Department for Applied Statistics
Johannes Kepler University Linz

IFAS Research Paper Series
2004-05

Simulation studies of the Austrian Microcensus

Helga Wagner and Doris Eckmair

Juli 2004

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. In this paper we describe the setup for a simulation study according to the sampling plan of the Austrian Microcensus (AMC) and present basic results.

Key words: generation of universes, complex 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) is 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 homogenity within respectively heterogenity 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.

One of the surveys investigated within the DACSEIS project is the Austrian Microcensus (AMC). In this paper we describe the generation of the AMC pseudo universe and present the results of a simulation study. As the sampling plan of the AMC is rather complex, a restriction to its basic properties was necessary. These are described in section 2. Section 3 deals with the generation of the AMC pseudo universe and in section 4 the implementation of the sampling procedure is described. In the simulation study different variance estimators were compared and the effects of modifications of the pseudo universe, i.e. a different sampling frame and a certain nonresponse-mechanism, were investigated. Results are presented in section 5 and a summary is given in section 6.

2 The Austrian Microcensus

The Austrian Microcensus is a quarterly survey of 1% of all Austrian households conducted by interviewers since 1967. The AMC 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.

The sampling frame for the current AMC is the Austrian Housing Census (HWZ = Häuser-und Wohnungszählung) which is performed with a period of 10 years. Sampling units are dwellings. These are selected for all AMC surveys to be conducted in the following 10 years. Quarterly one eighth of the sampling units is 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 are recorded.

The sample design of the AMC consists of two parts. The first, in the following called Part A, comprises mainly dwellings in larger urban municipalities, the other, called Part B, dwellings in small, rural communities. Sampling is carried out seperately for each of the nine federal states in these two parts, except for two federal states (Wien and Vorarlberg), which consist only of Part A dwellings.

In Part A dwellings are selected as a stratified random sample where strata are built according to several dwelling characteristics. As a combination of all strata variables would result in very small strata, these are pooled to give sample sizes of at least 10 dwellings per stratum, resulting in 100 - 150 strata per federal state. The sampling fraction is 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) is carried out. PSUs are communities or – in case of very small communities – groups of communities. Strata are defined on the number of dwellings and agrarian quota, resulting in 5 -16 strata per federal state. Within each stratum the sample size is allocated proportional to size (i.e. number of dwellings) of districts. PSUs are selected randomly within each stratum. On the second stage a prespecified number of dwellings are drawn from the selected PSUs as secondary sampling units (SSUs). Depending on the federal state the number of SSUs is 20 or 25.

Selection of dwellings is 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 AMC sampling plan. A more detailed description of the AMC sampling plan can be found in Haslinger (1996).

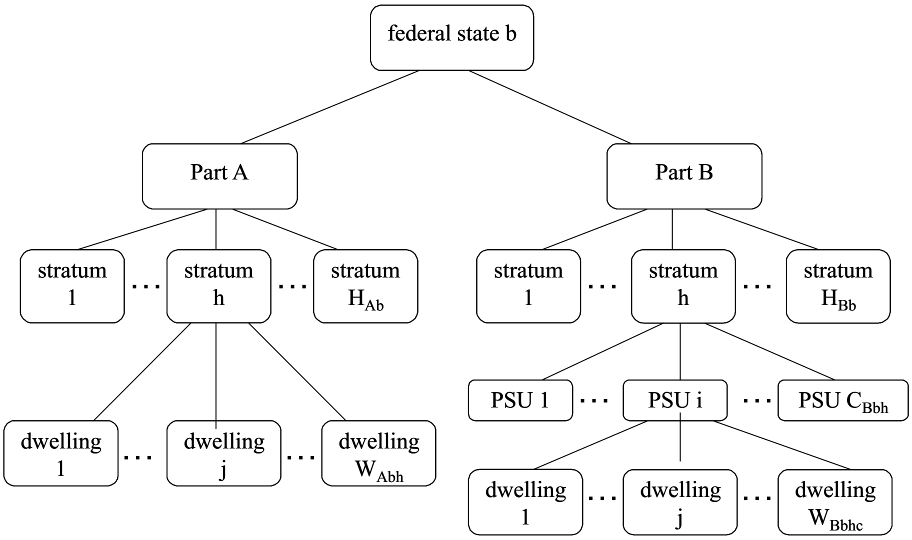

Figure 1: Structure of the AMC pseudo universe

3 The Generation of the AMC pseudo universe

The AMC pseudo universe was generated according to the general process for pseudo universes developed within the DACSEIS project. This process was applied to generate 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 the generation of the AMC pseudo universe 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 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. 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 the empirical distributions in the data are used as generation distributions.

Generation of the AMC 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 .

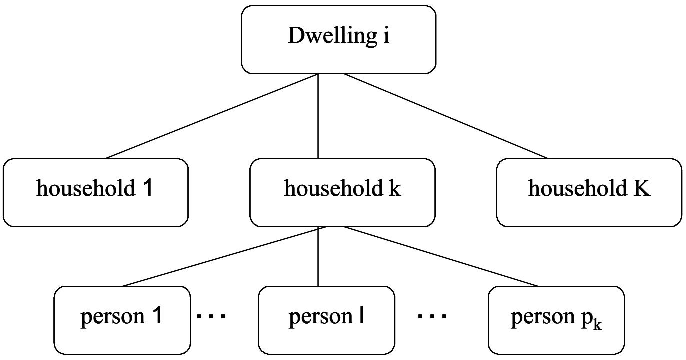

Figure 2: Hierarchical structure within dwellings

For the deterministic part of the generation process, the pseudo universe was 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 shows the number of strata per part and federal state. The total number of PSUs within Part B is 1,710. Each Part A stratum and each Part B PSU forms a separate generation group, meaning that the stochastic part of the generation process is identical within a generation group and different between generation groups. This leads to a total of 2,899 generation groups. The size of generation groups, i.e. the number of dwellings, is regarded as deterministic. Conditional on this size, dwellings are generated stochastically. For each dwelling values for the following variables were created:

Table 1: 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	—

K	number of households
p_k	number of persons in household k
x_{ilk}	personal characteristic i of person l in household k
	$i = 1, ..., 5$; $l = 1, .., p$
	(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 values of the variables are generated as random numbers according to the following modell

$$\begin{array}{rcl} K & \sim & P_K \\ p & \sim & P_p \\ (x_{11}...x_{1\,p}, x_{21}...x_{2\,p}) & \sim & P^p_{(x_{11}...x_{1\,p},x_{21}...x_{2\,p})} \\ (x_{3j}, x_{4j}, x_{5j}) & \sim & P^{(x_1^*,x_2)}_{(x_3,x_4,x_5)} \end{array}$$

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 this is the case for all individuals of one household - to a replication of entire households. To reduce the extent of replication of households, a modified variable x_1^*, i.e age measured in 5 year-categories, was used for the construction of conditional distributions.

3.2 The Generation of the AMC Pseudo Universe

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.

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

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 C 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 for 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 $\frac{w_{act}}{w_0}$, 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 in section 3.1, 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 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. 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 $\frac{w_{act}}{w_0}$ 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.

A comparison of one - and two dimensional marginal distributions of the personal characterstics in the AMC pseudo universe and the AMC data was carried out. Results which are presented in detail in Münnich and Schürle (2003) 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 Description of the Implemented 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 Simulation Studies

In the simulation studies the properties of various variance estimators were investigated. The whole simulation process, i.e. drawing an AMC sample from the pseudo universe and calculating estimates was repeated 10,000 times. The simulation studies were carried out on a Pentium P3 using C++ programs written by the second author.

One of the main purposes of variance estimation is to get at least approximate confidence intervals for parameters of the universe. 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

$$\hat{\tau} \pm u_{1-\alpha/2}\sqrt{\hat{V}(\hat{\tau})}.$$

Useful criteria for variance estimators therefore are bias, mean square error and the coverage of confidence intervals, i.e. the proportion of confidence intervals out of 10,000 that cover the true value τ.

5.1 Variance estimation of totals

5.1.1 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 nonresponse 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 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 $\frac{c}{C}$.

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

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

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 nonresponding 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 3 indicate that estimates of all three totals are rather close to their respective true values in the pseudo universe.

Table 3: Simulation results for the estimation of totals

	true value	mean of estimate	bias	MSE
τ_e	3 138 666	3 139 059.61	393.61	540 697 607
τ_u	304 581	304 589.37	8.37	47 414 511
N	7 463 802	7 462 737.51	-64.48	1 382 383 998

5.1.2 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.

In our simulation we investigated 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 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 $\frac{\hat{\tau}_b}{N_b} = \frac{\hat{\tau}}{N}$ 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 is 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

$$\begin{aligned}\hat{V}_4(\hat{\tau}) = & \sum_{bh} \frac{W_{Abh}^2}{w_{Abh}}(1 - \frac{w_{Abh}}{W_{Abh}})s_{Abh}^2 + \\ & + \sum_{bh} \frac{C_{Bbh}^2}{c_{Bbh}}(1 - \frac{c_{Bbh}}{C_{Bbh}})s_{Bbh}^2 + \frac{C_{Bbh}}{c_{Bbh}} \sum_{i=1}^{c_{Bbh}} \frac{W_{Bbhi}^2}{w_{Bbhi}}(1 - \frac{w_{Bbhi}}{W_{Bbhi}})s_{Bbhi}^2\end{aligned}$$

  where W_{xbh} and w_{xbh} 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 $\hat{V}_1$- $\hat{V}_3$, 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 4 gives the results of the simulation study for estimated standard errors $\hat{s}_i = \sqrt{\hat{V}_i}$. Figure 3 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 $\hat{s}_1 - \hat{s}_3$ 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.

The most serious consequence of underestimation of standard errors is that confidence intervals do not reach the nominal coverage. Actual coverages can be far to low for simple variance estimators as can be seen from table 4, 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 4 compares the noncoverage, 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 $\hat{s}_1$. Noncoverage is about 5% for confidence intervals based on $\hat{s}_4$, but the simple estimator $\hat{s}_1$ leads to actual noncoverage ranging from 10% to nearly 30%.

Table 4: Simulation results for standard error estimators of totals

total	estimated s.e.	mean	bias	MSE	coverage
τ_e	simulation	23 250.74			
	$\hat{s}_1$	13 553.91	-9 696.82	94 029 014.46	0.7484
	$\hat{s}_2$	14 819.01	-8 431.73	71 095 281.33	0.7880
	$\hat{s}_3$	14 825.90	-8 424.84	70 979 108.33	0.7882
	$\hat{s}_4$	24 422.63	1 171.89	1 842 767.35	0.9607
τ_u	simulation	6 886.16			
	$\hat{s}_1$	5 432.16	-1 454.00	2 117 528.80	0.8777
	$\hat{s}_2$	5 939.17	-946.98	900 801.94	0.9077
	$\hat{s}_3$	6 291.42	-594.74	358 965.50	0.9265
	$\hat{s}_4$	6 890.12	3.96	21 165.36	0.9514
N	simulation	37 182.23			
	$\hat{s}_4$	40 866.68	3 684.46	15 006 617.86	0.9688

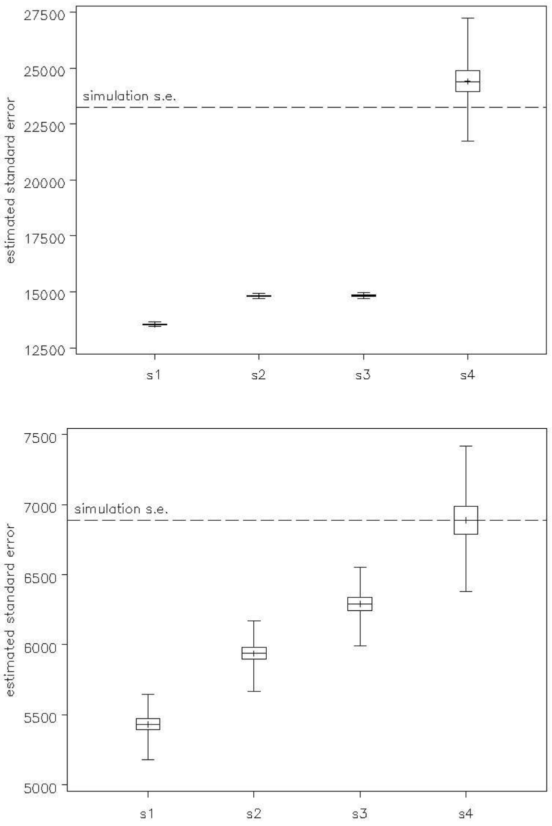

Figure 3: Boxplots for standard error estimates of τ_e (left) and τ_u (right)

5.2 Variance estimation of proportions

Ratios of two unknown population totals $R = \frac{\tau_1}{\tau_2}$ are estimated by the ratios of their respective estimates. In the simulation study we investigated the proportion

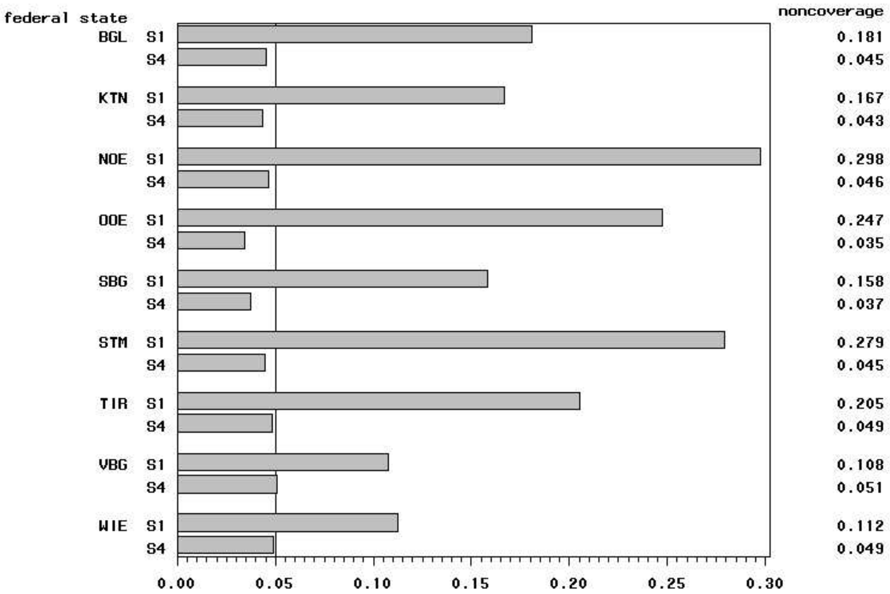

Figure 4: Percentage of confidence intervals not including the true value τ_{eb} for federal states

of persons employed at least one hour, i.e. $\hat{R} = \hat{p}_e = \hat{\tau}_e/\hat{N}$. Simulation results for this estimator given in table 5 show that the true value in the pseudo population is reproduced very well.

Table 5: Simulation results for the estimation of the proportion p_e

	true value	mean of the estimate	bias	MSE
p_e	0.42058	0.42063	5.5616E-5	4.8902E-6

Variance estimators for proportions can be derived from those for totals (see section 5.1). Simple variance estimators $\hat{V}_1(\hat{p})$ for proportions – $\hat{V}_3(\hat{p})$ are adaptions of $\hat{V}_1(\hat{\tau})$ – $\hat{V}_3(\hat{\tau})$, with estimated population size $\hat{N}$ instead of assuming N known, e.g.

$$\hat{V}_1(\hat{p}) = \frac{\hat{p}(1-\hat{p})}{n} \cdot (\frac{\hat{N}-n}{\hat{N}})$$

A more complex variance estimator for an arbitrary ratio R is based on $\hat{V}_4(\hat{\tau})$ using Taylor linearization

$$\hat{V}(\hat{R}) = \frac{1}{\hat{\tau}_2^2} \cdot \left(\hat{V}(\hat{\tau}_1) + \hat{R}^2\hat{V}(\hat{\tau}_2) - 2\hat{R}\hat{C}(\hat{\tau}_1, \hat{\tau}_2)\right)$$

where $\hat{C}(\hat{\tau}_1, \hat{\tau}_2)$ denotes the estimator of the covariance of $\hat{\tau}_1$ and $\hat{\tau}_2$, which takes into account clustering and stratification. The formula for $\hat{C}(\hat{\tau}_1, \hat{\tau}_2)$ is rather lenghty, its derivation can be found in Quatember (2003). Simulation results for the four

Table 6: Simulation results for the standard error estimation of p_e

	mean	bias	MSE	coverage
simulation	0.002211			
$\tilde{s}_1$	0.001816	-3.9458E-4	15.5717E-8	0.8919
$\tilde{s}_2$	0.001986	-2.2523E-4	5.0757E-8	0.9215
$\tilde{s}_3$	0.001987	-2.2415E-4	5.0270E-8	0.9217
$\tilde{s}_4$	0.002216	0.0515E-4	0.2524E-9	0.9495

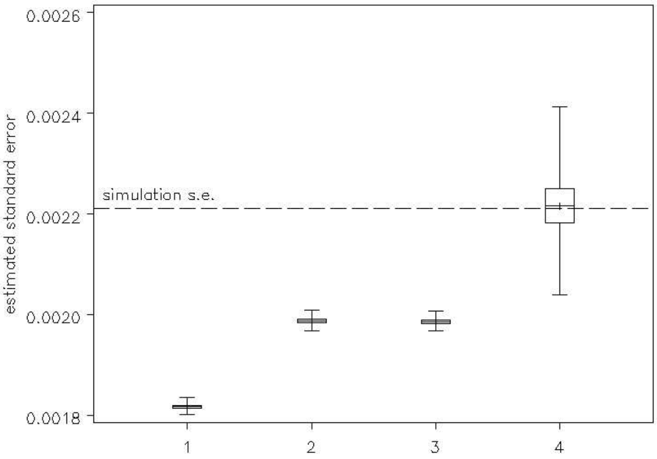

Figure 5: Boxplots for standard error estimates of the proportion p_e of persons employed at least one hour

different standard error estimators are summarized in table 6, their distribution is illustrated through boxplots in figure 5.

Again simple estimators underestimate the true standard error. Coverages for confidence intervals using simple standard error estimators are slightly higher than for totals, but again only the complex estimator leads to coverages of about 95%.

5.3 Effects of modification of pseudo-universes

Effects of the sampling frame or non-sampling errors, such as nonresponse, on variance estimation can be investigated by modifying the pseudo-universe. 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 nonresponse mechanism. It was created by implementing a unit nonresponse mechanism. For every dwelling a 0-1 random variable — 1 indicating response, 0 nonresponse of all individuals in this dwelling — was generated according to the nonresponse rate of the respective generation group in the AMC data. This is the only

available information about nonresponse in the AMC at present. Implementation of a more realistic nonresponse mechanism, e.g. nonresponse 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 7, results for the standard error estimators of $\hat{\tau}_e$ are presented in table 8. Obviously nonresponse is not quite adequately accounted for as $\hat{\tau}_e$ is more biased upwards in PU3 than in PU1 and PU2.

Table 7: Simulation results for the estimation of τ_e in modified pseudo universes

	true value	mean of estimate	bias	MSE
PU1	3 138 666	3 139 059.61	393.61	540 697 607
PU2	3 138 666	3 139 001.05	335.05	448 328 923
PU3	3 138 666	3 139 670.38	1 004.38	596 037 072

Table 8: Simulation results for standard error estimators in modified pseudo universes

Pseudo-Universe	PU1: all dwellings		PU2: inhabited dwellings		PU3: unit-nonresponse	
	mean	coverage	mean	coverage	mean	coverage
simulation	23 250.74		21 172.19		24 394.42	
$\hat{s}_1$	13 553.91	0.7484	13 554.69	0.7934	14362.06	0.7539
$\hat{s}_2$	14 819.01	0.7880	14 818.66	0.8306	15 807.40	0.7947
$\hat{s}_3$	14 825.90	0.7882	14 826.02	0.8307	15 814.99	0.7950
$\hat{s}_4$	24 422.63	0.9607	23 135.61	0.9675	25 343.08	0.9584

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 of standard error of about 8.9% for PU2 compared to PU1.

Implementation of nonresponse amounts to a reduction of the number of sampled *respondents*, thus leading to an increase of the sample standard error. The overall nonresponse 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 illustrated for the special case of a pseudo universe for the Austrian Microcensus. 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.

In the simulation study 10 000 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 of the simulation studies show that simple variance estimators though widely used in practice can severely underestimate the sampling error for the complex sampling design of the AMC, indicating a design effect greater than 1. 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. From the results we conclude that the additional effort for deriving and calculating complex variance estimators, regarding specifics of the sampling plan and the estimator of interest clearly is worth wile.

Acknowledgement

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. The work of the second author was supported by the DACSEIS Project.

References

Haslinger, A. (1996). Stichprobenplan des Mikrozensus ab 1994. *Statistische Nachrichten*, 4/1996:312–321.

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.*

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

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

Quatember, A. (2003). An Example for the Estimation of the Variance of a Ratio in a Complex Survey Design. Working paper.

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