,914,752

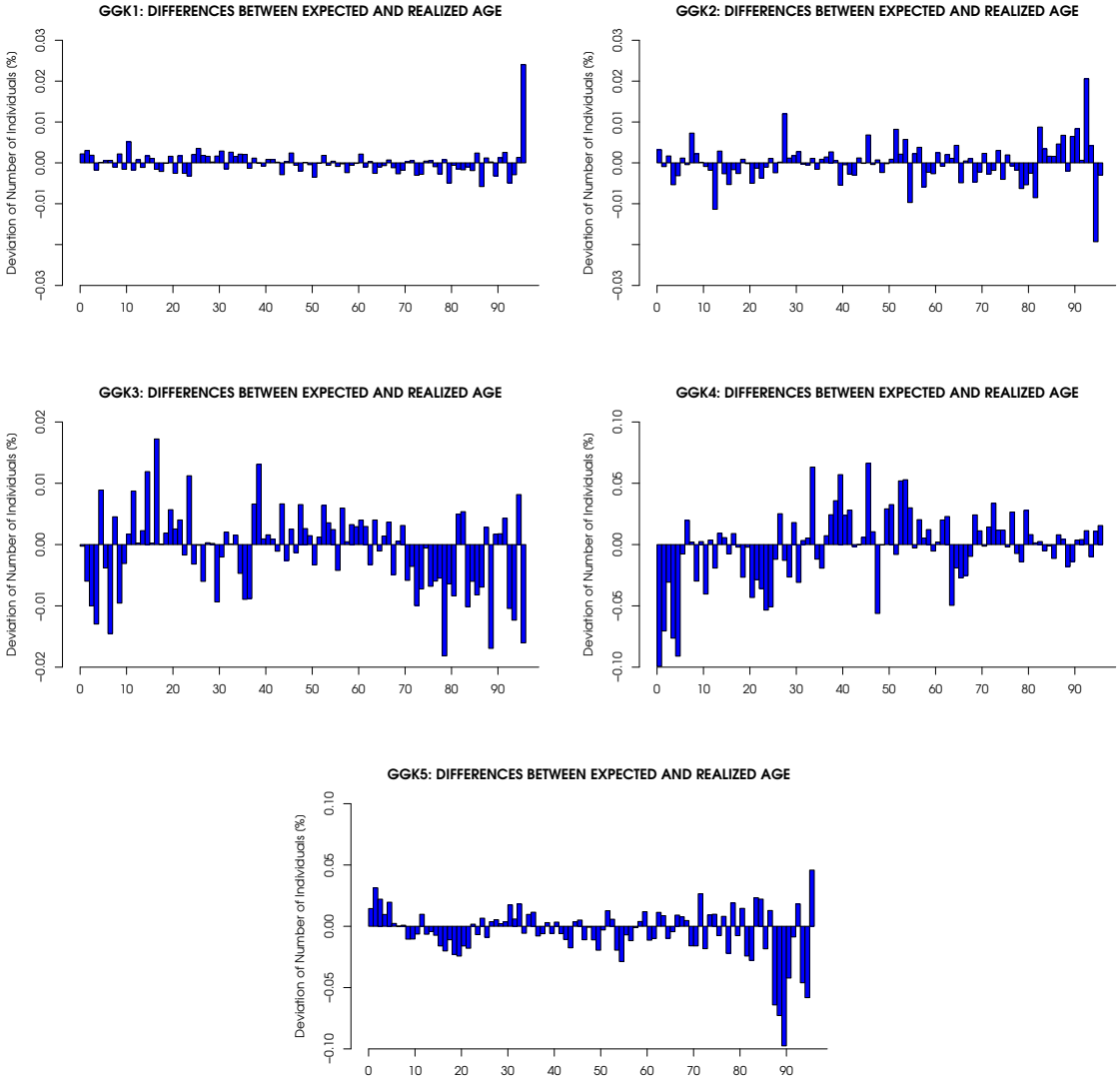

Figure 4.4: Differences between the expected and the realized marginal age frequency distributions within the German house size classes 1 to 5.

The results presented here show that the mechanism used for the simulation of the German pseudo universe is working well. The global figures show that the structure of the survey data is included in the pseudo universe. There are significant regional differences included in the universe. Especially across the different house size classes major heterogeneities are included.

4.1.3 Description of the Implemented Sampling Procedure

Within this subsection the implemented survey sampling procedure is described. This procedure was orientated mainly on the real Microcensus survey procedure. Because of the complexity of the real one, some simplifications were made.

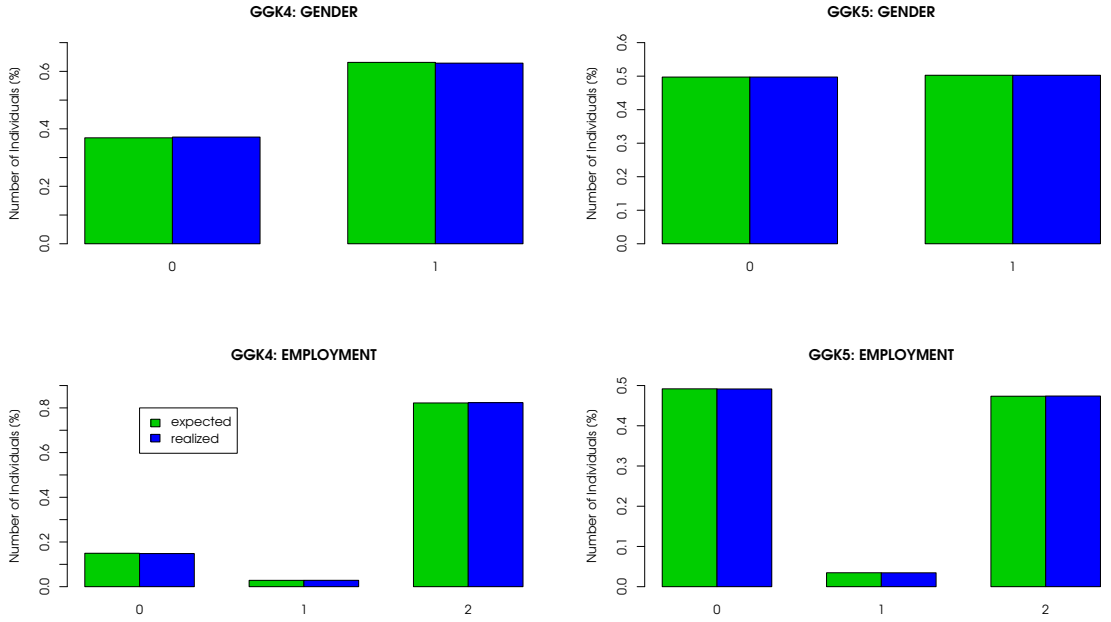

Figure 4.5: Marginal gender and employment frequency distributions within the German house size classes 4 and 5.

Within each federal state/regional class/house size combination, the selection areas are placed in a row. The ordering is done in a natural way, i.e. simply in the succession they were created. Then the sample areas are grouped with 100 sample areas in each group. The pooling is done by taking the first 100 sample areas as the first group, the second 100 areas as the second and so on. A simplification within the simulation is that the number of sample areas in each federal state/regional class/house size combination is divisible by 100 without remainder. So there are no problems with the last group. From each sample area group, one sample area is selected at random. All households and the respective persons in the households within the selected sample areas are included in the survey sample. By doing so, approximately 1% of the population is included in the sample.

4.2 The Austrian Microcensus Pseudo Universe (AMC)

4.2.1 Description of the Data and Application of the General Model

For the generation of the AMC pseudo universe AMC data of quarter 1 in 2001 were used. These comprise a total of 233 variables containing information on dwelling, household and personal characteristics. Except cases where missing values occur on the household or dwelling level, every record contains data of one individual. Relevant personal characteristics used for the generation process are displayed in Table 4.10.

Table 4.9: Contingency coefficients within the German house size classes 4 and 5.

data source		gender	ethnicity	dojs	empl.	rec
class 4 survey data	age	0.3897	0.3585	0.3583	0.5286	0.2814
	gender	-	0.1563	0.1301	0.2029	0.1264
	ethnicity	-	-	0.1569	0.1469	0.1406
	dojs	-	-	-	0.6579	0.6609
	empl.	-	-	-	-	0.6722
class 4 pseudo universe	age	0.3913	0.3607	0.3602	0.5266	0.2784
	gender	-	0.1570	0.1269	0.2017	0.1249
	ethnicity	-	-	0.1487	0.1426	0.1365
	dojs	-	-	-	0.6576	0.6586
	empl.	-	-	-	-	0.6711
class 5 survey data	age	0.1000	0.1139	0.1779	0.6189	0.1597
	gender	-	0.0136	0.0159	0.1391	0.0119
	ethnicity	-	-	0.0626	0.0875	0.0591
	dojs	-	-	-	0.6171	0.6241
	empl.	-	-	-	-	0.6671
class 5 pseudo universe	age	0.1002	0.1149	0.1772	0.6187	0.1589
	gender	-	0.0140	0.0154	0.1390	0.0121
	ethnicity	-	-	0.0620	0.0864	0.0588
	dojs	-	-	-	0.6172	0.6241
	empl.	-	-	-	-	0.6674

The generation of the Austrian Microcensus (=AMC) pseudo universe basically follows the proceeding described in section 2 with some modifications necessary due to specifics of the sampling plan. The AMC sampling plan is different for two parts of the universe of Austrian dwellings. The first, called Part A, comprises dwellings in larger urban municipalities, the other, called Part B, dwellings in small, rural communities. Only in Part A a stratified random sampling is done, whereas in Part B a two-stage sampling procedure with stratified random sampling of PSUs is carried out. The generation of the AMC pseudo universe thus should reflect homogeneity or heterogeneity within respectively between strata in Part A as well as PSUs in Part B.

According to the sampling plan the universe therefore was partitioned into federal states, within federal states in Parts A and B and within each part into strata. Strata of Part B additionally are partitioned into PSUs. Table 4.11 shows the number of strata per part and federal state. The total number of PSUs within Part B is 1,710. Generation of individuals was carried out separately for each Part A stratum and each Part B PSU. Within one generation group - that means a Part A stratum or a Part B PSU - the same generation distributions are used. As a consequence of the different sampling plans for Part A and B, every generation group of Part A (i.e. every stratum) but not of Part B (i.e. every PSU) is represented in the AMC data set. That means that AMC data are available for each generation group of Part A, but only for sample generation groups of Part B.

Table 4.10: Variables included in the AMC pseudo universe

Variables	possible outcomes	
age	0-99	age in years
gender	0	male
	1	female
nationality	0	Austria
	1	former Yugoslavia
	2	Turkey
	3	other
employment	0	employed at least one hour
	1	not employed
	2	not relevant / unknown
educational level	0	not completed compulsory school
	1	completed compulsory education
	2	completed apprenticeship
	3	medium secondary level
	4	secondary academic schools
	5	upper secondary level school
	6	post secondary school
	7	tertiary level school, not university
	8	university
	9	school aged child

Table 4.11: Partition of federal states of the AMC pseudo universe into Parts and strata

number	federal state	number of strata	
		Part A	Part B
1	Burgenland (BGL)	110	6
2	Kärnten (KTN)	132	7
3	Niederösterreich (NOE)	134	16
4	Oberösterreich (OOE)	134	12
5	Salzburg (SBG)	131	5
6	Steiermark (STM)	123	13
7	Tirol (TIR)	115	11
8	Vorarlberg (VBG)	146	—
9	Wien (WIE)	164	—

Sampling units in the AMC are dwellings, the sample frame is the stock of dwellings. Information on the number of dwellings in each Part A stratum and each Part B PSU in Austria was available from the HWZ 91, that is the Austrian Housing Census of 1991 (in Austria called: Häuser-und Wohnungszählung). Changes in the stock of dwellings are reflected in the AMC as aborts are reported and new dwellings are selected additionally

to the initial sample. Virtual households and persons were generated only for housings serving as permanent residence, for all other dwellings no households and individuals had to be generated.

Thus first of all the actual number for both types of dwellings had to be estimated for every generation group. Therefore the ratio of their respective number in the AMC data to their number in the initial sample was built for every AMC generation group. An estimate of the number of actual permanent residence housings and other dwellings was determined by multiplying their number in the HWZ with the respective ratio.

For virtual dwellings serving as permanent residence the number of households, the number of persons per household, and values for the personal characteristics of individuals had to be generated, for other dwellings the number of households and persons per household were set zero.

In a data processing step cases with missing values in one of the interesting variables were discarded and strata respectively PSUs are numbered consecutively within parts of federal states respectively strata.

Generation distributions according to the general model(see section 2) were built separately from this data set for each of the 1,557 generation groups represented therein, that is for each of 1,189 Part A strata and each of 368 sample PSUs of Part B.

For permanent residence dwellings in Part A the number of households in a dwelling was generated as random variable from the empirical distribution of number of households in these dwellings in the respective stratum of the AMC data set. Generation of households was carried out as for the German Microcensus Universe with a small modification regarding the creation of personal characteristics.

As strata are rather small, in many cases only one person with a given age and gender would be available. Thus generation of the additional personal characteristics educational level, nationality and employment according to their conditional distribution given age and gender would result in a "cloning" of this individual and - if this is the case for all individuals of one household - to a replication of entire households. To reduce the extent of replication of households, age measured in 5 year-categories was used for the construction of conditional distributions.

Generation of dwellings in Part B needed some modification as not every PSU is represented in the AMC. To generate a specific pseudo universe PSU therefore an AMC sample PSU of the same stratum was chosen at random and used as a model for the generation process. The number of actual permanent residence housings and other dwellings was estimated using their respective ratios in the model PSU. For the creation of permanent residence dwellings the generation distributions of the model PSU were used. So in every stratum of Part B several pseudo universe PSUs have the same generation distributions. Due to the random nature of the generation process and their different sizes these PSUs are not identical. Given the model PSU the proceeding for generation of dwellings was the same as for Part A dwellings.

In a last step all dwellings of a generation group, that means permanent residence and other dwellings, were pooled and arranged such that the positions of other dwellings where chosen at random and permanent residence dwellings were arranged according to their generation order.

4.2.2 Selected Results of the Simulation

Generation of the AMC pseudo universe was performed for small groups which are aggregated to form the total pseudo universe. Main interest therefore concerns the question whether the structure in the pseudo universe corresponds to that in the AMC data. The generated universe consists of 7,462,802 virtual individuals. Table 4.12 shows the distribution of households and persons in Part A and B of each federal state.

Table 4.12: Number of households and persons within federal states and Parts of the AMC pseudo universe

federal state	Part A: number of		Part B: number of	
	households	persons	households	persons
BGL	66,827	172,345	32,403	93,558
KTN	148,010	346,925	58,663	167,775
NOE	266,431	628,372	289,735	809,258
OOE	269,728	615,602	230,810	649,752
SBG	138,558	326,147	46,759	139,367
STM	223,145	479,157	209,143	622,631
TIR	133,373	316,693	101,172	307,652
VBG	121,053	321,451	—	—
WIE	745,942	1,466,117	—	—

Marginal Distributions of generated personal characteristics are presented in Figure 4.6. The frequencies realized in the pseudo universe are compared to so called "expected frequencies" which are computed from the empirical distributions within strata in the AMC data given the number of persons generated per stratum. Only for Part A these frequencies are expected from the generation distributions conditioned on the number of individuals per stratum. Distributions in Part B strata in fact are a mixture of the different generation distributions - that is the empirical distributions within model PSUs - where mixing proportions are the proportions of individuals in the respective model PSUs of one stratum. Nevertheless differences between realized and "expected" frequencies are of small order.

To assess whether correlation structures in the pseudo universe are similar to those in the AMC data contingency coefficients were computed and are shown in Table 4.13. Absolute differences are small, relative differences are large only for small contingeny coefficients. Therefore it can be concluded that the global structure of the AMC data is reproduced well in the pseudo universe.

To give a more detailed insight into the structure of the pseudo universe, distributions of personal characteristics are also analyzed for the federal states "Tirol" and "Wien". Figure 4.7 shows the marginal distributions for the variables age, gender and educational level. Differences between the federal states are obvious for all three variables: The population in Wien is older with a higher proportion of females and people with higher educational level, especially completed secondary school or university. Differences between realized and from the AMC data expected frequencies again are of small order.

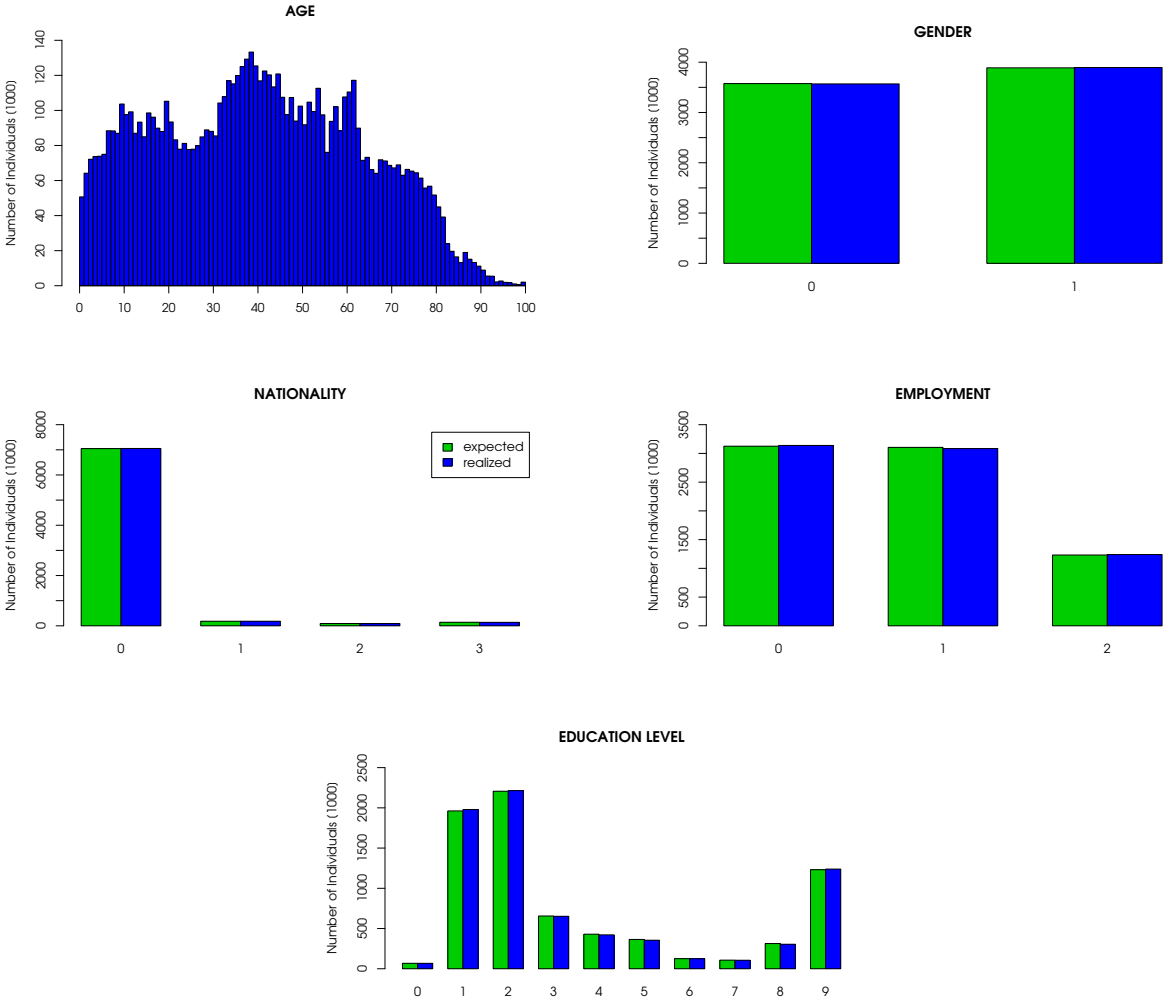

Figure 4.6: Marginal frequency distributions within the AMC pseudo universe and the AMC data

Table 4.14 shows the contingency coefficients for both federal states in the AMC data and the pseudo universe. Differences in the correlation structure between Tirol and Wien are distinct for two-dimensional marginals involving nationality as well as gender and education. These differences are reflected in the pseudo universe though some coefficients (age/gender in TIR, age/nationality in TIR and WIE) differ considerably from those in AMC data.

Heterogeneity should also be present between Parts A and B of the AMC sampling plan as partitioning is done according to the size of communities: whereas Part A comprises larger urban municipalities, Part B consists of rural communities. Besides comparison of the two parts is also of interest due to differences in the generation procedure: in Part A generation distributions can be determined from the AMC data for each generation group, i.e. each stratum, whereas in Part B this was the case only for part of the generation groups, i.e. PSUs which served as generation models. Figure 4.8 shows the marginal distributions of the variables age, gender and educational level. Differences between Part A and B are considerable, those between realized and expected frequencies again are of

Table 4.13: Contingency coefficients within the AMC data and the Austrian pseudo universe

source		gender	nat.	empl.	educ.
AMC	age	0.0992	0.1383	0.7622	0.7501
data	gender	-	0.0147	0.1676	0.2116
	nationality	-	-	0.0745	0.1472
	employment	-	-	-	0.7227
pseudo	age	0.1051	0.1423	0.7625	0.7437
universe	gender	-	0.0173	0.1626	0.2070
	nationality	-	-	0.0753	0.1559
	employment	-	-	-	0.7213
relative	age	5.9%	2.9%	0.0%	-0.9%
differences	gender	-	17.7%	-3.0%	-2.2%
	nationality	-	-	1.1%	5.9%
	employment	-	-	-	-0.2%

small order.

Differences in the correlation structures of Part A and B are shown in Table 4.15. For Part A pseudo universe coefficients reproduce those in the AMC data rather closely, for part B discrepancies are a little larger.

The results presented in this section show that the generation procedure is fairly successful in rebuilding the global structure as well as heterogeneity between federal states and parts of the AMC data in the generated pseudo universe.

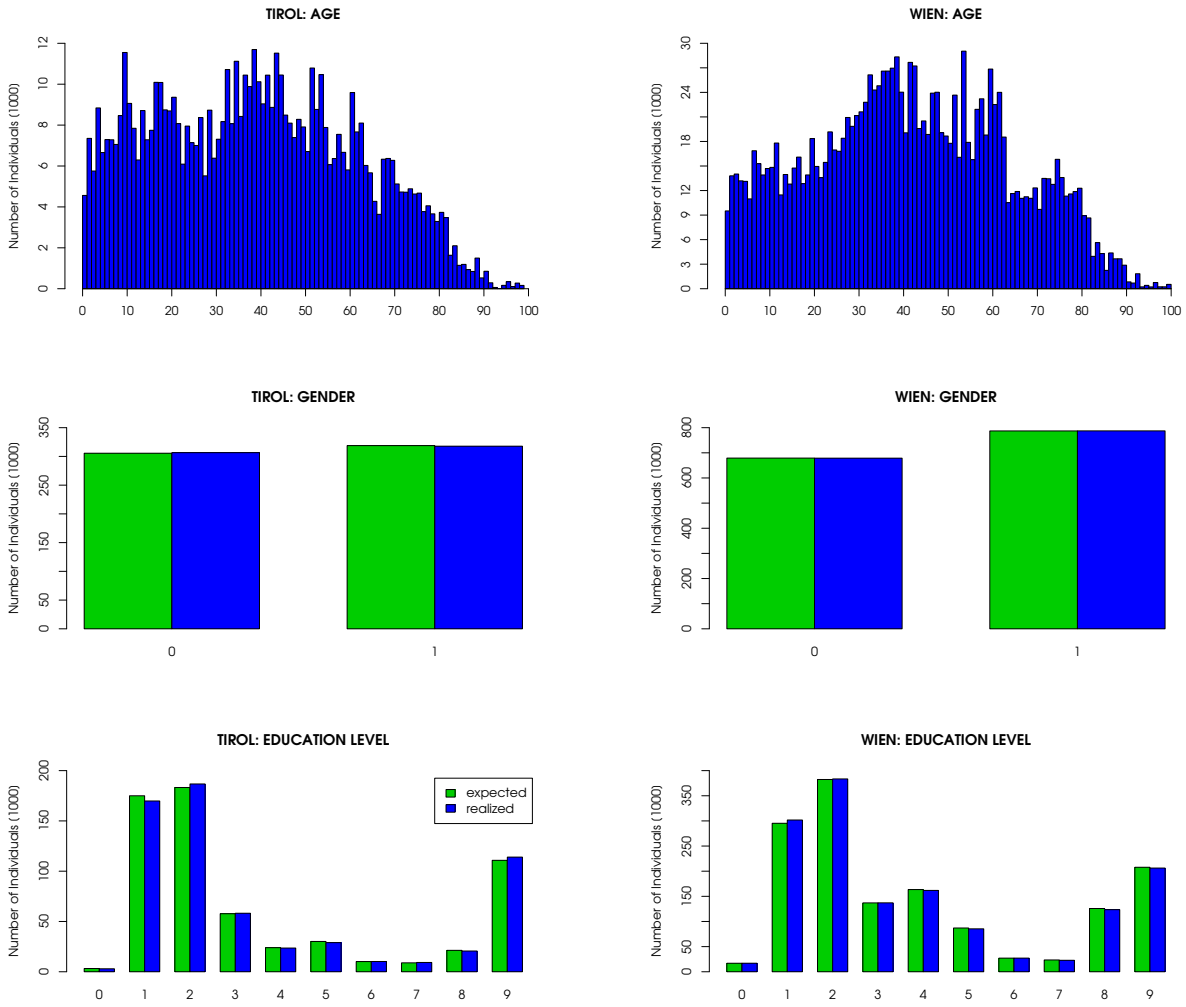

Figure 4.7: Marginal frequency distributions within federal states TIR and WIE of the AMC pseudo universe

4.2.3 Description of the Implemented Sampling Procedure

The sampling procedure for the simulation study imitates the procedure of the AMC but is not exactly identical.

In the AMC the proportional stratified sampling of Part A dwellings is realized by a systematic selection. Dwellings are sequentially ordered according to a specific ordering. The systematic selection is carried out with a deterministic starting value and a selection interval to obtain the desired sampling fraction.

This procedure cannot be replicated for the simulation studies as, given the ordering, it is purely deterministic. Moreover, not all variables used to determine the ordering in the AMC are generated in the pseudo universe. Therefore in the simulation studies a systematic sampling of dwellings in each stratum with a random starting value per stratum is carried out. Dwellings are ordered according to their dwelling number, which corresponds to the order of their generation for permanent residence dwellings.

Table 4.14: Contingency coefficients within the Austrian federal states TIR and WIE

source		gender	nat.	empl.	educ.
TIR	age	0.1450	0.2168	0.7543	0.7626
AMC	gender	-	0.0224	0.1900	0.2014
data	nationality	-	-	0.0557	0.1683
	employment	-	-	-	0.7230
TIR	age	0.1617	0.1903	0.7534	0.7555
pseudo	gender	-	0.0219	0.1863	0.1993
universe	nationality	-	-	0.0518	0.1638
	employment	-	-	-	0.7219
TIR	age	11.5%	-12.2%	-0.1%	-0.9%
relative	gender	-	-2.2%	-1.9%	-1.0%
differences	nationality	-	-	-7.0%	-2.7%
	employment	-	-	-	-0.2%
WIE	age	0.1452	0.2861	0.7662	0.7562
AMC	gender	-	0.0422	0.1430	0.1947
data	nationality	-	-	0.1168	0.2653
	employment	-	-	-	0.7165
WIE	age	0.1484	0.2555	0.7656	0.7474
pseudo	gender	-	0.0430	0.1447	0.1985
universe	nationality	-	-	0.1178	0.2640
	employment	-	-	-	0.7161
WIE	age	2.2%	-10.7%	-0.1%	-1.2%
relative	gender	-	1.9%	1.2%	2.0%
differences	nationality	-	-	0.9%	-0.5%
	employment	-	-	-	-0.1%

In Part B of the AMC PSUs are selected according to a proportional stratified sampling. The PSUs of one stratum are selected randomly with manual control to guarantee a uniform regional distribution of selected PSUs. In the second stage dwellings are selected systematically with a fixed starting value and a specifical ordering of the dwellings (according to dwelling criteria) within the PSU.

For the simulation studies the adequate number of PSUs is selected randomly. Within a selected PSU dwellings are ordered according to their dwelling number and selected systematically with a random starting value.

Furthermore different from the AMC sampling procedure only selection of dwellings for one interview wave, that is without rotations, is realized.

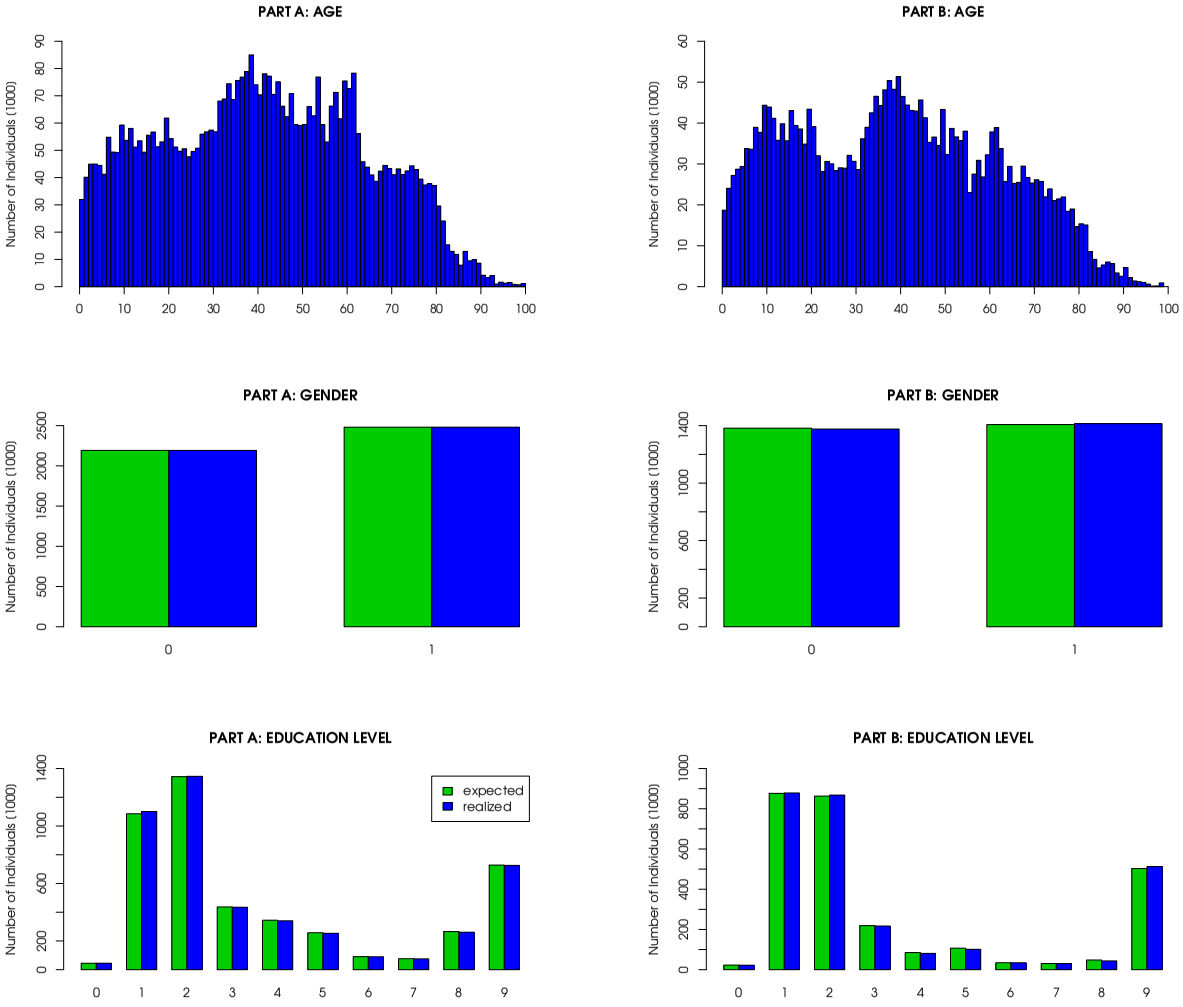

Figure 4.8: Marginal frequency distributions within Parts A and B of the AMC pseudo universe

Table 4.15: Contingency coefficients within Parts A and B of the AMC pseudo universe

source		gender	nat.	empl.	educ.
PART A	age	0.1036	0.1717	0.7640	0.7483
AMC	gender	-	0.0196	0.1578	0.2010
data	nationality	-	-	0.0927	0.1761
	employment	-	-	-	0.7206
PART A	age	0.1095	0.1779	0.7644	0.7432
pseudo	gender	-	0.0251	0.1521	0.1981
universe	nationality	-	-	0.0947	0.1876
	employment	-	-	-	0.7194
PART B	age	0.1095	0.1269	0.7598	0.7602
AMC	gender	-	0.0120	0.1831	0.2323
data	nationality	-	-	0.0550	0.1026
	employment	-	-	-	0.7280
PART B	age	0.1201	0.1235	0.7598	0.7529
pseudo	gender	-	0.0044	0.1791	0.2267
universe	nationality	-	-	0.0611	0.1041
	employment	-	-	-	0.7266

Chapter 5

Household and Budget Surveys (HBS)

5.1 The German Sample Survey of Income and Expenditure Pseudo Universe (EVS)

5.1.1 Description of the Data and Application of the General Model

For the simulation of the universe EVS survey data from 1998 is used. In opposition to the universes discussed previously the EVS data does not contain individual, but only household information. From the original 73890 survey households 55935 were obtained. This represents a fraction of $75,70\%$. The data covers 74 variables. The ones chosen for the simulation are presented in Table 5.1. The variable federal state was also used. The outcomes equal the ones from the German Microcensus. The variables net income and total expenditure are continuous. Therefore, they have to be treated separately. Quota variables needed for the sampling procedure are type, stasocio and net income.

Income and expenditure are highly influenced by the variable quarter. Hence, an individual universe for each quarter is created. Therefore, the survey data is separated along this variable. This leads to four data sets containing 12643, 14753, 14345, and 14194 households. Within each of the four pseudo universes 16 federal states are defined and all data sets are divided again on the basis of this criteria. A deterministic number of households is assigned to each of the federal states. The numbers are based on the German Microcensus 1995 and are presented in Table 5.2.

After finishing the deterministic part the random one is started. First of all the discrete variables nop, stasocio, and type are created. Therefore the three-dimensional frequency distributions of the variables within the quarter/federal state sub - datasets are calculated and used as probability distributions. By applying the methods discussed in Section 2.2 the attributes are generated individually for each household within the four universes. Based on the realizations the continuous variables are generated by using conditional distributions.

Table 5.1: Variables included in the German EVS pseudo universe.

Variables	possible outcomes	
quarter	1	January - March
	2	April - June
	3	July - September
	4	October - December
net income	continuous	DM per quarter
total expenditure	continuous	DM per quarter
number of persons (nop)	1 to 8	1 to 8 persons
	9	9 and more persons
type of household (type)	0	other
	1	mother/father alone + 1 child
	2	mother/father alone + 2 or more children
	3	couple with 1 child - spouse employed
	4	couple with 1 child - spouse unemployed
	5	couple with 2 or more children - spouse employed
	6	couple with 2 or more children - spouse unemployed
socio-economic status of the household (stasocio)	0	self employed
	1	civil servant or military
	2	employee
	3	worker
	4	unemployed, pensioner, student, others

Several problems appear when the continuous variables are to be simulated. The correlations within the survey data need to be considered. On the other hand, the number of survey units per quarter is relatively low. If differences between federal states are considered the units are additionally separated on the basis of this variable. And finally there are three discrete variables with $9 \cdot 7 \cdot 5 = 315$ different outcome - combinations. Hence, the number of survey units is not sufficient to estimate conditional distributions reflecting all dependencies. To avoid this problems, two simplifications are made. The first one is to include correlations between the discrete and the continuous variables explicitly only by using quarter, stasocio and type. The idea is that the income and expenditure of a household is mainly influenced by these three variables. The second simplification is that distributional assumptions are made.

For the variables income and expenditure a two-dimensional log normal distribution was chosen[1]. The respective parameters of the distributions were conditioned to the quarter of the universe and the realizations of stasocio and type. The means and variances as well as the covariances of log income and log expenditure were estimated from the survey data. When a (quarter, stasocio, type) - combination was realized less than 10 times within the data, then the realizations from a neighbor - quarter were added. This was done by pooling quarter 1 and quarter 2 or quarter 3 and 4 elements.

[1]This assumption is analog to the one made for creating the Swiss HBS universe (Section 5.2). It is motivated by investigations of the SFSO.

Table 5.2: Number of households included in the German EVS pseudo universes. (Kühnen, 2001, p. 12)

	federal state	number of households
1	Schleswig-Holstein (SWH)	1258500
2	Hamburg (HAM)	881400
3	Niedersachsen (NIE)	3434500
4	Bremen (BRE)	344600
5	Nordrhein-Westfalen (NRW)	8031800
6	Hessen (HES)	2707700
7	Rheinland-Pfalz (RLP)	1757600
8	Baden-Württemberg (BAW)	4701700
9	Bayern (BAY)	5339300
10	Saarland (SAL)	507100
11	Berlin - West (BER)	1180111
12	Brandenburg (BRA)	1073700
13	Mecklenburg-Vorpommern (MVP)	760800
14	Sachsen (SAC)	2030200
15	Sachsen-Anhalt (SAA)	1200600
16	Thüringen (THN)	1075600
	total	36285200

5.1.2 Selected Results of the Simulation

For the discrete variables the same analyses were performed as before. The expected and realized distributions as well as the respective p-values were determined. The contingency coefficients were also calculated. For the variables income and expenditure statistical values like mean, variance and correlation coefficient were calculated and compared.

Table 5.3: p-values when applying a χ^2 goodness of fit test to the expected and realized marginal frequency distributions within quarter 1 to quarter 4 EVS pseudo universes.

quarter	stasocio	type	nop
1	0.3483	0.9643	0.2637
2	0.6556	0.7052	0.8072
3	0.1704	0.8728	0.0956
4	0.4377	0.8661	0.6719

In Figure 5.1 the expected and realized marginal distributions of the variables nop, stasocio, and type within the quarter 1 and quarter 2 universes are presented. The results show that only small differences are given. Hence, the distributions within the survey data are re - created and included in the pseudo universes. This is also reflected by the p-values. For all four quarters, the values are presented in Table 5.3. Only within quarter 3 the value for type is slightly below 10%. The contingency coefficients within the four

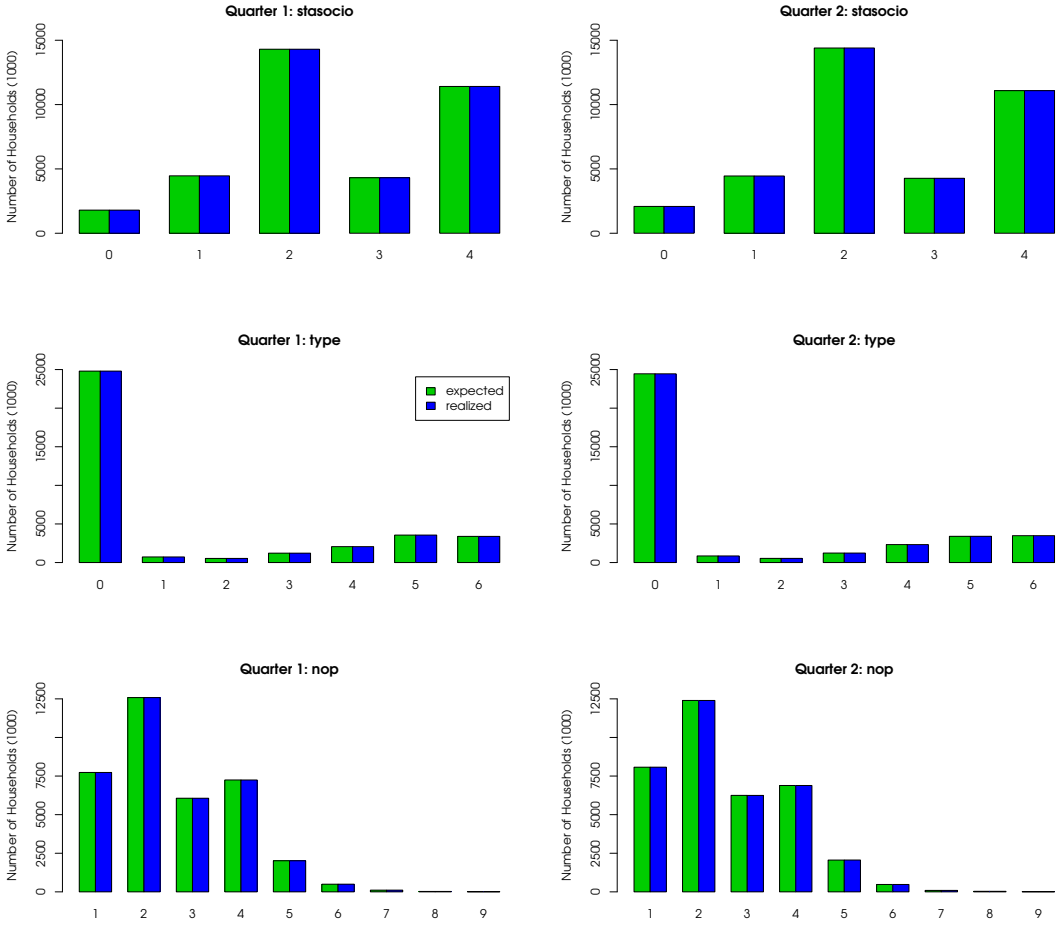

Figure 5.1: Marginal frequency distributions within the quarter 1 and quarter 2 EVS pseudo universes.

universes are presented in Table 5.4. They indicate that the correlation structure of the three variables is kept. The differences between expected and realized values are almost negligible.

Within Table 5.5 the realized correlation coefficient values for income and expenditure are presented. They show two facts. First of all the correlation between net income and total expenditure within the survey data is comparatively low. This is a result of the definition of the variables. All of the values are around 50%. Secondly, the coefficients within the universes are higher than the ones in the survey data. This indicates that the generation process overestimates the correlation between net income and total expenditure.

In Table 5.6 the means, medians, standard deviations as well as the 5% and 95% quantiles for income and expenditure are displayed. The results within the survey data and the universes do not differ heavily. The realized and expected mean and median values are close to each other, the standard deviations within the universes only differ slightly from the ones within the survey data. The realized quantiles within the universes are also close to the ones in the survey data. Hence, the used two - dimensional log normal distributions reproduces main characteristics of the survey distributions.

Table 5.4: Contingency coefficients within the quarter 1 to quarter 4 EVS pseudo universes.

data source		stasocio	type
quarter 1 survey data	nop	0.4288	0.7396
	stasocio	-	0.3774
quarter 1 pseudo universe	nop	0.4288	0.7396
	stasocio		0.3774
quarter 1 relative differences	nop	0	0
	stasocio	-	0
quarter 2 survey data	nop	0.4218	0.7481
	stasocio	-	0.3688
quarter 2 pseudo universe	nop	0.4217	0.7481
	stasocio	-	0.3687
quarter 2 relative differences	nop	−0.02%	0
	stasocio	-	−0.02%
quarter 3 survey data	nop	0.4279	0.7483
	stasocio	-	0.3791
quarter 3 pseudo universe	nop	0.4280	0.7483
	stasocio	-	0.3793
quarter 3 relative differences	nop	0.02%	0
	stasocio	-	0.05%
quarter 4 survey data	nop	0.4328	0.7557
	stasocio	-	0.3857
quarter 4 pseudo universe	nop	0.4328	0.7557
	stasocio	-	0.3857
quarter 4 relative differences	nop	0	0
	stasocio	-	0

Table 5.5: Pearson correlation coefficient values for income and expenditure within the quarter 1 to quarter 4 EVS survey data and pseudo universes.

quarter	pseudo universe	survey data
1	0.7274	0.5432
2	0.6229	0.5511
3	0.6442	0.5050
4	0.7014	0.5313

5.1.3 Description of the Implemented Sampling Procedure

The German Sample Survey of Income and Expenditure is a quota sample. For the technical implementation there are more simplifying assumptions necessary than before. The quotation is made on the basis of the variables stasocio, household type and net income. The frequencies of the several combinations are directly taken from the survey

Table 5.6: Statistics for income and expenditure within the quarter 1 to quarter 4 survey data and the pseudo universes.

		mean	median	std. dev.	5%	95%
q1 survey data	income	18381	16374	10791	5809	38072
	expenditure	29478	23440	30935	6948	66478
q1 pseudo universe	income	18750	16250	12452	5580	39659
	expenditure	28816	23785	21107	6876	67386
q2 survey data	income	18530	16408	11164	5535	38667
	expenditure	30671	23643	35392	6634	71375
q2 pseudo universe	income	18949	16184	15443	5530	40048
	expenditure	29671	23912	23591	6537	71610
q3 survey data	income	18414	16421	10834	5715	37382
	expenditure	30229	23844	34635	6980	67595
q3 pseudo universe	income	18806	16114	13876	5759	39482
	expenditure	29353	23908	22683	6909	69532
q4 survey data	income	20374	18132	12083	5986	42473
	expenditure	33359	26073	36820	7161	76396
q4 pseudo universe	income	20718	17862	14403	5950	43921
	expenditure	32294	26200	24792	6942	77792

data, because the data includes the original quotation.

The target number of households to be simulated within the federal states are taken from Kühnen (2001). Each quotation cell within a given federal state is multiplied by the respective number of target households. The resulting value is rounded to the next integer value. The number of target values are presented in Table 5.7.

The units within the pseudo universes are stratified according to the sample quotas. Within each quota - stratum a simple random sample without replacement is performed. This will enable to implement models about the response behavior of units within the different quota cells - independently of the other cells. After doing this the partial samples are aggregated to the EVS sample.

Table 5.7: Target number of households to be surveyed per federal state in the implemented German EVS drawing procedure.
(Kühnen, 2001, p. 12)

	federal state	number of households
1	Schleswig-Holstein (SWH)	2752
2	Hamburg (HAM)	2002
3	Niedersachsen (NIE)	6803
4	Bremen (BRE)	860
5	Nordrhein-Westfalen (NRW)	14614
6	Hessen (HES)	5496
7	Rheinland-Pfalz (RLP)	3719
8	Baden-Württemberg (BAW)	9026
9	Bayern (BAY)	10118
10	Saarland (SAL)	1213
11	Berlin - West (BER)	2434
12	Brandenburg (BRA)	2390
13	Mecklenburg-Vorpommern (MVP)	1750
14	Sachsen (SAC)	4241
15	Sachsen-Anhalt (SAA)	2644
16	Thüringen (THN)	2398
	total	73890

5.2 The Swiss HBS Pseudo Universe

5.2.1 Description of the Data and Application of the General Model

The data in use for the simulation of the Swiss data is survey data from the 1998 Household and Budget survey. The information within the data refers to households only. Included are $9,295$ households. Because the sampling units are households too, no individuals are created in the Swiss pseudo universe. The variable list in Table 5.8 shows that the variables income and expenditure are surveyed continuously. This means that the sampling methods discussed in Chapter 2 are not applicable without further work.

Within the survey data, the expenditure of each household is classified according to the 13 expenditure classes of the HBS nomenclature 1. This additional information is used to create not only one but 14 universes. This is done to allow for separate estimations concerning the different nomenclature 1 levels of expenditure. The first universe is generated on the basis of the total income and the total expenditure of all households within the survey data. Therefore all expenditures for a given household are cumulated. The remaining 13 universes cover the expenditure of the households at the 13 different expenditure classes of nomenclature 1 by considering only the respective expenditures from the survey data. Income is not included in those universes. The proceeding described subsequently was done for all of the 14 universes by using the respective data set.

Table 5.8: Variables included in the Swiss pseudo universe.

Variables	possible outcomes	
income	continuous	SFr. per month
expenditure	continuous	SFr. per month
type of household (type)	0	person alone
	1	mother/father alone + 1 child
	2	mother/father alone + 2 or more children
	3	couple
	4	couple with 1 child
	5	couple with 2 children
	6	couple with 3 or more children
	7	other
socio-economic status of the household (stasocio)	0	other
	1	salaried
	2	independent (without farmer)
	3	farmer
	4	unemployed
	5	pensioner

Table 5.9: Number of households included in the Swiss pseudo universes. (RENFER, 2001, p. 4)

Internal number	Stratum	Number of households
1	Plateau central	715,272
2	Region lemanique	592,590
3	Zurich	562,439
4	Suisse du Nord-Ouest	429,860
5	Suisse orientale	546,869
6	Suisse centrale	271,921
7	Tessin	150,281
	Total	3,179,231

The deterministic strata within the pseudo universe are built according to seven geographical regions. Hence, for the simulation of a given region, only the survey data out of the respective region is used. First of all, the number of households within the regions is assigned. Therefore the number of telephone connections in April 1998 is used (cf. RENFER, 2001, p. 4). This means that the number of households in the pseudo universe regions is deterministic, too. As mentioned above, no individuals are created. Hence, the next step will be the simulation of the household variables income, expenditure, type of household (type) and socio-economic status (stasocio). The main problem is that the variables expenditure and income are continuous.

The method of choice is to classify the income and expenditure data, i.e. to transform the continuous variables into discrete ones. Then the four-dimensional frequency distributions are easily determined by simply counting the weighted number of occurrence of each combination within the seven regions. The weights in use should account for non-response. After that the empirical distributions are taken as probability distributions and used for simulation. This is done again by applying a coding function, the alias method and a decoding function. After this, the realizations for income and expenditure are made continuous again. Therefore, for a given household, the income and expenditure are selected at random out of the values within the realized classes interval.

Two additional characteristics that have to be predefined. First of all, the categories for the classification of income and expenditure have to be fixed. Second, the distributions within the classes have to be defined. Those distributions are used for transforming the discrete income and expenditure into a continuous one. Because of its simpleness, an equal partition is taken within each class.

Table 5.10: Classification of expenditure and income.

Variable	Nomenclature 1	lowest value	class width	number of classes
Income	total	0	1,000	120
Expenditure	total	500	1,000	125
	1	0	10	370
	2	0	10	424
	3	0	100	404
	4	0	100	551
	5	0	100	501
	6	0	100	191
	7	0	100	475
	8	0	10	182
	9	0	1,000	113
	10	0	10	651
	11	0	10	900
	12	0	100	351
	13	0	1,000	109

Several ad hoc approaches for the classification were applied. The first one was to use *a lot of* classes. This classification is presented in Table 5.10. Two other approaches were to use only 5 categories and the logarithmic transformed income and expenditure for classification. For the first approach, similar class-widths were used. For the second, the classes were built that approximately the same number of households was included within each class. The income and expenditure distributions within the resulting universes were analyzed by the Swiss Federal Statistical Office (SFSO).

The analysis shows that the results of the universes created with only 5 income and expenditure classes are not very promising. The marginal distributions differ significantly from reality. Of course, this is strongly influenced by the equal distribution used. But also the correlations between the two variables are significantly reduced. This is because

the class widths are very big. This fact is independently of the distributions in use for re-transformation. Therefore it was resigned to try other distributions than the equal partition. Instead, the results of the universe with *a lot of* classes are close to what they should be. When using a lot of classes, applying an equal partition does only little harm. Because the class widths are very narrow, the correlation structures are also widely kept. On the other hand, there are a lot of empty combinations in the frequency distribution gained from the survey data. Because the approaches using only five classes seemed to produce unpredictable results, the approach using *a lot of* classes was taken.

Additionally to this approach, another generation process was implemented. Within this process, the variables income and expenditure are assumed to be two-dimensional log normal distributed. This is motivated by investigations of the SFSO. For all of the five stasocio combinations within the seven regions, individual estimates for the mean and the covariance matrix are calculated. The simulation process works as follows. First of all, the variables type and stasocio are simulated by using the two-dimensional frequency distributions from the survey data. In the next step, the variables income and expenditure are simulated by using a two-dimensional log normal distribution with a parameter set according to the realized region and stasocio combination. This is done for each household within the universe.

5.2.2 Selected Results of the Simulation

To do the same analysis for the Swiss pseudo universe as for the others, the classified distributions of expenditure and income were taken. This way it is possible to calculate the discrete marginal distributions, the p-values and the contingency coefficients. Those values give information, if the structure of the pseudo universe resembles the structure of the survey data before re-transformation of income and expenditure. To see if the correlation structure between income and expenditure is also kept after re-transformation, Pearson correlation coefficient values are calculated.

In Figure 5.2, the marginal distributions of type and stasocio are presented for the universe with the total expenditure and the expenditure for class 1 and 8 of nomenclature 1. The figures show two things. First of all, the expected and realized values almost equal each other. This means that the marginal distributions of type and stasocio in the pseudo universes are the desired ones. The other thing is that the marginal distributions do not differ between the three universes. Hence the differences lie only in the different expenditure classes of nomenclature 1 included in the respective universes. Because of the high number of classes, no figure including the marginal distributions of income and expenditure is presented. Instead, the hypothesis that the marginal distributions within the pseudo universe equal the expected marginal distributions was tested by applying a χ^2 goodness of fit test. The results are presented in Table 5.11.

Some p-values are very good and some are very low. This comes from the fact that the regions are simulated independently and are aggregated after that. Differences in the regions are cumulated and lead to the low values. The p-values for income and total expenditure indicate that the classified distributions within the pseudo universe and the expected ones are the same. Finally the contingency coefficients for the total expenditure

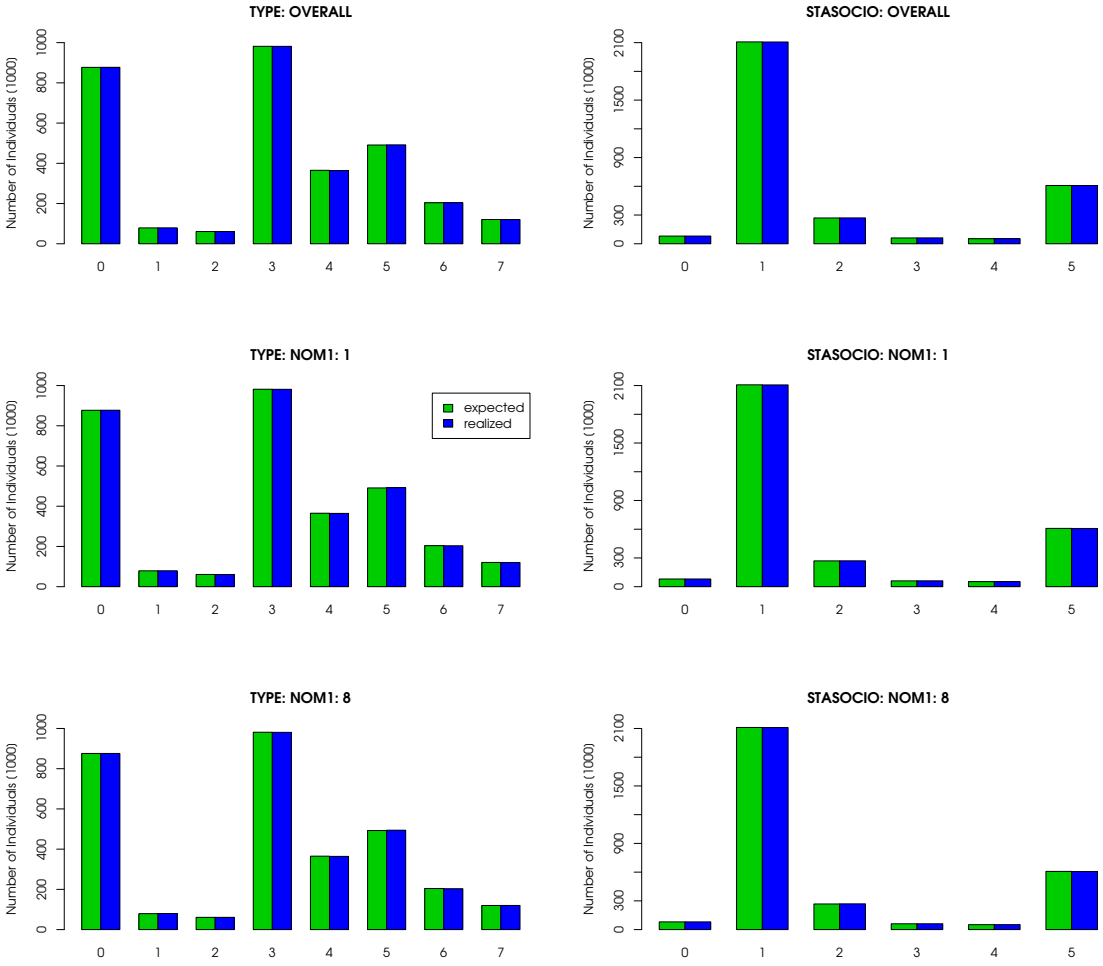

Figure 5.2: Marginal frequency distributions within selected Swiss *a lot of classes* pseudo universes.

Table 5.11: p-values when applying a χ^2 goodness of fit test to the expected and realized marginal distributions within the Swiss *a lot of classes* pseudo universe.

nomenclature 1	income	expenditure	type	stasocio
total	0.6653	0.3718	0.3534	0.0095
1	-	0.3848	0.2839	0.0083
8	-	≈ 0	0.0435	0.0119

universe show that the two-dimensional correlation structure within the survey data is included into the pseudo universe. The numbers are presented in Table 5.12.

The results in selected regions are regarded. Therefore Plateau central, Zurich and Tessin were chosen. In Figure 5.3 the marginal distributions for type and stasocio are presented. Especially for the variable type there are significant differences between the distributions. Hence, regional differences are included into the pseudo universe.

Table 5.12: Contingency coefficients within the Swiss *a lot of classes* pseudo universe.

source		expenditure	type	stasocio
survey data	income	0.9453	0.4924	0.4553
	expenditure	-	0.4958	0.4083
	type	-	-	0.3966
pseudo universe	income	0.9523	0.4922	0.4551
	expenditure	-	0.4955	0.4086
	type	-	-	0.3969
relative differences	income	0.7%	-0.04%	-0.04%
	expenditure	-	-0.06%	0.07%
	type	-	-	0.08%

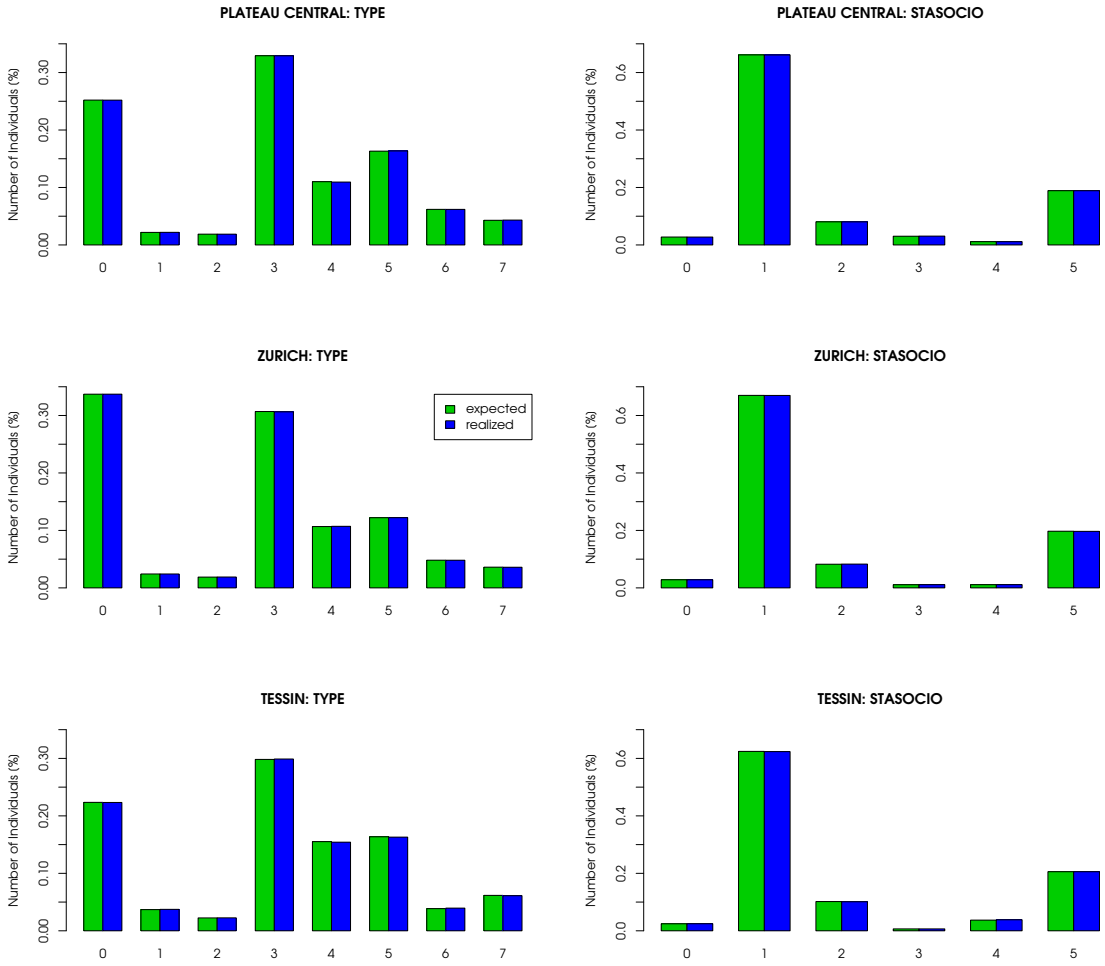

Figure 5.3: Marginal frequency distributions within the Swiss *a lot of classes* pseudo universe.

The p-values in Table 5.13 show that those differences are according to the survey data. It

Table 5.13: p-values when applying a χ^2 goodness of fit test to the expected and realized marginal distributions within the Swiss *a lot of classes* pseudo universe regions Plateau central, Zurich and Tessin.

Region	income	expenditure	type	stasocio
Plateau central	0.8057	0.8608	0.5578	0.4752
Zurich	0.8549	0.0752	0.9223	0.5995
Tessin	0.4898	0.9785	0.4583	0.2700

also becomes clear that the regional fit of expected and realized distributions is very good. This approves the proposition that the global differences are a result of the aggregation. In Table 5.14 the contingency coefficients for all four variables are presented. Those figures also show that the two dimensional correlations within the pseudo universe reflect the correlations within the survey data.

The results show that the classified multivariate structure of the Swiss survey data is included in the Swiss pseudo universe. The re-transformation that follows is done individually for income and expenditure, i.e. stochastically independent. To see if the two-dimensional correlations are kept, Pearson correlation coefficient values are calculated for the survey data as well as for the pseudo universe. The values are presented in Table 5.15.

It can be seen that the use of *a lot of* classes widely keeps the correlations. The resulting differences are almost negligible. Hence, the structure of the survey data is also kept after re-transformation. Within these tables, the correlation coefficients for the *log–normal* distribution are also presented. Those values are not as close to the survey data as the values for the *a lot of classes* universe. This is not surprising because the log–normal approach is not completely depending on the survey sample. On the other hand, the statistics in Table 5.16 show that characteristics like mean and median within the universes do not differ that much. Hence, the global characteristics are comparable.

5.2.3 Description of the Implemented Sampling Procedure

In reality, the Swiss Household and Budget survey consists of 12 waves, i.e twelve independent random samples, one per each month of a year. The Swiss pseudo universe represents a population at a specific point in time. This means that there is no time effect included in the universe. Therefore, within the sampling procedure implemented in the simulation, not twelve but only one sample is drawn. Instead, the size of the sample is twelve times the size of a real monthly sample (cf. Renfer, 2001, p. 9). The number of households being selected per region is shown in Table 5.17. Altogether, 9,000 households are sampled per sampling procedure.

In each stratum, a simple random sample without replacement is drawn. Therefore, a household within a stratum is drawn at random. By using a flag, it can be seen if the household is already included in the sample. If yes, then another one is selected, if no,

Table 5.14: Contingency Coefficients within the Swiss *a lot of classes* pseudo universe regions Plateau central, Zurich and Tessin.

source		expenditure	type	stasocio
Plateau central survey data	income	0.9097	0.5527	0.4932
	expenditure	-	0.5244	0.4161
	type	-	-	0.4597
Plateau central pseudo universe	income	0.9102	0.5527	0.4934
	expenditure	-	0.5247	0.4160
	type	-	-	0.4594
Plateau central relative differences	income	0.05%	0	0.04%
	expenditure	-	0.06%	-0.02%
	type	-	-	-0.07%
Zurich survey data	income	0.9946	0.5483	0.5062
	expenditure	-	0.5393	0.4977
	type	-	-	0.3713
Zurich pseudo universe	income	0.9465	0.5484	0.5060
	expenditure	-	0.5395	0.4954
	type	-	-	0.3723
Zurich relative differences	income	-4.8%	0.02%	-0.04%
	expenditure	-	0.04%	-0.5%
	type	-	-	0.3%
Tessin survey data	income	0.9307	0.5905	0.5950
	expenditure	-	0.5856	0.5903
	type	-	-	0.4328
Tessin pseudo universe	income	0.9309	0.5918	0.5964
	expenditure	-	0.5847	0.5930
	type	-	-	0.4331
Tessin relative differences	income	0.02%	0.2%	0.2%
	expenditure	-	-0.2%	0.5%
	type	-	-	0.07%

then the household is added to the sample and marked as already being included. This procedure is repeated until the desired number of households is reached.

Table 5.15: Pearson correlation coefficient values for income and expenditure within the Swiss survey data and the Swiss pseudo universes.

Region	pseudo universe		survey data
	log–normal	a lot of classes	
All	0.7625	0.6253	0.6298
Plateau central	0.7811	0.6815	0.6835
Region lemanique	0.7723	0.4711	0.4621
Zurich	0.7718	0.7431	0.7522
Suisse du Nord-Ouest	0.7407	0.6635	0.6720
Suisse orientale	0.7264	0.6138	0.6081
Suisse centrale	0.7665	0.7813	0.7779
Tessin	0.7903	0.5731	0.5737

Table 5.16: Statistics for income and expenditure within the pseudo universes and the survey data.

	mean	median	std. dev.
	income		
a lot of classes	8,367.41	7,346.48	5,736
log–normal	8,411.04	7,355.11	5,044
	expenditure		
a lot of classes	7,419.87	6,425.15	5,345
log–normal	7,425.63	6,501.01	4,307

Table 5.17: Number of households surveyed per stratum in the implemented Swiss survey procedure.

Number	Stratum	Sample size
1	Plateau central	2,004
2	Region lemanique	1,584
3	Zurich	1,572
4	Suisse du Nord-Ouest	1,188
5	Suisse orientale	1,164
6	Suisse centrale	732
7	Tessin	756
	Total	9,000

Chapter 6

Summary

The main task to be solved within workpackage 3 was based on finding best or at least appropriate universes for the surveys to be able to simulate the surveying process. This will allow to test the estimators and variance estimators in a practical environment to obtain *best practice recommendations* on the use of the methodology.

Since for the surveys of interest no register data or censuses but only samples were available, the universes had to be constructed from the given samples. Almost important in this process was the basic condition of data protection and anonymity of units. Therefore, several simplifications in the generation process had to be considered. Generally, the major aim of constructing adequate universes for the simulation process that include the frame of the surveys and the heterogeneity in the data as observed in the sample was met sufficiently. In some cases, where the simplifications seem have lead to weaker accuracy of the universes, additional efforts will be undertaken within the simulation study in regard of an applied sensitivity study under workpackage 1.

Additionally to the above described generation of DACSEIS universes, further emphasis during the simulation study will be laid on particular settings to allow for further investigations. These will comprise special settings for

1. the generation of a continuous Finnish LFS (3 waves) for variance estimation for change,
2. the generation of further variables needed to be used as auxiliary variables, e. g. for small area estimation methods,
3. the artificial coding of NUTS 3 – 5 level ares for small area estimation methods (this information couldn't be used in original due to disclosure control rules),
4. response influence for the German EVS for the investigation of raking methods,
5. and as a very important task, the generation of nonresponse behaviour for all surveys to allow the investigation of the methodology under nonresponse and weighting or imputation to correct for nonresponse.

These settings will be considered under workpackage 1 in strong cooperation with the corresponding methodological workpackages.

References

Devroye, L. (1986): *Non-Uniform Random Variate Generation*. New York: Springer.

Johnson, M. E. (1987): *Multivariate Statistical Simulation*. New York a.o.: John Wiley & Sons.

Kühnen, C. (2001): Das Stichprobenverfahren der Einkommens- und Verbrauchsstichprobe 1998. *Statistisches Bundesamt, Methodenberichte* **Heft 1**, 1–35.

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

Press, W. H., Teukolsky, S. A., Vetterling, W. T. and Flannery, B. P. (1992): *Numerical Recipes in C - The Art of Scientific Computing*. Second edition. Cambridge: Cambridge University Press.

Renfer, J.-P. (2001): *Description and process of the Household and Budget Survey of 1998 (HBS 1998)*. Swiss Federal Statistical Office. 1-19.

Statistics Finland (2001): *Continuous Community LFS Project: Implementation of Council Regulation 577/98*. Report. Statistics Finnland. 1-29.