Journal of Official Statistics
Vol. 6. No. 4, 1990, pp. 379–392
© Statistics Sweden

Small Domain Estimation of the Number of Employed in Different Weekly Work Hour Categories

Sixten Lundström[1]

Abstract: The paper discusses small domain estimation. More specifically, we report on investigations of synthetic estimation techniques for obtaining information on the number of persons in different "weekly work hour" categories for domains on the Swedish municipality level. On the basis of the results from the investigations reported on here, Statistics Sweden has in fact taken a synthetic estimation system into regular use, in which small domain estimates are produced by combining information from a newly established Annual Register of Employment (ARE) with information from the Labour Force Survey (LFS).

Key words: Model-dependent; synthetic estimator; simulation study.

1. Introduction

We start by presenting the background of Statistic Sweden's new system of employment statistics and deal in depth with the synthetic estimators.

Sweden has a long tradition of population registration. For 20 years, employment statistics have been supplemented by censuses in which additional information on individuals, households, and housing has been collected.

During the last few decades, censuses have been carried out every five years. Some of the census questions have varied, while others have been repeated from census to census. Since the 1960s, questions on employment have been asked in the censuses, and census data constituted the basis for small domain statistics on employment circumstances.

The short-term fluctuating aspects of employment, such as the rate of unemployment, have been charted with the aid of the Labour Force Survey (LFS) since the mid-sixties. The LFS is a sample survey that yields accurate statistics on a national and large domain regional basis, but not for small domains. However, the LFS not only collects information on rapidly varying employment variables but on almost all variables of interest in employment contexts.

Statistics Sweden started in 1984 to build up a new system for employment statistics, in which the main vehicle is the Annual Register of Employment (ARE), which is

[1] Senior Statistician, Statistical Research Unit, Statistics Sweden, S-701 89 Örebro, Sweden.
Acknowledgements: I would like to express my thanks to the members of the project group Jan Sävenborg, Björn Tegsjö, Staffan Wahlström and especially Claes Cassel and Göran Råbäck who made important contributions to this research. The Associate Editor Bengt Rosén and two referees provided many helpful suggestions.

described in more detail later. As a consequence of the new system and in order to minimize response burden in the census, questions on employment were omitted from the 1985 and 1990 censuses.

However, even if the new ARE contains many employment variables which can be reported on any level, it does not contain all employment variables which are of interest to local authorities.

One such variable, which was lost through omitting employment questions from the census, is the variable "hours worked per week". This variable is used in the definition of notions such as "full-time employed" (which according to Statistics Sweden's standards means that a person works at least 35 hours per week), "part-time employed," etc. Knowledge of the number of persons in different work-hour classes for small domains is of particular relevance for the planning of day care facilities for children.

Statistics Sweden has therefore been working on the development of a synthetic estimation system which will, we hope, enable reasonably good small domain statistics for work-hour classes. The basic idea is that estimates should be obtained by combining information from the ARE and the LFS (where, e.g., weekly work hours is measured). Our aim in the following is to report on findings from these investigations. First we give some further information on the ARE and the LFS.

2. Short Description of the ARE and the LFS

Ideally the ARE register has the following properties. For each calendar year it provides a complete register over all individuals in the country, in the age group 16-w, that were employed in November. For each individual in the register, the values of the following variables are recorded: name, sex, age, living address, work address, industry, and income.

The information in the ARE is collected from a number of sources. Each year employers send an income verification for each employee to the tax authorities. Statistics Sweden gets a copy of this tax register and merges it with five other registers, of which the Register of Enterprises and the Register of Total Population are the main sources.

The ARE register will annually yield employment statistics on different administrative levels, such as the whole country, counties, municipalities, and parts of municipalities. The administrative distribution is estimated for the resident population (night-time population) and for employed by place of work (day-time population). This division into night- and day-time population makes it possible to report commuting. The statistics from the ARE are published about one year after the income year.

Of course, there are many problems in such a complicated system of registers and thus, the statistics will suffer from errors. We present some of these errors in Section 4.4.1.

The LFS is performed every month based on a sample of 18,000 individuals. The sample system consists of three independent samples designed for three consecutive months. Each sample consists of eight panels and is rotated so that a sampled individual is in the sample for one month and out of sample two months. The individual participates for two years. Each sample is stratified according to county, sex, marital status, and citizenship and allocation is proportional to stratum size.

The LFS sample data file includes values on the variables mentioned for the ARE and many other variables.

3. Some Comments on the Distribution of Small Domain Sizes

The estimator selection procedure is mainly dependent on the distribution of the small domain sizes and the number of sample observations in each of the small domains. We give some information on these two aspects in this section.

When we discuss the possibility of using the LFS as a source of data we think that, in order to reduce the sampling error, the combined sample for three consecutive months should be used. Thus, we will have observations from outside the time period (November) and that will introduce an error, but we think that this error is much smaller than the gain in precision obtained by using the larger sample. The expected number of economically active persons in the combined sample is about 35,000.

The most common small domain in the synthetic estimation system is the municipality. In Sweden we have 279 municipalities of which three are very large, a few are rather large, and the majority are small. The expected number of observations for the combined LFS-sample is displayed in Figure 1.

Figure 1 shows that the number of expected observations are smaller than 100 for the majority of the municipalities. Thus, conventional estimators based on this nationwide sample do not give reliable estimates. This becomes even more certain when we know that the ultimate aim is to have estimates for parts of the municipalities.

If we classify the municipality sizes with the terms of Purcell and Kish (1979), we have 16 minor domains and 263 mini-domains.

4. Simulation Studies

4.1. Estimators

Small domain estimation has received considerable attention in recent years, which has resulted in many new estimators. Of course, we cannot compare all of them, and thus, we have restricted our study to the estimators we are most confident about and also, those that readily lend themselves to comparison. We also assume that the income

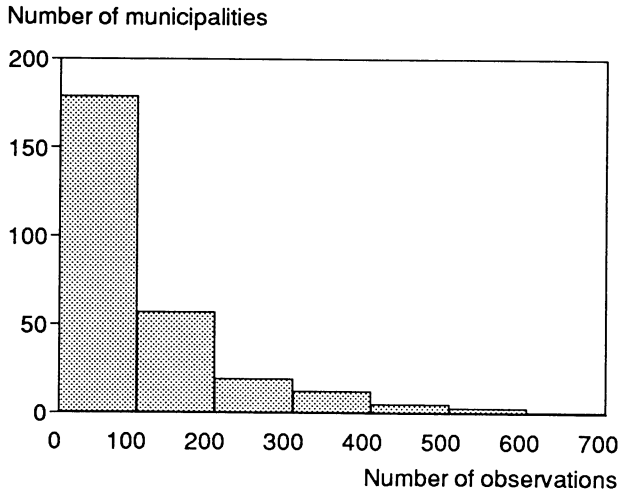

Fig. 1. Number of economically active persons in each municipality in the sample (except for the three largest municipalities)

variable would be the "best" auxiliary variable and this assumption affects the selection of some estimators. Some estimators depend on models and thus, are design biased, and the other estimators are at least approximately design unbiased.

The population consists of Q non-overlapping small domains labelled $q = 1, 2, \ldots, Q$. Moreover, the population is partitioned into another set of non-overlapping auxiliary groups, which we label $h = 1, 2, \ldots, H$. The cells in the corresponding cross-classification will be referred to by (h, q). Let N_{hq}, $N_{.q}$, and $N_{h.}$ denote the number of units in cell (h, q), in domain q and in group h, respectively. The corresponding sample counts are n_{hq}, $n_{.q}$, and $n_{h.}$ for a sample of size n.

Let y be a category indicator variable which is defined for the population, i.e., $y_i = 1$ if object i belongs to the category under consideration and $= 0$ otherwise. In our case the categories of interest are made up of different "weekly work hours". The parameters of interest are

$$T_q = \frac{100}{N_{.q}} \sum_{i=1}^{N_{.q}} y_i. \tag{1}$$

The first estimator that we discuss is somewhat inefficient and serves here mainly as a benchmark against which other estimators are compared. This estimator, the direct estimator (DIR), is given by

$$\hat{T}_{q\,\mathrm{DIR}} = 100 \sum_{i=1}^{n_{.q}} y_i / n_{.q}. \tag{2}$$

Remark: When $n_{.q} = 0$ we set $\hat{T}_{q\,\mathrm{DIR}} = 0$.

The well-known synthetic-count estimator (SYN/C) is defined as

$$\hat{T}_{q\,\mathrm{SYN/C}} = \frac{100}{N_{.q}} \sum_{h=1}^{H} N_{hq}\, \hat{\bar{Y}}_{h.}. \tag{3}$$

where

$$\hat{\bar{Y}}_{h.} = \sum_{i=1}^{n_{h.}} y_i / n_{h.}.$$

For the SYN/C estimator it is assumed that small domain means in a given class h are equal to their population counterparts. Any departure from this assumption will introduce bias and the greater departure, the greater the bias. The variance of the synthetic estimators are normally very small. The SYN/C estimator is discussed in Gonzalez (1973), Gonzalez and Hoza (1978), Levy (1979), and Schaible (1979).

The synthetic-ratio estimator (SYN/R) is given by

$$\hat{T}_{q\,\mathrm{SYN/R}} = \frac{100}{N_{.q}} \sum_{h=1}^{H} X_{hq} \left(\sum_{i=1}^{n_{h.}} y_i \Big/ \sum_{i=1}^{n_{h.}} x_i \right), \tag{4}$$

where x_i is the ith person's income and

$$X_{hq} = \sum_{i=1}^{N_{hq}} x_i.$$

This estimator is discussed in Choudhry and Hidiroglou (1987) and Hidiroglou, Morry, Dagum, Rao, and Särndal (1984).

The LOGIT estimator is given by

$$\hat{T}_{q\,\mathrm{LOGIT}} = \frac{100}{N_{.q}} \sum_{h=1}^{H} N_{hq}\, p_h^*. \tag{5}$$

Let p_h be the probability that a person in subgroup h belongs to a specific class of hours worked and that

$$u_h = \ln \frac{p_h}{1 - p_h}.$$

Introduce the regression equation

$$u_h = \alpha + \beta_1 x_1 + \cdots + \beta_m x_m + \varepsilon$$

where $\{x_1, \ldots, x_m\}$ denotes a set of dummy variables for the categorical variables and a variable for the mean income in each class. When estimating the coefficients, the observed proportion, which belongs to a specific class of hours in subgroup h, is used as an estimate of p_h. We estimate the regression coefficients using weighted least squares.

The predicted proportion in subgroup h is denoted by p_h^* and is computed from

$$p_h^* = e^{u_h^*}/(1 + e^{u_h^*})$$

and

$$u_h^* = \hat{\alpha} + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_m x_m.$$

Remark: In Section 4.4.3, the methodology study, the LOGIT estimator is used for different classes of hours worked. Throughout this study we consider hours worked a two-category problem: working more than a given number of hours or working less than that number.

We justify the use of the LOGIT estimator in the following way. It seems plausible to assume that as income increases, so does the probability of being employed full time. It also seems plausible that this probability increase is greatest at the middle of the income scale. We assume an S-probability curve and thus, a logit model is fairly obvious.

Särndal (1981) has, through the generalized regression method, developed a design-model based estimator (DM-estimator). Formally, it consists of two parts, the synthetic-count estimator (SYN/C) minus an estimator of the synthetic-count estimator bias

$$\hat{T}_{q\,\mathrm{DM}} = \hat{T}_{q\,\mathrm{SYN/C}} - \hat{B}_q, \tag{6}$$

where

$$\hat{B}_q = \frac{100}{N_{.q}} \frac{N}{n} \sum_{i=1}^{n} c_{iq} e_i$$

where $c_{iq} = 1$ if unit i belongs to small domain q; $= 0$ otherwise and

$$e_i = \hat{\bar{Y}}_{h.} - y_i$$

for units belonging to class h.

The estimator of the bias, $\hat{B}_q$, is often affected by a large sampling error and it is a good practice to dampen the effect of that part of the estimator. To minimize the mean square error of the estimator, (under the assumption that the two parts of the DM-estimator are uncorrelated) Cassel (1984) suggests that the bias term should be multiplied by a constant

$$\alpha_q = \frac{B_q^2}{B_q^2 + V(\hat{B}_q)},$$

where $V(\hat{B}_q)$ is the variance of $\hat{B}_q$.

An estimate of that constant is

$$\hat{\alpha}_q = 1 - \hat{V}(\hat{B}_q)/(\hat{B}_q)^2 \tag{7}$$

where

$$\hat{V}(\hat{B}_q) = \frac{100^2}{N_{.q}^2} N^2 \left(1 - \frac{n}{N}\right) \frac{1}{n-1} \times \sum_{i=1}^{n} (z_i - \bar{z})^2$$

and

$$z_i = c_{iq} e_i; \qquad \bar{z} = \frac{1}{n} \sum_{i=1}^{n} z_i.$$

Thus, the dampened DM-estimator (DDM) has the following form

$$\hat{T}_{q\,\mathrm{DDM}} = \hat{T}_{q\,\mathrm{SYN/C}} - \hat{\alpha}_q \hat{B}_q. \tag{8}$$

This estimator is, in contrast to $\hat{T}_{q\,\mathrm{DM}}$, design biased, which is the price of reducing the large sampling variability in the DM-estimator.

4.2. Performance measures

The basic measure of the performance of the estimators in this study is the mean square error (MSE), which has the following form

$$\mathrm{MSE}(\hat{T}_q) = V(\hat{T}_q) + B^2(\hat{T}_q), \tag{9}$$

where $V(\hat{T}_q)$ is the variance and $B(\hat{T}_q)$ is the bias of the estimator.

In order to have the correct dimension, we calculate the square root of the mean square error and obtain RMSE ($\hat{T}_q$). To

measure the size of the bias term as part of RMSE we compute the proportion of the (absolute) bias to the RMSE ($\hat{T}_q$)

$$PB_q = 100\,|B(\hat{T}_q)|/\text{RMSE}\,(\hat{T}_q). \quad (10)$$

To have a summary measure we also calculate the means (over the small domains) of the error components.

4.3. *The formation of auxiliary groups*

As is well known, the success of a synthetic estimator largely depends on the strength of the association between the study variable and the auxiliary variables. For example, the bias for the synthetic-count estimator (SYN/C) has the following form

$$B(\hat{T}_{q\,\text{SYN/C}}) = \frac{100}{N_{.q}} \sum_{h=1}^{H} N_{hq}(\hat{\bar{Y}}_{h.} - \hat{\bar{Y}}_{hq}), \quad (11)$$

where

$$\hat{\bar{Y}}_{hq} = \sum_{i=1}^{n_{hq}} y_i/n_{hq}.$$

Thus, if we find auxiliary groups that make $\hat{\bar{Y}}_{h.} \approx \hat{\bar{Y}}_{hq}$, then the bias for the SYN/C estimator will be small.

The auxiliary variables we chose to use are: sex, age, income, and industry. Income is defined as an individual's total annual salary and compensation (for illness) from the National Social Insurance Board. The fact that we estimate the employment circumstances in November and that we have income information for the year will cause some problems, which we discuss in Section 4.4.1. The industry variable consisted of 10 categories, mainly based on the one-degree level of the ISIC68 nomenclature.

Those variables seem, even from an unrestricted point of view, to be the most important. Perhaps, the education variable could be a strong competitor to the other variables. At Statistics Sweden we have an education register that can be merged with the LFS and the ARE. But such a merger would be rather expensive and moreover, it would delay the publication of the estimates. Another alternative would be to use the variable cohabiting status. However, the only information we have about this is marital status from the RTP, but that variable is not particularly informative because so many people cohabit without being married.

In the simulation study we used sex and income only as auxiliary variables, but thereafter, we refined (Section 6) the selected estimator by using all the variables mentioned. We also investigated different sets of income classes in the simulation study.

4.4. *Simulated experiments and their results*

4.4.1. Introduction

To measure the quality of model-dependent estimators one usually conducts simulation studies. Here, we conduct simulation studies for several estimators using different mini-populations and different sets of auxiliary information. When taking many repeated samples from a population it is important to have a population that is not too large. Yet we want our results to be realistic and applicable in practice. This problem will be discussed later.

We also had another type of problem with the simulation population due to the fact that the ARE suffered from errors that probably will be reduced in the future. The income reported in the tax register may pertain to only part of the year in which case the employer is instructed to provide information about the period of work. However, many employers do not provide the proper information and we also suspect that many others uncorrectly reported the entire year. The subject matter staff believes that the quality of the information will be higher in

the future, but in this methodology study, we had to deal with this problem. The following example illustrates our procedure when the period of work is known. For example, if we know that a person has worked only three months (where the LFS week is included), we multiply the income value by 4 in order to obtain the entire year's income. Lack of information about the period of work will, of course, give an incorrect relationship between the study variable and the auxiliary variable. The employers per se also present a problem. They belong to the group of economically active persons, but, due to the Swedish tax system, the relationship between the study variable and the auxiliary variable will be weak. We have decided to let all self-employed belong to the class of those working full time and to exclude them from the study.

To take those problems into consideration we carried out the simulation study in two experiments and finished the work by attempting to refine the selected estimator (Section 6).

Experiment 1

The aim of this experiment was to simulate a population that showed a realistic relationship between the study variable and the auxiliary information, i.e., we wanted to model a situation where the ARE-system has reached a level of good quality. Another aim was to gain some information on the discrepancies between the LFS and the census employment data.

To meet those aims we constructed the minipopulation in the following way. At first we merged the respondents of the LFS samples for October, November, and December 1980 and the respondents in outgoing panels for the other months. Then we tried to exclude the employers and those not working the whole year. (That screening was not entirely successful.) We matched that data file population with the 1980 census register in order to obtain census information about hours worked.

After these deletions the study population consisted of about 35,000 economically active persons. This population was too small to allow a study of estimates at the municipality level, so we used counties as small domains. The population register contained information about sex, income, hours worked (from the LFS and the census), and a county code.

Experiment 2

One shortcoming of experiment 1 was that we did not have municipalities as small domains, and thus, we wanted to use municipalities in experiment 2.

Another aim of this experiment was to prepare for the refinement of the selected estimator. To do this we wanted to have all the municipalities included in the population, but it would be troublesome and expensive to carry out the screening used in experiment 1. Thus, we wanted to use experiment 2 as a bridge from the "screened" to the "unscreened" version.

We studied two minipopulations in this experiment. The first one was called "the unscreened version" and consisted of the economically active persons in the municipalities in Örebro county. The other minipopulation – "the screened version" – emerged from "the unscreened version" but after a screening described in experiment 1. The same set of auxiliary groups was used in both experiment 1 and 2.

4.4.2. Sample sizes

In this methodology study we used a simple random sampling design. This gave us a sampling variability which was larger than if we had used the LFS-design. However, this

was not a serious problem that could stop us from performing a realistic comparison of the different estimators.

It was impossible to use a realistic sample size in the simulation study and therefore we had to make a reasonable reduction. In experiments 1 and 2, we used two sample sizes, 1,000 and 2,500, respectively.

The model-dependent estimators SYN/C, SYN/R and LOGIT use the total sample information in each h-group and due to the fact that the number of groups is modest, the sampling variability is small even for small samples. The DIR-estimator is based on the observations in each small domain and thus, it will usually suffer from a large variance. The DDM-estimator has a variance which is between those of the model-dependent estimators and the DIR-estimator.

Figure 1 in Section 3 shows the expected sample size in the municipalities when we use three LFS samples (about 35,000 economically active persons). We are indeed interested in even smaller domains. In experiment 1 we use Swedish counties as small domains and a sample size of 2,500 gives roughly the same picture as in Figure 1. The sample size 1,000 corresponds to a situation where we are interested in parts of municipalities. In experiment 2 we use municipalities in Örebro county as small domains. The expected number of observations in Örebro county is about 1,000 when $n = 35{,}000$.

The model-dependent estimators depend on the total sample size and therefore, the small sample sizes in the simulation study generate too much sampling variability. We should bear that in mind when judging the different estimators.

4.4.3. Results

For each sample r ($r = 1, \ldots, R$) in the Monte Carlo simulation study, we calculated the estimator value $\hat{T}_{q(r)}$ and finally

$$B(\hat{T}_q)_{\text{sim}} = \left(\sum_{r=1}^{R} \hat{T}_{q(r)}/R \right) - T_q \qquad (12)$$

and the simulation variance of $\hat{T}_q$.

From these computations, we calculated all of the error components.

In *experiment 1* we used $R = 500$ and the auxiliary information consisted of data on sex combined with income classes. We carried out several simulations for different income classes and for different classes of hours worked. When $n = 2{,}500$ and $H = 16$ (eight income classes) we obtained the following simulation estimates of RMSE for the class "more than 34 hours worked."

Table 1 indicates that there is no over-all best estimator. We can see that the design unbiased estimator is not an acceptable alternative. The largest county (01 Stockholm) is the only one where the design unbiased estimator has an error of the same magnitude as the other estimators. The RMSE for the model-dependent estimators vary greatly over the counties. All of the model-dependent estimators have for the counties Gotland (09) and Norrbotten (25) large RMSE. Those counties have an unusual industrial structure. Thus, we hope to have better results when including the industry variable in Section 6. According to the mean of the RMSE, the SYN/C estimator exhibits superior performance. Moreover, it does not have any extremely bad estimates except for the counties mentioned earlier.

In Table 2 the mean value of simulated estimates of RMSE for the same small domains and for the other two classes of hours worked are presented.

The estimators SYN/C and SYN/R have nearly the same mean RMSE for the two classes of hours worked.

In Table 3 we present simulation estimates of the proportion of the (absolute)

Table 1. Simulation estimates of Root Mean Square Error for different estimators, when estimating the proportion (in percent) of economically active persons in the class "more than 34 hours worked" in each county (T_q-value is in the interval 68–80). Sample size: 2,500.

County	SYN/C	SYN/R	LOGIT	DDM	DIR
01	1.31	4.33	1.11	1.44	2.23
03	2.00	2.61	1.86	2.36	4.52
04	1.51	1.07	1.18	2.14	4.59
05	1.24	1.05	1.09	2.08	4.81
06	1.05	1.23	1.29	2.09	4.85
07	1.09	1.56	1.18	1.93	5.11
08	1.16	1.02	1.09	2.27	5.04
09	3.74	4.27	4.36	3.43	5.16
10	1.05	2.03	1.76	2.21	4.88
11	1.15	1.63	1.51	2.19	5.03
12	1.09	1.17	1.03	1.64	3.38
13	1.05	1.19	1.01	2.09	5.04
14	1.35	2.03	1.08	1.60	3.33
15	2.29	2.47	2.85	2.36	4.32
16	1.88	2.66	2.58	2.35	5.01
17	1.54	2.79	1.66	2.25	5.17
18	1.46	1.83	2.07	2.08	4.93
19	1.03	1.05	1.24	2.12	4.67
20	2.03	2.80	2.71	2.56	4.54
21	1.03	1.22	1.58	2.26	5.06
22	1.51	2.25	1.71	2.33	4.78
23	2.14	2.55	3.08	2.85	5.25
24	1.26	1.26	1.03	2.42	5.12
25	4.65	3.43	3.57	4.01	5.12
Mean	1.65	2.06	1.82	2.29	4.66

bias to the RMSE, i.e., PB_q, for the estimates in Table 1.

The simulation estimate of PB_q shows that the DIR-estimator has a rather small bias compared to the RMSE. The table also shows that the DDM-estimator is placed between the pure model-dependent estimators and the design unbiased estimators. The bias is the dominating part of the RMSE for the model-dependent estimators, even for this small sample.

In *experiment 2* we studied the municipalities in Örebro county. For the screened version we have carried out the same type of selections (not working the whole year, etc.) as in experiment 1. The result presented in

Table 2. Mean (over counties) of simulation estimates of Root Mean Square Error (RMSE) for different estimators, when estimating the proportion (in percent) of economically active persons in different classes of hours worked. Sample size: 2,500.

Class of hours worked	SYN/C	SYN/R	LOGIT	DDM	DIR
1–19	1.07	1.08	1.28	1.58	2.70
20–34	1.19	1.19	1.63	2.08	4.16

Table 4 is based on $R = 500$ samples of size 1,000.

We can make nearly the same comments about this table as we did for Table 1. We have also studied estimates for the other classes of hours worked and here too (compare with the results in Table 2) the SYN/C and SYN/R are nearly equal according to the mean RMSE.

Up to now we have used only sex and income as auxiliary information, but in practice we will also have age and industry. Nevertheless, we think that the auxiliary information used is adequate for choosing the best type of estimator.

The simulation studies have shown us that a design unbiased estimator gives estimates of unacceptable quality. Therefore, we have to use a model-dependent estimator. Among these the SYN/C estimator has the smallest mean RMSE and does not have weaknesses not found in the other estimators. Moreover, it has an uncomplicated mathematical form and, as we show in the following, it is also easy to study in refinement work.

As a lead-in to the refinement work, we also studied the unscreened version of the municipalities in Örebro county. We found that the RMSE usually was, as expected, larger for the unscreened version than the screened version, but still we thought that the latter type of data could be used to refine the estimator. Thus, we could use data for

Table 3. Simulated estimates of PB_q for different estimators, when estimating the proportion of economically active persons in the class "more than 34 hours worked" in each county. Sample size: 2,500.

County	SYN/C	SYN/R	LOGIT	DDM	DIR
01	59.2	96.1	21.0	31.8	6.3
03	86.4	91.4	83.7	44.7	9.0
04	74.7	22.8	49.1	32.5	2.0
05	57.9	21.2	30.1	17.7	2.7
06	42.4	54.5	62.0	4.7	9.5
07	43.8	74.8	50.8	14.2	3.2
08	52.1	17.3	30.8	23.9	4.4
09	96.2	97.0	97.1	47.0	2.3
10	27.5	85.7	80.8	1.6	0.8
11	47.0	77.4	73.3	1.6	0.8
12	37.0	41.5	16.2	9.1	6.3
13	41.2	51.2	5.1	7.6	2.8
14	65.1	83.8	21.0	32.3	3.3
15	90.5	90.8	93.4	40.5	2.8
16	85.3	92.4	91.5	37.7	2.9
17	77.4	93.0	78.8	27.6	0.8
18	73.5	82.6	86.3	26.9	2.2
19	20.8	7.8	51.1	10.5	2.0
20	86.6	92.7	92.2	32.3	8.1
21	23.6	52.8	74.5	0.6	6.4
22	74.2	88.8	79.0	18.9	1.0
23	88.3	91.5	93.9	32.7	0.0
24	60.4	59.4	13.8	17.1	2.1
25	97.5	94.9	95.3	47.5	2.2
Mean	62.9	69.2	61.3	23.9	3.9

Table 4. Simulation estimates of Root Mean Square Error (RMSE) for different estimators, when estimating the proportion (in percent) of economically active persons in the class "more than 34 hours worked" in each municipality in Örebro county (the screened version). Sample size: 1000.

Municipality in Örebro county	SYN/C	SYN/R	LOGIT	DDM	DIR
1860	2.57	3.06	4.75	3.66	7.82
1861	1.42	2.61	2.17	2.43	5.92
1862	1.36	1.82	2.11	2.70	6.96
1863	1.35	1.95	1.73	2.85	7.65
1864	1.82	2.80	3.37	3.97	9.38
1880	1.51	2.61	1.72	1.55	2.87
1881	1.96	2.79	3.29	2.90	5.91
1882	1.71	2.46	2.78	2.89	7.34
1883	1.35	1.96	2.21	1.70	3.92
1884	1.79	1.87	1.68	3.00	7.81
1885	1.40	1.96	2.21	2.12	4.95
Mean	1.66	2.35	2.55	2.71	6.41

each municipality from the 1975 and 1980 censuses to refine the estimator.

6. Refinement of the SYN/C Estimator

The results presented in Table 3 show that bias is the dominating part of the RMSE of the SYN/C estimator. Thus, a relevant approach to reduce the RMSE is to reduce the bias. In this section we will discuss different ways of making that reduction. Mainly, we address the problem of choosing between alternate combinations of associated variables.

As we have stated before, the only readily available and relevant variables in both the LFS and the ARE are sex, age, industry, and income. Data from these variables for each municipality from the 1980 census are used as test data. An important feature of the SYN/C estimator is that it performs well in the long run and in order to test its long run performance we also checked this estimator on 1975 census data.

The bias term is defined in (11). We also want to have a measure of the sampling variability and based on the assumption of simple random sampling, it is possible to derive the formulas of the (expected) variance of the SYN/C estimator, viz.

$$\mathrm{E}[\mathrm{Var}(\hat{T}_{q\,\mathrm{SYN/C}})] \approx \left(\frac{100}{N_{.q}}\right)^2 \times \sum_{h=1}^{H} N_{hq}^2 \frac{\bar{Y}_{h.}(1-\bar{Y}_{h.})}{nW_h} K_h \qquad (13)$$

where

$$K_h = 1 + \frac{1-W_h}{nW_h} \quad \text{and} \quad W_h = \frac{N_{h.}}{N}.$$

Remark: The formula for the expected variance is derived in the same way as in Cochran (1977, sec. 5A.9).

In this study the parameters in (11) and (13) are know and thus, the (expected) variance and the bias for a given sample size, n, can be computed. In Table 5 the results are summarized by a mean value (over municipalities) of relative expected RMSE. That

Table 5. Mean value (over municipalities) of rel – ERMSE. Sample size 35,000.

Auxiliary information	Hours worked			
	1980 Census		1975 Census	
	20-w	35-w	20-w	35-w
SEX*AGE2*INC4*IND5	0.91	1.54	1.07	1.29
SEX*INC4*IND5	0.92	1.55	1.10	1.34
AGE2*INC4*IND5	0.89	1.94	1.05	1.72
INC4*IND5	0.91	1.90	1.09	1.76

measure has the following form

$$\text{Rel} - \text{ERMSE} = 100\ \text{ERMSE}/T_q \tag{14}$$

where $\text{ERMSE} = \{\text{E}[\text{Var}(\hat{T}_{q\text{SYN/C}})] + B^2(\hat{T}_{q\text{SYN/C}})\}^{1/2}$.

In practice, the estimator is based on three consecutive LFS samples, which aggregates to a sample size 35,000 and to make this study as realistic as possible; we used the same sample size here.

We have computed this measure for many different sets of auxiliary groups based on the variables sex, age, income, and industry. This work resulted in a particular set for each variable denoted SEX, AGE2, INC4, and IND5 (the number denotes the number of classes). We do not describe the work that led to this result.

Table 5 indicates that we get the best result when we use all of the four associated variables. Excluding AGE2 does not have any significant effect on the mean rel – ERMSE. On the other hand, we can see that SEX is an important auxiliary variable, especially when estimating the percentage economically active persons working more than 34 hours/week. Table 5 also shows that the "optimal" (raw 1) SYN/C estimator performs well in both periods.

To estimate the total effect of the auxiliary information, we calculated the mean rel – ERMSE for a synthetic estimator using no auxiliary information (i.e., the nationwide estimate is used in each municipality). For the class 35-w hours per week and using data from the 1980 census we received the value 2.22, which can be compared to 1.54 for the "optimal" estimator. It is also interesting to compare this to the DIR-estimator. If the sample size is 35,000 the mean rel-ERMSE is 7.7 for the DIR-estimator.

We have also tried several other refinement alternatives, which we now describe briefly.

If the bias is stable over time, we could use the bias calculated from the 1980 census data and subtract it from the SYN/C estimate. We tried this adjustment on the estimator for the 1975 census and the resulting mean rel-ERMSE was for the classes 20-w and 35-w hours/week, 1.07 and 1.12, respectively. Thus, it is not a good alternative to base an estimator on the bias measured in a previous census.

One way of reducing the bias (model error) in a model-dependent estimator is to group the small areas in such a way that the assumptions underlying the estimators are more likely to be met. It is true that the sampling variability will increase, but perhaps not to the same degree as the bias decreases (Purcell 1979). We have tried to group the municipalities but found that, after five years, the groupings have no significant effect on the bias.

7. Implementation of the SYN/C Estimator

The simple optimal SYN/C estimator is found to be the best estimator. But will it provide estimates of acceptable quality? That question has not been answered by our methodological study and to help answer that question we presented the results of the simulation studies for some of our statistical users.

Besides the errors studied in the preceding sections we also gave the users some information on other factors that affect the bias of the optimal SYN/C estimators. Users want a continuation of the series of census estimates, but when using the LFS and the ARE as data sources we introduce discrepancies between the previous series and the new series due to different definitions of the study variable, different data collection methods, etc. For the minipopulation in experiment 1 we had information on the study variable both from the census and the LFS. The net difference (in %) for the classes of number of hours worked per week 35-w, 20–34, and 1–19, is 0.2, 8.3 and −24.6, respectively. Thus, there are rather large discrepancies between the two sources concerning these classes.

As stated in Section 4.4.1 even the selection of economically active persons will be problematic and generate errors.

The users have either these synthetic estimates or no statistical information at all on hours worked. Thus, it cannot be expected that they will flatly reject the method. Nor did they react to the fact that this is a new type of estimator, because they sometimes have to make their own "synthetic" estimates when statistics are not available.

After the users gave their opinions about the estimator the subject matter unit put the estimator into practice. Now the users can order several tables on hours worked, for instance a table on sex∗age∗industry∗classes of hours worked for the municipality. Even smaller domains can be found in the tables. The users receive both the tables and a short description of the method, its shortcomings, and a reference to the report on the methodological study.

8. References

Cassell, C.-M. (1984). Optimal Selection of ∂. Unpublished memo, Statistics Sweden (in Swedish).

Choudhry, G.H. and Hidiroglou, M.A. (1987). Small Area Estimation: Some Investigations at Statistics Canada. Bulletin of the International Statistical Institute, 52 (4), 451–468.

Cochran, W.G. (1977). Sampling Techniques, third edition. New York: John Wiley, 134–135.

Gonzalez, M.E. (1973). Use and Evaluation of Synthetic Estimates. Proceedings of the Social Statistics Section, American Statistical Association, 33–36.

Gonzalez, M.E. and Hoza, C. (1978). Small Area Estimation with Application to Unemployment and Housing Estimates. Journal of the American Statistical Association, 73, 7–15.

Hidiroglou, M.A., Morry, M., Dagum, E.B., Rao, J.N.K., and Särndal, C.E. (1984). Evaluation of Alternative Small Area Estimators Using Administrative Records. Proceedings of the Section on Survey Research Methods, American Statistical Association, 307–313.

Levy, P.S. (1979). Small Area Estimation — Synthetic and Other Procedures, 1968–1978. Synthetic Estimates for Small Areas, ed. C. Steinberg (National Institute on Drug Abuse Research Monograph 24), Washington, DC: U.S. Government Printing Office, 4–19.

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

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

Schaible, W.L. (1979). A Composite Estimator for Small Area Statistics. Synthetic Estimates for Small Areas, ed. C. Steinberg (National Institute on Drug Abuse Research Monograph 24), Washington, DC: U.S. Government Printing Office, 36–83.

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.

Received January 1988
Revised July 1990