Variance Estimation for the Scientific Use File of the German Microcensus

Ulrich Rendtel and Bernhard Schimpl-Neimanns*

Paper prepared for the International Conference on Quality in Official Statistics, Stockholm, May 14-15, 2001

Summary. One aspect of the quality of official statistical data that is becoming more important is their utility for social science research. In Germany, scientific use files of the microcensus are released to the scientific community in the form of so-called factually anonymous microdata. However, usage of such files for the purpose of statistical analyses is not unproblematic. This paper focuses on the problems of variance estimation, arising from both the specific sampling design of the German microcensus and the procedure by which scientific use files are generated from the original data.

The scientific use file of the microcensus 1996 is the first to provide sampling information. However, it does not provide all relevant information. After a short presentation of the sampling design of the microcensus and the selection procedure of the scientific use file in this paper, a solution is presented that uses the available design information in an efficient way. The variance estimates based on the scientific use file are compared with results from the German Federal Statistical Office. Furthermore, we develop the regression estimator (group mean model) for the post-stratification of the microcensus estimates. The results show that variance reduction by post-stratification is not relevant for the scientific use file. Large differences between the post-stratified data and the unadjusted data, however, pose the question as to which of the two population estimates is biassed. We also investigate the behaviour of the linear regression of the design effect, which is frequently used as a tool for variance estimation. Generally, the variance estimation, by means of design effects, produces reasonable results. But in detail, we find a considerable amount of over- and underestimation.

Our results indicate that variables for stratum, clustering, and post-stratification should be released with the data if there are no confidentiality concerns. This would considerably improve variance estimation based on scientific use files.

Key words: Variance estimation; Stratified multistage survey data; Post-stratification; Scientific Use File, Microcensus.

1 Introduction

In empirical social research, the official statistical information that is conducted in multipurpose surveys is of considerable importance as a data basis. Microdata from sources such as the EC Labour Force Survey or the US Current Population Survey contain a very large number of cases, and response rates are normally higher than in surveys conducted by the scientific community itself. Economic and social structures and processes can therefore be analysed with a high degree of differentiation, offering the possibility of even making reliable statements about small populations. The main objective of this paper is to assess how researchers interested in variance estimation as an indicator of the quality of estimates can util-

* Addresses:
Prof. Dr. Ulrich Rendtel, J.W. Goethe Universität, Fachbereich Wirtschaftswissenschaften, Institut für Statistik und Mathematik, Merton Str. 17, D-60054 Frankfurt/M., E-Mail: rendtel@em.uni-frankfurt.de
Bernhard Schimpl-Neimanns, Zentrum für Umfragen, Methoden und Analysen (ZUMA), Abteilung Mikrodaten, Postfach 12 21 55, D-68072 Mannheim, E-Mail: schimpl-neimanns@zuma-mannheim.de

ize the available design information.

Since official statistical surveys are often based on a stratified multistage sampling design, a researcher needs relevant design information in order to assess the quality of the estimates. Identifiers for strata and clusters are usually related to regional information, which is subject to data confidentiality. This frequently leads to restrictions for the design information in so-called public use or scientific use files (see Eltinge 1999). From the point of view of research, particular attention needs to be given to the limitations that this places on variance estimation. To assess such limitations, variance estimations based on the scientific use file of the German microcensus are compared with results from the Federal Statistical Office.

Since 1996, the scientific use files of the microcensus have contained some design information. Earlier microcensuses require the evaluations have to be performed on the assumption of simple random sampling. The estimated variance must be corrected using published design effects to adjust for the sampling design. This method is frequently suggested for surveys like the UK Labour Force Survey (ONS 1998) and the US Current Population Survey (BLS/BOC 2000). To find out whether this method gives proper results in research practice, the behaviour of the linear regression of the design effect of the microcensus is investigated in this paper.

When non-sampling or systematic errors (e.g. nonresponse) are present in the survey, the variance is an incomplete measure of the quality of the estimate. The sampled units are frequently weighted to adapt the sample distributions to "known" marginal population totals to account for undercoverage and nonresponse. If such weights and information on nonresponse are released with the scientific use file, researchers can estimate variances using post-stratification weights. Among other things, comparisons of unweighted and weighted estimates can reveal non-sampling errors.

The rest of this paper is organized as follows: the second section contains a brief description of the sampling design of the German microcensus and the selection procedure of the scientific use file. The relevant estimation assumptions that result from the information available for the scientific use file are stated in section 3. Section 4 presents empirical results for variance estimates and compares them with the corresponding values from the statistical office. Section 5 shows the implementation of post-stratified variance estimates following Särndal et al. (1992). A comparison of direct variance estimations and approximations using design effects is given in section 6. Section 7 summarizes the most significant results and suggests further improvements on the release of design information for scientific use files of the Ger-

man microcensus. Furthermore, we address the issue in which respect the experiences based on the German microcensus could be relevant to variance estimation with other scientific use files.

2 The German Microcensus and its Scientific Use File

The basic programme of the German microcensus (MC), which is conducted annually with a sampling fraction of 1 percent, provides figures on issues such as demographic, social and employment structure, families, and households. Since the respondents are obliged to provide information to the basic programme, the response rate is around 97 percent of all households.

The MC is released as a scientific use file to the scientific community according to the Federal Statistics Act of 1987. The act stipulates that a factually anonymized subset of the data can be passed on to universities or other independent research institutes in Germany if the particulars of those surveyed can be (re)identified only by a disproportionately large investment of time, cost, and labour (factual anonymity) (see Müller et al. 1991). The scientific use file (SUF) of the microcensus consists of a 70 percent subsample and covers around 500,000 people in over 230,000 households. In order to determine variances based on the SUF, it is necessary to describe the sampling design of the MC in advance.

2.1 The Sampling Design of the German Microcensus

Since 1990, a new sampling design has been used for the MC (see Krug et al. 1999: 304pp.; Meyer 1994). The survey is designed as a single-stage cluster sample, where the primary sampling units (clusters, sampling districts) are composed of neighbouring buildings or parts thereof. All households within a cluster are sampled. The formation of the clusters is arranged by grouping the buildings into four different classes: small, medium-sized, and large buildings, as well as institutions. Within these building classes, detailed regional stratification is used for the formation of the clusters. The average number of households within a cluster is quite low, ranging from 7 to around 15, resulting in an average of about 9 dwellings per sampling district. The formation of the regional clusters means that a high administrative burden that usually occurs only in long intervals together with the population census. Aggregate data of the 1987 Population Census served as a sampling frame for the microcensus in the former territory of the Federal Republic, and in the new Länder the frame is based on the Central Residents' Register of 1991.

In order to account for changes in the building stock based on the statistics on building activity, a new buildings stratum is formed. Clusters are formed within this stratum according to the number of dwellings of the building (building size group) and regional criteria.

Except for the new building stratum, the sampling of the clusters is similar to simple random sampling from strata. However, the strata were made so small that only one cluster per stratum was chosen. The formation of the strata used the same criteria that were used for the construction of the clusters; i.e. regional information within building classes. The strata size was 100 for all strata. Hence, selecting one cluster per stratum yields a uniform sampling rate of one percent.

A different selection rule was used for the selection of clusters from the new buildings class. Here, the clusters were sequenced for each building size group and a systematic sampling with an interval size 100 was used. The sequencing used the same regional criteria that were used for the constitution of the clusters.

2.2 The Selection of the Scientific Use File from the Microcensus

The SUF is a 70 percent subsample of the MC households and was sampled by some kind of systematic sampling.

For the purpose of subsampling, the MC households had to be sequenced. The criteria used for this were somewhat different from the strata criteria in the MC. The sequencing variable was set up by three broad regional identifiers.[1] Within these strata, the household size was used. Then the households were sorted according to their cluster numbering, which reflects their local stratification. Within each cluster, the internal ordering of the households was used. The households were numbered according to this sequence. The last digit of this numbering was used for the selection. 7 digits were taken for the selection into the SUF, resulting in a subsampling rate of 70 percent.

3 Estimation Assumptions

Strictly speaking, the SUF resulted from a 2-phase sampling procedure in which households were taken by a cluster sampling at the first stage and the SUF-households resulted from systematic sampling at the second stage. It is assumed that 2-stage sampling is a good approximation.

[1] State, administrative region, and classification of community size.

To ensure sufficient confidentiality, most regional criteria are not available with the SUF. The sequencing number used in the second sampling stage is also not preserved. However, even if these variables were known, there would be no possibility to calculate unbiassed variance estimates for population totals and means. This is also true for the original data of the MC and is due to the sampling design of the MC. The selection of just one primary sampling unit per stratum is the reason why approximations must be used for variance computations. The German Federal Statistical Office solves this problem by defining larger areas as strata,[2] ignoring the smaller regional structures on which the sampling procedure is actually based. The assumption of simple random sampling of clusters within these enlarged strata leads to an overestimation of sampling variances. The same applies to systematic sampling (see Wolter 1985: 282).

In the SUF, the following sampling information can be used: building size class, federal state, classification of community size, and the cluster identifier (i.e. an indicator for households belonging to the same cluster). The omission of some of the detailed regional stratification variables tends to inflate the variance estimate based on the SUF compared to the original MC.

Concerning the ordering of the households at the second stage of the selection of the subsample, the sequencing – except for the ordering by the household size – more or less reflects the original strata. To a large extent, this is due to the use of the original cluster number, which reflects the stratification of the MC.[3] Since the last digit of the sequencing number used in the second sampling stage was dropped, it is not possible to draw conclusions from the variance of the selected households on the basis of the variance component caused by the last digit sampling. Additionally, one must assume a constant sampling rate of 70 percent of households per cluster. This is not always be true for the systematic sampling, but we can expect that this simplifying assumption will not significantly affect estimations.

Following Särndal et al. (1992), the definitions used in the subsequent sessions are as follows. Let H be the number of strata and N_h the number of sampled clusters (primary sampling units; PSU's) in the hth stratum of the population. N_i is the number of households in the ith PSU. Let Y be a variable for which data has been collected, with $y_{h,i,k}$ being the value observed for the kth household in the ith PSU in the hth stratum.

[2] The stratification by the size of the buildings is combined with a regional stratification. Cities larger than 200,000 inhabitants and other spatial units with more than 250,000 inhabitants constitute a stratum.
[3] We note that the cluster and the household identification numbers in the SUF were re-sorted.

For a 2-stage stratified sampling design we have the following entities: let U_I be the population of N_I PSU's, and $U_{I,h}$ the set of PSU's in the hth stratum with $N_{I,h}$ elements. Let U_i be the set of all households in the ith PSU (size = N_i). Finally, let U be the set of all households in the survey area (size = N).

Let s_I be the sample of the PSU's of size n_I that is spread over H strata as $s_{I,h}$ of size $n_{I,h}$. Let s_i be the sample of the households in the ith PSU of size n_i, and let s be the sample of all households of size n.

The inclusion probability of the ith PSU is defined as $\pi_{I,i}$. Then $\pi_{k|i}$ refers to the inclusion probability of the kth household in the ith PSU, given that the PSU has been drawn. For the MC holds $\pi_{I,i} = 0.01$, and for the SUF we assume $\pi_{k|i} = 0.7$.

4 Estimates of Totals and Standard Errors for the Scientific Use File

Let us look at the various population parameters that are typical for Labour Force Surveys. There is usually interest in estimates of population totals, means, or ratios at the time the survey is taken. Even though the MC has a rotating panel design[4] we do not consider change estimates, since the SUF does not contain identifiers for the rotation group.[5]

Allowing for element inclusion probabilities, but ignoring the post-stratification weights for the moment (see section 5), a total of the characteristic y is estimated by:

(1) $$\hat{t} = \sum_{h=1}^{H} \hat{t}_h = \frac{100}{0.7} \sum_{k \in s} y_k$$

where $\hat{t}_h$ is the estimated total of stratum h.

According to Särndal et al. (1992: 142), if we assume in two-stage sampling a design consisting of simple random sampling without replacement in both stages (SI, SI), the variance is:

(2) $$V_{SI,SI}\left(\hat{t}_h\right) = N_{I,h}^2 \frac{1-f_h}{n_{I,h}} S_{U_{I,h}}^2 + \frac{N_{I,h}}{n_{I,h}} \sum_{i \in U_{I,h}} N_i^2 \frac{1-f_i}{n_i} S_{U_i}^2$$

where $f_h = \frac{n_{I,h}}{N_{I,h}} = 0.01$, $f_i = \frac{n_i}{N_i} = 0.7$

[4] A sampling district stays in the sample for four years and one-fourth of the sampling districts is replaced every year. According to the principle of an area sample, households moving away from the sampling districts will not be interviewed further. But households moving into the sampling district are included in the sample.

Replacing the variances $S^2_{U_{I,h}}$ and $S^2_{U_i}$ by their respective sample estimates gives:

(3) $$S^2_{s_{I,h}} = \frac{1}{n_{I,h}-1}\sum_{i\in s_{I,h}}\left(\hat{t}_i - \hat{\bar{t}}_{s_{I,h}}\right)^2$$

= variance between PSU in the *h*th stratum

(4) $$S^2_{s_i} = \frac{1}{n_i - 1}\sum_{k\in s_i}\left(y_k - \bar{y}_{s_i}\right)^2$$

= variance within the *i*th PSU

The second term of (2) involves the computation of the within variance for each of the approx. 40,000 PSU's. A computationally simpler method would be to use only the first term with the estimated variance of the PSU totals (see Särndal et al. 1992: 139pp.). Provided that the first-stage sampling fraction is small, as it is in the case of the MC, this approximation is very accurate.[6] This point is demonstrated in Table 1.

For selected characteristics, Table 1 compares the estimated total, standard error, and the coefficient of variation as well as the design effect (see section 6) based on the SUF with the results of the German Federal Statistical Office for the MC (Statistisches Bundesamt 1998b). First of all, the inspection of the coefficient of variation shows a very high precision even for characteristics with a rather small number of cases. For example, for employed women whose net income per month is less than 600 DM, the coefficient of variation is about 1 percent.

A striking difference concerns the totals based on the SUF vs. the MC data, where the former are always smaller than the latter. The MC totals are far beyond the conventional 95 percent confidence interval of the SUF totals. This bias is due to the fact that the German Federal Statistical Office uses a nonresponse adjustment weight to compensate for the nonresponse of about 2.5 percent of the households.[7] But the SUF does not contain this fine-grained compensation weight variable. If we alternatively use a constant weight factor of 1.025, all corresponding MC estimates are included within the confidence limits of the SUF estimates.

[5] See Holmes and Skinner (2000) for change estimates for the UK Labour Force Survey.
[6] This simplified estimator is available in generalised software packages as STATA and SAS.
[7] The nonresponse adjustment weight is computed at the level of 401 regional subgroups for 19 different combinations of socio-demographic characteristics.

Table 1: Estimated Totals for Selected Characteristics (Unit: thousands)

Characteristic	Total		Standard Error				Coefficient of Variation (%)			Design Effect	
	SUF	MC	SUF			MC	SUF	MC	Approx.	SUF	MC
			Between	Within	Overall						
Foreign labour force[1)]	2,283.7	2,330.6	29.4	1.2	29.4	27.3	1.29	1.17	0.92	1.66	1.83
Unemployed persons (ILO-Definition)[1)]	2,976.7	3,034.2	24.0	1.2	24.1	21.4	0.81	0.71	0.81	1.20	1.26
Main source of livelihood: employment[1)]	29,607.9	30,285.0	81.4	3.2	81.4	77.7	0.28	0.26	0.30	1.63	1.85
Employed women, net income per month less than 600 DM[1)]	1,545.4	1,585.4	15.8	0.8	15.9	13.9	1.03	0.88	1.11	1.08	1.12
Private households, 1 Person, female[2)]	7,010.1	7,259.6	33.1	1.5	33.1	32.0	0.47	0.44	0.54	1.19	1.35

Source: SUF: Scientific Use File of the German Microcensus 1996 (70 percent subsample). MC: Microcensus 1996 (totals multiplied by nonresponse compensation weight (inflation factor), Source: Statistisches Bundesamt 1998b). Approx. value: Coefficient of variation based on published design effects for the microcensus 1990 (Statistisches Bundesamt 1998a: 17). Subpopulations: 1) Population on main residence; 2) Private households.

Compared with the MC, the estimated variances based on the SUF are always larger. This was to be expected regarding the smaller sample size. The German Federal Statistical Office recommended that users of the SUF increase the standard error published for some characteristics by a factor of 1.2, which refers to the assumption of the SUF as a simple random sample of the MC ($(1/0.7)^{½} = 1.195$). However, based on the evaluation of more than 450 characteristics of the standard tabulation of the statistical office, we evaluated an increase of the standard error of the SUF to about 1.09 on average. That this loss of precision in respect to the sample size is just half as much as expected, is related to the reduction of the cluster effect, which, in turn, results from the 70 percent subsample of all households belonging to a MC cluster.

In summary, we conclude that users can properly and easily evaluate variance estimates based on the SUF. However, estimation would be improved if research could use the household nonresponse compensation weight.

5 The Variance of Population Estimates after the Adjustment to the Results of the Current Updating of the Population

The German statistical offices publish results for the microcensus fitted to benchmark data derived from the current updating of the population. For this purpose, a weight variable on the

basis of regional "known" marginal totals for six post-strata[8] is formed, which mainly reflects the ratio of the nominal value to the sample value (see Heidenreich 1994). For the post-stratified estimator, the users of the scientific use file can work with the weight variables for persons and households contained in the data file.

Formally, the use of weights can be interpreted as a regression estimate. The estimate used in this case is based on the Group Mean Model that can be depicted as follows (see Särndal et al. 1992: 324). The population U can be partitioned into G disjoint subsets U_g, $g \in \{1,\ldots,G\}$ of post-strata. The estimation for the population means in group g is:

$$(5) \qquad \hat{B}_g = \frac{\sum_{k \in s_g} y_k / \pi_k}{\sum_{k \in s_g} 1/\pi_k} = \frac{1}{\hat{N}_g} \sum_{k \in s_g} \frac{y_k}{\pi_k}$$

The post-stratified regression estimate $\hat{t}_{reg}$ for the Group Mean Model is the following:

$$(6) \qquad \hat{t}_{reg} = \sum_{g=1}^{G} N_g \hat{B}_g = \sum_{g=1}^{G} \sum_{k \in s_g} \frac{N_g}{\hat{N}_g} \cdot \frac{y_k}{\pi_k} = \sum_{g=1}^{G} \sum_{k \in s_g} w_g \frac{y_k}{\pi_k}$$

The SUF contains the weight w_k, which assumes for each person $k \in s_g$ the corresponding value $N_g / \hat{N}_g$.

One important characteristic of the regression estimate is the fact that it returns the known marginal values of each sample (see Särndal et al. 1992: 324). Consequently, the variance of $\hat{t}_{reg}$ for these characteristics is zero. However, the SUF is only a 70 percent subsample of the MC and has not been separately adjusted to the current updating of the population. For this reason, the variance estimation will also provide positive values for the adjustment characteristics.

A Taylor approximation of $\hat{t}$ is used for the derivation of the variance $V(\hat{t}_{reg})$. The linear portion of the Taylor approximation is given by the following concomitant variable u_k (see Särndal et al. 1992: 331):

$$(7) \qquad u_k = w_k (y_k - \bar{y}_{s_g}) \qquad k \in s_g$$

Thus, y_k needs only to be substituted by u_k in the equations (2) to (4).

[8] The post-strata are: sex in combination with citizenship (Germans/Foreigners) in regional units of at least 500,000 inhabitants. When the results for soldiers and men liable to military service are adjusted, the registers of

The influence of the post-stratification adjustment on the estimation of totals and their variance for the characteristics of Table 1 is depicted in Table 2. It illustrates that the relative standard error is only slightly reduced by the adjustment. For characteristics that are closely correlated to the post-strata (for example, lines 1 and 5 in Table 2), the decrease is steep. However, in some cases, there is a remarkable difference between the estimates in the order of magnitude of several millions. This shows the problematic nature of using the post-stratification weights. They hardly ever lead to a decrease of the variance. Instead, they cover a bias. Either the MC, and hence the SUF, provides biassed population estimates, or the current updating of the population produces biassed estimations on its part.

Table 2: Estimated Population Totals (thousands) and Coefficient of Variation (percent) for selected Characteristics with and without Post-stratification Adjustment

Characteristic	Total		Coefficient of Variation	
	adjusted	unadjusted	adjusted	unadjusted
Foreign labour force[1)]	3,609.7	2,283.7	0.72	1.29
Unemployed persons (ILO-Definition)[1)]	3,490.3	2,976.7	0.79	0.81
Main source of livelihood: employment[1)]	33,806.1	29,607.9	0.20	0.28
Employed women, net income per month <600 DM[1)]	1,740.7	1,545.4	1.00	1.03
Private households, 1 Person, female[2)]	7,896,6	7,010.1	0.27	0.47

Source: see Table 1

Another potential source of this bias are problems linked to the realisation of the sample; for example, actuality of the sampling plan, accessibility of households and unit-nonresponse. It is presumed that the undercoverage of soldiers and men liable to military service, which is reflected in a weight of about 1.6, is caused by non-accessibility.[9] However, the current updating of the population is not free of errors, either. Particularly with regard to foreigners, it is presumed that moves are insufficiently covered and that the results of the current updating of the population are therefore too high. The post-stratification weighting leads to the transfer of errors from the current updating of the population to the microcensus. This adjustment results in a weighting factor of 1.1 for the Germans and even 1.5 for the foreigners regarding the SUF. One the one hand, a correction of such extent cannot be derived from the results of the field work (see Heidenreich 1994: 116). On the other hand, the comparison between the results of the current updating of the population and the census from 1987 show much smaller deviations (Heidenreich 1989: 328; Jäger 1992: 105pp.). Thus, the considerable differences

the Ministry of the Interior and the Ministry of Defence are used on the level of administrative districts and on the level of the federal states.

[9] According to the sampling plan, the PSU's do not comprise barracks. Soldiers and men liable to military service, hence, cannot be interviewed at their barracks, but only at their main or secondary place of residence.

between the microcensus and the current updating of the population cannot be traced exclusively to the errors of the current updating of the population. However, research on the source of error and the differences between the microcensus and other population statistics is missing.

6 Design Effects

So far, the users of the SUF who want to estimate the variance of totals have been dependent on the results of the design effect of the sampling error published by the German Federal Statistical Office. The design effect is the ratio of the design-based standard error to the standard error of a simple random sample. The core of this approach is a linear regression of the design effects $k(p_d)$ on the proportion p of the population with the characteristics of interest.

The number of households and persons per PSU in the SUF is reduced by the selection of the subsample compared to the MC. This also reduces the cluster effect. It can therefore be assumed that the published design effects are not assignable to the SUF. This is why we want to see if the selection of the SUF can lead to different design effects. We also want to evaluate the goodness of fit of the linear approximation of the design effects.

The German Federal Statistical Office uses a separate simple linear regression of $k(p_d)$ on p for three groups of characteristics (population and economically active persons (B/E), foreigners and people employed in agriculture (A/L) and households (H)):

For given values of a constant a and a slope b, the following approximation for the coefficient of variation in dependence on p is obtained:

$$(8) \qquad cv = k(p)cv_{SI} = (a+bp)\sqrt{\frac{1-f}{n-1}\,\frac{1-p}{p}}$$

To sum up, Table 3 compares the regression coefficients for each group obtained on the basis of the SUF with the published values for the microcensus 1990 (MC90) and the computation for the microcensus 1996 (MC96).[10] Generally, the slopes for the SUF run more flatly than those of the MC96 (see also Figure 1). Referring to the more than 450 characteristics of the standard tabulation programme of the German Federal Statistical Office, the decrease amounts to around 10 percent and is equal for the three groups of characteristics B/E, A/L, and H. The decrease of the design effect is caused by the reduction of the cluster-effect,

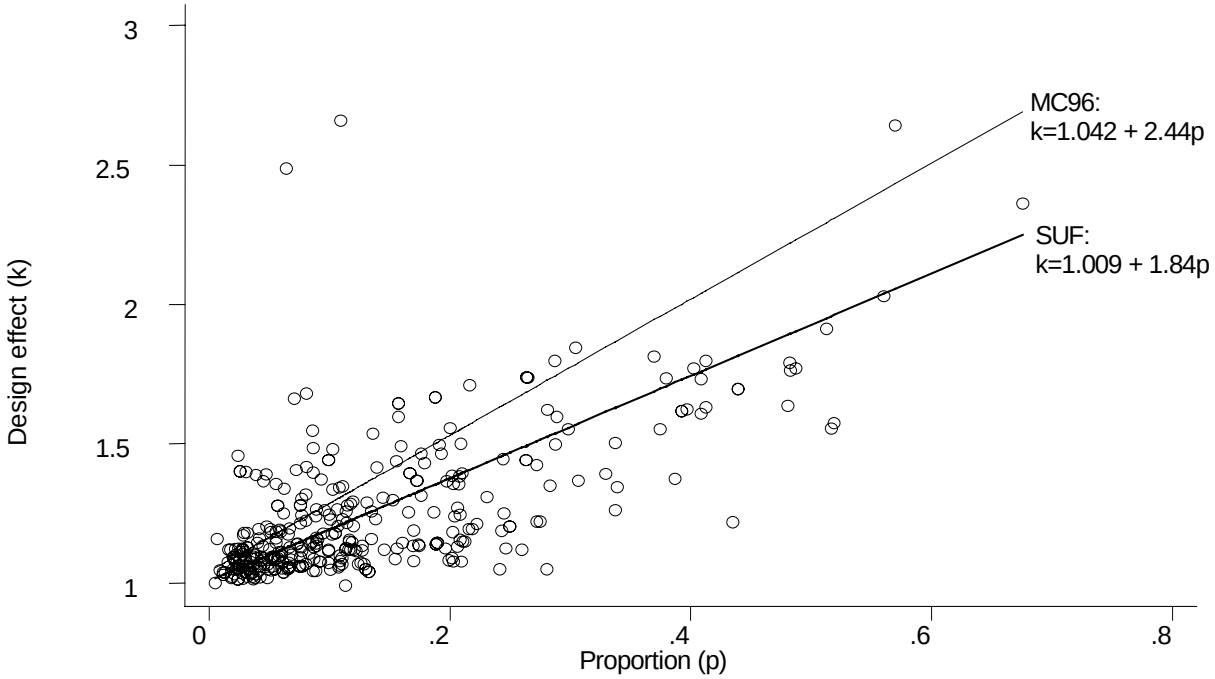
[10] Results on the sampling error for the German Microcensus 1990 (Statistisches Bundesamt 1998a: 22). Own calculations based on unpublished results on sampling error for the German microcensus 1996 (Statistisches Bundesamt 1998b).

which, in turn, is due to the selection of the subsample. So far, the users of the SUF have been dependent on the design effects of the MC90. Without regard to the fact that those coefficients are actually obsolete and only refer to the former Federal Republic of Germany, Table 3 shows that the values of the MC90 are a convenient approximation for the coefficient of variation of the SUF.

Table 3: Comparison of Linear Regression Coefficients of the Design Effect on the Estimated Proportion of the Population

Group	Data Base	Constant	Slope
Population and Employed Persons (B/E)	SUF	1.009	1.84
	MC96	1.042	2.44
	MC90	1.136	1.61
Foreigners and Employed Persons in Agriculture (A/L)	SUF	1.088	21.69
	MC96	1.162	25.47
	MC90	1.169	25.04
Households (H)	SUF	0.988	1.01
	MC96	1.009	1.60
	MC90	1.119	1.14

Figure 1: Regression of the Design Effect (k) on the Percentage of the Population for 346 Characteristics of the Type of Population and Employed (B/E)



It still must be checked, how well the linear regression matches the calculated design effects.[11] Figure 1 depicts the regression of the design effect for the characteristics of the standard tabulation programme for the population and economically active persons (B/E). Generally, the linear approximation provides a useful model for the description of the design effect. However, in some cases there are significant deviations of the respective design effects from the estimated values. Hence, in some cases, the use of design effects leads to considerable over- and underestimations of the variance.

7 Concluding remarks

In this paper we have investigated some possibilities of variance estimation based on the available design information for the scientific use file of the German microcensus 1996. In closing, we conclude that the methods, as described in the paper, can be easily applied and give reasonable results, compared with those of the German Federal Statistical Office for the original microcensus data. With respect to the design effect, formerly imperative as a tool for variance estimation, we have shown that this method is connected with considerable over- and underestimation of the standard error. Users of the scientific use file no longer depend on this approximation method since the release of anonymised design information. Consequently, scientific use files should ideally be released with the design information relevant for an appropriate variance estimation; such as variables for stratum and cluster, as well as nonresponse and post-stratification weights. The reasons for the calculation of direct estimates of sampling errors using design information that comes with the file appear to be true not only for German microdata, but also for a number of international surveys.

The estimation methods discussed in this paper have been classically restricted to the sampling error that assumes an ideal realisation of the sampling plan and an exact collection of information in the survey etc. While neglecting non-sampling errors that are always present in survey practise, our analyses have pointed to such systematic sources of error. Particularly the large differences between estimated totals with and without post-stratification adjustment on the current updating of the population point out that there is great need for systematic research on the quality of statistical data. Compared to the difference of the totals, the reduction of the standard error, which is due to the weighting adjustment, is only of minor importance. All in all, the weighting adjustment hardly leads to a reliable gain in precision; rather, it hides

[11] So far, only the respective regression coefficients for the MC90 have been published. A verification of the goodness of fit of the simple regression model has not been documented, except for the evidence that the variances of the estimated design effects from the calculated ones average less than 15 to 20 percent (Statistisches Bundesamt 1998a: 17).

a bias. Given the state of information, it remains an open question whether the microcensus produces biassed population estimates or the benchmark data provides biassed estimates.

Against the background that in Germany the next census will not be conducted as a classical enumeration of the population, and different sources of data will be united in a register-based system instead – among others the results of the microcensus – questions about data quality are of central importance. However, it is not only for the statistical offices to act. To enable methodological studies on the part of academic research, it is desirable to provide more information on data collection procedure and survey design in the data file.

Acknowledgements

This paper has its origin in a stay of the first author as a guest professor at ZUMA in October 1999. We thank Wolf Bihler (German Federal Statistical Office), Ralf Münnich (University of Tübingen) as well as Siegfried Gabler, Sabine Häder and Michael Wiedenbeck (ZUMA) for stimulating discussions. We extend our thanks to Ulrich Pötter and Götz Rohwer for valuable suggestions on a previous version of the paper. We also thank Noelle Crist-See and Jörg Müller for their assistance in preparing this English version.

References

BLS/BOC (Bureau of Labor Statistics/U.S. Census Bureau). 2000. *Current Population Survey. Technical Paper 63. Design and Methodology*. <http://www.census.gov/prod/2000pubs/tp63.pdf>.

Eltinge, J.L. 1999. *Use of Stratum Mixing to Reduce Primary-Unit-Level Identification Risk in Public-Use Survey Datasets*. Paper presented at American Statistical Association, Committee on Privacy and Confidentiality. Session at the 1999 Joint Statistical Meeting. <http://www.fcsm.gov/papers/eltinge.pdf>.

Heidenreich, H.-J. 1989. Erwerbstätigkeit im April 1988. *Wirtschaft und Statistik* (6): 327-339.

Heidenreich, H.-J. 1994. Hochrechnung des Mikrozensus ab 1990. Pp. 112-123 in *Gewichtung in der Umfragepraxis*, edited by S. Gabler, J.H.P. Hoffmeyer-Zlotnik and D. Krebs. Opladen: Westdeutscher Verlag.

Holmes, D.J. and Skinner, C.J. 2000. *Variance Estimation for Labour Force Survey Estimates of Level and Change*. Government Statistical Service Methodology Series No. 21. London: Office for National Statistics.

Jäger, M. 1992. Im Westen was Neues? - Im Osten was Besseres? Möglichkeiten der Nutzung von Daten der Einwohnermelderegister für statistische Zwecke. Pp. 103-124 in *Volkszählung 2000 - oder was sonst?* Band 21 der Schriftenreihe Forum der Bundesstatistik, edited by Statistisches Bundesamt. Stuttgart: Metzler-Poeschel.

Krug, W., Nourney, M. and Schmidt, J. 1999. *Wirtschafts- und Sozialstatistik. Gewinnung von Daten*. München: Oldenbourg.

Lüttinger, P. and Riede, T. 1997. Der Mikrozensus: amtliche Daten für die Sozialforschung. *ZUMA-Nachrichten* Nr. 41: 19-43. <http://www.gesis.org/Dauerbeobachtung/Mikrodaten/documents/doc/zn41.pdf>.

Meyer, K. 1994. Zum Auswahlplan des Mikrozensus ab 1990. Pp. 106-111 in *Gewichtung in der Umfragepraxis*, edited by S. Gabler, J.H.P. Hoffmeyer-Zlotnik and D. Krebs. Opladen: Westdeutscher Verlag.

Müller, W., Blien, U., Knoche, P. and Wirth, H. 1991. *Die faktische Anonymität von Mikrodaten*. Band 19 der Schriftenreihe Forum der Bundesstatistik, edited by Statistisches Bundesamt. Stuttgart: Metzler-Poeschel.

ONS (Office for National Statistics). 1998. *Labour Force Survey User Guide, Volume 1: Background and Methodology*. London: Office for National Statistics.

Rendtel, U., Schimpl-Neimanns, B. 2001. Die Berechnung der Varianz von Populationsschätzern im Scientific Use File des Mikrozensus ab 1996. *ZUMA-Nachrichten* Nr. 48: 85-116. <http://www.gesis.org/Publikationen/Zeitschriften/ZUMA_Nachrichten/documents/pdfs/zn48_10-bernhard.pdf>.

Särndal, C.-E., Swensson, B. and Wretman, J. 1992. *Model Assisted Survey Sampling*. New York: Springer.

Schimpl-Neimanns, B., Rendtel, U., 2001: *SAS-, SPSS- und STATA-Programme zur Berechnung der Varianz von Populationsschätzern im Mikrozensus ab 1996*. ZUMA-Methodenbericht 2001/04. <http://www.gesis.org/Publikationen/Berichte/ZUMA_Methodenberichte/documents/pdfs/tb01_04.pdf>

Statistisches Bundesamt. 1998a. *Fachserie 1, Bevölkerung und Erwerbstätigkeit. Reihe 4.1.1, Stand und Entwicklung der Erwerbstätigkeit 1996 (Ergebnisse des Mikrozensus)*. Stuttgart: Metzler-Poeschel.

Statistisches Bundesamt. 1998b. *Fehlerrechnung Mikrozensus 1996 (nach Kompensation der bekannten Ausfälle)*. Wiesbaden (unpublished tables).

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