SAMU

The System for Co-ordination of Samples from the Business Register at Statistics Sweden

Esbjörn Ohlsson

SCB

R&D Report
Statistics Sweden
Research - Methods - Development
1992:18

INLEDNING

TILL

R & D report : research, methods, development / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1988-2004. – Nr. 1988:1-2004:2.
Häri ingår Abstracts : sammanfattningar av metodrapporter från SCB med egen numrering.

Föregångare:

Metodinformation : preliminär rapport från Statistiska centralbyrån. – Stockholm : Statistiska centralbyrån. – 1984-1986. – Nr 1984:1-1986:8.

U/ADB / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1986-1987. – Nr E24-E26

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

Efterföljare:

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

R & D Report 1992:18. SAMU : The system for co-ordination of samples from the business register at Statistics Sweden / Esbjörn Ohlsson.
Digitaliserad av Statistiska centralbyrån (SCB) 2016.

urn:nbn:se:scb-1992-X101OP9218

SAMU

The System for Co-ordination of Samples from the Business Register at Statistics Sweden

Esbjörn Ohlsson

R&D Report
Statistics Sweden
Research - Methods - Development
1992:18

Från trycket December 1992
Producent Statistiska centralbyrån, utvecklingsavdelningen
Ansvarig utgivare Lars Lyberg
Förfrågningar Esbjörn Ohlsson, 08/783 45 13

ISSN 0283-8680

SCB STATISTICS SWEDEN

Dept. of Enterprise Statistics
Dr. Esbjörn Ohlsson

December 1992

SAMU

The System for Co-ordination of Samples from the Business Register at Statistics Sweden

- A Methodological Description -

ABSTRACT. Most of the samples from the Business Register at Statistics Sweden are drawn in the so called SAMU system. A main purpose of the SAMU is to co-ordinate these samples, in order to give an even distribution of the response burden among the businesses. The co-ordination is obtained by using the so called JALES technique, which is based on the use of random numbers that are permanently associated with the sampling units. The surveys may have different design, as regards population delimitation, stratification and allocation. In the strata, samples are drawn with equal probabilities or, less common, with probabilities proportional to size. The system ensures that subsequent samples for the same survey are overlapping, though each sample is drawn from an up-to-date version of the register. The samples are partially rotated each year. Another aim of the SAMU is to promote the use of similar definitions of population units and compatible population delimitation for the surveys. In this paper we give a description of the SAMU system from a methodological point of view.

CONTENTS

SCB STATISTICS SWEDEN

Dept. of Enterprise Statistics
Dr. Esbjörn Ohlsson

December 1992

SAMU

The System for Co-ordination of Samples from the Business Register at Statistics Sweden

- A Methodological Description -

---

1. INTRODUCTION AND OUTLINE OF THE PAPER

A majority of the business surveys at Statistics Sweden use sample frames extracted from the *Business Register*. With few exceptions, the samples for these surveys are drawn in the so called SAMU system ("SAMordnade Urval", in English "co-ordinated samples"). The aim of this paper is to give a description of the SAMU system from a methodological point of view.

A main purpose of the SAMU is to co-ordinate samples for different surveys, in order to give an even distribution of the response burden among the businesses. The system ensures that subsequent samples for the same survey are overlapping, though each sample is drawn from an up-to-date version of the register. Another purpose of the SAMU is to promote the use of similar definitions of population units and compatible population delimitation for the surveys. Such standardisation facilitates comparisons of survey results; in particular, it is vital for the completeness of the National Accounts.

In Section 2 we describe the so called JALES technique that is used to co-ordinate stratified simple random samples in the SAMU. In Section 3 we present a technique for pps (probability proportional to size) sampling, called sequential Poisson sampling, that is used for a few surveys in the SAMU. Section 4 contains a discussion of the sample rotation technique of the SAMU. In Section 5 we give some details on the actual implementation of the mentioned techniques, as well as brief descriptions of the sampling frame and the facilities for stratification and allocation in the SAMU system. Section 6 gives an international overview of similar systems.

2. THE EQUAL PROBABILITY SAMPLING TECHNIQUE IN THE SAMU

In the SAMU, samples are drawn by the so called JALES technique, developed at Statistics Sweden in the early 70's by Johan Atmer and Lars-Erik Sjöberg (for whom 'JALES' is an acronym). It is described (in Swedish) by Atmer, Thulin and Bäcklund (1975), while Thulin (1976) discusses the implementation of the JALES technique in the SAMU. In this section we discuss the properties of the JALES technique.

2.1. Requirements on a technique for co-ordination of business samples

In most of our business surveys, estimates are required for subgroups of the population, defined by type of activity (industry) and in a few cases also geographical region. A primary stratification is performed in accordance with these subgroups. Most surveys use a further stratification by some measure of the size of the businesses. A frequently used size measure is 'number of employees', which is highly correlated with most variables of interest. In the ultimate strata, samples are drawn by simple random sampling without replacement (srswor). Instead of using size stratification and srswor, a few samples are drawn with probabilities proportional to size (pps sampling), see Section 3.

We want a large overlap between the samples at subsequent occasions (positive co-ordination over time), since this increases the precision in estimates of change over time. On the other hand, the business population is subject to rapid changes, due to births, deaths, splits, mergers, changes in size or in type of activity, etc. The Business Register is regularly up-dated according to such changes. To maintain a good quality of the surveys, the samples must be up-dated, too. Before the introduction of the SAMU system, the up-dating was made each year by sampling from additional 'nova' strata containing the new units. Such a procedure is, however, both inconvenient and inefficient. After a few years, the multitude of strata becomes unmanageable and an entirely new, independent sample has to be drawn. The JALES technique, on the other hand, offers a simple solution to the problem of drawing subsequent samples that are both up-to-date and overlapping.

If all the samples from the Business Register were independent, the response burden would be unevenly spread among the businesses. Since an ultimate stratum often contains only a moderate number of units, it could well happen that some units were included in several samples while others were in none. The respondent annoyance this would cause could partially be reduced by sample rotation. More important than to rotate the sample is to minimize the overlap between samples for different surveys (negative co-ordination in space). This will spread the burden at each sampling occasion, while rotation only gives an even distribution over a longer period. A few surveys, however, are required to have a large sample overlap (positive co-ordination), so as to enable comparisons of variables at the micro level. In the SAMU, with the JALES technique, we can obtain either positive or negative co-ordination of samples, even if they use different stratifications. Without destroying this co-ordination, the samples in the SAMU are rotated once a year, cf. Section 4.

2.2. Sequential srswor

Here we shall describe a particular technique for drawing a simple random sample without replacement (srswor) of size n from a population of size N. To each unit in the register, we associate a random number, uniformly distributed over the interval (0,1). Let X_i denote the random number for unit i. The X_i's should be mutually independent. Next we order the population in ascending order of the X_i's. The n first units on the list constitute the desired srswor.

Following Fan, Muller and Rezucha (1962), who seem to be the first to describe this technique, we will call this type of selection "sequential". It is intuitively clear that the result is an srswor. A formal proof of this fact is given in Appendix 1.

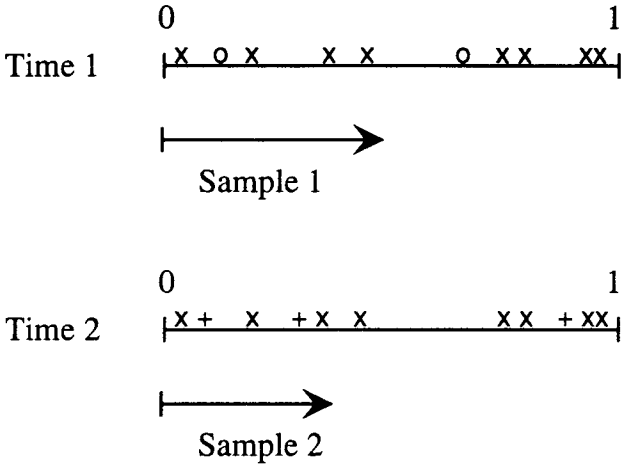
2.3. The JALES technique

The basic idea in the JALES technique is to let the random numbers in sequential srswor (Section 2.2) be *permanently* associated with the units. For units persisting in the register, which we will call *persistants*, we use the same random number on each sampling occasion (the *permanent random number*, PRN). New businesses, *births*, are assigned new PRN, independent of the already existing ones. *Deaths* (closed-down businesses) are withdrawn from the register together with their random numbers.

On each sampling occasion, we take a new sample by sequential srswor, using the PRN as random numbers. This way we always get an srswor from the up-to-date register. Nevertheless, we get a large amount of overlap with the latest sample since persistants have the same random numbers (PRN) on both occasions. We can not be sure that persistants stay in the sample, though, since we might get more births in our new sample than we have deaths in the old sample. This may simply be due to the random selection or there may actually be more births than deaths in the population. Though persistants may leave or enter the new sample, they will most frequently stay, yielding the desired

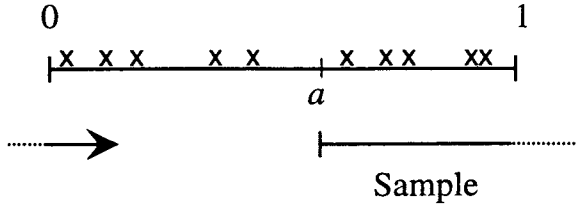
overlap (cf. Figure 1). Of course, a condition is that births and deaths are not *too* numerous; in the SAMU frame they are, on the average, less than 15%.

Figure 1. Overlap of subsequent samples. PRN's are denoted by:
x = *persistants,* o = *deaths,* + = *births.*



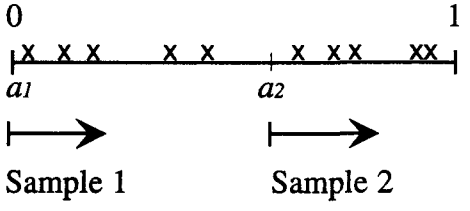
Returning to sequential srswor, we note that, by the symmetry of the uniform distribution, we might just as well take the *last n* units to obtain an srswor. Indeed, selecting the first n units to the left, or to the right, of *any* fixed point a in (0,1) will yield an srswor. If there are not enough (n) points to the right (left) of our starting point a, we simply continue the selection to the right (left) of the point 0 (the point 1), as exemplified in Figure 2.

Figure 2. Sampling from an arbitrary point a.



In order to reduce the overlap between two surveys, with desired sample sizes n_1 and n_2, choose two constants a_1 and a_2 in (0,1). Then take the units with the n_1 PRN's closest to the right (or left) of a_1 as the first sample and the ones with the n_2 PRN's to the right (or left) of a_2 as the second sample. If a_1, a_2 and the sampling directions are chosen properly, the result will be a negative co-ordination of the samples (see Figure 3 for an example). If the population is large enough, that is $N >> n_1 + n_2$, we can choose a_1 and a_2 so that the samples will most probably be disjoint. On the other hand, when $N < n_1 + n_2$ we can not possibly make the samples disjoint, but we can still reduce their overlap.

Figure 3. Negative co-ordination of samples.



Similarly, any number of samples can be negatively co-ordinated, if N is large enough. The best *positive* co-ordination of two surveys is, of course, obtained by using the same starting point and direction for both.

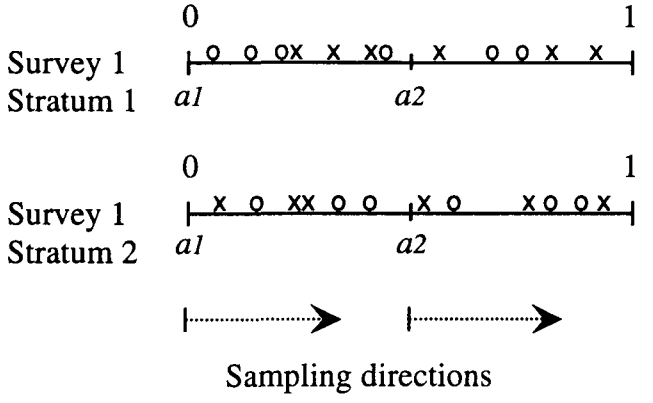
2.4. Co-ordination of samples with different design

In practice we have several strata and draw a sequential srswor in each of them. For a particular survey we use *the same* direction and starting point a in all the strata. Now suppose that two surveys use different stratifications of the same population. If the starting points are distinct, the surveys will still be negatively co-ordinated. This is so since a "small" (or whatever) random number in one stratum is likely to be "small" in

another stratum, too (cf. Figure 4). The extent to which the negative co-ordination is attained depends, naturally, on the sampling fractions in the strata.

Figure 4. Co-ordination of samples with different stratifications.

x = *PRN for unit in Stratum 1 of Survey 2;* o = *PRN for unit in Stratum 2 of Survey 2;*
a1 =staring point of Survey 1; *a2 = starting point of Survey 2.*



Similarly, the co-ordinated surveys do not have to use the same population delimitation.

By the same token, *positive* co-ordination of two samples can be obtained even if they have different stratifications. Applied to subsequent samples for a single survey, this means that we may redesign the survey, as regards population definition, stratification and/or allocation, and still have an overlap between the old and new sample. Note also that a unit in the old sample that changes stratum (due to changes in size or activity) has still got a large probability of being included in a new sample, since we use the same starting point in all the strata.

2.5. Sequentially deleting out-of-scope units

Sometimes the register contains a substantial number of units that are out of scope for a particular survey. Suppose that these units can be detected to a cost that is reasonable for our sample, but is too large to identify them all in the register.

In such a case, an ordinary srswor from the register will have a large variability in the *effective* sample size. With sequential srswor, on the other hand, we can simply continue down the list of random numbers until we have achieved a predetermined number n of in-scope units for our sample. By the independence of the PRN's, this "net" sample will have the same probability distribution as if the out-of-scopes had never been there. Hence, the result is an srswor of size n from the population of in-scope units. Similarly, we can stratify our sample according to variables not present in the register, as pointed out by Fan et al. (1962).

In the SAMU, this kind of technique is only used in connection with pps sampling, see Section 3.

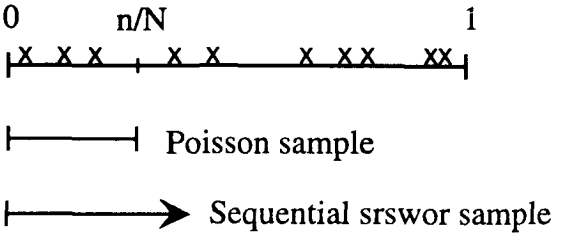
2.6. Interval (Poisson) sampling

In their description of the JALES technique, Atmer et al. (1975) mentions that, as an alternative to using sequential srswor, one may select all the units with PRN falling in a pre-set interval in (0,1). The resulting sample is the equal probability version of so called Poisson sampling. (It may be noted that this version is called "Bernoulli sampling" by Särndal, Swensson and Wretman, 1992). See Section 3 for a general description of Poisson sampling. Brewer, Early and Joyce (1972) suggested the use of Poisson sampling with permanent random numbers for co-ordination purposes, apparently independent of the contemporary development of the JALES technique and the SAMU.

Poisson sampling yields a random sample size, m say, which in the equal probability case follows a binomial probability distribution. If we want a sample of size n from a

population of size N, the interval should have length n / N. Then each unit will have inclusion probability n / N and m will have expectation n. Figure 5 illustrates the relation between Poisson sampling and sequential srswor.

Figure 5. Poisson sampling and sequential srswor (n=4).



The random sample size is a disadvantage of this procedure. If the desired sample sizes in the strata are small, the variability in sample size may cause severe deviations from an optimal allocation. We may even have a non-negligible probability of getting a sample of size 0 in some stratum. It can readily be shown that he probability that $m=0$ is less than e^{-n}. When the sampling fraction n / N is small, e^{-n} is a good approximation to this probability. With large n, the variation in the actual sample size m is of less importance from an efficiency point of view. However, from a practical point of view, the random sample size can be a problem for a statistician who has undertaken to carry out a survey with a stipulated sample size.

Another problem with Poisson sampling is that the ordinary, unbiased Horvitz-Thompson estimator is known to have very poor precision in connection with Poisson sampling (Sunter, 1977a, Brewer and Hanif, 1983, and Ohlsson, 1990). As an alternative, it is conventional to use the sample mean as an estimator of the population mean in this case. Because of the random sample size, this estimator is a ratio of random variables; hence it is only approximately unbiased and the variance is only approximately known, see, e.g., Brewer and Hanif (1983).

It seems preferable in this case to do the estimation *conditional* on the observed sample size. Let us consider the probability distribution of a Poisson sample, conditional on the outcome of m. It is readily seen that this is just the probability distribution of ordinary srswor with (fixed) size m (in case $m \neq 0$). Hence, conditioned on m, we can do any kind of estimation simply as if we had drawn an srswor.

Co-ordinated interval (Poisson) samples behave similarly to co-ordinated sequential srswor's (as described in Section 2); here we shall only indicate the differences.

In theory, we could keep the Poisson sampling interval for a survey fixed from year to year, making the overlap *complete* among persistants. In practice, N will change between the years, forcing us to use (at least slightly) different sampling fractions n/N for different years, even if n is unchanged. Hence the overlap of persisting units will, in reality, not be complete, but it is likely to be larger than for sequential srswor.

With Poisson sampling we are able to make sure that two or more samples are non-overlapping, if the sum of their selection interval lengths does not exceed 1. As mentioned in Section 2.3, with sequential srswor we can only make overlaps improbable, not absolutely impossible.

We conclude that Poisson sampling gives better positive and negative co-ordination than sequential srswor. This advantage of Poisson sampling should be balanced against the disadvantage of the random sample size. Note also that it is quite possible to combine the two methods, so that an srswor sample is co-ordinated with a Poisson sample, by using the same PRN's in both cases.

Poisson sampling is used in New Zealand (see Section 6.2), but has not been used in the SAMU system so far.

3. PPS SAMPLING IN THE SAMU

In this section we describe sequential Poisson sampling, which is the procedure for probability proportional to size (pps) sampling associated with the SAMU. This procedure is mainly used to sample outlets for the Swedish Consumer Price Index (CPI), see Ohlsson (1990). This section is not necessary for the understanding of the rest of the paper.

Suppose we want a sample of size n, in which unit i is included with probability proportional to p_i. The variate p_i may be arbitrary; we will think of it as some measure of the size of unit i. For simplicity, we assume that the p_i's are normed so that

$$\sum_{i=1}^{N} p_i = 1 \ .$$

If we let π_i denote the probability of including unit i in the sample, our request is to have

$$\pi_i = np_i . \tag{1}$$

A procedure that fulfils (1) will be called a procedure for (strict) *pps sampling* (sampling with probabilities proportional to size).

As a preparation for the presentation of sequential Poisson sampling we now describe how the well known pps sampling procedure called *Poisson sampling* can be carried out in a PRN context. Recall that X_i is the PRN of unit i. Choose a starting point a in (0,1). Next pass through the units in the register, one by one, and apply the following rule: Include unit i in the sample if, and only if,

$$a < X_i \le a + np_i \ , \tag{2}$$

with an obvious "wrap around" adjustment, in case $a+np_i > 1$, cf. Figure 2. The result is a Poisson sample having random sample size m with expectation n. Obviously, (1) is fulfilled, i.e. Poisson sampling is strictly pps. In the equal probability case, with $p_i = 1/N$, the sample will consist of all units with PRN in an interval of length n/N, as noted in Section 2.6.

The discussion on the disadvantages with the random sample size in Section 3 is equally relevant for the pps case. The estimate e^{-n} for the probability that m=0 still applies, but is now a bit more pessimistic.

The idea to co-ordinate Poisson samples through the use of PRN was introduced by Brewer et al. (1972). The properties of co-ordinated equal probability Poisson samples, discussed in Section 2.6, carry over to the pps case with minor modifications. Poisson sampling is not used in the SAMU.

Sequential Poisson sampling was introduced by Ohlsson (1990) as a way to generalize sequential srswor to the pps case. It can also be considered as an attempt to obtain a fixed sample size alteration of (pps) Poisson sampling. For each i, we introduce the normed random numbers

$$\xi_i = \frac{X_i}{Np_i} \tag{3}$$

Note that unit i is included in a Poisson sample with a=0 if and only if

$$0 < \xi_i \leq \frac{n}{N}. \tag{4}$$

In *sequential* Poisson sampling, on the other hand, we sort the population by ξ_i and then select the n first units on the sorted list. When the p_i's are equal, $\xi_i = X_i$ and sequential Poisson sampling reduces to sequential srswor. Note that the co-ordination of sequential Poisson samples is through the PRN's, X_i, while the ξ_i's differ from time to

time and from survey to survey. If we want a starting point a other than 0, we subtract a from X_i before applying (3).

Unfortunately, sequential Poisson sampling is not a *strict* pps procedure, as shown in Ohlsson (1990). Exact expressions for the inclusion probabilities of sequential Poisson sampling are not readily obtained, so the (exact) Horvitz-Thompson estimator can not be used. The suggestion is to construct the estimator as if the procedure really was strict pps, see Ohlsson (1990). That paper also refers a simulation study on data from the Swedish Consumer Price Index and the survey of Financial Accounts of Enterprises. In the studied cases, the inclusion probabilities are very close to the desired ones in (1). Furthermore, the suggested estimator performed slightly better with sequential Poisson sampling than the conventional (ratio) estimator did with Poisson sampling, as concerns bias and variance.

Co-ordinated sequential Poisson samples behave similarly to co-ordinated sequential srswor's (Section 2). As in the equal probability case, sequential Poisson sampling will often give less overlap of subsequent samples than ordinary Poisson sampling does.

Sequential Poisson sampling was developed for the Swedish Consumer Price Index (CPI), where it has replaced Poisson sampling from 1989 and on. A major reason for the change of sampling procedure for the CPI was a desire to have a fixed sample size of relevant establishments (shops, restaurants, etc.). This is obtained by using sequential Poisson sampling which, like sequential srswor, has the property to allow sequential deletion of out-of-scopes, as described in Section 2.5.

In the SAMU, the sequential Poisson sample for the CPI is co-ordinated with the stratified srswor's of the other surveys, through the use of the unique PRN's. To summarize: In the SAMU we co-ordinate samples with quite different design as regards population delimitation, stratification, allocation and inclusion probabilities (equal or pps).

Finally we mention another pps procedure that can be used in connection with PRN, called *collocated sampling* (see Brewer et al., 1972 and Brewer et al.,1984). This procedure is strict pps, and has a random sample size with less variance than Poisson sampling. For a comparison of collocated sampling to sequential Poisson sampling, see Ohlsson (1993). Collocated sampling is not used in the SAMU.

4. THE SAMPLE ROTATION METHOD OF THE SAMU

The Random Rotation Group method (RRG) was developed for and introduced in the SAMU system in 1989 (Ullberg, Segelberg and Ohlsson, 1990). In this section we describe this method and compare it to another technique, which we call *the constant shift method.*

For the SAMU it has been decided that samples should be rotated once a year, and that units preferably should be out of sample after, at most, 5 years. For the sake of simplicity, our discussion of rotation techniques below will be in terms of this rotation rate. We consider the rotation successful only if rotated units are not immediately included in another sample. Note that the largest units will almost surely be in at least one of our samples, and hence can not be rotated successfully. For medium-sized units, the sampling fractions are too large to enable an entirely successful rotation. Hence, the main aim of the rotation is to reduce the response burden for small units. Note also that the substitution of some of the sample due to births and deaths has very little, or no, rotation effect on persistants.

4.1. The constant shift method

The following procedure was suggested by Brewer et al. (1972) for rotation of a Poisson sample; it could equally well be applied to the sequential samples in the SAMU. Between the years, shift the starting points of all surveys to the right, by a constant

value. Let us say that the constant shift is 0.02. Then the expected rotation among units with inclusion probability 0.10 will be 20 percent, and after 5 years they are out of sample. For units with larger inclusion probabilities we will have less rotation. Small units with inclusion probabilities less than 0.02 will stay in sample for only one year. For the rotation to be successful we must of course have a large enough distance between the starting points for negatively co-ordinated surveys, else the units may just rotate out of one survey and into another.

For a single survey, we might be able to adjust the amount of shift (0.02 above) so that the overall rotation fraction is 20 percent (or whatever fraction we want). The varying number of years in sample for the units is hardly acceptable from the respondents point of view, though. Furthermore, the proper amount of shift will differ from survey to survey. We can not, however, use individual shifts for our surveys since this would destroy both the negative and positive co-ordination after a few years. For these reasons, the constant shift method is not used in the SAMU. It may also be noted that the method was abandoned by the Australian Bureau of Statistics in 1982.

4.2. The random rotation group method

Let us first note that instead of shifting all the starting points of our surveys 0.02 to the right, as in 5.1, we could shift all the PRN 0.02 to the left, with equal effect. Shifting PRN's is preferable when monitoring a complex system, since we will then always have the starting points in the same place.

In RRG, each unit in the register is randomly designated to one out of five rotation groups. To be more specific: for each unit we perform a multinomial trial giving a 20 per cent probability to each of the rotation groups. The rotation group number is permanently associated with the unit. Births are assigned a rotation group this way as they enter the sampling frame. After the first year, the units in rotation group 1 are shifted 0.10 to the left. Next year those in group 2 are shifted 0.10 to the left, etc.

Among the vast majority of units, that have selection probabilities less than 0.10, we will get an expected rotation rate of 20 percent each year and the unit can expect to be out of sample after 5 years. This holds true for any survey in the system, irrespective of sampling design. Among units with larger inclusion probabilities, rotation will be slower. Exclusion of the largest units from rotation in the SAMU has been considered. For the sake of simplicity, it was decided that all units in the SAMU should be subject to the RRG shift in PRN, though.

Note that a collection of independent uniform random variables that are shifted this way, remain such a collection. Hence, the RRG rotation does not change the probabilistic features of our sampling procedures. In particular, the equal probability sample will still be an srswor.

The RRG method was designed as a tool for co-ordinated rotation of all the SAMU samples. Unfortunately, this means that individual rotation schemes for the surveys are not possible. Another disadvantage is that we can not guarantee the units to be out of samples after (at most) 5 years. We can only say that if the inclusion probability is considerably less than 0.10, then the unit is very likely to be out of sample in the prescribed time.

5. FURTHER DETAILS ON THE SAMU SYSTEM

In this section we describe the sampling frame and the tools for stratification and allocation in the SAMU. We also give some details on the implementation of the JALES and RRG techniques in the SAMU. Additional information can be found in Ullberg and Segelberg (1989, in Swedish).

5.1. The Business Register and the SAMU sampling frame

The sampling frame used in the SAMU is based on the *Business Register* at Statistics Sweden, called the Central Register of Enterprises and Local Units (in Swedish "Centrala Företags och ArbetsställeRegistret", CFAR). In principle, this register contains records on all (active) businesses, authorities and organisations in Sweden, and their local units.

There are two levels in the register. The enterprise level (legal units) consists of all juridical persons in Sweden. It also contains physical persons who are employers, are registered for value added tax (VAT), have a registered firm and/or pay tax as businesses (rather than as employees). The main source of information on legal units is the National Tax Board.

Local units are the addresses where an enterprise is operating. For each enterprise in the register, there is at least one local unit. New multiple-location enterprises are detected with the aid of telephone directories and addresses on employers' tax lists. The register information on all known multiple-location enterprises is updated twice a year on the basis of special questionnaires. Out of more than half a million enterprises, only some ten thousand have multiple locations.

For the vast majority of enterprises, which have just one local unit, there is no regular questionnaire. There are several other sources for updating, though, one being the annual census of manufacturers with at least 10 employees (in Swedish "Industristatistiken").

Though the Business Register is kept by Statistics Sweden, it is not only used for our own surveys. The information in the register is also the basis for the National Register of Enterprises and Local Units, BASUN. The BASUN is used both by authorities and the private sector and is entirely financed by commercial revenue. Because of the

multi-purpose use of the Business Register, the units are not in all respects defined so that they are feasible as sampling (and reporting) units. Therefore, the Business Register has to be modified for the SAMU. A special register, the SR (in Swedish "Statistikregistret"), holds a modified version of the Business Register for the SAMU sampling frame. The main type of modification is that some of the units in the SR are combines of legal units. At present, local units are not modified. The SAMU sampling frame is a copy of the SR at the sampling occasion.

In the SAMU frame, an additional level of units is created, so called functional sampling units FU (in Swedish "funktionella urvalsobjekt"). These are best described through an example. Suppose we have a survey of manufacturers at the local unit level that uses samples drawn at the enterprise level. Now an enterprise in a non-manufacturer sector of the frame may have a few local units classified as manufacturers. These local units are aggregated to an FU. The FU's constructed in this way are added to the manufacturing industry before sampling. A point is that the size of the FU is the aggregated size of the relevant local units, not the size of the "mother" enterprise.

There are discussions on the possibility of introducing additional levels of sampling units in the SR and the SAMU.

5.2. Stratification and selection of population

All surveys in the SAMU are stratified according to industry (type of activity), as given by the SNI code (in Swedish "Svensk NäringsgrensIndelning"). From 1969 to 1993/94 SNI is based on the ISIC (the United Nations International Standard Industrial Classification of all economic activities). In the SAMU there are three hierarchical levels of stratification by industry to choose among. All three levels are aggregates of SNI codes. The intention is that the surveys should stay within this stratification, for the sake of commensurability. A few surveys, however, use their own grouping of SNI-codes. From 1993/94 a revised version of SNI will be used which is based on the

NACE Rev. 1 (the European Communities' revised "Nomenclature générale des Activités economiques dans les Communautés Européennes").

There are several other qualitative variables for stratification and selection, including region and variables related to ownership (e.g., incorporated business, governmental agency). It is also possible to add any feasible selection variable individually for a particular survey.

Most surveys use some measure of size for further stratification, the most common being 'number of employees'. This essentially integer-valued variable from the Business Register is defined as the number of employees at the last investigation (the tax pay-roll of last December; in the multiple-location case alternatively last questionnaire). Employees with a negligible income are excluded from the count and each employee is counted just once. For many surveys it would be more relevant to count the accumulated work-load during, say, last year, i.e. to give each employee a weight equal to her/his proportion of full-time work. At present, such a measure can unfortunately not be obtained.

A few surveys use other size measures, e.g. turnover as obtained from the VAT register. The surveys are free to add any size measure they may have to the survey frame.

5.3. Sample allocation

The SAMU has a built-in allocation system that can perform Neyman allocation (see e.g. Cochran, 1977) on any of the size variables. First a precision for each industrial stratum, in terms of relative confidence bounds, is chosen. The SAMU then gives the corresponding Neyman allocation over the size strata within the industrial strata. This allocation can be manually adjusted in an arbitrary way. In particular most surveys adjust the sample sizes so that no stratum contains less than 5 sampled units. Though

this is nominally a fixed precision allocation system, the allocation is usually adjusted to fit with a fixed budget.

In practice, the allocation variable is always the same as the size stratification variable. By force, the variation inside the strata of this variable is usually less than for the target variables of the survey. Hence, the chosen precisions can not be used as predictors of the precision of the survey. This fact can make it harder to do a proper allocation.

Mechanical use of the default Neyman allocation system often causes slight changes in allocations between the years. This increases the uncontrolled rotation in and out of sample, which is annoying for respondents. On the other hand, a small change in allocation usually has very little impact on the efficiency of our estimates.

Indeed, the involuntary rotation of persistants, due to deaths, births, re-classifications and re-allocation is a major problem of the system today.

5.4. The samples in the SAMU

The SAMU is used for about 15 annual and sub-annual surveys at Statistics Sweden; a few samples are drawn for external use. Some 35-40 thousand out of half a million enterprises are included in at least one SAMU sample (disregarding the external samples and the CPI sample). Of the sampled enterprises, only some 25 thousand are bothered with questionnaires. Most samples are drawn in December; samples for a few annual surveys are drawn in May, though. With PRN, there is no problem to co-ordinate the May samples with those of December. It should be noted that the samples in December year t are used for annual surveys covering year t and for sub-annual surveys covering year $t+1$.

The nine-digit PRN's of the local units in the SAMU frame are generated by a pseudo-random number generator and kept in a separate register. The period of this generator is chosen large enough so that we do not obtain any ties. For simplicity, the presum-

ably unimportant ninth digit of the PRN is also used to determine the rotation group. A single-location enterprise is given the same PRN as its local unit. A muliple-location enterprise is initially assigned the PRN of its largest local unit. In this case, the PRN is kept over the years even if the corresponding local unit is no longer the largest one. Note that more than one location can not be linked to the "mother" enterprise without destroying the independence of the random numbers. Hence, the SAMU is less efficient in co-ordinating samples drawn at different levels of the frame. For the vast majority of single-location enterprises this causes no problem. However, this forces us to use cluster (or two-stage) sampling of locations, at the enterprise level, whenever we want co-ordination across the levels.

The samples in the ultimate strata are sequential srswor's, except the sample for the Consumer Price Index and a few related samples which are drawn by sequential Poisson Sampling.

The surveys are grouped into "blocks" for which we use the same starting point and the same sampling direction. The starting points and sampling directions of the present six blocks in the SAMU are shown in Figure A1 of Appendix 2. This appendix also contains a list of the surveys. It should be noted that the sampling fractions can rate from less than 1% up to 100% for different strata of the same survey. Hence, the arrows in Figure A1 are just indicators of sampling directions, not of extensions of sampling intervals.

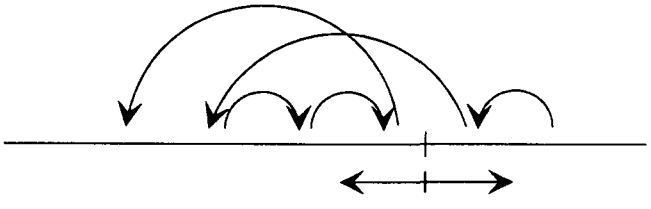
Figure A1 should not be seen as an ideal placing of samples, but rather as the result of a system that has gone through 20 years of additions and adjustments. In particular, we have found no explanation for the strange relation between Blocks 2 and 3. Note that there is no risk for overlap between Blocks 5 and 6, since Block 5 consists of surveys of the service sector, while the surveys in Block 6 are restricted to manufacturers only. The Consumer Price Index has an entirely different kind of data collection than the

other surveys; hence it is not so important to keep it negatively co-ordinated with the other surveys. There is no intention to make external surveys negatively co-ordinated with our own surveys. The reason behind the rest of the positive co-ordinations in Blocks 1, 5 and 6 is to make it possible to compare data on the micro level and/or use common questionnaires.

5.5. Rotation in practice

Since 1989, samples in the SAMU are rotated by the RRG method (Section 5.2). Once a year, before the sampling in December, the PRN's for the part of the population that is in the rotation group in question are shifted 0.10. Because of the complex pattern of starting points and directions (Figure A1) the shifting direction can not consequently be left (or right). Figure 6 below shows how we can avoid that units are rotated out of one sample and into another when two surveys have the same starting point in the interval (0,1).

Figure 6. Transformation of PRN's when two negatively co-ordinated surveys have the same starting point.



In accordance with the idea behind Figure 6, the first (decimal) digit of the PRN in the rotation group in question is shifted as follows:

For SNI 1-5: $7 \leftarrow 0 \leftarrow 1 \leftarrow 2 \leftarrow 3 \leftarrow 4 \leftarrow 6 \leftarrow 5 \leftarrow 9 \leftarrow 8 \leftarrow 7$

For SNI 6-9: $9 \leftarrow 0 \leftarrow 1 \leftarrow 2 \leftarrow 3 \leftarrow 4 \leftarrow 6 \leftarrow 5 \leftarrow 7 \leftarrow 8 \leftarrow 9$

For the sake of the rotation scheme, it would be preferable to have all the samples drawn in the same direction.

Ullberg et al. (1990) contains an evaluation of the new rotation technique. The net effect of rotating 20 percent of the PRN's varies from survey to survey and from stratum to stratum. On the average, only around 15 percent of the smallest units that were in some sample 1988 rotated out without being selected in another survey. For units in the class '20-49 employees', the efficient rotation was just 2 percent. The reason that we do not reach 20 percent rotation is that several sampling fractions are quite large even in these strata of rather small businesses .

Suppose we want to move the starting point (and/or sampling direction) in (0,1) for a sample, without substantially decreasing the overlap with earlier samples. This can be done in five years as follows. Year 1 we draw 1/5 of the sample at the new place, among units with the rotation group of that year (group 1, say); 4/5 of the sample are drawn at the old place in the other groups. Next year 2/5 of the sample are drawn at the new place in group 1 and 2, etc. This technique is presently used in the SAMU for part of the sample for the annual survey of salaries.

5.6. Variance estimation

Estimation based on a single sample from the SAMU causes no new problems, since the samples are ordinary stratified srswor's. Estimates of the change between different years are based on two samples, though. Let us say that we want to estimate the ratio of the current total, Y_t, to the total for the same period last year, Y_{t-1}. The conventional linearised variance formula of such a ratio contains a covariance term $\mathrm{Cov}(\hat{Y}_{t-1}, \hat{Y}_t)$, where ^ denotes estimate. This covariance depends on the amount of overlap between the samples at time t and $t-1$, and requires quite intricate calculations. The details are given in Garås (1989).

6. AN INTERNATIONAL OVERVIEW

Besides Sweden, permanent random number techniques are used for business surveys by statistical agencies in Australia, New Zealand and France (and perhaps in some other countries as well). It may be noted that in all these three cases, PRN is used exclusively for *equal* probability sampling.

6.1. The Australian Synchronised sampling system

The PRN technique used by the Australian Bureau of Statistics (ABS), called *Synchronised sampling* differs from the techniques introduced so far. Synchronised sampling replaced collocated sampling in 1982. It is described in detail in Hinde and Young (1984) and Australian Bureau of Statistics (1985). A shorter description and a comparison with the SAMU system can be found in Ohlsson (1993).

Synchronised sampling yields fixed size samples. The system protects against uncontrolled rotation in and out of the sample, which is a great disadvantage of the SAMU system. Synchronised sampling is as flexible as the SAMU in allowing surveys to have different stratifications and still be negatively co-ordinated. Positive co-ordination of surveys with different design is not readily obtained with this system, though. In particular, it is hard to get a large overlap of subsequent samples for the same survey when it is redesigned. With the other PRN techniques, a unit that changes stratum is likely to remain in sample (or remain out of sample); this is not necessarily the case with the Australian technique.

A rotation technique is used which permits each stratum of each survey to rotate at its own rate, and ensures that (small) units only have to stay in sample for a pre-set number of years. A problem with Synchronised sampling is that the system is quite complicated to administrate as compared to the SAMU system. Every stratum has its own selection interval and rotation pattern, to be adjusted each year. Another problem is that

the probability distribution of the sample is not known. In particular we can not be sure that the sample is an srswor. See Hinde and Young (1984) for details.

6.2. Poisson sampling in New Zealand

In 1989, the Department of Statistics in New Zealand started to use equal probability Poisson sampling with PRN. Surveys using the Business Directory as a frame are gradually introduced into the system, which by autumn 1992 included five annual and sub-annual surveys. The problem with variable sample sizes is handled by choosing large expected sample sizes. So far, the samples have not been rotated. For more information, see Templeton (1990).

6.3. Sequential srswor in France

A few years ago, a PRN technique for business surveys was introduced at the INSEE (Institut National de la Statistique et des Études Économiques) in France, see Cotton (1989). The sampling technique is sequential srswor. Negative co-ordination is obtained by transforming the random numbers after the drawing of each sample, rather than by sampling from different starting points. The technique is used to co-ordinate samples for different surveys drawn the same year *or* to co-ordinate subsequent samples for a single survey. It is not used to create an entire system for co-ordination of several surveys over time, as in the other countries mentioned above. The reason is that the sampling frame is updated by using information from the samples. The result of such updating can give selection bias in any PRN system.

APPENDIX 1 - A proof that the sequential technique yields an srswor

Again we consider the sequential technique by Fan et al. (1962) which was described in Section 2.2. We shall give a strict proof of the fact that this method yields an srswor. Let s be an arbitrary collection of n units from a population of size N. Let $\Pr(s)$ denote the probability that the sequential technique results in the sample$s$, when applied to this population. We shall prove that

$$\Pr(s) = \frac{1}{\binom{N}{n}} . \tag{11}$$

The proof is basically the same as the one given by Sunter (1977b); we have corrected some minor errors, though.

Let Y be the largest of the PRN corresponding to the n units in s, and let $f(x)$ be the probability density of Y. Conditioning on the outcome of Y we get

$$\Pr(s) = \int_0^1 \Pr(s|Y = x) f(x) dx .$$

The conditional probability of selecting s given that $Y=x$ is just the probability that the N-n units not in s all have PRN's larger than x and we get

$$\Pr(s) = \int_0^1 (1-x)^{N-n} n x^{n-1} dx = n\,B(n, N-n+1) = n\frac{(n-1)!(N-n)!}{N!} , \tag{12}$$

where we have used a well-known property of the Beta-function $B(\cdot,\cdot)$. Since the right-hand side of (12) equals that of (11) the proof is now complete.

APPENDIX 2 - Co-ordinated samples and their starting points in the SAMU

Figure A1. Starting points and sampling directions in the interval (0,1) for the six blocks of surveys in the SAMU.

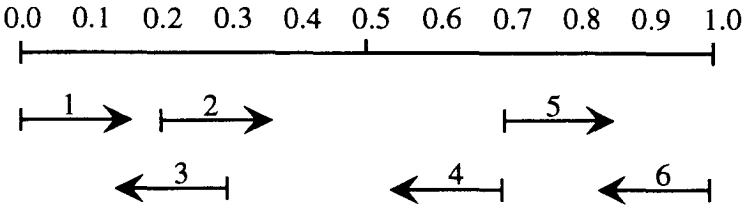


Table A1. Surveys in the six blocks of the SAMU.

Block 1. Financial Accounts of Enterprises, ISIC 1-5, annual
Fixed Capital Formation, sub-annual
Preliminary Company Profits, annual
Capacity Utilization of Mining and Manufacturing, quarterly
Consumer Price Index, monthly
(One external survey)

Block 2. Salaries, annual

Block 3. Expected Exports, sub-annual

Block 4. Employment, quarterly
Wages and Employment, monthly
Salaries, quarterly
(One external survey)

Block 5. Service Industries (sales and costs), annual
Financial Accounts of Enterprises, ISIC 6-9, annual
Wholesale and Retail Trade (sales and stocks), sub-annual

Block 6. Deliveries and Orders, ISIC 2-3, monthly
Stocks, ISIC 2-3, quarterly
(Two external surveys)

REFERENCES

Atmer, J., Thulin, G. and Bäcklund, S. (1975). "Co-ordination of samples with the JALES technique" *Statistisk Tidskrift,* Vol. 13, pp. 443-450. (In Swedish with English summary.)

Australian Bureau of Statistics (1985). "ABS Computing Network Systems Manual" Vol. VIII.

Brewer, K.R.W., Early, L.J. and Hanif, M. (1984). "Poisson, modified Poisson and collocated sampling" *Journal of Statististical Planning and Inference,* Vol. 10, pp. 15-30.

Brewer, K.R.W., Early, L.J. and Joyce, S.F. (1972). "Selecting several samples from a single population" *Australian Journal of Statististics,* Vol. 14, pp. 231-239.

Brewer, K.R.W. and Hanif, M. (1983). *Sampling with unequal probabilities.* Springer, New York.

Cochran, W.G. (1977). *Sampling Techniques, 3rd Ed.* Wiley, New York.

Cotton, F. (1989). "Use of SIRENE for enterprise and establishment statistical surveys" 4th International Roundtable on Business Survey Frames.

Fan, C.T., Muller, M.E. and Rezucha, I. (1962). "Development of sampling plans by using sequential (item by item) techniques and digital computers" *Journal of the American Statistical Association,* Vol. 57, pp.387-402.

Garås, T. (1989) "Förändringsestimatorer vid dynamiska populationer" F-Metod nr 21, Statistics Sweden. (In Swedish.)

Hinde, R. and Young, D. (1984). "Synchronized sampling and overlap control manual" Australian Bureau of Statistics.

Ohlsson, E. (1990). "Sequential sampling from a business register and its application to the Swedish Consumer Price Index" Stockholm, Sweden: Statistics Sweden R&D Report 1990:6.

Ohlsson, E. (1993). "Co-ordination of several samples using permanent random numbers". To appear in the monograph from the International Conference on Establishment Surveys, Buffalo 1993, Wiley.

Sunter, A.B. (1977a). "Respons burden, sample rotation, and classification renewal in economic surveys" *International Statistical Review*, Vol. 45, pp. 209-222.

Sunter, A.B. (1977b). "List sequential sampling with equal or unequal probabilities without replacement" *Applied Statistics,* Vol. 26, pp. 261-268.

Sunter, A.B. (1986). "Implicit longitudinal sampling from administrative files: A useful technique" *Journal of Official Statististics,* Vol. 2, pp. 161-168.

Särndal, C.E., Swensson, B. and Wretman, J.H. (1992). *Model assisted survey sampling.* Springer, New York.

Templeton, R. (1990). "Poisson meets the New Zealand Business Directory" *The New Zealand Statistician,* Vol. 25, pp. 2-9.

Thulin, G. (1976). "Co-ordination of samples in enterprise statistics (SAMU). Implementation of the JALES technique" *Statistisk Tidskrift* Vol. 14, pp. 85-101. (In Swedish with English summary.)

Ullberg, A., Segelberg, U. and Ohlsson, E. (1990). "Rotationssystemet i SAMU fr o m 1989" F-Metod nr 31, Statistics Sweden. (In Swedish.)

Ullberg, A. and Segelberg, U. (1989). "SAMU-handbok 1989" Statistics Sweden. (In Swedish.)

R & D Reports är en för U/ADB och U/STM gemensam publikationsserie, som fr o m 1988-01-01 ersätter de tidigare "gula" och "gröna" serierna. I serien ingår även **Abstracts** (sammanfattning av metodrapporter från SCB).

R & D Reports Statistics Sweden are published by the Department of Research & Development within Statistics Sweden. Reports dealing with statistical methods have green (grön) covers. Reports dealing with EDP methods have yellow (gul) covers. In addition, abstracts are published three times a year (light brown/beige covers).

Reports published during 1992:

1992:1 (grön)	Industrins konkurrenskraft och produktivitet i fokus - en utvärdering av statistiken **(Margareta Ringquist)**
1992:2 (grön)	Automated Coding of Survey Responses: An International Review **(Lars Lyberg and Pat Dean)**
1992:3 (grön)	TABELLER ,... TABELLER ,... TABELLER ,... - Variation och Förnyelse **(Per Nilsson)**
1992:4 (grön)	Basurval vid SCB? Studier av reskostnadseffekter vid övergång till basurval **(Elisabet Berglund)**
1992:5 (beige)	Abstracts I - sammanfattning av metodrapporter från SCB
1992:6 (grön)	Utvärdering av framskrivningsförfarande för UVAV-statistik **(Kerstin Forssén & Bengt Rosén)**
1992:7 (grön)	Cross-Classified Sampling for the Consumer Price Index **(Esbjörn Ohlsson)**
1992:8 (grön)	Bortfallsbarometern nr 7 **(Mats Bergdahl, Pär Brundell, Anders Lindberg, Håkan Lindén, Peter Lundquist, Monica Rennermalm)**
1992:9 (beige)	Abstracts II - sammanfattning av metodrapporter från SCB
1992:10 (gul)	Organizing the Metainformation Systems of a Statistical Office **(Bo Sundgren)**
1992:11 (grön)	CLAN - ett SAS-program för skattningar av medelfel **(Claes Andersson, Lennart Nordberg)**

1992:12 (grön)	KVALITETSRAPPORTEN - Utveckling av kvaliteten för SCBs statistikproduktion **(Jan Eklöf, Per Nilsson)**
1992:13 (grön)	The Use of Registers as Auxiliary Information in the Swedish Labour Force Survey **(Jan Hörngren)**
1992:14 (beige)	Abstracts III - sammanfattning av metodrapporter från SCB
1992:15 (grön)	Operationalising a Hedonic Index in an Official Price Index Program: personal computers in the Swedish import price index **(Jörgen Dalén)**
1992:16 (gul)	Some Properties of Statistical Information: Pragmatics, Semantics, and Syntactics **(Bo Sundgren)**
1992:17 (gul)	Statistical Metainformation Systems - Pragmatics, Semantics, Syntactics **(Bo Sundgren)**

Kvarvarande **beige** och **gröna** exemplar av ovanstående promemorior kan rekvireras från Inga-Lill Pettersson, U/LEDN, SCB, 115 81 STOCKHOLM, eller per telefon 08-783 49 56.

Kvarvarande **gula** exemplar kan rekvireras från Ingvar Andersson, U/LEDN, SCB, 115 81 STOCKHOLM, eller per telefon 08-783 41 47.

SCB-Tryck, Örebro 1996.06