DACSEIS
research paper series
No. 4

On the Simulation of Complex Universes in the Case of Applying the German Microcensus

Ralf Münnich and Josef Schürle

2003

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/`

On the Simulation of Complex Universes in the Case of Applying the German Microcensus

Ralf Münnich and Josef Schürle

University of Tübingen
Department of Statistics, Econometrics, and Operations Research
Mohlstraße 36, D-72074 Tübingen
e-mail: {ralf.muennich, josef.schuerle}@uni-tuebingen.de

Abstract: The aim of the DACSEIS project is to deliver recommendations on the use of variance estimators under complex survey designs in the presence of non-response. Since mathematical comparisons on the efficiency of variance estimation methods in this field are generally unavailable or lead to irrelevant results, adequate simulation studies have to be carried out that are based on realistic data sets. To be able to carry out a simulation study in the frame of complex designs one has to draw samples from a universe respecting for the true sampling design. However, in many cases, no data or only outdated data are available for the universe which leads to the need of adequately generating a micro data set from the sample.

Within this paper, a procedure of generating the universe for the German microcensus, which is a 1% sample of the population living in Germany, will be presented. The procedure allows for an adequate consideration of the individual information on a limited data set and can therefore be used as a basis for the simulations on variance estimation methods on the German microcensus data.

Keywords: Complex Survey, Monte-Carlo simulation, Sampling Design

1 Introduction

The aim of developing *current best* practice recommendations for variance estimation methods requires many items starting from the methodology to be considered and ending up in the correct environment which should be as close as possible to real settings including a set of *interesting variations*. A general comparison under these widely spread settings

makes it indispensable to elaborate the methodology in a larger Monte Carlo simulation study. As a basis for the simulation study, adequate universes have to be used. However, only in a manageable amount of cases universe data are available, e. g. censuses or register data. In other cases, adequate universes have to be constructed from available survey data.

This paper aims to show how universes were constructed in the case of the DACSEIS study using the data of the German Microcensus as an example. The generation process of the DACSEIS universes, however, should satisfy the following conditions:

- The concrete sizes of regions and strata should be considered precisely;
- the heterogeneity of household and individual distributions should reflect the real settings including cross correlations on region and stratum level;
- marginal distributions of all variables should be considered as they were known;
- disclosure control rules must not be violated.

All these conditions have to be regarded in the context of the methodology to be considered for the simulation study. Since the accuracy of estimates is the main research goal, estimators and variance estimators respecting for non-response, weighting, and imputation are to be considered. Therefore, sampling and non-sampling errors are the main interest of the study for which the use of adequate sampling schemes have to be included in the study to enable the investigation of the survey sampling techniques in a close-to-reality framework.

The consideration of the first three tasks seems easy to fulfil. However, more emphasis has to be put on the disclosure control rules. To avoid any possible identification of individuals, universe generation mechanisms like replication or mass imputation methods cannot be used since the former individuals from the sample are still present in the universe with the same dataset. Therefore, full stochastic generation mechanisms based on conditional distributions seem preferable.

However, the number of variables and the sensitivity of information of selected variables may also play an important role in this context. The more variables are selected for simulation the more difficult it is to apply the before mentioned mechanisms to generate a universe for the simulation study that satisfies the four conditions above.

Within the next section, the settings of the German Microcensus will be described including the underlying universe and the sampling scheme. The general universe generation procedure will be described in section 3. First, an overview to the general mechanism for individual and household surveys will be given, and second, the application to the German dataset which only contains discrete variables with a finite number of outcomes. The outcome of the procedure with respect to the German data will be shown in the fourth section. Finally, the results will be summarised with special emphasis on the DACSEIS simulation study.

2 The German Microcensus

The German Microcensus (GMC) is a 1% sample of the population living in Germany conducted by interviewers each year since 1957. The main aim of this survey is to gain

information on the structure of the population, the labour market including the labour participation and the housing situation. The questionnaire consists of a mandatory core program and a voluntary supplementary program. The full survey program for the years 1996 to 2004 is determined in the Microcensus law from January 17th 1996. The participation of sampling units is limited to a maximum four years in a row.

According to the description of the German Federal Statistical Office, the German Microcensus is conducted as a one-stage stratified area sample, where certain sampling districts are drawn in which all households and persons are interviewed (cf. `http://www.destatis.de/micro/e/micro_c1.htm`). Interpreted as a household and individual sample, the German Microcensus is a stratified cluster sample where the clusters are areas. Within each selected cluster all inhabitants and therefore all households are selected. The selection of clusters is described below.

The stratification of the population is done by three levels in the following way. The universe of the inhabitants in Germany is on the first level split up in 214 regional classes, which are in the context of the federal states (Bundesländer) and districts (RK1 .. RK214). The regional classes are arranged such that a minimum of 200.000 inhabitants is collected in each class. On the second level, the classes are build with respect to the size of houses (GGK). Five different classes are distinguished according to the following scheme:

GGK 1 Small houses with 1 to 4 apartments;

GGK 2 Middle size houses with 5 to 10 apartments;

GGK 3 Large houses with minimum 11 apartments;

GGK 4 Institutions with no apartments or when the number of persons exceeds $4 \times (4 + \text{number of apartments})$;

GGK 5 New buildings.

Within these $214 \times 5 = 1070$ strata, the sampling units are generated to achieve homogeneous clusters of approximately the same size. These sampling units are constructed regarding the size of houses in the following way:

GGK 1 About 12 apartments with a maximum of 70 persons;

GGK 2 One building each sampling unit;

GGK 3 6 to 9 apartments with classification per floor;

GGK 4 About 15 persons by initial of surname;

GGK 5 Selection of sampling units is based on the size of the new houses with respect to the former classification.

The sampling units (AWB) are arranged sequentially within each stratum of the universe as shown in figure 1. To achieve a one percent sample on households and individuals, the sampling units are pooled in zones each consisting of 100 sampling units where randomly one sampling unit is drawn. The number of zones $Z_{i,j}$ in each combination of regional

stratum i and house size class j may vary considerably since this number highly depends on the distribution of the households, individuals and their clustering within the sampling units. The organisation of the latter is under the responsibility of the statistical offices of the federal states. In fact, since the cluster size is one major critical issue for the efficiency of estimates gained by cluster sampling, the sizes were rearranged 1990 to achieve clusters of moderate size (cf. Meyer, 1994).

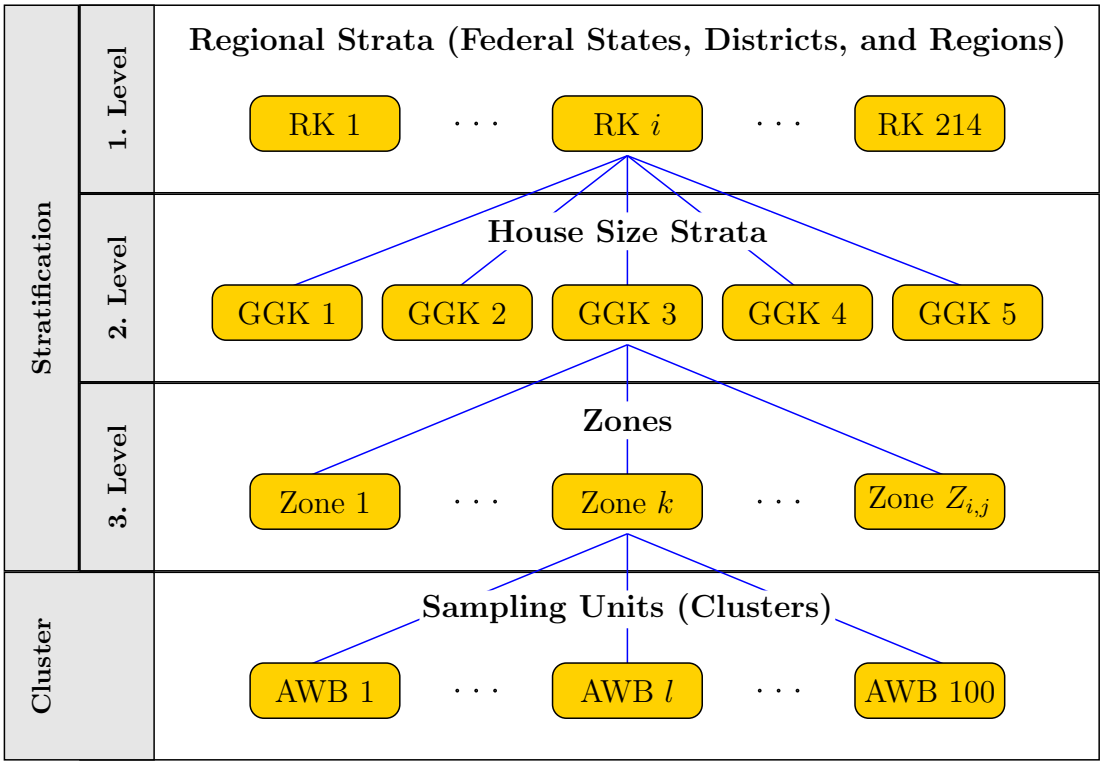

Figure 1: Overview to the German Microcensus design

According to the Microcensus law of 1996, households may only participate in the survey for 4 years in a row. This will be achieved by merging four zones in a so-called block and replacing every year the sampling units within the four zones consequently such that after four years all sampling units were replaced. These four zones of a block are sometimes called rotations quarters and play an important role for estimation of changes or longitudinal studies (cf. Heidenreich, 2002, or Zühlke, 2003).

The special design of the Microcensus is constructed to allow for a proportional design with respect to very small regional subclassifications. This may enable the user to gain information from small areas, such as estimates on NUTS 5 area level (*Nomenclature d'unités territoriales statistiques*; cf. `http://europa.eu.int/comm/eurostat/ramon/nuts/home_regions_en.html`). Nevertheless, further studies on the efficiency of adequate estimators have to be considered and will follow in the context of the DACSEIS project. A more detailed description of the German Microcensus can be found in Esser et al. (1989), Heidenreich (1994), Meyer (1994), Quatember (2002), and StBA (1999) or at `http://www.destatis.de/themen/e/thm_microzen.htm` and `http://www.gesis.org/Dauerbeobachtung/Mikrodaten/Daten/Abteilungsdaten/Mikrozensen/mz.htm`.

3 Generation of a German Microcensus Pseudo Universe

3.1 Generation of Random Variables

Within this subsection, the methods used for generating random variables are briefly described. Those methods are commonly 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), Kronmal and Peterson Jr. (1979) and Press et al. (1992) 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 known 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

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

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 (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 $f(x)$. A detailed description of how the two-point distributions are calculated is given in Kronmal and Peterson Jr. (1979). For each of the m distributions, the two outcomes as well as their probabilities are determined. This is called the setup phase. After the setup is completed, random variables 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 the number of random variables to be generated. 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 number 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 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

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

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

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

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 has been 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.

3.2 Description of a General Simulation Model for Creating Pseudo Universes

Before the generation process for the German Microcensus pseudo universe is described, the problem is treated more generally by regarding an arbitrary survey process. A model is described which allows for the generation of a pseudo universe adapted to this survey process. In the next subsection, the general model which is outlined in Figure 2 will be applied to the German Microcensus survey process.

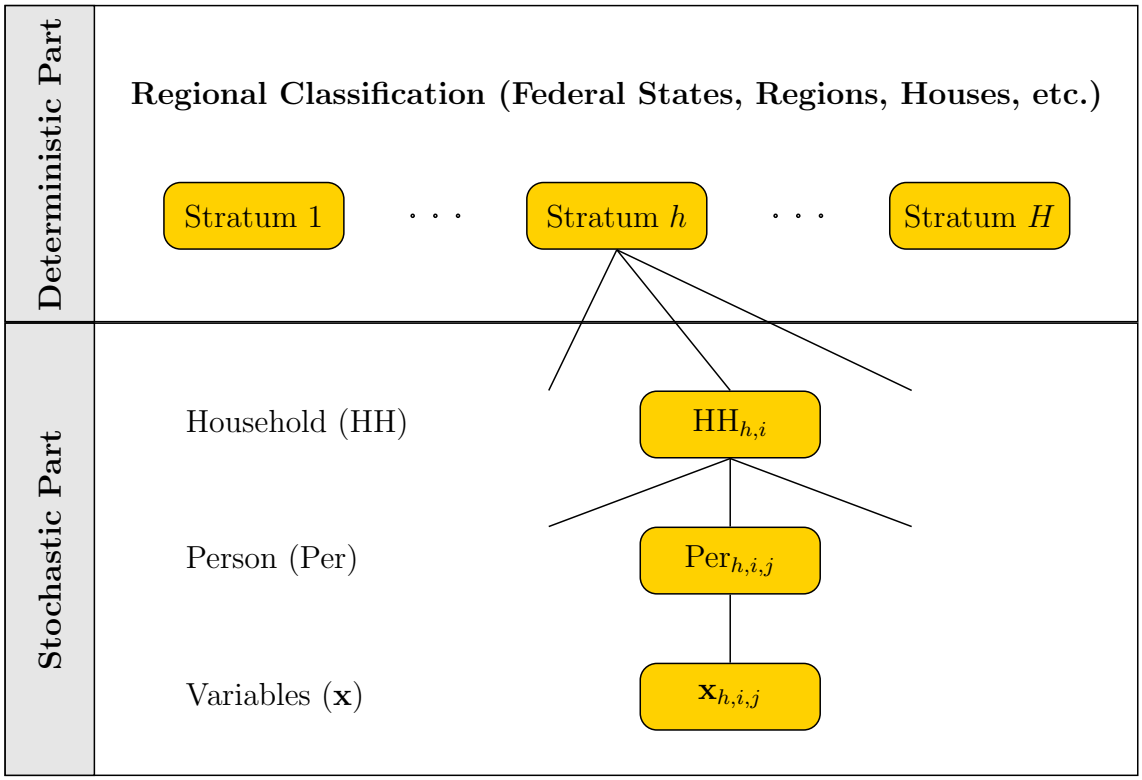

Figure 2: Main simulation principle.

Before a survey is performed, the respective universe is generally 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 realisation 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 realisations for the k variables of a given person in a 5-person household have to be conditioned by the realisation of the 4 other persons. The exact realisation 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 reasonable 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 realise 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 randomly drawing a realistic k-person household from a given real 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.

3.3 Application of the General Model to Create a German Microcensus Pseudo Universe

For the simulation the German pseudo universe, real GMC 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 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.

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	more than 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 1: Variables included in the German pseudo universe.

As described in the previous subsection, a partition of the universe has to be done. This partition 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 sampling units. This number is taken to be deterministic. The respective frequencies are obtained by taking 100 times the number of sampling units within the appropriate regional class and house size class combination in the GMC data. This is done because approximately one-hundredth of the selection sampling units are selected during the survey process. In Table 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 GMC data is used to get the distributions needed for the simulation.

Within each surveyed sampling unit, complete households are sampled. Therefore household structures are needed within the universe. To create them, first of all the number of households within each sampling unit is created at random. The distributions for the number of households per sampling unit are extracted from the GMC 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 sampling unit and the household number. Next the number of households within each sampling unit is counted for each regional class and house size class combination. The absolute numbers are divided by the number of sampling units within the respective regional class and house size class combination. The resulting $1,070$ relative frequency distributions are taken as probability distributions for the number of households per sampling unit in the $1,070$ strata. On the basis of this distributions, the number of households in each sampling unit within the $1,070$ classes are simulated at random by using the alias method.

	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

Table 2: Partition of the federal states into regional classes and number of classes (noc) within each federal state.

The next step is the simulation of the number of persons in the households. Therefore the number of persons per household in the GMC data is counted and the resulting frequency distribution is taken as a 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.

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 GMC 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.

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

Table 3: Number of households and persons within the federal states of the German pseudo universe.

To design the remaining variables of interest conditional distributions are calculated. First of all, the multivariate realisations 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 realisation in the GMC 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.

An overview to the data simulation routine can be drawn from figure 3.

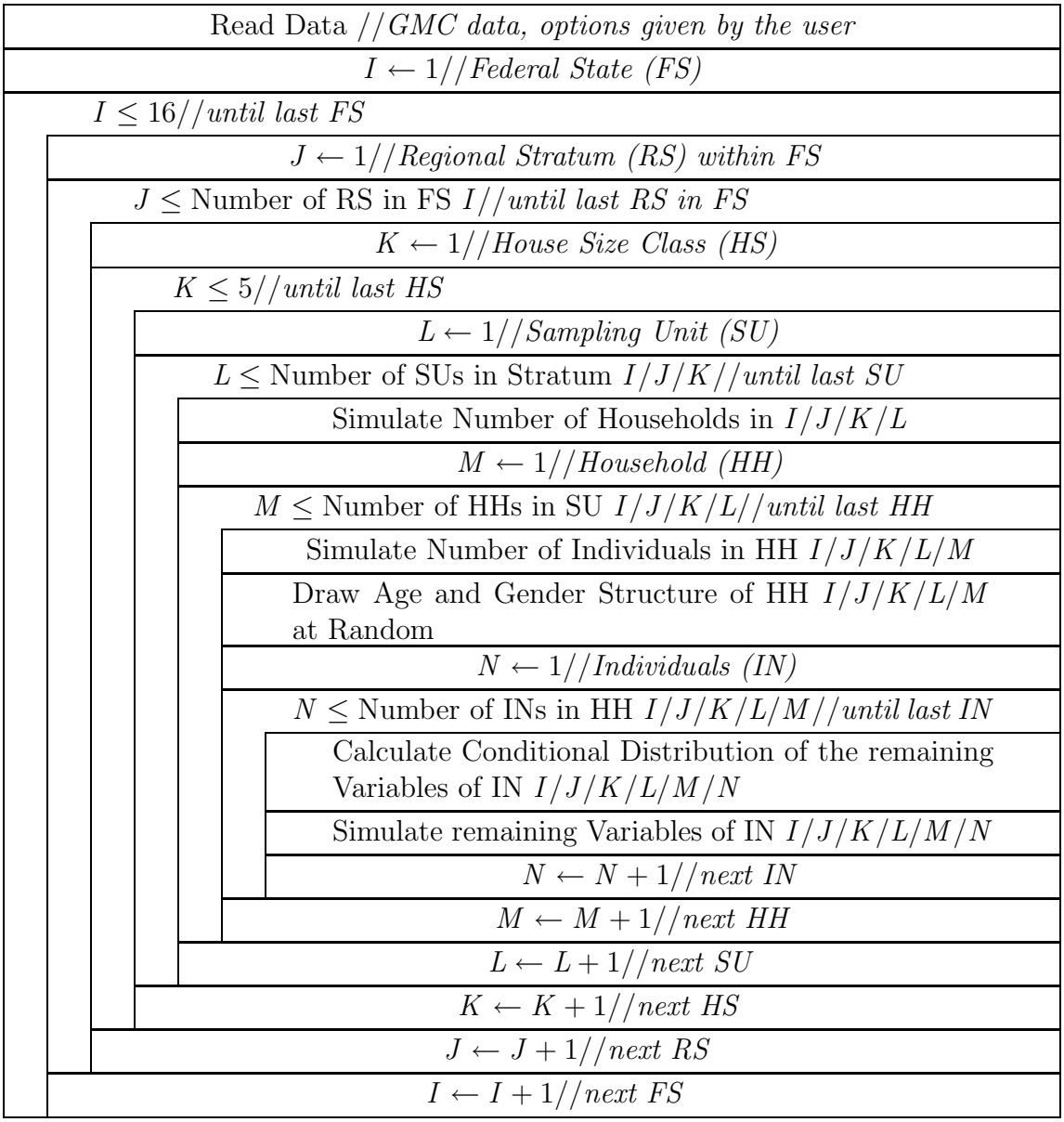

Figure 3: Overview to the data generation procedure in the German Microcensus

4 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 are subsequently aggregated to the whole universe. Therefore it is of main interest if the realised 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 GMC data.

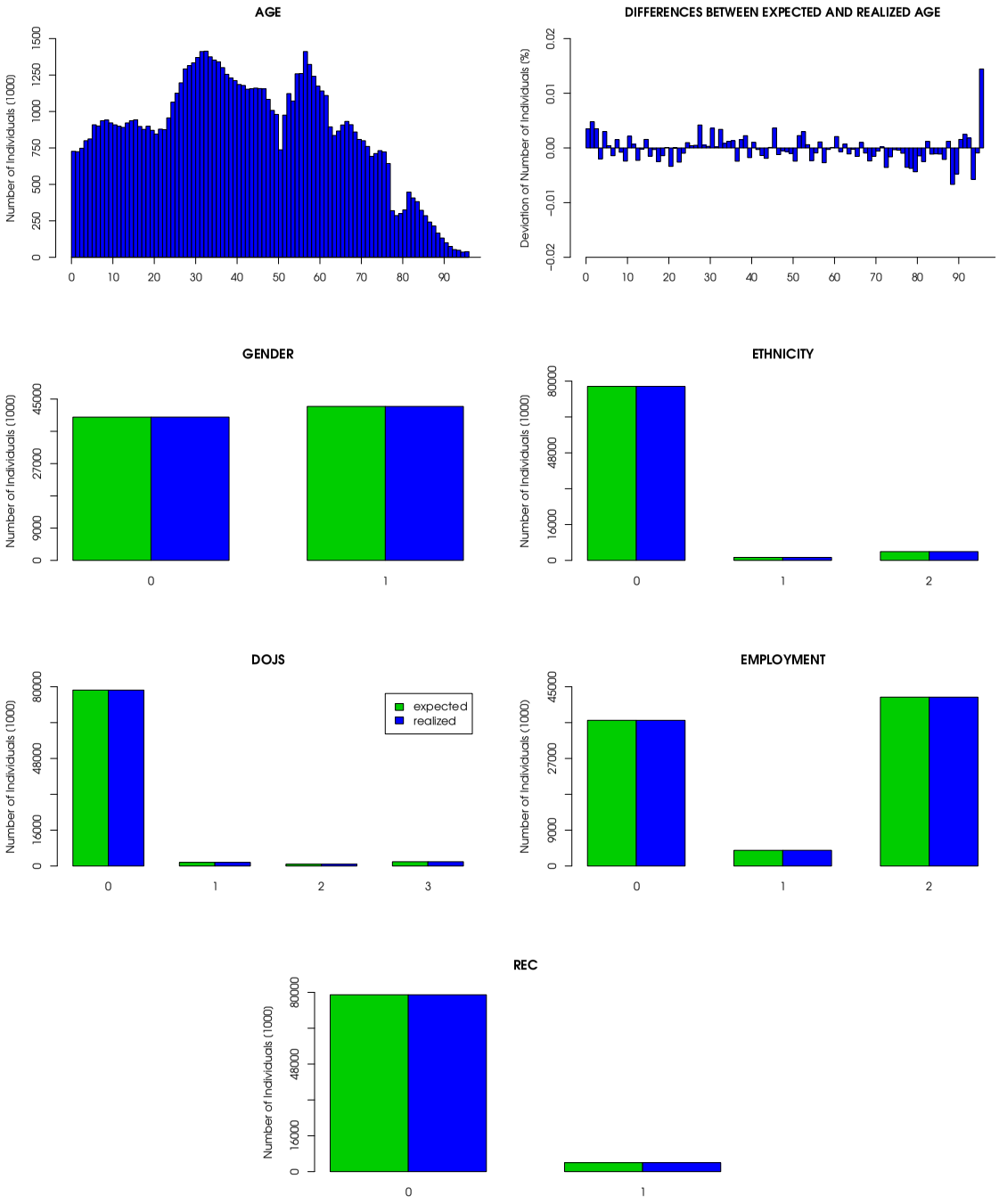

Figure 4: Marginal frequency distributions within the German pseudo universe.

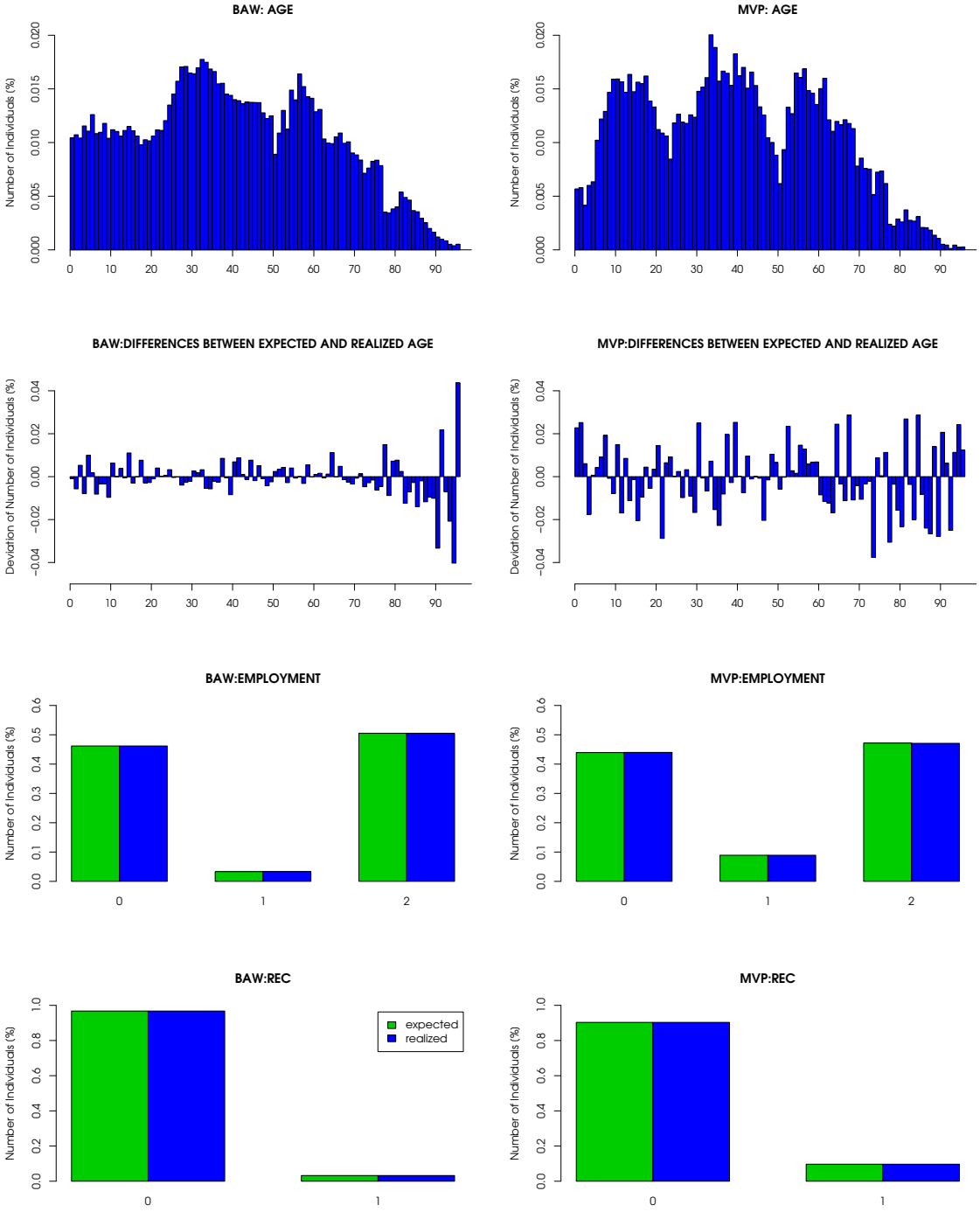

Figure 5: Marginal frequency distributions within the pseudo universe federal states BAW and MVP

First of all, the global figures are treated. In Table 3, the number of households and person within the pseudo universe are displayed. In total, 82,914,752 people are created. Within Figure 4, the expected and realised marginal distributions of the 6 variables of interest are presented. It can be seen that the numbers expected on the basis of the GMC data almost equal the numbers realised 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.

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

Table 4: p-values when applying a χ^2 goodness of fit test to the expected and realised GMC marginal distributions.

The two-dimensional correlations within the pseudo universe equal the respective ones in the GMC data. This is shown by the contingency coefficients displayed in Table 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 GMC data used for simulation.

source		gender	ethn.	dojs	empl.	rec
GMC 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 5: Contingency coefficients within the German pseudo universe.

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 which effects they have. In Figure 5 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 resulting are similar to those in the GMC data is shown by the p-values when the expected and realised marginal distributions are tested with a χ^2 goodness of fit test (Table 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.

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

Table 6: p-values when applying a χ^2 goodness of fit test to the expected and realised marginal frequency distributions within BAW and MVP.

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 7 also show that the correlations reflect the values within the GMC data. Hence, the heterogeneity within the German pseudo universe is a result of the heterogeneity within the GMC data.

data source		gender	ethnicity	dojs	empl.	rec
BAW GMC 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 GMC 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

Table 7: Contingency coefficients within the German federal states BAW and MVP.

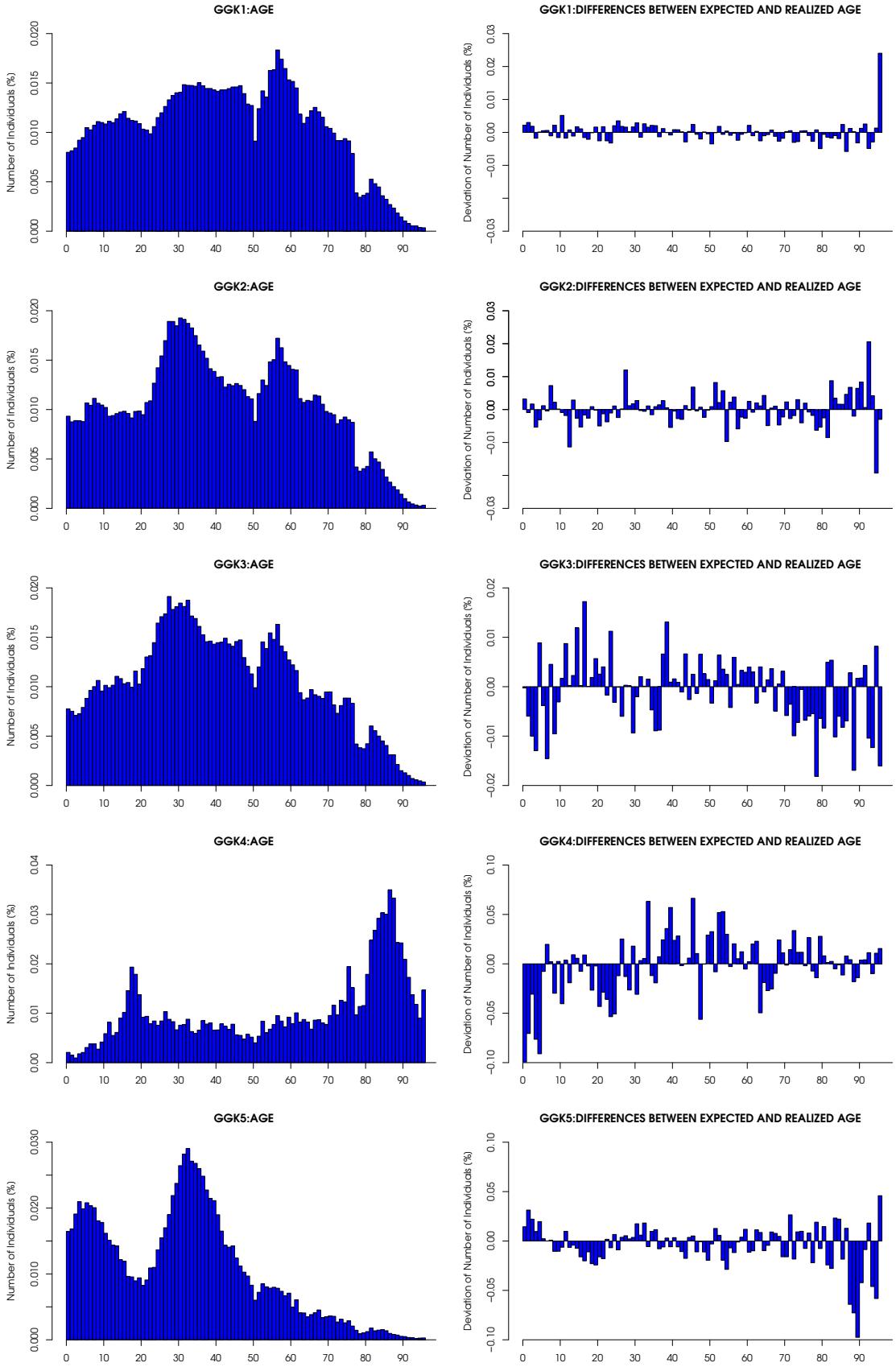

Figure 6: Realised and differences between the expected and the realised marginal age frequency distributions within the pseudo universe house size classes 1 to 5.

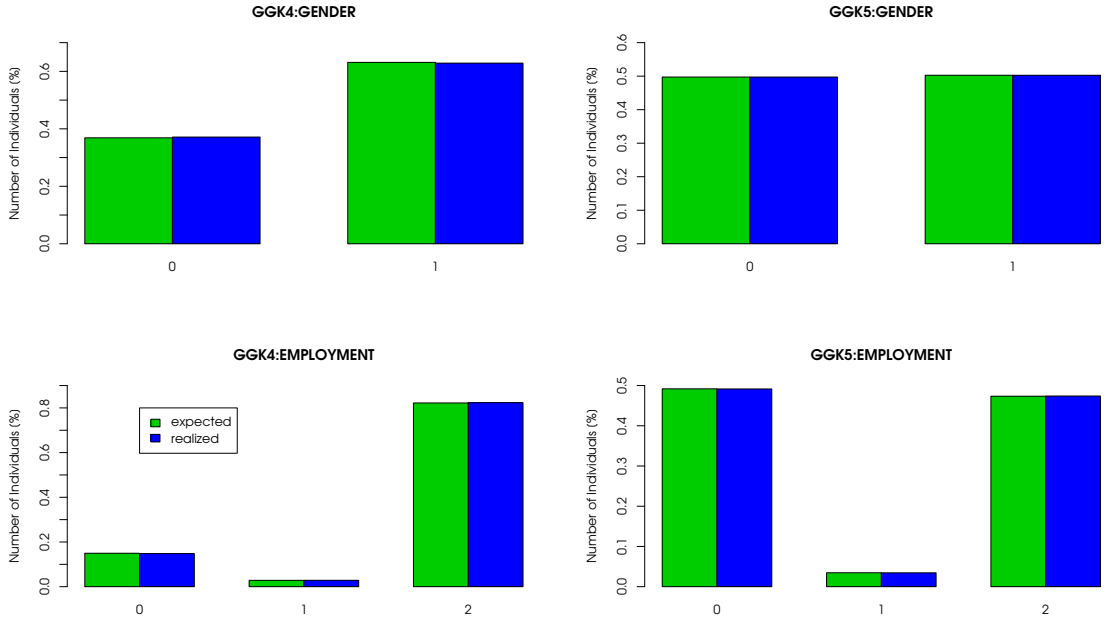

Figure 7: Marginal gender and employment frequency distributions within the pseudo universe house size classes 4 and 5.

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

Table 8: Number of households and persons within the five house size classes of the German pseudo universe.

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 8. In Figure 6, 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 7. The completely different age structures lead to significantly different gender and employment realisations. Within

data source		gender	ethnicity	dojs	empl.	rec
class 4 GMC 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 4 relative differences	age	0.04%	0.61%	0.53%	-0.38%	-1.07%
	gender	-	0.45%	-2.46%	-0.59%	-1.19%
	ethnicity	-	-	-5.23%	-2.93%	-2.92%
	dojs	-	-	-	-0.05%	-0.35%
	empl.	-	-	-	-	-0.16%
class 5 GMC 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
class 5 relative differences	age	0.02%	0.88%	-0.39%	-0.03%	-0.5%
	gender	-	2.94%	-3.13%	-0.07%	1.68%
	ethnicity	-	-	-0.96%	-1.26%	-0.51%
	dojs	-	-	-	0.02%	0%
	empl.	-	-	-	-	0.04%

Table 9: Contingency coefficients within the German house size classes 4 and 5.

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 9. Especially the two-dimensional correlations between age and the other 5 variables differ significantly.

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 GMC 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.

5 Summary

One of the central aims of the DACSEIS research project is to elaborate *best* practice recommendations on the use of adequate variance estimation methods in the context of measuring the accuracy of estimates. The accuracy is one important aspect of data quality (cf. Eurostat report from GRÜNEWALD and LINDEN, 2001, and the references therein). Additionally to the accuracy of the estimates, the influence of non-response in the surveys as well as the techniques to correct for the non-response, i. e. weighting and imputation, on the estimators and their efficiency is of special interest. To achieve this aim, especially in the context of *real* data, adequate universes to allow for practically oriented simulation studies have to be available. Since only a few sources of real universes are available from which in practice generally no samples are drawn, synthetic universes have to be generated in order to have best possible data as a basis for a large simulation study. These universes must be best possible in the sense of adequately showing the joint distributions with regards to microdata use while taking into consideration disclosure control rules.

Within the simulation study *true* – in the sense of size and design – samples are drawn repeatedly in order to gain the distributions of estimators and variance estimators as well as respective measures to enable a comparative study. The results will be summarised in the recommended practice manual (for an overview to the DACSEIS project see MÜNNICH and WIEGERT, 2001).

The aim of this paper was to elaborate the generation mechanism of the universes for the DACSEIS study considering the different aspects according to the introduction. As a prototype, the data of the German microcensus were applied. This generalised mechanism was applied to all labour force oriented survey data within the DACSEIS project, i. e. the Dutch and Finnish labour force survey as well as the Austrian and German Microcensus. Within these surveys only categorical with a finite number of outcomes have to be considered. Some modifications with respect to continuous data in the Swiss household and budget survey and the German income and expenditure survey had to be made by estimating parameters of adequate continuous distributions. Details will be available in the final report of workpackage 3 of the DACSEIS project.

The simulation set-up based on the before described universe generation mechanism at microdata level now easily allows to include further investigations, e. g. the inclusion of non-response in the data or frame imperfections. These aspects will be taken into consideration within the DACSEIS simulation study.

Acknowledgement

The authors want to thank all of the members of the DACSEIS team for their valuable comments. Further, we would like to thank the evaluators for their intensive discussions during evaluations, especially with respect to an adequate simulation of nonresponse. Special thanks go to Wolf Bihler, DESTATIS, for his support during the implementation of the German Microcensus.

References

Devroye, Luc (1986): Non-Uniform Random Variate Generation, New York et al.: Springer.

Esser, H.; Müller W.; Schäffer K.-A., H.; Grohmann (1989): Mikrozensus im Wandel: Untersuchungen und Empfehlungen zur inhaltlichen und methodischen Gestaltung, *Forum der Bundesstatistik*, 11, Metzler-Poeschel: Stuttgart.

Heidenreich, H.-J. (1994): Hochrechnung im Mikrozensus ab 1990, Gewichtung in der Umfragepraxis, Westdeutscher Verlag: Opladen.

Heidenreich, Hans-Joachim (2002): Längsschnittdaten aus dem Mikrozensus: Basis für neue Analysemöglichkeiten. *In:* Allgemeines Statistisches Archiv, 86 (2), 213–231.

Johnson, Mark E. (1987): Multivariate Statistical Simulation, New York et al.: John Wiley & Sons.

Kronmal, Richard A.; **Peterson Jr., Arthur V.** (1979): On the Alias Method for Generating Random Variables From a Discrete Distribution. *In:* The American Statistician, 33, 214–218.

Grünewald, Werner; Linden, Håkan (2001): Quality Measurement - Eurostat Experiences. In: Statistics Canada (Ed.), Proceedings of Statistics Canada Symposium 2001.

Meyer, K. (1994): Hochrechnung im Mikrozensus ab 1990, Gewichtung in der Umfragepraxis, Westdeutscher Verlag: Opladen.

Münnich, Ralf; **Wiegert, Rolf** (2001): The DACSEIS Project, DACSEIS research paper series 1.
`http://w210.ub.uni-tuebingen.de/dbt/volltexte/2001/428`

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

Quatember, Andreas (2002): A comparison of the five Labour Force Surveys of the DACSEIS project from a sampling point of view. DACSEIS research paper series 3.
`http://w210.ub.uni-tuebingen.de/dbt/volltexte/2002/547`

Statistisches Bundesamt (StBA; Ed.) (1999): Fachserie 1: Bevölkerung und Erwerbstätigkeit, Band Reihe 4.1.1 von *Stand und Entwicklung der Erwerbstätigkeit.*

Zühlke, Sylvia (2003): Systematische Ausfälle im Mikrozensus-Panel: Ausmaß und Auswirkungen auf die Qualität von Arbeitsmarktanalysen. *In:* Allgemeines Statistisches Archiv, 87 (1), 39–58.

The following papers are already published in the

DACSEIS research paper series

No.1 **Münnich, Ralf; Wiegert, Rolf (2001)**

The DACSEIS Project

http://w210.ub.uni-tuebingen.de/dbt/volltexte/2001/428

No.2 **Zhang, Li-Chun (2002)**

A method of weighting adjustment for survey data subject to nonignorable nonresponse

http://w210.ub.uni-tuebingen.de/dbt/volltexte/2002/451

No.3 **Quatember, Andreas (2002)**

A comparison of the five Labour Force Surveys of the DACSEIS project from a sampling theory point of view

http://w210.ub.uni-tuebingen.de/dbt/volltexte/2002/547