SCB STATISTISKA CENTRALBYRÅN

1986-10-14

PROMEMORIOR FRÅN U/STM

NR 25

TWO EVALUATION STUDIES OF SMALL AREA ESTIMATION METHODS: THE CASE OF ESTIMATING POPULATION CHARACTERISTICS IN SWEDISH MUNICIPALITIES FOR THE INTERCENSAL PERIOD.

AV SIXTEN LUNDSTRÖM

INLEDNING

TILL

Promemorior från U/STM / Statistiska centralbyrån. – Stockholm : Statistiska centralbyrån, 1986. – Nr 25-28.

Föregångare:

Promemorior från P/STM / Statistiska centralbyrån. – Stockholm : Statistiska centralbyrån, 1978-1986. – Nr 1-24.

Efterföljare:

R & D report : research, methods, development, U/STM / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1987. – Nr 29-41.

R & D report : research, methods, development / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1988-2004. – Nr. 1988:1-2004:2.

Research and development : methodology reports from Statistics Sweden. – Stockholm : Statistiska centralbyrån. – 2006-. – Nr 2006:1-.

Promemorior från U/STM 1986:25. Two evaluation studies of small area estimation methods: the case of estimating population characteristics in Swedish municipalities for the intercensal period / Sixten Lundström.
Digitaliserad av Statistiska centralbyrån (SCB) 2016.

urn:nbn:se:scb-PM-USTM-1986-25

SCB STATISTISKA CENTRALBYRÅN

1986-10-14

PROMEMORIOR FRÅN U/STM

NR 25

TWO EVALUATION STUDIES OF SMALL AREA ESTIMATION METHODS: THE CASE OF ESTIMATING POPULATION CHARACTERISTICS IN SWEDISH MUNICIPALITIES FOR THE INTERCENSAL PERIOD.

AV SIXTEN LUNDSTRÖM

Preface

The Register of the Total Population (RTP) gives yearly population statistics for municipalities and parts of municipalities, but does not provide any information on households. The census (FoB) is the only investigation that can give the desired regional breakdowns, but the census is conducted only once every fifth year. Municipal planners, in particular, are in need of timely data on households more often than the census can provide.

This report presents two studies on methods of providing intercensal data on households (the work was done within U/STM's project "Sampling and Estimation"). The first study's purpose was to develop an estimator of the number of cohabiting persons in each municipality. In the other study, efforts have been made to develop an estimator for the number of households of different sizes in each municipality. Both studies use data from the most recent census, from the RTP, and from a sample survey (the Labor Force Survey in the one study and the Survey of Living Conditions in the other).

The work on the estimator of the number of cohabiting persons is finished, a computer program has been developed, and the first set of estimates have been sent out. The work on the other estimator is not yet finished, but it is our intention that it too will be put into use.

The two articles that comprise this report describe the completed studies and are partially overlapping in that the first article is summarized in chapter four of the second article. These reports were written for and presented at two different conferences. The first "An Evaluation of Small Area Estimation Methods: The Case of Estimating the Number of Non-Married Cohabiting Persons in Swedish Municipalities" was presented at a conference in Ottawa, Canada. It will be published in Small Area Statistics: An International Symposium, Ottawa, May 22-24, 1985, edited by J.N.K. Rao, C.E. Särndal, and M.P. Singh. The other article, "Estimating Population Characteristics and Households in Swedish Municipalities Using Survey and Register Data" was presented as an invited paper at the US Census Bureau's Second Annual Research Conference and can also be found in the conference's Proceedings.

An Evaluation of Small Area Estimation Methods: The Case of Estimating the Number of Non-married Cohabiting Persons in Swedish Municipalities.

CONTENTS

1985-08-20

An Evaluation of Small Area Estimation Methods: The Case of Estimating the Number of Non-married Cohabiting Persons in Swedish Municipalities

Sixten Lundström, Statistics Sweden

1 Abstract

The Swedish Population and Housing Census gives, every five years and for every municipality, the number of non-married cohabiting persons. Such information, however, is almost completely lacking for the years between the censuses, despite a strong demand. The paper presents, in that context, an evaluation of some alternative small area estimation techniques presented in the statistical literature.

By means of Monte Carlo simulations, asymptotically design unbiased estimators (direct, poststratified and generalized regression estimators) and model-dependent, design biased estimators (synthetic, SPREE and composite estimators) are compared with respect to the square root of the relative mean square error, the standard error and the design bias.

2 Introduction

With five years intervals, the Swedish Population and Housing Census provides the quantity $N_{.iq}$ which denotes the number of not married ("not married" refers to unmarried, divorced, widows and widowers)

people in municipality q (q=1,...,277) belonging to class i (i=1,2) with respect to cohabitational status (cohabiting, not cohabiting). Despite a strong demand, such information for years between censuses is almost completely lacking.

It is true that the Register of the Total Population (RTP) provides the number of married persons, but since nearly 11 per cent of the adults are cohabiting without being married the RTP-information is of limited value. Moreover, the Survey on Living Conditions (SLC) provides a yearly national estimate, but the sample size is too small (3.400 persons) to give an acceptable estimate for each municipality. The expected number of observations is less than 10 for about 70 per cent of the 277 municipalities.

It is not feasible to include the characteristic of cohabitation in the RTP, and it is too costly to increase the sample size in the SLC to yield the statistics we seek. Therefore we try to develop estimators that combine data from different sources. In the literature we find several examples on rather successful attempts to use such model-dependent estimators as the synthetic estimator and the SPREE estimator.

Most small area estimators in use introduce a third classification of the population into H mutually exclusive and exhaustive classes, which can be based on age, sex, income, etc; they are labelled h=1,...,H. The population will thereby, in our case, be completely cross-classified into 277x2xH cells with unknown cell sizes N_{hiq}, representing the number of not married people in municipality q with cohabitational status i, belonging to sex- and age-group h.

The statistical problem is to estimate the quantities $N_{.iq} = \sum_h N_{hiq}$.

The computerized RTP register provides each month current information on the number of not married people in municipality q belonging to sex. and age-group h, $N_{h.q}$ ($= \sum_i N_{hiq}$). Moreover, we have current sample information from the SLC, which can be used to form estimates of $N_{hi.}$ ($= \sum_q N_{hiq}$). At the intercensal period we also know N'_{hiq}, which denotes the number of not married people in cell hig at the latest census.

The available data are rather extensive, and conditions are thus favourable for producing small area statistics.

In order to measure both the sampling error and the bias, Monte Carlo simulations are carried out. (Some of the findings of this paper have been published in Lundström (1984).)

3 Description of the Monte Carlo simulations

3.1 The design of the study

The Monte Carlo simulations are designed with the purpose of studying estimators for the totals $N_{.iq}$ at the 1980 Census period. Thus, we know the parameter values and can for each sample repetition compare the estimates with the parameters and compute different quality measures.

The latest complete data, N'_{hiq}, will be retrieved from the 1975 Census. This implies that the time between the computation of N'_{hiq} and the estimates $\hat{N}_{.iq}$ is at most five years. The current information $N_{h.q}$ is in the study taken from the 1980 Census - not from the RTP. However, this will not affect the results because the difference between the RTP and the Census is negligibe with respect to this quantity.

The sample information in the simulation study is obtained in the following way:

The interval (0,1) is divided into parts where the part corresponding to the cell hiq has the length $N_{hiq}/N_{...}$. Then $n_{...}$ random numbers are drawn from a variable uniformly distributed on (0,1). Each random number is located in one and only one cell and n_{hiq} is obtained.

The SLC is based on a stratified sample with systematic sampling within strata. The population is divided into only two strata according to age. Among older people a larger fraction is drawn than among younger people. Thus the sampling design of the simulation study is different from that of the SLC, but this will not to any large extent affect the evaluation.

There are some differences between the SLC and the census in the definition of study variable, reference period and data collection methods, which lead to a deviation between the expected SLC-value and the census-value. These differences are not reflected in the simulation study.

3.2 Estimators under study

In a first part of the study six estimators are examined. Three of them are approximately design unbiased (ADU), while the others are model-dependent and hence design biased. In a second part we also examine a composite estimator, which is a weighted function of an ADU estimator and a model-dependent estimator.

It should be observed that the estimators to be studied utilize different amounts of auxiliary information. Since an estimator which incorporates strong auxiliary information is expected to outperform an estimator which only incorporates weak such information (or none at all), some of the estimators under study could a priori have been excluded. However, we think it is of interest to measure the effect of different types and amounts of information.

(i) If neither the information N'_{hiq} nor the associated variable (age/sex) is used, the Horvitz-Thompson estimator is an obvious (but poor) candidate:

$$HT = \frac{N_{...}}{n_{...}} n_{.iq} \qquad (1)$$

(ii) By making use of the known quantities $N_{h.q}$ the following poststratified estimator is close at hand:

$$PST = \sum_h \frac{N_{h.q}}{n_{h.q}} n_{hiq} \qquad (2)$$

Obviously, this estimator is not defined for every possible sample, viz. when $n_{h.q} = 0$. One solution often recommended in literature is to merge two or more poststrata. In the present simulation study, however, we have adopted a computationally simpler method: if, for some h, a zero count ($n_{h.q} = 0$) is realized, the corresponding term is dropped when summing over h. This rule makes the PST estimator biased, especially for small sample sizes.

(iii) In a generalized regression approach Särndal (1981) has developed an ADU estimator:

$$DM = \sum_h \{\frac{N_{h.q}}{n_{h..}} n_{hi.} + \frac{N_{...}}{n_{...}} (n_{hiq} - n_{h.q} \frac{n_{hi.}}{n_{h..}})\} \qquad (3)$$

(The estimator takes both the design and the model into consideration, therefore the abbreviation DM.)

This estimator is in effect the synthetic estimator (see SYNT, below) corrected för design bias (SYNT minus estimated design bias), and it can also be written:

$$DM = \sum_h \{\frac{N_{h.q}}{n_{h..}} n_{hi.} + \hat{N}_{h.q} (\frac{n_{hiq}}{n_{h.q}} - \frac{n_{hi.}}{n_{h..}})\} \text{ , where} \qquad (4)$$

$$\hat{N}_{h.q} = \frac{N_{...}}{n_{...}} n_{h.q}$$

Remark 1. If we in (4) exchange $\hat{N}_{h.q}$ for $N_{h.q}$, the DM estimator is transformed into the PST estimator. Hence, the larger the sample size, the smaller the difference between the two estimators.

Remark 2. With small sample sizes the DM estimator can give negative estimates. Such estimates would hardly be accepted in a real-life application, since they are not in the parameter space. (If negative estimates are replaced by e.g. zero, the DM estimator will be biased but have smaller variance.) In the present simulation study, however, negative estimates are accepted.

The above three estimators are all asymptotically design unbiased. We will now turn to some estimators which lack this appealing large sample property. However, they have other merits, which make them strong competitors. These latter estimators can all be considered as special cases of a class of estimators (Structure Preserving Estimates - SPREE) proposed by Purcell (1979) in a categorical data analysis approach. One feature of this approach is the implicit assumption of a super-population model for the behavior of small domain frequencies over time.

In short, a SPREE estimator is defined through an adjustment of data from a previous point of time to given current marginal totals, while at the same time as far as possible preserving the structure of interaction between the variables as established at the previous point of time. Special cases will be spelled out in more detail below.

(iv) When at least two current margins are known, the SPREE estimators are the result of an iterative proportional fitting (IPF) procedure. In the present context this means that we have access to the current margins $\hat{N}_{hi.}$ $(= \frac{N_{h..}}{n_{h..}} n_{hi.})$ and $N_{h.q}$ (as well as to the complete data set N'_{hiq} from the previous point of time). In this situation the IPF algorithm can be written:

The known previous data N'_{hiq} are taken as initial proxies, i.e.

$$\hat{N}^{(0)}_{hiq} = N'_{hiq} \tag{5}$$

At the k^{th} iteration we compute

$$_1\hat{N}^{(k)}_{hiq} = \frac{\hat{N}^{(k-1)}_{hiq}}{\hat{N}^{(k-1)}_{hi.}} \hat{N}_{hi.} \quad , \text{ where } \hat{N}^{(k-1)}_{hi.} = \sum_q \hat{N}^{(k-1)}_{hiq}$$

and

$$\hat{N}^{(k)}_{hiq} = \frac{_1\hat{N}^{(k)}_{hiq}}{_1\hat{N}^{(k)}_{h.q}} N_{h.q} \quad , \text{ where } _1\hat{N}^{(k)}_{h.q} = \sum_i {}_1\hat{N}^{(k)}_{hiq} \tag{6}$$

The iterative process is continued until some convergence criterion is satisfied (assume that this will happen when $k=k_0$); finally the SPREE estimate is calculated as

$$SPR = \sum_h \hat{N}^{(k_0)}_{hiq} \tag{7}$$

(v) If the complete previous data, N'_{hiq}, and the current margin $\hat{N}_{hi.}$ but not the $N_{h.q}$ are known we obtain the SPREE estimator

$$SYNT1 = \sum_h \frac{N'_{hiq}}{N'_{hi.}} \hat{N}_{hi.} \quad , \text{where } \hat{N}_{hi.} = \frac{N_{h..}}{n_{h..}} n_{hi.} \tag{8}$$

This estimator will result from an IPF-procedure where only step 1 in each cycle is performed. It can also be derived directly as a solution of the minimizing problem associated with SPREE estimation.

(vi) If only the current information $\hat{N}_{hi.}$ and $N_{h.q}$ is known, the SPREE estimator is given by

$$SYNT = \sum_h \frac{N_{h.q}}{N_{h..}} \hat{N}_{hi.} \tag{9}$$

This estimator is the result of the IPF-procedure when all initial proxies have the same value.

3.3 Results

The SPREE estimators suffer from sampling error and design bias and therefore the quality measure must include these two quantities. The quality measure we have chosen to estimate in the simulation study is the square root of the relative mean square error

$$(\text{rel-MSE})^{1/2} = \{ E(100 \frac{\hat{N}_{.iq} - N_{.iq}}{N_{.iq}})^2 \}^{1/2} \tag{10}$$

We also estimate the relative standard error

$$\text{rel -s.e.} = \{ E(100 \frac{\hat{N}_{.iq} - E(\hat{N}_{.iq})}{N_{.iq}})^2 \}^{1/2} \qquad (11)$$

and the relative bias

$$\text{rel-bias} = E(100 \frac{\hat{N}_{.iq} - N_{.iq}}{N_{.iq}}) \qquad (12)$$

These quantities are associated in the following way

$$\text{rel-MSE} = (\text{rel -s.e.})^2 + (\text{rel-bias})^2 \qquad (13)$$

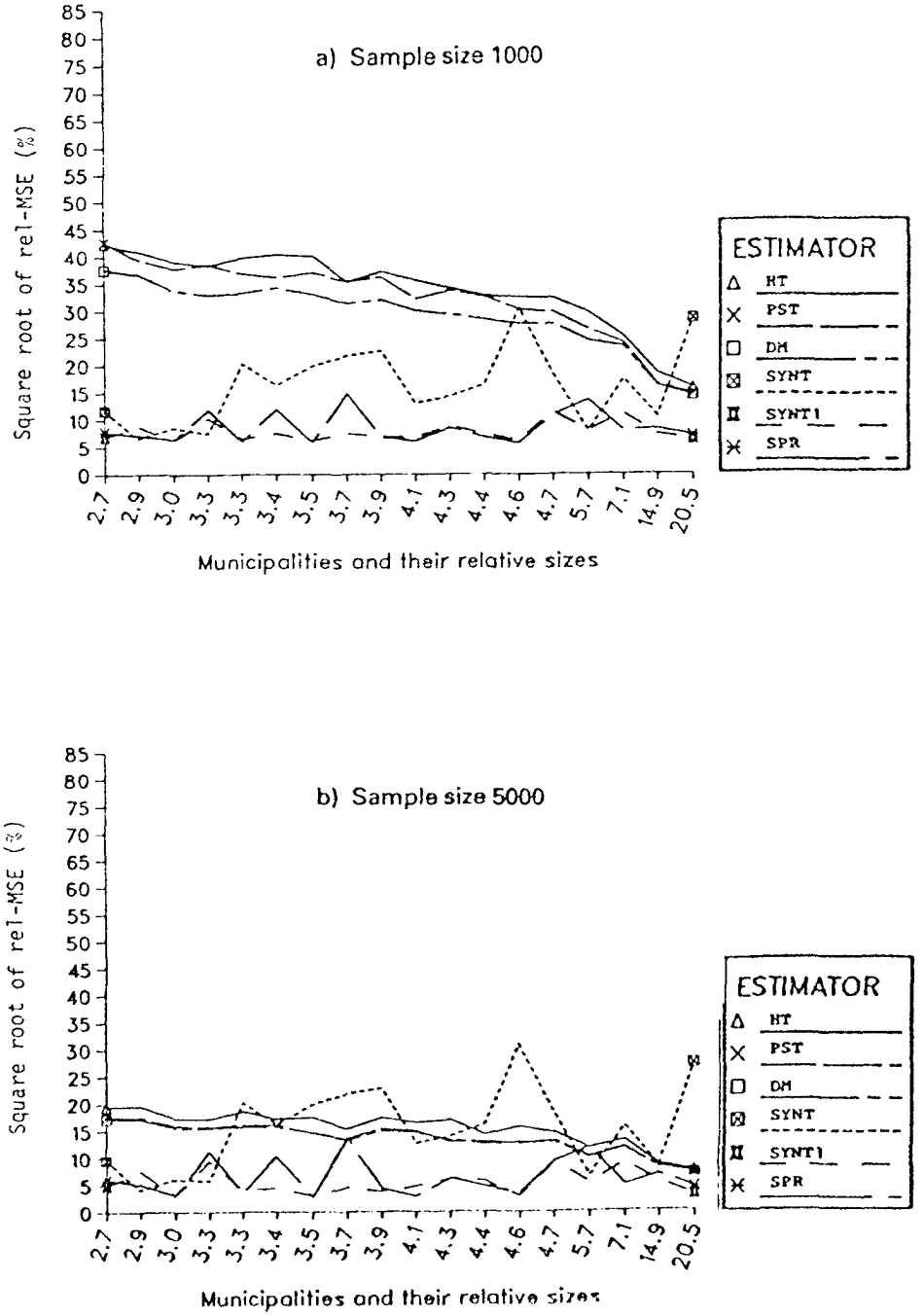
The variable of interest consists of only two categories. Since we know the sum $N_{..q}$ ($= N_{.1q} + N_{.2q}$) we only present the results for the estimation of $N_{.1q}$, the number of non-married cohabiting people in each municipality.

The budget does not allow the inclusion of all 277 municipalities. Instead we examine several minipopulations. The largest of these consists of 55 municipalities. Our two Monte Carlo simulations consist of: a) the selection of 400 samples, each of size $n... = 1\ 000$; b) the selection of 200 samples of size $n... = 5\ 000$.

The results expressed in terms of the square root of rel-MSE are presented in figures 1a-1b where the minipopulation consists of the 18 largest municipalities; simulations for other minipopulations show a similar quality picture.

Figure 1a-1b

Square root of rel-MSE. Minipopulation: The 18 largest municipalities.



The figures show that the ADU estimators have a larger $(\text{rel-MSE})^{1/2}$ than the model-dependent estimators, when the sample size is 1000. Only the usual synthetic estimator, SYNT, has for some municipalities a larger $(\text{rel-MSE})^{1/2}$ than the ADU estimators.

When comparing the biased estimators, the importance of the auxiliary information appears quite clearly. If the previous information N'_{hiq} is lacking (i.e. the SYNT estimator) the problem with a large bias is obvious. The SPR estimator uses the current data, $N_{h.q}$, in contrast to the SYNT1 estimator, but the figures (and also table 1) show that $N_{h.q}$ has no clear positive effect on the results.

The annual survey SLC mentioned earlier in this paper has a sample which contains about 1000 not married persons from the 18 largest municipalities. If we want estimates for these municipalities, and if we believe that $(\text{rel-MSE})^{1/2}$ is a relevant measure of the quality, then figure 1a indicates that we ought to choose SYNT1 or SPR.

When the sample size is increased, the ADU estimators and the biased estimators are brought closer to each other. At the sample size 1000, the DM estimator has a smaller $(\text{rel-MSE})^{1/2}$ than PST, but at the sample size 5000 the difference is negligible.

The biased estimators change just slightly when we increase the sample size from 1000 to 5000, which means that the bias is the dominating error even in the case with the smaller sample size.

Also, simulations have been carried out for two different samples of 18 municipalities (denoted "Sample 1" and "Sample 2") and for the 55 largest municipalities. In table 1 we present the average $(\text{rel-MSE})^{1/2}$ when the sample size is 1000.

Table 1

Average root of rel-MSE for different minipopulations when the sample size is 1000.

	Estimator					
Minipopulation	HT	PST	DM	SYNT	SYNT1	SPR
The 18 largest	33.9	32.2	29.3	16.3	7.8	8.5
"Sample 1"	39.8	37.1	34.9	21.2	13.2	11.3
"Sample 2"	33.7	30.9	28.4	12.4	11.5	11.7
The 55 largest	57.1	51.9	50.1	14.3	9.3	9.6

Table 1 shows that if you are looking for an approximately design unbiased estimator, the DM estimator will be a good choice. If your only demand is to have an estimator with a small average mean square error and you only have the current information $\hat{N}_{hi.}$ and $N_{h.q}$ then you should choose SYNT.

If you also have access to previous data, N'_{hiq}, the choice is less straightforward. One advantage with SPR compared to SYNT1 is that SPR directly provides estimates which add up to the current marginal totals $N_{h.q}$. On the other hand, SPR is more complicated to calculate than SYNT1.

The size of the ratio between the bias and the standard error is decisive for the possibility of computing confidence intervals. In table 2 below this ratio for the biased (the relative bias for the HT and DM estimators is never important) estimators is displayed.

Table 2

Ratio between the absolute value of the bias and the standard error for the biased estimators when the sample size is 1000. Minipopulation: "Sample 1".

Municipality	Estimator PST	SYNT	SYNT1	SPREE	Relative size (%) of the municipality
1	.32	4.09	1.70	1.27	0.6
2	.13	6.93	2.55	.52	1.0
3	.29	.27	6.60	2.69	1.0
4	.30	1.44	4.82	4.08	1.4
5	.39	5.36	1.09	.88	1.6
6	.19	2.16	.88	.48	1.9
7	.17	4.90	2.93	1.94	2.7
8	.09	.32	1.54	1.89	2.8
9	.01	1.42	3.63	4.25	3.1
10	.04	5.57	.81	.28	3.8
11	.01	.90	2.08	.72	4.6
12	.07	6.10	3.54	2.96	5.1
13	.00	4.64	.91	1.66	6.9
14	.06	4.88	1.61	.41	8.1
15	.02	1.14	.42	.74	9.3
16	.01	.83	.58	.60	11.5
17	.00	2.67	.34	.79	15.5
18	.01	1.29	.89	1.66	19.0

Table 2 shows that PST has a bias which, in general, can not be neglected when confidence intervals are computed. The bias is caused by the rule used in the present simulation study when $n_{h.q} = 0$. However, the model-dependent estimators are much more affected by the bias, and hence a conventionally computed confidence interval will be quite misleading as a quality measure.

4 Monte Carlo simulations for composite estimators

The ADU estimators are approximately design unbiased but suffer from large sampling errors, while the model-dependent estimators are design biased with small sampling errors. In the literature there are several examples of attempts to construct composite estimators aiming to combine the strenghts of each group of estimators (e.g. Schaible, Brock and Schnack (1977)). They have in some cases performed well.

The estimators SYNT1 and SPR have a smaller relative mean square error than the other estimators examined in section 3. Therefore it seems reasonable to choose one of those two. However, due to wider general applicability we will concentrate on estimators based on current data only. We have chosen the poststratified estimator, PST, and the synthetic estimator, SYNT.

One simple type of composite estimator can be written

$$\hat{N}_{.iq} = C_q PST + (1-Cq)SYNT \qquad (14)$$

where C_q is an a priori fixed weight.

It is easily shown that the weight which minimizes the mean square error of (14) is

$$C_q^* = \frac{E\{ (SYNT-N_{.iq})(SYNT-PST) \}}{E(PST-SYNT)^2} \qquad (15)$$

which may be rewritten as

$$C_q^* = \frac{MSE(SYNT) - E\{ (SYNT-N_{.iq})(PST-N_{.iq}) \}}{MSE(PST) + MSE(SYNT) - 2E\{ (SYNT-N_{.iq})(PST-N_{.iq}) \}} \qquad (16)$$

Obviously, the relevant quantities in expression (16) are not easily assessed; this is especially so for the expected cross-product. However, if this latter quantity is negligible relative to MSE(SYNT), the optimal weight simplifies (we mainly follow suggestions in Schaible (1979)) to

$$C_q^{**} = \frac{1}{1 + MSE(PST)/MSE(SYNT)} \qquad (17)$$

The problem of determining the optimal weight is reduced to the assessment of the ratio MSE(PST)/MSE(SYNT).

Another type of composite estimates is

$$\hat{N}_{.iq} = \hat{C}_q PST + (1-\hat{C}_q)SYNT \qquad (18)$$

where $\hat{C}_q$ is a sample dependent weight. In an effort to arrive at a simple weight we have used the following crude lines of argument:

Suppose $MSE(PSI\,|n_{..q}) \doteq a/n_{..q}$, where a is a constant.
Further, suppose that we have $MSE(SYNT\,|n_{..q}) \doteq b$, where b is a constant. If so, we might - with (17) in mind - try the weight

$$\tilde{C}_q = \frac{1}{1 + \frac{a}{bn_{..q}}} = \frac{n_{..q}}{n_{..q} + a/b} \qquad (19)$$

If a/b can be approximated from e.g. simulation studies on populations similar to the one under study we have found a sample dependent weight.

The simulation study is designed in the same way as the study reported in section 3. However, we have only carried out one Monte Carlo simulation consisting of a selection of 400 samples, each of size $n...= 1000$. The minipopulation consists of a random sample of 18 municipalities that will be denoted "Sample 3" in the following.

In the simulation study we examine the composite estimator $\hat{N}_{.iq}$, using three different sets of weights.

a) C^*_q = optimal values for the current population

b) C_q = optimal values at the 1975 census; and

c) $\tilde{C}_q = 1/(\frac{105}{n_{..q}} + 1)$

where the value 105 (= a/b) is determined from the simulation study for the minipopulation "Sample 1" reported in section 3.

In the first step the optimal C^*_q-values in a) are estimated by unbiased estimation of the numerator and denominator of (15) by using the 400 estimates.

The same procedure is then repeated for the minipopulation from the 1975 census in order to estimate the C_q-values in (b).

In the second step $(\text{rel-MSE})^{1/2}$ for the composite estimators are estimated in the same way as for the estimators reported in section 3. The $\tilde{C}_q$-values in alternative (c) are calculated for each sample.

The results from the Monte Carlo study are presented in table 3.

Since the weight in alternative c) varies from one sample to another we have calculated the quantity $1/\{1 + 105/E(n_{..q})\}$ in order to give the reader some idea of the size of the weights actually used.

Table 3

Weights in the composite estimators and square root of rel-MSE.

Minipopulation: "Sample 3".

Municipality	Optimum weight (a)	Optimum weight in the 1975 census (b)	$\frac{1}{1+\frac{105}{E(n_{..q})}}$	(rel-MSE)$^{1/2}$ for the following estimators: Composite (a)	(b)	(c)	PST	SYNT	Rel. size (%) of the municipality
1	-0.04	0.07	0.05	9.2	11.9	11.0	75.4	9.5	0.6
2	-0.09	-0.02	0.05	14.3	14.8	16.6	74.4	15.2	0.6
3	-0.05	0.03	0.05	14.3	15.5	16.2	70.9	14.8	0.7
4	-0.03	-0.03	0.07	24.2	24.2	24.5	66.8	24.2	0.8
5	-0.06	0.07	0.07	13.5	15.7	15.5	67.0	14.0	0.8
6	-0.02	0.10	0.07	28.5	28.8	28.4	58.5	28.5	0.8
7	0.04	-0.05	0.07	27.7	27.6	27.7	61.4	27.6	0.8
8	0.07	-0.02	0.07	27.2	27.4	27.0	59.3	27.3	0.8
9	-0.03	-0.03	0.08	24.7	24.7	25.0	61.2	24.7	0.9
10	-0.02	-0.01	0.08	20.5	20.5	21.2	62.8	20.6	0.9
11	-0.03	-0.01	0.09	21.2	21.1	21.3	57.6	21.1	1.1
12	0.29	0.30	0.10	35.3	35.3	36.7	48.2	38.5	1.1
13	0.00	0.15	0.14	14.0	15.5	15.2	52.9	14.0	1.7
14	0.00	0.01	0.16	10.4	10.4	12.5	45.7	10.4	2.0
15	0.03	0.09	0.18	12.4	12.6	14.0	46.7	12.4	2.2
16	0.20	0.27	0.29	14.7	14.9	15.1	31.2	16.1	4.4
17	0.72	0.79	0.41	16.0	16.3	19.2	19.5	30.4	7.4
18	0.88	0.89	0.87	78.5	8.5	8.5	8.7	14.7	2.4

We notice that the synthetic estimator has a smaller $(\text{rel-MSE})^{1/2}$ than the poststratified estimator in all of the municipalities but the two largest. All three composite estimators also provides estimates that, reffering to $(\text{rel-MSE})^{1/2}$, to a large degree resemble that of the synthetic estimator in all municipalities but the two largest, where the composite estimator is superior.

The composite estimator based on (b)-weights is for most municipalities a better estimator than the estimator based on (c)-weights. However, notice that the use of alternative (c) provides a smaller $(\text{rel-MSE})^{1/2}$ than both PST and SYNT for six municipalities.

Looking at the weights and the $(\text{rel-MSE})^{1/2}$ we find that the composite estimator is rather insensitive to deviations from the optimum weight.

5 Conclusions and plans for future work

The results of the Monte Carlo simulations for small area estimation of the number of non-married cohabiting persons in municipalities have led to the following conclusions:

a) the model-dependent estimators are superior to the ADU estimators for "common" sample sizes with respect to $(\text{rel-MSE})^{1/2}$;

b) among the model-dependent estimators, SYNT1 and SPR show good potential;

c) the DM estimator seems to be superior to the alternative ADU estimators included in the study;

d) when only current information is available, the composite estimator seems to be a good choice.

Recently a question on cohabitional status has been included in the Swedish Labour Force Survey (LFS). The LFS is based on a much larger sample than the SLC, and therefore the conditions are more favourable for the ADU estimators. On the other hand, the larger sample size can also be used to reduce the mean square error of the model-dependent estimators. Ongoing work, not reported here, shows that to a large extent it is possible to reduce the bias of the model-dependent estimates by clustering the municipalities in homogeneous groups according to previous data. Estimates can then be calculated for the municipalities in each group. This will increase the standard error (you use only the part of the sample belonging to the group of municipalities), but this increase will - at least in some groups - be much smaller than the reduction of the bias. We are at present working on the problem of finding a suitable clustering strategy.

Finally, we intend to calculate and publish the small area estimates.

References

Drew, J.D., Singh, M.P. and Choudhry, G.H. (1982). Evaluation of small area estimation techniques for the Canadian Labour Force Survey. 1982 Proceedings of the Section on Survey Research Methods, American Statistical Association, pp 545-550.

Gonzales, M.E. (1973). Use and evaluation of synthetic estimates. 1973 Proceedings of the Social Statistics Section, American Statistical Association, pp 33-36.

Lundström, S. (1984). Estimation for small domains: Two studies using combined data from censuses, surveys and registers. Statistical Review 1984:2, Statistics Sweden, pp 119-126.

Purcell, N.J. (1979). Efficient small domain estimation: A categorical data analysis approach. Unpublished Ph.D. thesis, University of Michigan, Ann Arbor, Michigan.

Purcell, N.J. and Kish, L. (1979). Estimation for small domains. Biometrics 35, pp 365-384.

Schaible, W.L., Brock, D.B. and Schnack, G.A. (1977). An empirical comparision of the simple inflation, synthetic and composite estimators for small area statistics. 1977 Proceedings of the Social Statistics Section, American Statistical Association, 1017-1021.

Schaible, W.L. (1979). A composite estimator for small area statistics. Synthetic estimates for small areas. Synthetic Estimates for Small Areas, NIDA Research Monograph 24, pp 36-53.

Särndal, C.E. (1981). Framework for inference in survey sampling with applications to small area estimation and adjustment for nonresponse. Bulletin of the International Statistical Institute, 49 (1), pp 494-513.

Särndal, C.E. (1984). Design-consistent versus model-dependent estimation for small domains, JASA 79, pp 624-631.

Estimation Population Characteristics and Households in Swedish Municipalities Using Survey and Register Data

CONTENTS

1986-01-08

ESTIMATING POPULATION CHARACTERISTICS AND HOUSEHOLDS IN SWEDISH MUNICIPALITIES USING SURVEY AND REGISTER DATA

Sixten Lundström, Statistics Sweden

1 INTRODUCTION

The Swedish Population and Housing Census collects and reports population and household information every five years and for each municipality. Despite a strong demand for such information fresh data are nearly completely lacking for years between the censuses. Statistics Sweden does however conduct a number of sample surveys producing population and household data on an annual basis, but the samples are not large enough to yield acceptable estimates even for the counties. Our computerized Register of the Total Population (RTP) gives very accurate information for individuals, but it cannot identify households. Also, it contains only few variables and thus, it cannot provide the desired information even for individuals.

Marital status is one variable in the RTP, but as several persons in Sweden are cohabiting without being married the statistics deriving from the RTP-variable is quite irrelevant as description of actual cohabitation frequencies. There is a demand for information on the number of cohabiting persons in each municipality even for the years between the censuses and therefore we have, in that context, tried to develop an estimator. We have carried out Monte Carlo simulation studies in order to measure both the sampling error and the bias. On the basis of these studies we then refined the selected estimator and finally Statistics Sweden published the estimates.

In this paper we also present a study concerning the problem of estimating the number of households of different sizes in each municipality. The methodological work is, however, still in its infancy and more research is needed before a method can be put into practice.

2 DESCRIPTION OF THE AVAILABLE DATA SOURCES

The population and housing census gives, every five years (November 1st is reference time point) and for every municipality, both the number of cohabiting persons and the number of households of different sizes. For years between censuses we have mainly three data sources containing parts of the requested information which we describe below.

The Register of the Total Population (RTP) covers the whole population (the covering errors are very small) and contains variables such as sex, age, marital status and income. In the RTP one can also bring together persons married to each other and also connect children to their mother. One cannot, however, bring together persons cohabiting without being married (living as a married couple). Also, children that live at home and are more than 17 years old are registrered as single households.

Statistics Sweden publishes monthly demografic information from the RTP.

The Survey on Living Conditions contains, among many other variables, the study variables cohabitational status and household size. It provides every year estimates for the whole nation, but the sample size is too small (it contains approximately 3.400 non-married persons) to give acceptable estimates for each municipality. The expected number of observations is less than 10 for about 70 per cent of the 277 municipalities.

During each quarter a fourth of the sample is collected and therefore we have no fixed reference time point. Also, the census measure a more permanent living compared to the SLC. Moreover different data collection methods are used in the two data sources: In the census data are collected by mail and in the SLC mainly by personal visits. These are the main reasons why the estimates from the two data sources differ.

For example, among unmarried persons in November 1980 the SLC overestimated the number of cohabiting non-married persons compared to the census by about 20 per cent.

The Labour Force Survey (LFS) provides every month tabulations of labour force status by age, sex, marital status, broad occupational categories, and other characteristics. Since 1983 it also collects data about cohabitational status.

The sample includes each month about 18.000 persons drawn from the population aged 16-74 years. The sample is rather large, particullary if one combines samples from several months, but in spite of that it is not sufficient to provide accurate estimates for municipalities.

It is not realistic to expect any large change in the census, the RTP, the SLC or the LFS in terms of content, sample size and periodicity in the near future. Hence, none of the data sources alone will be sufficient to produce accurate estimates, but perhaps a method for combining their data could give acceptable estimates?

3 DISCUSSION OF SMALL AREA ESTIMATION METHODS

3.1 Notation and properties

For the particular problem of estimating, for each municipality, the number of non-married cohabiting persons and the number of households

of different sizes, respectively, some notation will be introduced. To this end, suppose that the population of size N consists of Q mutually exclusive and exhaustive small areas labelled by q = 1,...,Q. For each small area 'q', units are classified into H mutually exclusive and exhaustive associated variable classes. Moreover, suppose that the study variables categories, i = 1,...I, split the population along a third "dimension". This labelling gives a three-way cross-classification into HIQ cells with N_{hiq} population units in the hiq-th cell, and a sample count n_{hiq}. Aggregation across a subscript is indicated by replacing that subscript by a dot '.'; e.g. $N_{..q} = \sum_{hi}^{HI} \sum N_{hiq}$ is the population size for the q-th area. The sample aggregates n..q are defined similarly.

The small areas are, in both applications, equal to municipalities, but the associated variable and, of course, the study variable differ.

The ultimate aim of the two studies presented in this paper is to develop estimators for $N_{.iq}$ ($= \sum_{h}^{H} N_{hiq}$). To our help we have information about the number of persons in cell hiq in the latest census, which we denote N'_{hiq}. In addition, RTP provides the current number of persons in municipality q belonging to category h of the associated variable, i.e. $N_{h.q}$, and our sample survey gives a nation-wide estimate of the number of persons in cell hi; we denote it $\hat{N}_{hi.}$.

Small area estimators are usually model-dependent and thus, we are faced with potentially biased estimates; the bias arising whenever the assump-

tion about the model is not satisfied. One appropriate measure of the accuracy of the small area estimator $\hat{N}_{.iq}$ will then be the mean square error, i.e.

$$MSE\ (\hat{N}_{.iq}) = Var\ (\hat{N}_{.iq}) + B^2\ (\hat{N}_{.iq}), \qquad (1)$$

where $Var\ (\hat{N}_{.iq})$ is the variance and $B\ (\hat{N}_{.iq})$ the bias of the estimator $\hat{N}_{.iq}$

The usual probability (design) estimators, as e.g. the expansion estimator and the stratified or the poststratified estimator, are basically unbiased and in addition they have the appealing property that one can calculate their accuracy (e.g. by a confidence intervall) of the estimate from the sample. However, we know that the variances are unfortunately too large in the present context.

A model-dependent estimator, on the other hand, usually has small sampling variability but occasionally it suffers from a large bias.

When developing a model-dependent small area estimator the key issues are to find useful data sources, to select "optimal" associated variables and "optimal" categorizations of them and of course to seek a most effective estimation method when relying on these data. There is usually a conflict inherent in the attempt to minimize the mean square error because a decrease of the bias often leads to an increase of the variance.

One serious disadvantage of the model-dependent estimators is that one cannot obtain a reliable estimate of the mean square error based on the available data. It is therefore very important to evaluate the estimator before it is put into practice. In this paper we present attempts to investigate the properties of different small area estimators in two different cases; estimation of cohabitation frequencies and estimating the number of households of different sizes. For the latter case it must be stressed, however, that the research reported here is preliminary and many issues require further investigation before the method can be put into practice.

3.2 Some small area estimators

The efforts to investigate and develop small area estimators have increased greatly during the last decade. A comprehensive review of research in small area estimation is given by Purcell and Kish (1979).

Other examples of the research are given in Steinberg ed. (1979) and the forthcoming conference proceedings edited by Rao, Platek, Särndal and Singh. Several of the estimators and ideas investigated in the work presented in this paper emanate from Purcell (1979) and Schaible (1979).

In a categorical data analysis approach, Purcell (1979) develops a group of estimators which he denotes Structure Preserving Estimates (SPREE). Briefly, the SPREE methods consist of adjusting some known previous data to known current marginal totals, while in some way preserving, as far

as possible, the interaction structure between the variables as established in the previous data (minimizing a weighted sum of squared differences between the previous data and the estimates subject to the current marginal constraints). The varying degree of access to previous data and marginal constraints result in different estimators. In the present context we have complete previous data, N'_{hiq} (representing the association structure) and two current margins, $\hat{N}_{hi.}$ and $N_{h.q}$ (representing the allocation structure). To be able to solve the minimization problem an iterative procedure denoted Iterative Proportional Fitting (IPF) has to be used. The IPF-algorithm is described below.

At the initial step the starting values are put equal to the known previous data, i.e.

$$\hat{N}^{(0)}_{hiq} = N'_{hiq} \tag{2}$$

At the k^{th} iteration we compute

$$\hat{1}N^{(k)}_{hiq} = \frac{\hat{N}^{(k-1)}_{hiq}}{\hat{N}^{(k-1)}_{hi.}} \hat{N}_{hi.}, \text{ where } \hat{N}^{(k-1)}_{hi.} = \sum_q \hat{N}^{(k-1)}_{hiq} \tag{3}$$

and

$$\hat{N}^{(k)}_{hiq} = \frac{\hat{1}N^{(k)}_{hiq}}{\hat{1}N^{(k)}_{h.q}} N_{h.q}, \text{ where } \hat{1}N^{(k)}_{h.q} = \sum_i \hat{1}N^{(k)}_{hiq} \tag{4}$$

The iterative process is continued until some convergence criterion is satisfied (assume that this will happen when $k = k_0$); finally the SPREE estimate is calculated[1] as

$$SPR = \sum_h \hat{N}_{hiq}^{(k_0)} \qquad (5)$$

The SPREE estimator is model-dependent where the model consists of a set of assumptions about preserved association interactions. The theory behind the SPREE estimators is rather complicated and the details are therefore excluded from this presentation. Readers interested in the theory are recommended to study Purcell (1979).

If we do not have access to all the data mentioned above we can obtain other SPREE estimators. It is reasonable to expect that an estimator

---

1) The SPREE estimates are calculated by an APL-program.

using all information will be superior to the other estimators but, because of their wider applicability we also investigate other estimators.

If the complete previous data, N'_{hiq}, and the current margin $\hat{N}_{hi.}$ (but not the $N_{h.q}$) are known we obtain the SPREE estimator

$$SYNT1 = \sum_h \frac{N'_{hiq}}{N'_{hi.}} \hat{N}_{hi.} \qquad (6)$$

If only the current margins $\hat{N}_{hi.}$ and $N_{h.q}$ are known, the SPREE estimator is

$$SYNT = \sum_h \frac{N_{h.q}}{N_{h..}} \hat{N}_{hi.} \qquad (7)$$

Both SYNT and SYNT1 are model-dependent estimators and are sometimes denoted "synthetic estimators".

In a generalized regression approach Särndal (1981) developed an estimator which is asymptotically design unbiased,

$$DM = \sum_h \{\frac{N_{h.q}}{N_{h..}} \hat{N}_{hi.} + \frac{N_{...}}{n_{...}}(n_{hiq} - n_{h.q} \frac{n_{hi.}}{n_{h..}})\} \qquad (8)$$

The first term equals the SYNT estimator and the second one is an estimator of the bias of the SYNT estimator. The bias estimate will of course suffer from a large sampling variability and therefore the estimator more resembles the "classical" estimators, like the poststratified estimator, than the estimators described above.

Another appealing group of estimators is composite estimators, which usually consist of a weighted sum of a classical (design unbiased) estimator and a model-dependent estimator. A great problem associated with these estimators, though, is to find good weights.

4 ESTIMATING THE NUMBER OF NON-MARRIED COHABITING PERSONS IN EACH MUNICIPALITY

4.1 Summary of a Monte Carlo simulation study

This section is based on an article by the present author which will be published in SMALL AREA STATISTICS: AN INTERNATIONAL SYMPOSIUM, OTTAWA MAY 22-24, 1985 (Editors: J.N.K. Rao, Richard Platek, C.E. Särndal, and M.P. Singh), Wiley & Sons. The methodological work presented in the following sections is based on the results in this article and we therefore confine ourselves to describe the way of conducting the simulation study and to present the main findings.

The purpose of the study was to compare different small area estimation methods when estimating the number of non-married persons in municipality q belonging to cohabitational status i (cohabiting, not cohabiting), $N_{.iq}$, for years between censuses.

The associated variable, available from the RTP, that we thought would be most related to cohabitational status was sex and age in combination. We used the following age categorization; 16-24, 25-34, 35-44, 45-54 and 55-74 years.

The effects of both sampling error and design bias should be covered and therefore Monte Carlo simulations were carried out.

The study was designed to estimate the totals $N_{.iq}$, for the 1980 Census period. The previous data, N'_{hiq} were caught from the 1975 Census and $N_{h.q}$ from the Census 1980 (RTP and the census yield the same quantity). The sample information, n_{hiq}, was based on repeated simple random samples of size n from the 1980 Census. Thus, conditions were more favourable than in reality because we did not take into consideration the conceptual differences between the census and the available surveys (SLC and LFS). The surveys are, on the other hand, based on stratified samples and therefore will yield estimates with smaller sampling variability than the estimates in the study but this is shown to have no serious effect.

The Monte Carlo simulations consisted of: (i) a selection of 400 samples, each of size n = 1000; (ii) a selection of 200 samples of size n = 5000. The budget did not allow the inclusion of all 277 municipalities but instead we studied several minipopulations. The largest of these consisted of 55 municipalities. We estimated among other things the square root of the relative mean square error

$$(\text{rel-MSE})^{1/2} = \{ E(100 \frac{\hat{N}_{.iq} - N_{.iq}}{N_{.iq}})^2 \}^{1/2} \qquad (9)$$

The three SPREE-estimators[1], SPR, SYNT and SYNT1 were included in the study and also the DM estimator. To be able to compare with some well-known unbiased estimators a Horvitz-Thompson and a poststratified estimator were also included in the study. The Horvitz-Thompson estimator was

$$HT = \frac{N_{...}}{n_{...}} n_{.iq} \qquad (10)$$

and the poststratified estimator

$$PST = \sum_h \frac{N_{h.q}}{n_{h.q}} n_{hiq} \qquad (11)$$

---

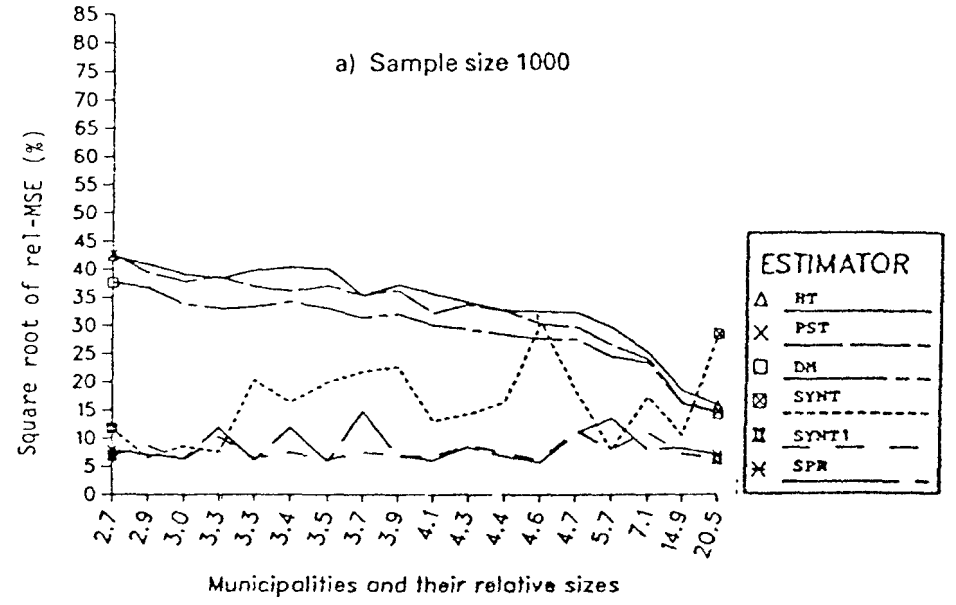
1) We used the following estimator of $N_{hi.}$: $\hat{N}_{hi.} = \frac{N_{h..}}{n_{h..}} n_{hi.}$

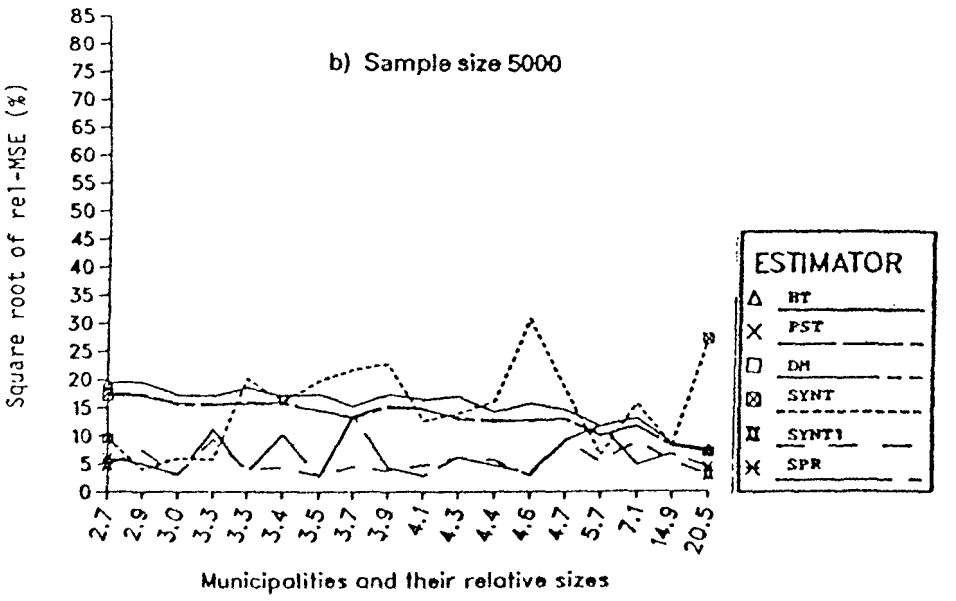
The variable of interest consists of only two categories: cohabiting, not cohabiting. Since we know the sum $N_{..q}$ ($= N_{.1q} + N_{.2q}$) we only present the results for the estimation of $N_{.1q}$, the number of non-married cohabiting people in each municipality.

The results expressed in terms of the square root of rel-MSE are presented in figures 1a-1b (will be found in the author's paper in Rao et.al. together with other results) where the minipopulation consists of the 18 largest municipalities; simulations for other minipopulations show a similiar picture as regards the quality of the estimates.

Figure 1a-1b

Square root of rel-MSE. Minipopulation: The 18 largest municipalities.





The figures show that the asymptotically design unbiased (ADU) estimators, i.e. the HT, the PST and the DM estimators have a larger $(\text{rel-MSE})^{\frac{1}{2}}$ than the model-dependent estimators, when the sample size is 1000. The SYNT estimator, however, has for some municipalities a larger error than the ADU estimators. When the sample size is increased, the ADU estimators and the biased estimators are brougth closer to each other. The biased estimators change just slightly when we increase the sample size from 1000 to 5000, which means that the bias is the dominating error even in the case with the smaller size.

Among the biased estimators the SYNT1 and the SPR estimators are superior to the SYNT estimator. It is difficult to order SYNT1 and the SPR estimator, which means that the current information $N_{h.q}$ has no significant effect on the result.

In another Monte Carlo simulation study the research concentracted on a composite estimator of the form

$$\hat{N}_{.iq} = C_q \text{ PST} + (1-C_q) \text{ SYNT}, \qquad (12)$$

where C_q, in one alternative, is an a priori fixed weight and in another alternative a sample dependent weight.

The study shows that the composite estimator, $\hat{N}_{.iq}$, can be a good estimator when one only has access to current data. However, the SPR

and the SYNT1 estimator are still the best candidates when all the data described in section 2 are available.

The SYNT1 estimate is easier to calculate than the SPR estimate but we think that the SPR estimate is more plausible, more relying on common sence and therefore more convincing to the user of statistics.

We have also calculated the mean absolute relative bias over all the 277 municipalities and then found that the SPR estimator has a smaller value than the SYNT1 estimator (5.96 compared to 6.25).

4.2 Refinements of the SPR estimator

We know from the simulation study that the bias (model error) is the dominating error for the SPR estimator and thus, the refinement efforts should be offered on bias reduction methods.

Implicitly, we have assumed in the simulation study that the municipalities are sufficiently homogeneous with respect to cohabiting status across the subgroups defined by sex/age, to justify the SPR estimator. Analysis of census data shows that this is not an accurate assumption and therefore we have tried to group the municipalities in some better way so that if the estimation is done within groups, then the assumptions underlying the estimates are more likely to be met. Obviously, the greater the number of groups we divide the municipalities into, the

greater we can make the homogenity within each group. However, there is a limit to the number of groups that should be formed, since the sampling error increases as the size of the group to which they belong decreases.

It is very difficult to obtain an optimal solution to the group formation problem. Firstly, it is hard to derive an analytic form for the variance and the bias of the SPR estimator and secondly, it is complicated to cluster the municipalities into homogeneous groups without obtaining some small clusters providing a large sampling error.

The Monte Carlo simulation study showed that the SPR estimator and SYNT1 estimator resemble each other not only with respect to the mean square error but also with respect to the variance and the bias. The SYNT1 estimator has an easier form than the SPR estimator and therefore we decided to study the effect of the grouping on the SYNT1 estimator and in that indirect way refining the SPR estimator.

The variance of the SYNT1 estimator has the following approximate form

$$\mathrm{Var}(\mathrm{SYNT1}) \doteq \sum_{h} \left(\frac{N'_{hiq}}{N'_{hi.}}\right)^2 \mathrm{Var}(\hat{N}_{hi.}) \qquad (13)$$

and the bias is

$$B(SYNT1) = \sum_h N'_{hiq}\left(\frac{N_{hi.}}{N'_{hi.}} - \frac{N_{hiq}}{N'_{hiq}}\right) \qquad (14)$$

Remark. The dot subscript is here used to denote summation over municipalities in the group.

We utilized the same data as in the simulation study, viz. the 1975 census provided N'_{hiq} and the 1980 census N_{hiq}. We calculated the expected number of observations from LFS in each group of municipalities and used this sample size in the computation of Var $(\hat{N}_{hi.})$.

It is not apparent which clustering techniques and which input data that should be used. As an initial tool we utilized a cluster analysis program in SAS (Statistical Analysis System) and tried different input data. After each clustering we calculated the mean square error, moved some municipalities from one cluster to another, calculated the mean square error again, and so on. We wanted to have a large mean effect and moreover no serious deterioration of the estimate for any municipality. When using $N_{h1q}/N_{.1q}$ as input data we attached the best result, which was a reduction of the mean (over municipalities) of the square root of rel-MSE by about 25 percent compared to the unclustered alternative. The municipalities were then clustered into four groups.

We have also tried to change the categorization of the age variable but this has not had any significant effect on the mean square error.

4.3 Publishing the estimates

The estimators we use at Statistics Sweden are basicly unbiased and usually we also calculate a confidence intervall and publish it in connection with the estimate. With the SPR estimator we have a quite new situation, where we know that the estimator is biased and where we cannot provide any probability statement about the accuracy. Therefore, we have entertained apprehensions about publishing the data. However, in June 1985 we distributed the results to the municipality planners and presented it as an "experiment". We told them how the estimators were constructed, which experiences we had about their shortcomings and we asked them to call us if they discovered some peculiarities. Up to now only a few of the planners have called us, but perhaps they will react when they obtain the Census 85 data.

5 ESTIMATING THE NUMBER OF HOUSEHOLDS OF DIFFERENT SIZES IN EACH MUNICIPALITY

5.1 Introduction

There is a strong demand for information about the number of households of different sizes in each municipality. The census yields every five

years such data but for years between two censuses there are no updated estimates available. The Survey on Living Conditions (SLC) provides nation-wide estimates on an annual basis but the sample size is too small to give acceptable estimates.

In this section we present an attempt to combine data from the latest census, the SLC, and the population register, the RTP, in order to estimate the number of households of different sizes in each municipality.

The study, and particularly the selection of the estimator, is mainly based on the results in the Monte Carlo simulation study reported in section 4. The available auxiliary information and the results in that study speak in favour of a SPREE estimator of a SPR type. The kernel of the estimating procedure is the following:

The variable of interest is "type of household", where the categories are constructed from "number of adults" and "number of children" (see Appendix 2). The unit "adult" (older than 17 years) is used in the main part of the estimation process, but in the last step the SPREE estimates are transformed into household data. For each variable of interest category, we know the number of adults and, accordingly, we know the number of times a particular household is represented in the estimates based on individuals. For example, the number of households with two adults and one child is estimated by dividing the estimate based on

individuals by 2. The reason why we compute the SPREE estimates by using individual values is that both RTP and SLC are based on individuals.

The main purpose of the study is to develop a SPREE estimator performing well in the above context. The task rises several questions: Which associated variables are the best? How should they be categorized? Is there any gain by clustering the municipalities?

Given a set of potential associated variables, there is good reason to reduce it to a smaller 'best' combination. The use of several variables, each assuming a moderate number of categories, can result in a large number of cells, which may cause difficulties in obtaining reliable sample estimates even at the national level.

Here we start with a pragmatic approach in selecting associated variables and their categories. We choose the variables suspected to be related to the variable of interest and which are readily available, viz. sex, age, marital status, and number of children (see Appendix 1).

The questions above and the selected type of estimator request a simulation study, but we have not been able to conduct such a study because of restricted resourses. Hence, we have to neglect the sampling variability and concentrate on the bias (model error). We know from the simulation study that, under those constraints (the small areas consist

of municipalities and the survey is based on a sample of "ordinary" size), the dominating error is the model error. However, in this study we use a rather large number of subgroups (H) and therefore, the sampling variability can be serious.

All three data sets are to be used simultaneously when calculating the SPREE estimate. Moreover, the calculations are rather comprehensive which means that a large computer is required and therefore only the municipalities in one particular county (Älvsborgs län) are included in the study.

Another restriction of the study is that we do not pay regard to the differences in the definition of the variable of interest and reference period between the SLC and the census.

5.2 Descripton of the study

We study a SPREE estimator with its association structure represented by N'_{hiq} and its allocation structure by $\{\hat{N}_{hi.}; N_{h.q}\}$, i.e. a SPR type estimator. In the applied situation, we take N'_{hiq} from the previous census, $\hat{N}_{hi.}$ from SLC and $N_{h.q}$ from RTP.

The study was designed for the purpose of examining the SPREE estimator $\hat{N}_{.iq}$ for the totals $N_{.iq}$ (= true values) for the 1980 Census period. Thus, we know the parameter values and can compare them with the estimates and compute quality measures.

The latest complete data, N'_{hiq}, will be retrieved from the 1975 Census. This implies that the time between the computation of N'_{hiq} and the estimates $\hat{N}_{.iq}$ is at most five years. In the study the current information $\hat{N}_{hi.}$ is taken from the 1980 Census - not from SLC.

In Table A a summary of the evaluations is displayed. The associated variables and their categories used in these SPREE estimates are described in Appendix 1. We also show the effect of excluding the age variable from the SPREE estimator in Table A.

The only available household data for municipalities in the intercensal period are those from the previous census and from the RTP. As mentioned above, we already know that there is a deviation between the definition of RTP-families and that of census-households, but due to the fact that some planners use the RTP information as substitute for census information, we think that it is also relevant to compare with RTP estimates, when evaluating the SPREE estimator.

The quality measure has the following form:

$$P_q = 100 \frac{\sum_i |S_{iq} - C80|}{\sum_i C80}, \qquad (15)$$

where, S_{iq} denotes the Census 1975 (C75), the RTP, or SPREE values, respectively, for household type i for municipality q

and

C80 denotes the Census 1980 values.

TABLE A

P_q values for different data sources.

Municipality q	S_{iq} = C75	RTP	SPREE	SPREE without the age variable
Dals-Ed	10.8	68.3	7.4	7.0
Färgelanda	11.1	85.8	6.9	8.0
Ale	10.6	55.7	3.6	5.2
Lerum	12.1	58.3	2.1	4.6
Vårgårda	12.0	73.6	3.5	3.6
Tranemo	12.0	66.6	4.9	5.7
Bengtsfors	10.8	60.8	4.6	5.0
Mellerud	9.5	62.9	4.5	3.3
Lilla Edet	9.1	64.8	3.8	4.4
Mark	10.1	67.9	2.1	2.2
Svenljunga	9.1	74.7	6.1	7.3
Herrljunga	7.5	70.9	3.8	5.0
Vänersborg	10.1	52.6	4.2	4.3
Trollhättan	8.7	52.4	1.9	1.6
Alingsås	9.9	57.8	3.6	3.0
Borås	8.4	53.9	1.5	1.6
Ulricehamn	8.2	60.8	2.8	3.6
Åmål	11.3	57.7	3.9	2.3

The table shows that, with respect to the measure P_q, the SPREE estimator provides much better information than the Census 1975 and the RTP. The RTP in particular gives quite misleading results. Using five years old data as current estimates also seems to be a rather doubtful practice.

When excluding the age variable, we obtain larger P_q values for nearly all municipalities. However, we know that when reducing the number of associated variables, we also reduce the sampling variability in $\hat{N}_{hi.}$. This study gives no answer as to which combination of associated variables is optimal, but the work will continue.

Up to now we have only presented an average measure (over categories) of the errors, but in Appendix 2 we also display the results for each category and for two "typical" municipalities. (In Lundström (1984) the results for every municipality are given.) Examination of the tables shows that the SPREE estimator performs well for most categories.

On an annual basis, local planners make prognoses about matters such as the demand for dwellings, and they base them to a great extent on what happened in the past. Thus, it is interesting to discover whether the SPREE estimator measures the sign (positive, negative or unaltered) of the trend in a correct manner. The number of categories (out of 16) where the SPREE estimator has been successful is (the municipalities are arranged in the same order as in Table A):

14, 12, 13, 15, 15, 15, 14, 11, 11, 9, 14, 14, 12, 12, 12, 15, 13, 12. Thus, the estimator also acts well with respect to that measure; moreover, when examining the cases where the estimator exhibits an incorrect sign of the trend one finds that they primairly consider small counts.

5.3 Conclusions and plans for the future

The findings of the study are promising and thus encourage us to continue to refine the SPREE estimator. We will be faced mainly with the problem of minimizing a measure such as "mean square error", that is, finding the minimum sum of the variance and the square of the bias. We know that the fewer the associated variables and categories are, the smaller the variance (of $\hat{N}_{hi.}$). At the same time, this probably increases the bias. One way of decreasing the bias is to cluster the municipalities so that the model, which the SPREE estimator is based on, becomes more relevant for each cluster. But here we are faced with another conflict between the variability and the bias: the estimate $\hat{N}_{hi.}$ is computed on the cluster level, which implies a larger variance.

As mentioned above the described SPREE estimator places substantial demands on computer resources. In the study where we confined ourselves to a minipopulation, we had no problem although we had restricted resources but we will perhaps meet problems in a full-scale investigation.

In this context, it would be advantageous to use a cluster approach since the computation of the SPREE estimator can be done for each cluster.

Finally, if the described future work turns out successfully, Statistics Sweden intends to calculate and publish the municipality estimates.

References

Lundström, S (1984). Estimating the Number of Households of Different Types in Each Municipality - Pilot Study in Älvsborgs county. (In Swedish)

Purcell, N. J. (1979). Efficient Estimation for Small Domains: A Categorical Data Analysis Approach. Unpublished Ph.D. dissertation, University of Michigan, Ann Arbor, Michigan.

Purcell, N. J. and Kish, L. (1979). Estimation for Small Domains. Biometrics, 35, 365-384.

Rao, J. N. K., Platek, R., Särndal, C. E., and Singh, M.P. ed. Small Area Statistics: An International Symphosium, Ottawa, May 22-24, 1985. Forthcoming in John Wiley & Sons, Inc. Publishers.

Schaible, W. L. (1979). A Composite Estimator for Small Area Statistics. In: Synthetic Estimates for Small Areas, ed. J. Steinberg, NIDA Research Monograph 24. Rockville, Maryland: National Institute on Drug Abuse, 36-53.

Steinberg, J. ed. (1979). Synthetic Estimates for Small Areas, NIDA Research Monograph 24. Rockville, Maryland: National Institute on Drug Abuse.

Särndal, C.E. (1981). Framework for Inference in Survey Sampling with Applications to Small Area Estimation and Adjustment for Nonresponse. Bulletin of the International Statistical Institute, 49 (1), 494-513.

Appendix 1

The Associated Variables and their Categories

Variable	Categories
Sex	Male, female
Age	18-24, 25-34, 35-44, 45-64, 65-
Marital status	Cohabiting married person, others
Number of children	0, 1, 2, 3- for cohabiting married person
	0, 1, 2- for others

Results for each category of two "typical" municipalities

Municipality: Ale.

Type of household Number of ...		Information from ...				Difference between ...	
						RTP and	SPREE and
adults	children	C75	C80	RTP*	SPREE	C80	C80
1	0	1 337	1 659	4 538	1 636	2 879	-23
1	1	115	166	407	163	241	-3
1	2	71	105	285	100	146	-5
1	3-	26	34		24		-10
2	0	1 918	1 971	1 941	2 053	-30	82
2	1	996	891	989	932	98	41
2	2	1 528	1 611	1 630	1 632	19	21
2	3-	620	585	609	607	24	22
3	0	438	441	0	435	-441	-6
3	1	190	247	0	245	-247	-2
3	2	87	126	0	104	-126	-22
3	3-	34	38	0	29	-38	-9
4-	0	104	117	0	108	-117	-9
4-	1	30	64	0	38	-64	-26
4-	2	14	24	0	18	-24	-6
4-	3-	13	17	0	16	-17	-1

*) We were unable to split the RTP data into the two categories "1 2" and "1 3-".

2

Appendix 2

Municipality: Mark.

Type of household		Information from ...				Difference between ...	
Number of ...						RTP and	SPREE and
adults	children	C75	C80	RTP*	SPREE	C80	C80
1	0	2 448	2 863	8 124	2 840	5 261	-23
1	1	125	213	501	179	288	-34
1	2	55	74	285	76	181	2
1	3-	24	30		26		-4
2	0	3 440	3 627	3 743	3 604	116	-23
2	1	1 116	1 011	1 224	1 042	213	31
2	2	1 436	1 617	1 665	1 592	48	-25
2	3-	544	538	565	535	27	-3
3	0	986	941	0	934	-941	-7
3	1	315	384	0	344	-384	-40
3	2	141	151	0	150	-151	-1
3	3-	66	46	0	45	-46	-1
4-	0	319	285	0	315	-285	30
4-	1	104	113	0	114	-113	1
4-	2	62	51	0	75	-51	24
4-	3-	21	20	0	23	-20	3

*) We were unable to split the RTP data into the two categories "1 2" and "1 3-".

Tidigare nummer av Promemorior från U/STM:

NR

1 Bayesianska idéer vid planeringen av sample surveys. Lars Lyberg (1978-11-01)

2 Litteraturförteckning över artiklar om kontingenstabeller. Anders Andersson (1978-11-07)

3 En presentation av Box-Jenkins metod för analys och prognos av tidsserier. Åke Holmén (1979-12-20)

4 Handledning i AID-analys. Anders Norberg (1980-10-22)

5 Utredning angående statistisk analysverksamhet vid SCB: Slutrapport. P/STM, Analysprojektet (1980-10-31)

6 Metoder för evalvering av noggrannheten i SCBs statistik. En översikt. Jörgen Dalén (1981-03-02)

7 Effektiva strategier för estimation av förändringar och nivåer vid föränderlig population. Gösta Forsman och Tomas Garås (1982-11-01)

8 How large must the sample size be? Nominal confidence levels versus actual coverage probabilities in simple random sampling. Jörgen Dalén (1983-02-14)

9 Regression analysis and ratio analysis for domains. A randomization theory approach. Eva Elvers, Carl Erik Särndal, Jan Wretman och Göran Örnberg (1983-06-20)

10 Current survey research at Statistics Sweden. Lars Lyberg, Bengt Swensson och Jan Håkan Wretman (1983-09-01)

11 Utjämningsmetoder vid nivåkorrigering av tidsserier med tillämpning på nationalräkenskapsdata. Lars-Otto Sjöberg (1984-01-11)

12 Regressionsanalys för f d statistikstuderande. Harry Lütjohann (1984-02-01)

13 Estimating Gini and Entropy inequality parameters. Fredrik Nygård och Arne Sandström (1985-01-09)

14 Income inequality measures based on sample surveys. Fredrik Nygård och Arne Sandström (1985-05-20)

15 Granskning och evalvering av surveymodeller, tiden före 1960. Gösta Forsman (1985-05-30)

16 Variance estimators of the Gini coefficient - simple random sampling. Arne Sandström, Jan Wretman och Bertil Waldén (Memo, Februari 1985)

17 Variance estimators of the Gini coefficient - probability sampling. Arne Sandström, Jan Wretman och Bertil Waldén (1985-07-05)

18 Reconciling tables and margins using least-squares. Harry Lütjohann (1985-08-01)

19 Ersättningens och uppgiftslämnarbördans betydelse för kvaliteten i undersökningarna om hushållens utgifter. Håkan L. Lindström (1985-11-29)

20 A general view of estimation for two phases of selection. Carl-Erik Särndal och Bengt Swensson (1985-12-05)

21 On the use of automated coding at Statistics Sweden. Lars Lyberg (1986-01-16)

22 Quality Control of Coding Operations at Statistics Sweden. Lars Lyberg (1986-03-20)

23 A General View of Nonresponse Bias in Some Sample Surveys of the Swedish Population. Håkan L Lindström (1986-05-16)

24 Nonresponse rates in 1970 - 1985 in surveys of Individuals and Households. Håkan L. Lindström och Pat Dean (1986-06-06)

Kvarvarande exemplar av ovanstående promemorior kan rekvireras från Elseliv Lindfors, U/STM, SCB, 115 81 Stockholm, eller per telefon 08 7834178

★Garnisonstryckeriet★Stockholm★1986★