AUSTRIAN JOURNAL OF STATISTICS
Volume 35 (2006), Number 4, 419–435

Simulation Studies for Complex Sampling Designs

Helga Wagner[1] and Doris Eckmair[2]
[1]Johannes Kepler Universität Linz
[2]Allgemeine Sparkasse OÖ Bank AG

Abstract: Choosing the appropriate variance estimation method in complex surveys is a difficult task since there exist a variety of techniques which usually cannot be compared mathematically. A relatively easy way to accomplish such a comparison is on the basis of simulation studies. Though simulation studies are widely used in statistics, they are not a standard tool for investigating properties of estimators in complex survey sampling designs. In this paper we describe the setup for a simulation study according to the sampling plan of the Austrian Microcensus (AMC), used 1994–2003 which is an example for a very complex sampling plan. To illustrate the proceeding we conducted a simulation study comparing basic variance estimators. Results of the study reveal the extent to which simple variance estimators may underestimate the true sampling error in close to reality situations.

Zusammenfassung: Die Wahl geeigneter Methoden zur Varianzschätzung in Erhebungen mit komplexen Stichprobenplänen ist schwierig, da es eine Reihe verschiedener Verfahren gibt, die i.a. nicht theoretisch vergleichbar sind. Relativ einfach kann jedoch ein derartiger Vergleich auf Basis von Simulationsstudien durchgeführt werden. Diese werden zwar in der Statistik häufig eingesetzt, für komplexe Stichprobenerhebungen war das jedoch bislang noch nicht der Fall. In diesem Artikel beschreiben wir ein Setup für Simulationsstudien am Beispiel des überaus komplexen Stichprobenplanes, der für den österreichischen Mikrozensus 1994–2003 verwendet wurde. Zur Illustration des Vorgehens dient eine Simulationstudie, in der einfache Varianzschätzer verglichen werden. Ihre Ergebnisse zeigen, in welchem Ausmaß einfache Varianzschätzer den wahren Stichprobenfehler in der Realität nahekommenden Situationen unterschätzen können.

Keywords: Generation of Universes, Sampling Design, Variance Estimation.

1 Introduction

Variance estimation for complex surveys is a challenging problem. Although a variety of techniques for variance estimation exists – see e.g. Wolter (1985) – theoretical comparisons of the properties of different estimators are at most feasible for rather simple sampling designs.

DACSEIS (= Data quality in complex Surveys within the New European Information Society) was a project within the IST program of the European Commission which investigates variance estimation methods for complex surveys. One of its main tasks is the realization of simulation studies to compare different variance estimation techniques for several national surveys (the outline of the project is given in Münnich and Wiegert, 2001, the investigated surveys are described in Quatember, 2002).

A basic prerequisite for simulation studies are adequate universes from which samples according to a specific sampling plan can be drawn repeatedly. As data of the relevant national universes – i.e. census data – are in general not available for simulation studies, pseudo universes have to be constructed from survey data. These pseudo universes should allow sampling according to the sampling plan of interest, be close to the respective national universe regarding distributions of interesting variables and not violate disclosure control rules, see Münnich and Schürle (2003). To meet these requirements the structure of the universe according to the sampling plan has to be rebuilt, sizes of strata and clusters should be correct and homogeneity within respectively heterogeneity between strata and clusters should be replicated in the pseudo universes. To avoid possible identification of individuals the generation process has to be at least partly stochastic.

Once generated, a pseudo universe can easily be modified to study different aspects of the sampling scheme, for instance the effect of a different sampling frame or of a particular non-response mechanism. Samples from a pseudo universe can provide all estimates of interest and their simulation distribution gives detailed insight in their performance.

As the DACSEIS project started in 2001, for Austria the sampling plan of the AMC, which was used until 2003, was investigated. It turned out to be the most complex sampling plan studied within the project and thus can serve as an exemplar for a complex sampling design. Since 2004 the AMC is carried out according to a different, simpler sampling plan.

In this paper we describe the generation of a pseudo universe, appropriate for the former AMC sampling plan. Due to the complexity of the sampling plan a restriction to its basic properties was necessary. These are described in Section 2. Section 3 deals with the generation of the pseudo universe and in Section 4 the implementation of the sampling procedure is described. The simulation study presented in Section 5 illustrates the application of the method for several basic variance estimators and modifications of the pseudo universe, i.e. a different sampling frame and a certain non-response-mechanism. The summary given in Section 6, concludes the paper.

2 The Sampling Design

The Austrian Microcensus is a quarterly survey of Austrian households and is conducted by interviewers since 1967. It is intended to provide information on the structure of the Austrian population, families, households and dwellings. The questionnaire contains a mandatory core program and a voluntary supplementary program. Until 2003 1%, at present 0.68% of the Austrian households are selected.

The sampling frame for AMC used 1994-2003, was the Austrian Housing Census (HWZ = Häuser- und Wohnungszählung), performed with a period of 10 years. Sampling units were dwellings. These were selected for all AMC surveys to be conducted in the following 10 years. Quarterly one eighth of the sampling units was replaced, thus limiting the participation of sampling units to a maximum of 2 years in row. For each sample dwelling, characteristics of all households and persons living therein were recorded.

The sample design of the former AMC consisted of two parts. The first, in the following called Part A, comprised mainly dwellings in larger urban municipalities, the other,

called Part B, dwellings in small, rural communities. Sampling was carried out separately for each of the nine federal states in these two parts, except for two federal states (Wien and Vorarlberg), which consisted only of Part A dwellings.

In Part A dwellings were selected as a stratified random sample where strata were built according to several dwelling characteristics, such as kind of dwelling, period of construction, floor space etc. As a combination of all strata variables would result in very small strata, these were pooled to give sample sizes of at least 10 dwellings per stratum, resulting in 100 – 150 strata per federal state. The sampling fraction was different for each of the nine federal states.

In Part B a two-stage sampling procedure with stratified random sampling of primary sampling units (PSUs) was carried out. PSUs are communities or – in case of very small communities – groups of communities. For PSUs their number of dwellings and the agrarian quota were used as stratification variables, with different values defining the strata in each federal state. Number of strata per federal state ranged from 5 to 16.

Within each stratum the sample size was allocated proportional to size (i.e. number of dwellings) of districts. PSUs were selected randomly within each stratum. On the second stage a specified number of dwellings were drawn from the selected PSUs as secondary sampling units (SSUs). Depending on the federal state the number of SSUs was 20 or 25.

Selection of dwellings was carried out systematically according to a list sequential selection with a fixed starting value within each federal state of Part A as well as selected PSUs of part B.

Figure 1 illustrates the hierarchical structure of the universe according to the former AMC sampling plan. A more detailed description of this sampling plan can be found in Haslinger (1996).

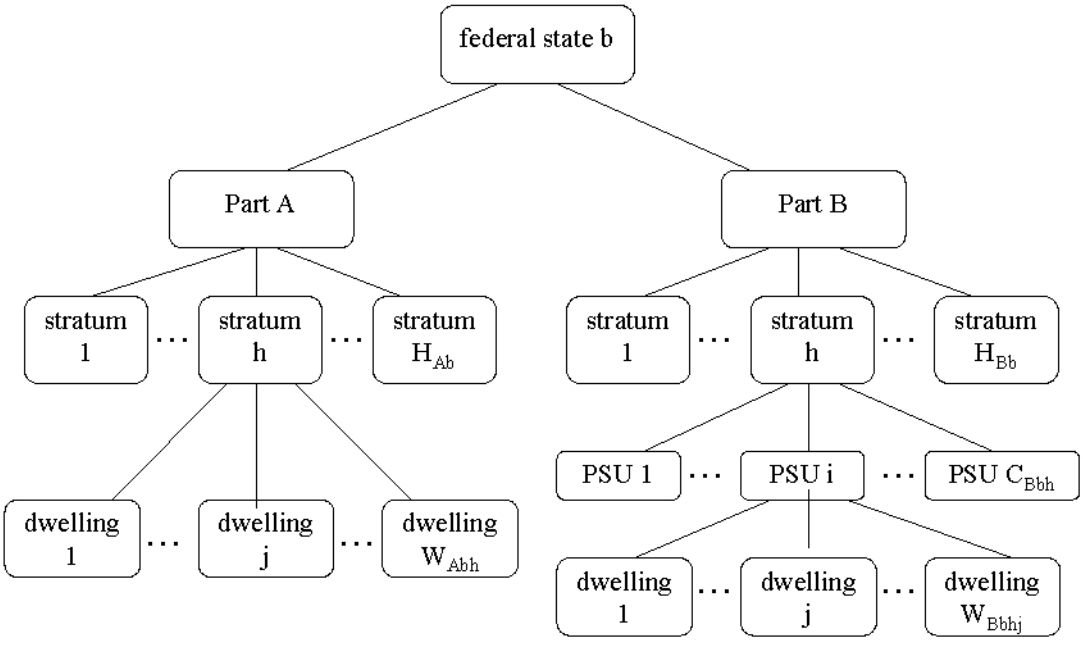

Figure 1: Structure of the universe

3 The Generation of the AMC Pseudo Universe

The pseudo universe for simulation studies according to the former AMC sampling plan, which we refer to as AMC pseudo universe, was generated following the general process for generation of pseudo universes developed within the DACSEIS project. This process was applied to build pseudo universes for different labour force and Microcensus surveys and is described in detail in Münnich and Schürle (2003). Principles of this generation process are exemplified for the AMC pseudo universe in Section 3.1, more technical details of its generation are given in Section 3.2.

3.1 Basic Principles for the Generation of the AMC Pseudo Universe

To generate a pseudo universe for simulation studies various aspects have to be regarded. For the AMC pseudo universe e.g., these are:

- The pseudo universe should have the same structure as the real Austrian universe in all relevant aspects of the sampling plan, i.e. reflect the hierarchical structure defined by federal states, strata, PSUs, dwellings, households and persons.
- The generated pseudo universe should be close to reality regarding the distribution of interesting variables. Especially all features which have an effect on the variance of estimators have to be regarded. Thus the generation of the AMC pseudo universe should reflect homogeneity or heterogeneity within respectively between strata in Part A as well as PSUs in Part B.
- As the intended simulation studies are CPU-time as well as storage consuming the pseudo universe should be as small as possible regarding the number of variables. A restriction to only a few variables impedes the identification of individuals and is thus advantageous also with a view to disclosure control. So apart from structure variables of the sampling plan only five personal characteristics were generated for pseudo individuals.

One of the main principles of the generation process in the DACSEIS project is to rebuild the structure of the universe concerning strata and clusters. This part of the generation process is deterministic as information on numbers and sizes of strata and clusters is available from the sampling plan. The pseudo universe thus has the same structure as the real universe from the viewpoint of the sampling plan.

The generation of sampling units – these are dwellings or households in most surveys – is carried out stochastically. A main problem in this context is to find a compromise between neglecting and maintaining correlation structures within sampling units (see also Münnich and Schürle, 2003). Creating individuals independently could lead to unrealistic results, e.g. a household consisting of children only, whereas taking into account all correlations would amount to sampling from high dimensional densities. Therefore to simplify sampling, age and gender structure are drawn from the data, i.e. from real households or dwellings, and values of the remaining variables are generated independently for each individuum conditional on age and gender. To obtain a close to reality situation empirical distributions from the data are used as generation distributions. Different generation distributions are used in strata and clusters. All dwellings in one stratum respectively cluster

share the same generation distribution – they are referred to as generation groups in the following. This proceeding accounts for

- homogeneity within groups by using the empirical distribution from this group
- heterogeneity between groups by using different distributions.

Generation of the AMC pseudo universe was deterministic concerning the structure down to the hierarchical level of sampling units as illustrated in Figure 1 and stochastic for sampling units, i.e. dwellings in the AMC. The hierarchical structure within dwellings is illustrated in Figure 2.

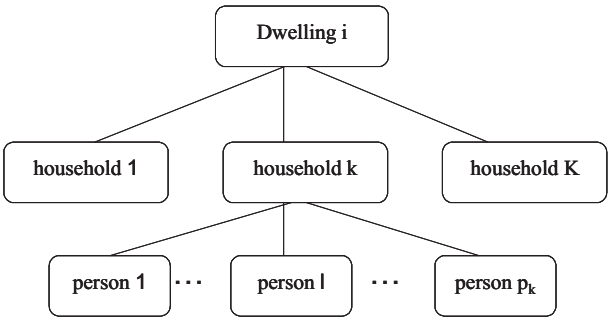

Figure 2: Hierarchical structure within dwellings

Deterministic Part of the Generation Process: Concerning the deterministic part of the generation process, the pseudo universe has the same structure as the universe shown in Figure 1, i.e. it is partitioned into federal states, within federal states in Parts A and B, and within each part into strata according to the sampling plan. Strata of Part B were additionally partitioned into PSUs. Table 1 gives the number of strata per part and federal state. The total number of PSUs within Part B is 1710, leading to a total of 2899 generation groups. The sizes of the generation groups, i.e. the number of dwellings, are regarded as deterministic.

Table 1: Partition of federal states of the AMC pseudo universe into parts and strata

		Number of Strata	
Number	Federal State	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	–
	Total	1189	70

Table 2: Personal characteristics included in the AMC pseudo universe

Variables	Possible Outcomes	
x_1: age	0–99	age in years
x_2: gender	0	male
	1	female
x_3: nationality	0	Austria
	1	former Yugoslavia
	2	Turkey
	3	other
x_4: employment	0	employed at least one hour
	1	not employed
	2	not relevant/unknown
x_5: educational level	0	not completed compulsory school
	1	completed compulsory education
	2	completed apprenticeship
	3	medium secondary level
	4	secondary academic school
	5	upper secondary level school
	6	post secondary school
	7	tertiary level school, not university
	8	university
	9	child of school age

Stochastic Part of the Generation Process: Conditional on the size, dwellings, households, persons, and personal characteristics are generated stochastically. The stochastic part of the generation process is identical within a generation group and different between generation groups, as empirical distributions from AMC data (of a Part A stratum or a Part B PSU) serve as generation distributions.

For each dwelling, values for the following variables were created:

K	number of households
p_k	number of persons in household k
x_{ilk}	personal characteristic i of person l in household k, $i = 1, \ldots, 5$; $l = 1, \ldots, p_k$ (Personal characteristics and possible outcomes are displayed in Table 2)

To describe the stochastic part of the generation process more formally let P_y denote the empirical distribution of y and P_y^x the conditional empirical distribution of y given x within a generation group. The hierarchical structure within a dwelling is generated to the following model

$$\begin{aligned} K &\sim P_K \\ p_k &\sim P_p\,, \qquad k = 1, \ldots, K\,. \end{aligned}$$

The personal characteristics are generated for each household separately. Let p denote the number of persons – we drop the index k from now on – in a given household, then first,

for all persons in this household values for variables x_1 and x_2, i.e. age and gender are generated in one step as

$$(x_{11}, \ldots, x_{1p}, x_{21}, \ldots, x_{2p}) \sim P^p_{(x_{11},\ldots,x_{1p},x_{21},\ldots,x_{2p})}$$

Here $P^p_{(x_{11},\ldots,x_{1p},x_{21},\ldots,x_{2p})}$ denotes the joint empirical distribution of x_1 and x_2 for all persons, determined from all households consisting of p persons in the generation group. For a 3 person household, e.g. values for age and gender are generated as random numbers from the appropriate 6-dimensional empirical distribution of all households of 3 persons.

The remaining personal characteristics x_3, x_4, x_5, i.e. nationality, employment and educational level are generated independently for each person from their joint distribution in the generation group, given the values of x_1 and x_2. As generation groups are rather small, in many cases only *one* person with a given age and gender would exist in the AMC data. 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 all individuals of one household are the only person with a specific age of their gender in the generation group – to a replication of entire households. To reduce the extent of replications, a modified variable x_1^*, i.e. age measured in 5 year-categories, was used for the construction of conditional distributions. Thus for each person values of x_3, x_4, x_5 are actually generated as random numbers

$$(x_{3l}, x_{4l}, x_{5l}) \sim P^{(x_1^*,x_2)}_{(x_3,x_4,x_5)}\,, \qquad l = 1, \ldots, p\,.$$

3.2 The AMC Pseudo Universe

Generation from AMC Data: 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 2.

In the generation process sizes of strata and PSUs, i.e. the number of dwellings W in a stratum of Part A or PSU of Part B and the number of PSUs in a stratum of part B were considered deterministic. Whereas the latter remains constant and therefore is known from the sampling plan, the number of dwellings is subject to change in the course of time and had to be estimated.

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

The actual number for both types of dwellings in each generation group represented in the AMC, was estimated by multiplying the number W of dwellings of each type in the HWZ with the change ratio w_{act}/w_0, that is their number in the actual AMC data w_{act} divided by the number of the first AMC sample w_0.

For each virtual dwelling serving as permanent residence the number of households, the number of persons per household, and values for the personal characteristics of individuals were generated as random numbers according to the model described before, for other dwellings the number of households and persons per household were set to zero.

Generation distributions according to the general model were built separately from this data set for each of the 1557 generation groups represented therein, i.e., for each of 1189 Part A strata and each of 368 sample PSUs of Part B. 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.

Generation of dwellings in Part B therefore needed some modification as not every PSU is represented in the AMC. To generate a specific PSU of the pseudo universe therefore a sample PSU of the AMC in 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 change ratios w_{act}/w_0 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 PSUs in the pseudo universe share 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.

Comparing Pseudo Universe and AMC: Generation of the AMC pseudo universe was performed for small groups which are aggregated to form the total pseudo universe. It is therefore of interest whether the structure in the pseudo universe corresponds to that in the AMC data. Figure 3 shows marginal distributions of the generated personal characteristics age, gender and educational level, for age also relative differences are given. 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. 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 3. Absolute differences are small, relative differences are large only for small contingency coefficients. Therefore it can be concluded that the global structure of the AMC data is reproduced well in the pseudo universe.

A more detailed insight into the structure of the pseudo universe is given in Figure 4 showing the marginal distributions for the variables age, gender, and educational level

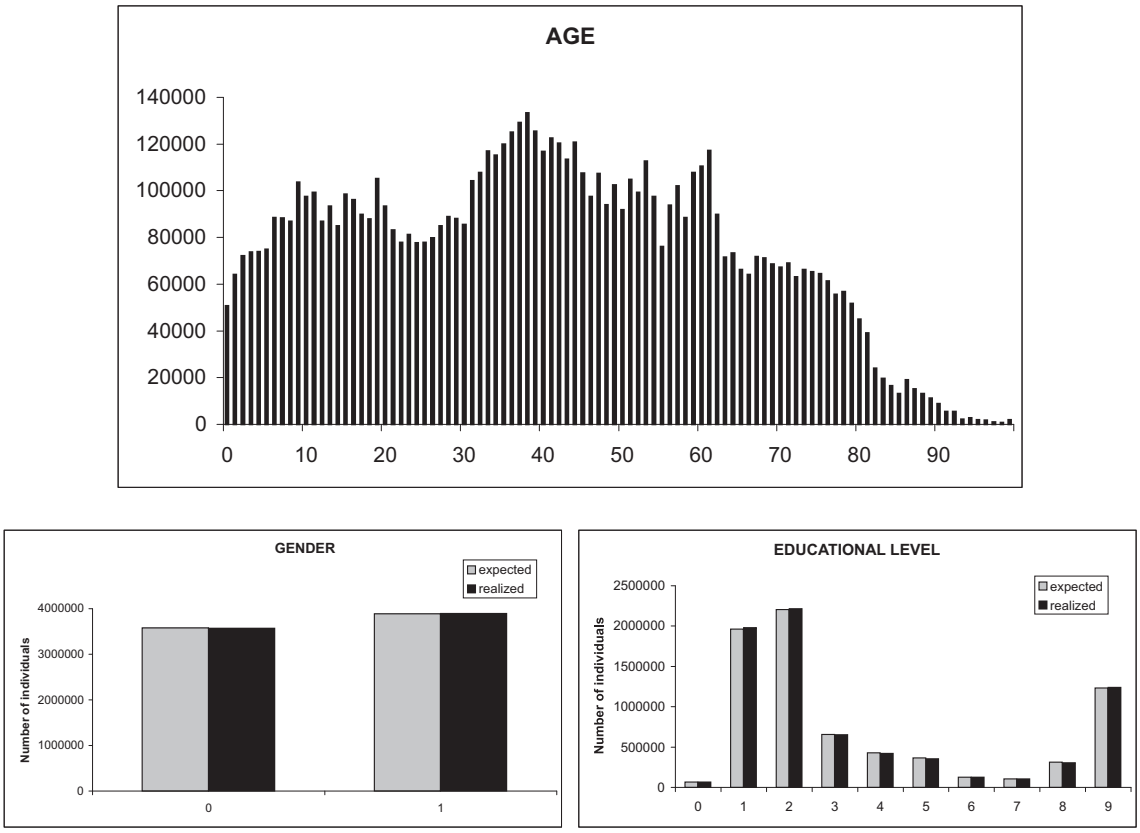

Figure 3: Marginal frequency distributions within the AMC pseudo universe and the data

Table 3: 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%

for federal states Tirol and Wien. Differences between the 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 expected frequencies from the AMC data again are of small order.

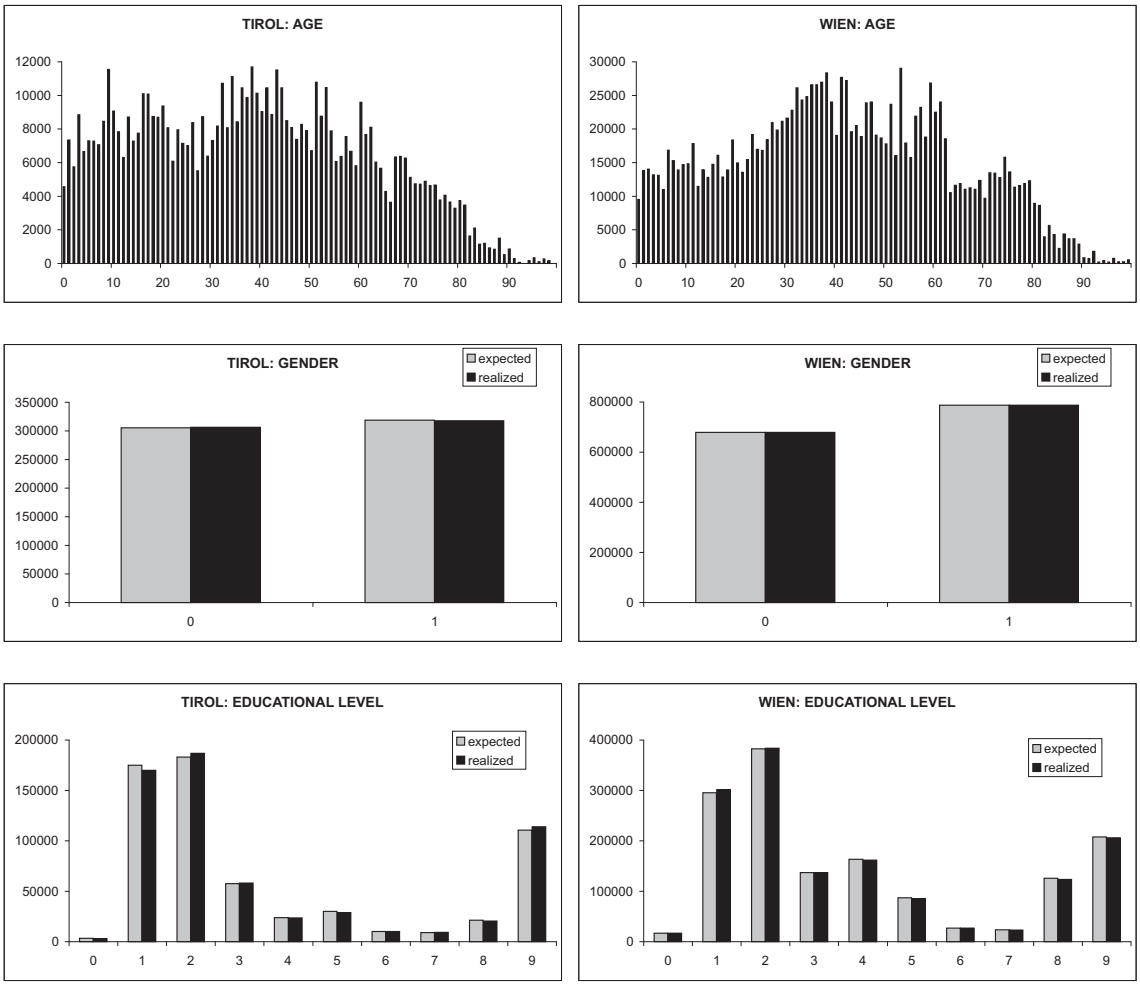

Figure 4: Marginal frequency distributions within federal states TIR and WIE of the AMC pseudo universe

Further results are presented for federal states and for Part A and B in Münnich and Schürle (2003). They 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.

4 The Sampling Procedure

The sampling procedure for the simulation study imitates that of the AMC but is not exactly identical to it, compare Quatember (2002). For the AMC the proportional stratified sampling of Part A dwellings is realized by a systematic selection. Dwellings are ordered sequentially 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.

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 in the simulation studies.

5 A Simulation Study

Carrying out simulation studies to gain insight in the properties of estimators is straightforward once the pseudo universe is generated and the sampling procedure is implemented. We demonstrate this in an exemplary simulation study comparing four direct variance estimators. The whole simulation process, i.e. drawing an AMC sample from the pseudo universe and calculating estimates was repeated 10000 times. The simulation studies were carried out on a Pentium P3 using C++ programs written by the second author.

Useful criteria for comparing variance estimators are bias, mean square error and – as one of the main purposes of variance estimation is to get at least approximate confidence intervals for parameters of the universe – the coverage of confidence intervals.

5.1 Variance Estimation of Totals

Estimation of Totals: An interesting total $\tau = \sum_{k \in U} y_k$ of a universe U is usually estimated by the Horvitz-Thompson estimator

$$\hat{\tau} = \sum_{k \in S} y_k \frac{1}{\pi_k},$$

where π_k is the inclusion probability of unit k into the sample S. Published total estimates for the AMC differ from the Horvitz-Thompson estimator as the weights used differ slightly from the inverse inclusion probabilities and additionally non-response is accounted for.

In the following let A and B denote part A and B, b the federal state, h the stratum and i the PSU, C and c the number of PSUs in the universe, respectively the sample, and W and w the number of dwellings in the universe, respectively the sample. Totals of personal characteristics of the Austrian population are estimated from the AMC data by combining

total estimates $\hat{\tau}_{Abh}$ for strata in part A and $\hat{\tau}_{Bbhi}$ for PSUs in Part B as

$$\hat{\tau} = \sum_b \sum_h \hat{\tau}_{Abh} + \sum_b \sum_h g_{Bbh} \sum_{i=1}^{c_{Bbh}} \hat{\tau}_{Bbhi} \,.$$

The inflation factor g_{Bbh} is defined as

$$g_{Bbh} = \frac{\sum_{i=1}^{C_{Bbh}} W_{Bbhi}}{\sum_{i=1}^{c_{Bbh}} W_{Bbhi}} \,,$$

its inverse is the proportion of dwellings in sample PSUs of all dwellings in a stratum of Part B. It differs from the weight of the Horvitz-Thompson estimator as the inclusion probability of PSUs is c/C.

Totals of strata in Part A estimated as weighted sample totals T, i.e.

$$\hat{\tau}_{Abh} = \frac{W_{Abh}}{w_{Abh}} \frac{w^{(1)}_{Abh} + w^{(2)}_{Abh}}{w^{(1)}_{Abh}} \cdot T_{Abh} \,.$$

Here, $w = w^{(1)} + w^{(2)} + w^{(3)}$ is the number of dwellings in the sample, where $w^{(1)}$ is the number of interviewed, $w^{(2)}$ the number of non-responding and $w^{(3)}$ the number of non-inhabited dwellings. The inverse of the inflation factor for T_{Abh} is the sampling fraction of dwellings multiplied with the proportion of responding dwellings.

Estimation of totals in PSUs of Part B is analogous with indices $Bbhi$ instead of Abh.

In the simulation study three different totals (τ_e = number of persons with employment status = employed at least one hour, τ_u = number of persons with educational level = university, N = population size) were estimated using this estimator. Results summarized in Table 4 indicate that estimates of all three totals are rather close to their respective true values in the pseudo universe.

Table 4: Simulation results for the estimation of totals

	True Value	Mean Estimate	Bias	MSE
τ_e	3138666	3139059.61	393.61	540697607
τ_u	304581	304589.37	8.37	47414511
N	7462802	7462737.51	–64.48	1382383998

Different Variance Estimators: Variance estimation is a difficult task as the variance of an estimator depends on the variation of the characteristic of interest in the universe as well as on the sampling design. Appropriate variance estimators taking into account the specifics of a sampling plan have to be derived for each sampling plan individually. Thus in practical applications often simple variance estimators are used.

In our simulation we compare the performance of four different direct variance estimators. $\hat{V}_1$, $\hat{V}_2$, and $\hat{V}_3$ are simple variance estimators mostly neglecting the complex sampling design of the AMC, whereas the complex variance estimator $\hat{V}_4$ takes into account the effects of stratification and clustering. For the following definitions of these variance estimators let N and N_b denote the number of individuals in the universe respectively federal state b and n and n_b their number in the sample.

- $\hat{V}_1$ is the appropriate variance estimator under simple random sampling without replacement, i.e.
$$\hat{V}_1(\hat{\tau}) = \frac{(N-n)(N-\hat{\tau})\hat{\tau}}{Nn}\,.$$
- Taking into account the different sampling fractions per federal states, but still assuming simple random sampling leads to the variance estimator
$$\hat{V}_2(\hat{\tau}) = \sum_b \frac{(N_b-n_b)(N_b-\hat{\tau}_b)\hat{\tau}_b}{N_b n_b}\,.$$
- Assuming $\hat{\tau}_b/N_b = \hat{\tau}/N$ gives the variance estimator
$$\hat{V}_3(\hat{\tau}) = \sum_b \frac{(N_b-n_b)(N-\hat{\tau})N_b\hat{\tau}}{N^2 n_b}\,.$$
This variance estimator usually was published with the results of the AMC, see Haslinger (1996).
- The variance of the Horvitz-Thompson estimator takes into account also stratification and clustering and is given by
$$\begin{aligned}\hat{V}_4(\hat{\tau}) = &\sum_{bh} \frac{W_{Abh}^2}{w_{Abh}}\left(1-\frac{w_{Abh}}{W_{Abh}}\right)s_{Abh}^2 \\ &+\sum_{bh}\frac{C_{Bbh}^2}{c_{Bbh}}\left(1-\frac{c_{Bbh}}{C_{Bbh}}\right)s_{Bbh}^2 + \frac{C_{Bbh}}{c_{Bbh}}\sum_{i=1}^{c_{Bbh}}\frac{W_{Bbhi}^2}{w_{Bbhi}}\left(1-\frac{w_{Bbhi}}{W_{Bbhi}}\right)s_{Bbhi}^2\,,\end{aligned}$$
where W_{xbh} and w_{xbh}, $x \in \{A, B\}$, are the number of dwellings in stratum xbh, in the universe respectively the sample, C_{2bh} and c_{2bh} are the number of PSUs in stratum Bbh in the universe and the sample, s_{Abh}^2 and s_{Bbhi}^2 are sample variance in stratum Abh and PSU $Bbhi$ and s_{Bbh}^2 is the variance between PSUs in stratum h.

In contrast to the simple estimators $\hat{V}_1$, $\hat{V}_2$, and $\hat{V}_3$, knowledge on the population size N is not required for $\hat{V}_4$. It is therefore a useful variance estimator for estimators of a population size.

Table 5 gives the results of the simulation study for estimated standard errors $\hat{s}_i = \sqrt{\hat{V}_i}$ and Figure 5 shows boxplots for the distributions of all 4 variance estimators for the totals τ_e and τ_u. The reference value for the performance of standard error estimators is the standard error in the simulation study. Obviously $\hat{s}_1$, $\hat{s}_2$, and $\hat{s}_3$ underestimate the true standard error in the simulation. As the total estimator $\hat{\tau}$ is more precise than the Horvitz-Thompson estimator, $\hat{s}_4$ is slightly biased upwards.

A $(1-\alpha)$-confidence interval for a total τ based on the normal approximation is obtained from an asymptotically unbiased estimate $\hat{\tau}$ and a variance estimate $\hat{V}(\hat{\tau})$ as
$$\hat{\tau} \pm u_{1-\alpha/2}\sqrt{\hat{V}(\hat{\tau})}\,.$$
The most serious consequence of underestimation of standard errors is that confidence intervals do not reach the nominal coverage. Actual coverages can be far too low for

Table 5: Simulation results for standard error estimators of totals

Total	Estimated s.e.	Mean	Bias	MSE	Coverage
τ_e	Simulation	23250.74			
	$\hat{s}_1$	13553.91	–9696.82	94029014.46	0.7484
	$\hat{s}_2$	14819.01	–8431.73	71095281.33	0.7880
	$\hat{s}_3$	14825.90	–8424.84	70979108.33	0.7882
	$\hat{s}_4$	24422.63	1171.89	1842767.35	0.9607
τ_u	Simulation	6886.16			
	$\hat{s}_1$	5432.16	–1454.00	2117528.80	0.8777
	$\hat{s}_2$	5939.17	–946.98	900801.94	0.9077
	$\hat{s}_3$	6291.42	–594.74	358965.50	0.9265
	$\hat{s}_4$	6890.12	3.96	21165.36	0.9514
N	Simulation	37182.23			
	$\hat{s}_4$	40866.68	3684.46	15006617.86	0.9688

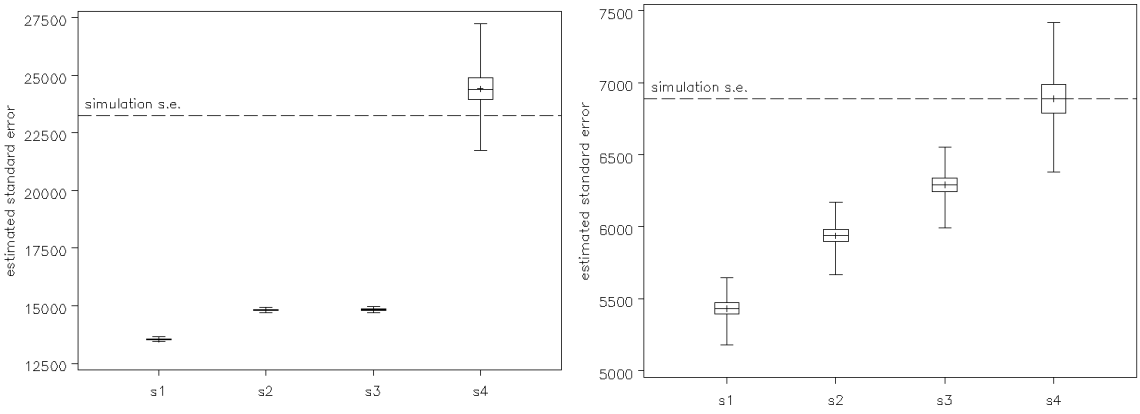

Figure 5: Boxplots for standard error estimates of τ_e (left) and τ_u (right)

simple variance estimators as can be seen from Table 5, only confidence intervals based on $\hat{s}_4$ reach the nominal confidence level.

Consequences are even worse for smaller areas, e.g. federal states. Figure 6 compares the non-coverage, i.e. the proportions of 95%-confidence intervals *not* covering the true value τ_e^b of federal state b for $\hat{s}_1$ and $\hat{s}_4$. Note that for federal states $\hat{s}_2$ and $\hat{s}_3$ coincide with $\hat{s}_1$. Non-coverage is about 5% for confidence intervals based on $\hat{s}_4$, but the simple estimator $\hat{s}_1$ leads to actual non-coverage ranging from 10% to nearly 30%.

The results indicate a design effect greater than 1 and show that simple variance estimators – though widely used in practice – can severely underestimate the sampling error for the complex sampling design. Confidence intervals based on these estimators have an actual coverage far below the nominal level. Complex variance estimation, taking into account stratification and clustering of the sampling plan results in less biased estimates and confidence intervals attaining the nominal coverage.

That simple variance estimators may be downward biased is well known from theoretical results – this simulation study allows to specify the extent of this bias in a close to reality situation.

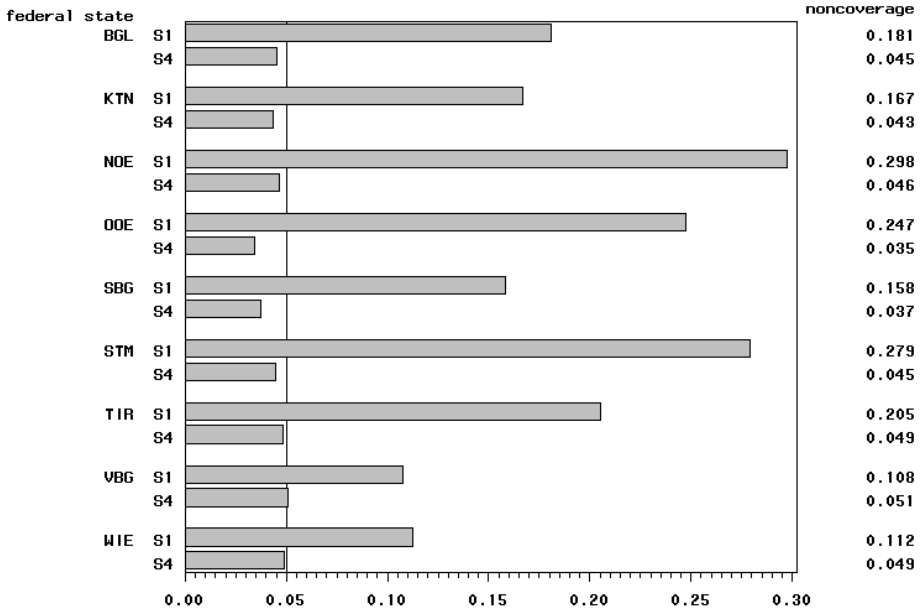

Figure 6: Percentage of confidence intervals not including the true τ_{eb} for federal states

5.2 Effects of Modification of Pseudo Universes

Effects of the sampling frame or non-sampling errors, such as non-response, on variance estimation can be investigated rather easily by modifying the pseudo-universe. To illustrate this point, two modifications of the first pseudo-universe, in the following called PU1, were realized:

1. Pseudo-universe PU2 comprises only inhabited dwellings and thus implies a different sampling frame. It was obtained from PU1 by removing all uninhabited dwellings.
2. Pseudo universe PU3 allows to study the effects of a certain non-response mechanism. It was created by implementing a unit non-response mechanism. For every dwelling a 0-1 random variable – 1 indicating response, 0 non-response of all individuals in this dwelling – was generated according to the non-response rate of the respective generation group in the AMC data. This was the only available information about non-response in the AMC. Implementation of a more realistic non-response mechanism, e.g. non-response probabilities depending on number of households or inhabitants of a dwelling would require further information which is not available from the AMC data.

Simulation results for the estimation of the total τ_e in the 3 different universes are given in Table 6, results for the standard error estimators of $\hat{\tau}_e$ are presented in Table 7. Obviously non-response is not quite adequately accounted for as $\hat{\tau}_e$ is more biased upwards in PU3 than in PU1 and PU2.

Standard errors are lower for PU2 as sampling from a universe without uninhabited dwellings for the AMC sampling plan implies a higher number of sampled *individuals*. The percentage of uninhabited dwellings is 14.5% in PU1 leading to a reduction in the standard error of about 8.9% for PU2 compared to PU1.

Table 6: Simulation results for the estimation of τ_e in modified pseudo universes

	True Value	Mean Estimate	Bias	MSE
PU1	3138666	3139059.61	393.61	540697607
PU2	3138666	3139001.05	335.05	448328923
PU3	3138666	3139670.38	1004.38	596037072

Table 7: Simulation results for standard error estimators in modified pseudo universes

Pseudo Universe	PU1: all dwellings		PU2: inhabited dwellings		PU3: unit non-response	
	Mean	Coverage	Mean	Coverage	Mean	Coverage
Simulation	23250.74		21172.19		24394.42	
$\hat{s}_1$	13553.91	0.7484	13554.69	0.7934	14362.06	0.7539
$\hat{s}_2$	14819.01	0.7880	14818.66	0.8306	15807.40	0.7947
$\hat{s}_3$	14825.90	0.7882	14826.02	0.8307	15814.99	0.7950
$\hat{s}_4$	24422.63	0.9607	23135.61	0.9675	25343.08	0.9584

Implementation of non-response amounts to a reduction of the number of sampled *respondents*, thus leading to an increase of the sample standard error. The overall non-response rate is 12.2% of inhabited dwellings which leads to an increase of the standard error of 4.9% in PU3 compared to PU1.

Results for standard error estimations are similar to those presented above and again show the better performance of the complex estimator $\hat{s}_4$. In each of the pseudo universes only $\hat{s}_4$ does not underestimate the true standard error and thus leads to confidence intervals attaining the nominal coverage.

6 Summary

For surveys with complex sampling designs variance estimators cannot be compared on theoretical grounds. The aim of this paper was to show that a comparison via simulation is feasible also for large universes.

Usually real universes are not available, therefore as a first step synthetic universes have to be generated. In the DACSEIS project a generation process for pseudo universes, consisting of a deterministic part and a stochastic part was developed. This process is described for the special case of a pseudo universe for the sampling plan of Austrian Microcensus, 1994–2003. The structure of the universe concerning stratification and clustering is rebuilt according to the sampling plan, assuming number and sizes of strata and clusters as deterministic. Sampling units of the AMC are dwellings. In the pseudo universe synthetic dwellings, including households and individuals living therein were generated stochastically.

To illustrate the application of the simulation setup, a simulation study is presented where 10000 samples according to the AMC sampling plan were drawn and estimates and direct variance estimates were calculated for each sample. True values of interesting quantities in the pseudo universe are known and the simulation distribution of e.g. a

total estimator provides a reference value for the performance of different variance estimators. Results show – not unexpectedly but nevertheless often ignored in applied work – that simple variance estimators underestimate the true sampling error and give a drastic impression of the extent of this bias.

The simulation setup described above is currently used for assessing properties of different methods for variance estimation under non-response, for first results see Quatember (2005).

Acknowledgements

We want to thank all members of the DACSEIS team, especially Andreas Quatember, Josef Schürle and Ralf Münnich, University of Tübingen, for many helpful discussions and Alois Haslinger, Statistics Austria, for helpful comments regarding the specifics of the AMC.

References

Haslinger, A. (1996). Stichprobenplan des Mikrozensus ab 1994. *Statistische Nachrichten*, 312-324.

Münnich, R., and Schürle, J. (2003). *On the Simulation of Complex Universes in the Case of Applying the German Microcensus.* DACSEIS Research Paper Series 5. Tübingen.

Münnich, R., and Wiegert, R. (2001). *The DACSEIS Project.* DACSEIS Research Paper Series 1. Tübingen.

Quatember, A. (2002). *Workpackage 2: Analysis of National Surveys.* DACSEIS Deliverable 2.1 and 2.2. Tübingen.

Quatember, A. (2005). Nonresponse in Bevölkerungsumfragen – Österrreich-Ergebnisse einer Simulationsstudie im Rahmen des EU-Projekts DACSEIS. *Austrian Journal of Statistics*, *34*, 263-281.

Wolter, K. M. (1985). *Introduction to Variance Estimation*. New York: Springer.

Authors' address:

Helga Wagner
Department of Applied Statistics and Econometrics
Johannes Kepler Universität Linz
Altenbergerstrasse 69
A-4040 Linz, Austria
E-mail: Helga.Wagner@jku.at

Doris Eckmair
Allgemeine Sparkasse OÖ Bank AG
Abteilung Risikomanagement
Sparkassenplatz 2
A-4020 Linz, Austria
E-mail: Doris.Eckmair@sparkasse-ooe.at