Coordination of business surveys

Li-Chun Zhang[1]

1 Introduction

In coordination of business surveys we have two main concerns: an even distribution of response burden over time among comparable units, and good statistical properties of the samples. An even distribution of response burden is an important factor in user satisfaction and, thus, a key quality indicator of statistical production. But it needs to be balanced against another main quality measure, namely accuracy. Thus, for instance, the larger business units will overall have to participate in many more surveys because of their statistical importance, while to avoid unnecessary over-burdening may assume a high priority among the smaller business units. Subjected to the requirement of accuracy, our task in sample coordination is to define as well as to carry out the exact conditions for an even distribution of response burden. The aim is to make survey participation transparent, fair and predictable.

In this paper we present the newly developed Norwegian system of coordinated samples (Norsamu). Norsamu is based on a set of explicit principles and conditions for coordination. These are the principle of sample rotation and that of long-run even distribution of response burden, and the condition for quarantine and that of mutually exclusive survey participation as well as the condition for overlapping samples. The key elements of Norsamu include the domains of coordination, the two counters of survey participation, and the various sample rotation schemes. To implement these conditions there are two fundamental requirements: proper adminstration of the frame and standardized sampling designs. We explain how the frame is maintained and updated over time, in order to handle frequent births, deaths, merging and division that occur among the business units. On the one hand, the samples that are drawn at different times of the year should be able to make use of the most updated frame, while on the other hand, one needs to be able to keep track of which units are comparable to each other over time.

2 Sample coordination requires coordination of frame and design

In most business surveys the population is stratified by economic activity (NACE) and a chosen size variable. Table 1 illustrates the situation within a given NACE domain, where there are 8 size domains (I - VIII) and 6 different surveys (A - F) to be coordinated. It provides a clear picture of the task at hand. For instance, it becomes immediately clear that *some* business units will have to take part in at least 2 surveys in domain IV, and *some* will have to take part in at least 3 surveys in domain V, and so on. These facts can not be changed by the method of coordination, without modifications of the given sampling designs.

Table 1 shows that sample coordination fundamentally requires coordination of the frames and the sampling designs. A *common* frame must be put into place, and it must be identified for each unit which target survey populations it belongs to. There must be a common size classification in order to tell unequivocally which units are compare to each other. The advantage of a common

[1]Statistics Norway, Kongensgt. 6, PB 8131 Dep, N-0033 Oslo, Norway. E-mail: lcz@ssb.no

Table 1: Illustration of sample coordination within a given NACE domain

Sample Size		Size Domain by Number of Employees								
Type	Survey	I	II	III	IV	V	VI	VII	VIII	Total
Structural	A	0	0	373	289	177	91	23	10	963
Structural	B	0	0	116	241	295	181	30	10	873
Structural	C	0	0	745	385	118	91	15	10	1364
Structural	D	0	167	466	964	590	181	30	10	2408
Short-Term	E	442	2673	3726	3854	1179	181	30	10	12095
Short-Term	F	0	0	112	193	177	91	30	10	613
	Total	442	2840	5538	5926	2536	816	158	60	18316
Population Size		4420	10690	7451	3854	1179	181	30	10	27815
Sampling Fraction (%)		Stratum by Number of Employees								
Type	Survey	I	II	III	IV	V	VI	VII	VIII	Total
Structural	A	0	0	5	7	15	50	77	100	3
Structural	B	0	0	2	6	25	100	100	100	3
Structural	C	0	0	10	10	10	50	50	100	5
Structural	D	0	2	6	25	50	100	100	100	9
Short-Term	E	10	25	50	100	100	100	100	100	43
Short-Term	F	0	0	2	5	5	50	100	100	2
	Total	10	27	74	154	215	450	527	600	66

standardized frame is enhanced for compatible sampling designs. Therefore, stratified simple random sampling that conforms to the standard NACE-size classification is the default sampling design in Norsamu. Notice that one does not need to specify the sampling design at the level shown in Table 1, but the stratum division must coincide with the NACE-size classification at the most detailed level that is allowed for. For example, in Table 1, domain III - V actually belong to a single design stratum for survey C, which is evident from the constant sampling fraction. Indeed, the 8 domains in Table 1 belong to the *minimum partition* of the frame, where each part is referred to as a *domain of coordination (DOC)*. Formally, the minimum partition and the DOCs must have the following property: (a) the DOCs constitute a partition of the frame, (b) within each DOC there is a constant sampling fraction for any given survey, and (c) no other partition of the frame can have fewer parts than the number of DOCs. Now, when a design stratum is a union of several DOCs, the specified stratum sample size can be scaled down to the relevant DOC sample sizes proportionally, which then share the same specified sampling fraction as in Table 1. In this way the different sampling designs are put in coordination.

3 Conditions of coordination

Norsamu adopts the following two principles of coordination:

-*Principle of sample rotation (R)*: For all comparable units, a common rotation rate applies in all structural business surveys, and a common rotation rate applies in all short-term business surveys.

-*Principle of long-run even distribution of response burden (L)*: Over the time the expected ratio

between in- and out-of-sample periods should be equal among comparable units.

The goal of sample rotation is to achieve the long-run even distribution of response burden. Exceptions are however necessary at any given moment, as long as there are some units that have to participate in the surveys while not the others. The following conditions of coordination are used to regulate the response burden *in a given DOC* at any given time point:

-*Condition of mutually exclusive survey participation (M)*: Surveys are put under the condition of mutually exclusive survey participation if the corresponding samples are not permitted to overlap.

-*Condition of quarantine (Q)*: Surveys are put under the condition of quarantine if participation in one of them is followed by a guaranteed minimum-length resting period from all of them.

Notice that, from time to time, the condition (Q) implies that certain units are made ineligible to all surveys, while the condition (M) implies that certain units may be ineligible to some surveys. As such both do cause design inconsistency. But it is a price that one pays in order to improve the fairness and predictability of survey participation. Notice also that the condition (M) is desirable only if the response burden is thought to be more evenly distributed when two different units participate in two different surveys, than it is for the same unit to participate in both. Exceptions are conceivable. Hence, the following condition of overlapping samples:

-*Condition of overlapping samples (O)*: Sample overlap among surveys is permitted only if it is *designed* to reduce the response burden and/or the estimation uncertainty.

4 Administration of sample rotation

Sample rotation takes place on a yearly basis in Norsamu, so we calculate the response burden by the number of years a unit participates in a survey.

To keep track of survey participation for a given business unit we construct two counters for each target survey population to which it belongs. Counter (A) records participation over time, and is updated as follows: it is increased by 1 for each year the unit participates in the survey of concern, whereas it is decreased by 1 for each year the unit rests from the same survey. Counter (B) records the current *consecutive* years of participation or resting, and is updated as follows. After each sample rotation, counter (B) in increased by 1 if the unit is selected in both the current and previous samples, it is decreased by 1 if the unit is not in the current sample nor in the previous one, it is set to 1 if the unit is in the current sample but not in the previous one, it is set to -1 if the unit is in the previous sample but not in the current sample.

There are two basic choices when it comes to sample rotation of a given survey. Under *simple* rotation, the units are rotated in and out of the sample only based on the history for the given survey. On each occasion, one rotates out the units that have the largest (positive) values of counter (B), and rotates in the units that have the smallest values first according to counter (A) and then counter (B). Under rotation with *quarantine*, a set of surveys that are put under the same quarantine system are administrated jointly. For sample rotation in a given survey, the units are rotated out of the sample in the same way as under simple rotation. Next, all the out-of-sample units that have not been given the guaranteed minimum-length resting period are identified according to counter (B). The rest out-of-sample units are then sorted according to the

sum of counter (A) across the quarantine system and possibly counter (B) for the given survey, and the first ones in the queue are rotated into the sample.

Obviously, long-run even distribution of response burden is achieved for each and all surveys that employ simple rotation. The main problem with a system that only uses simple rotation is the lack of control over the maximum number of survey participation in a given year. For instance, it is possible for a unit in DOC III (Table 1) to be selected for all the 6 surveys in one year, while there are always units that are not involved in any of the surveys in the same year. Rotation with quarantine can be used to regulate the upper limit of burden in this respect. For instance, it is possible to put DOC III under a single quarantine system. Subjected to the condition (M), i.e. maximum one survey per year, the units will have a one-year guaranteed minimum-length of resting period. The accumulated response burden, i.e. both across the surveys and over the time, is evened out more quickly under the quarantine system. But the participation is more unbalanced across the surveys in the sense that over time some units will participate more often in a certain survey while some others will do so in another survey, although which units for which survey here is entirely a matter of chance. Another aspect of this imbalance is that under the quarantine system it takes longer time for all the units to be selected to a given survey at all, which may be a potential concern for sample representativeness.

Typically in Norsamu, simple rotation is applied to surveys with either a relatively high sampling fraction (say, 50% or more) within a given DOC, or a clearly more time-consuming questionnaire such that even participation in the corresponding survey is important for the perceived fairness. As such simple rotation is more frequently used among medium to large business units, while small business units are often put under some quarantine system. Notice that it is possible to combine the two rotation schemes within a single DOC such that some surveys are set for simple rotation on their own, while the others are put under one or several mutually exclusive quarantine systems. For instance, again in DOC III (Table 1), one may choose simple rotation for survey E due to its relatively high sampling fraction, while putting the rest surveys under a single quarantine system with a guaranteed 4-year minimum resting period.

Table 2: Illustration of a coordination table for Norsamu

DOC	Population	Sample Size	Quarantine System	Quarantine Length	
Nr.	Size	1 $\cdots$ K	1 $\cdots$ K	1 $\cdots$ K	$\cdots$
1	N_1				
2	N_2	$n_{d,k}$	$I_{d,k}$	$Q_{d,k}$	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\cdots$
D	N_D				

All the information of coordination in Norsamu are stored in a *coordination table*. Table 2 provides an illustration, where the common frame is divided into D domains of coordination, each of the size N_d. There are up to K surveys to be administered in each DOC. The domain sample size is given by $n_{d,k}$ for the k-th survey in d-th DOC. The quarantine indicator $I_{d,k}$ takes a common value for all the surveys that are put under the same quarantine system, and it is unique for a survey that is put on simple rotation. The corresponding $Q_{d,k}$ specifies then the guaranteed

minimum-length of resting period, and so on. The coordination table is updated over time for changes and provides a continuous overview of Norsamu.

5 Updating of Norsamu

The most important frame updating procedures in Norsamu are as follows:

- The common frame is formed once a year.
- A birth unit is admitted to the appropriate DOC as it occurs, on which occasion a 'history' of survey participation, i.e. the set of counters (A) and (B), is randomly selected from the existing units in the DOC and assigned to the birth unit.
- A unit is kept in the same DOC despite possible changes in classification variables during the year. It is moved to the appropriate DOC when the common frame is updated, where it is treated as a birth unit to the new DOC.

Table 3: Illustration of changes in frame

	t_2				Transition t_1 - t_2			Transition t_2 - t_1		
t_1	I	II	III	Death	I	II	III	I	II	III
I	383590	1890	402	17457	0.994	0.005	0.001	0.9942	0.133	0.035
II	2052	11308	595	21	0.147	0.810	0.043	0.0053	0.798	0.052
III	187	968	10371	11	0.016	0.084	0.900	0.0005	0.068	0.912
Birth	28684	202	102							

The freezing of the DOC in case a unit changes status is necessary in order to keep a consistent account of which units are comparable with each other throughout a year. However, for a number of reasons, samples are drawn at different time points in a year. A *conversion program* is needed to allow one to use the most updated frame information. Table 3 illustrates the situation. Suppose two time points during the years, denoted by t_1 and t_2, where t_1 is the time point of the common frame and t_2 refers to some time point afterwards. The population is divided into 3 subpopulations, denoted by I - III. The first block in Table 3 gives the counts by cross-classification according to the status at t_1 and t_2. The margin Birth and Death correspond respectively to birth and death units between t_1 and t_2. The second block gives the domain-transition probabilities from t_1 to t_2 (i.e. with row sums equal to 1), and the third blocks gives those from t_2 to t_1 (i.e. with column sums equal to 1). Suppose one would like to drawn a sample at time point t_2, with stratum sample sizes (n_I, n_{II}, n_{III}). Using transition probabilities from t_2 to t_1, we obtain $n'_I = 0.9942 \cdot n_I + 0.133 \cdot n_{II} + 0.035 \cdot n_{III}$, and similarly for n'_{II} and n'_{III}. It can now be verified that $n_I = 0.994 \cdot n'_I + 0.005 \cdot n'_{II} + 0.001 \cdot n'_{III}$, and similarly for n_{II} and n_{III}. In other words, if one draws a stratified sample with the stratum sample sizes $(n'_I, n'_{II}, n'_{III})$ according to the frame at t_1, then the expected stratum sample sizes are (n_I, n_{II}, n_{III}) according to the frame at t_2, just as specified. Notice that the same conversion program can also be used in other situations where the population is subjected to two different classifications, such as changes in NACE codes.