UNITED NATIONS

E

Economic and Social Council

Distr.
GENERAL

ECE/CES/2009/31
22 April 2009

ENGLISH ONLY

ECONOMIC COMMISSION FOR EUROPE **STATISTICAL COMMISSION**

CONFERENCE OF EUROPEAN STATISTICIANS

Fifty-seventh plenary session
Geneva, 8-10 June 2009
Item 4 of the provisional agenda

SEMINAR ON STRATEGIC ISSUES IN BUSINESS STATISTICS

SESSION I: REDUCING RESPONDENT BURDEN

DEVELOPMENT OF A SAMPLE COORDINATION SYSTEM FOR ENTERPRISE SURVEYS AT THE FEDERAL STATISTICAL OFFICE OF SWITZERLAND

Note by the Federal Statistical Office of Switzerland

I. INTRODUCTION

1. The general program of enterprise statistics is one component of the modernization strategy being currently implemented at the Federal Statistical Office. This program aims at reforming the enterprise statistics and at developing a coherent and integrated system of enterprise surveys.

2. Burden reduction is one of the goals of the general program of enterprise statistics. Burden is defined here as the number of surveys in which an enterprise has to participate over a given period, e.g. one calendar year. As a rough measure of burden we can consider the overall sample size of the surveys carried out during one year, divided by the total number of enterprises. This quantity can be interpreted as the mean expected burden per enterprise. More refined measures of burden would be obtained by considering enterprises in different types of economic activity, size classes, etc. Burden reduction can be achieved through a better use of administrative data. Indeed, given the surveys which have to be conducted during one year, burden reduction can

only be achieved through reduced sample sizes. This entails a loss of efficiency, which may possibly be compensated by more efficient sampling plans and estimation procedures, provided that strong auxiliary information is available. Due to their importance, the large enterprises have to participate in all surveys. For large enterprises burden reduction cannot be achieved through sample size reduction but through profiling, e.g. personalized contacts with the enterprises, centralized administration of all the surveys in which they have to participate, etc.

3. Once all measures have been taken to reduce the burden, a National Statistical Institute (NSI) has the further possibility of spreading this burden evenly over the units in the population. This is the goal of a sample coordination system.

4. This paper addresses the following issues:

(a) Performance of a sample coordination system

The development of a sample coordination system is a fairly complex endeavour and one may legitimately ask what advantages such a system brings over independent selection of samples;

(b) Link with the statistical business register

A sample coordination system is based on a common sampling frame for all surveys. The business register is the basic tool for constructing and updating the sampling frame. A close coordination of sampling frame and business register is important;

(c) Impact on surveys

The introduction of a sample coordination system has consequences on the planning and sampling design of surveys.

5. The paper is organized as follows. In Section II we define the concepts of negative and positive coordination. In Section III we discuss the relationship between sampling frame and business register. In Section IV we describe the selection algorithm chosen by the Federal Statistical Office (FSO) for sample coordination. In Section V we address the issue of how to measure the performance of a sample coordination system. In Section VI finally we discuss some of the impacts that a sample coordination system can have on the surveys.

II. SAMPLE COORDINATION

6. The current practice is to design and carry out each survey independently of the other surveys. Each survey has its own sampling frame. Often special units are defined, in order to facilitate data collection. Furthermore, the sample is selected according to a highly optimized sampling plan, based on a very detailed stratification. For each survey, the sample is selected independently of the samples already selected for other surveys. Thus the overlaps of the samples for different surveys are not controlled.

7. The problem of sample coordination is to select several samples according to given sampling plans, while controlling the sizes of the intersections of the samples. In practice the samples are selected sequentially. Thus, when selecting a sample at a given occasion, according to a given sampling plan, we want to control the sizes of its overlaps with the already selected samples.

8. Spreading the response burden corresponds to negative coordination, in which case the sizes of the intersections have to be minimized. If the total size of the samples is less than the size of the population, then it is possible to select disjoint random samples. Otherwise some overlap is unavoidable.

9. Most of the enterprise surveys are repeated surveys, where the same units have to be observed on several occasions. In this case a certain degree of overlap between samples is required. This can be achieved through a rotating panel: at each new selection of a sample some of the units are kept in the sample, some are eliminated and replaced by new units. In the most extreme case of positive coordination, as many units as possible are retained.

III. LINK WITH THE BUSINESS REGISTER

10. When doing sample coordination, each selection of a sample has to take into account the samples which have already been selected. To be able to do this, we need a common sampling frame for all surveys. Furthermore, this sampling frame must keep track of the earlier surveys. Thus for each survey, the definition of the reference population, the stratification and the selected sample must be archived. The sampling frame has a limited validity, e.g. one year, and must be updated regularly. The new version of the frame must be compared with the old version, to identify units which have disappeared or have been created. Changes of structure must be identified and properly handled.

11. The sampling frame is built from the business register, but is not just a copy of it. Rather, the business register must be seen as the fundamental tool for constructing and updating the sampling frame. The business register provides a list of different types of units (legal unit, enterprise, local unit), together with links between them. Information is available on these units, such as address, economic activity, legal form, employment, turnover, etc. This information is updated on a regular basis, and longitudinal links between units must be available, allowing identifying demographic events and changes of structure (concentration and fragmentation).

12. The sampling frame should contain information which is as up to date as possible. The choice of the reference time at which to construct or to update the sampling frame is therefore important. To illustrate the problem we show in Table 1 a simplified view of when information will be available in the FSO business register for turnover and employment, once the general program of enterprise statistics GUS (from the German "Gesamtprogramm Unternehmensstatistik") will be implemented. For example, in December 2011, turnover is available for 2010 and the first quarter of 2011, employment for 2009 is available from administrative sources for the units existing up to the end of 2009 and, from updating surveys, for the new units entering in the business register in 2010 and in the first two quarters of 2011.

Table 1
Updating information for the business register

Calendar time		Availability of the auxiliary information										
Year	Month	Turnover					Employment		Employment, new units			
2010	1	2008	1q09				2007		2008	1q09	2q09	
	2	2008	1q09				2007		2008	1q09	2q09	3q09
	3	2008	1q09	2q09			2007		2008	1q09	2q09	3q09
	4	2008	1q09	2q09			2007		2008	1q09	2q09	3q09
	5	2008	1q09	2q09			2007		2009 [a]			
	6	2008	1q09	2q09	3q09		2007		2009			
	7	2008	1q09	2q09	3q09		2007		2009			
	8	2008	1q09	2q09	3q09		2007		2009	1q10		
	9	2009 [a]					2007		2009	1q10		
	10	2009					2008		2009	1q10		
	11	2009					2008		2009	1q10	2q10	
	12	2009	1q10				2008		2009	1q10	2q10	
2011	1	2009	1q10				2008		2009	1q10	2q10	
	2	2009	1q10				2008		2009	1q10	2q10	3q10
	3	2009	1q10	2q10			2008		2009	1q10	2q10	3q10
	4	2009	1q10	2q10			2008		2009	1q10	2q10	3q10
	5	2009	1q10	2q10			2008		2010 [a]			
	6	2009	1q10	2q10	3q10		2008		2010			
	7	2009	1q10	2q10	3q10		2008		2010			
	8	2009	1q10	2q10	3q10		2008		2010	1q11		
	9	2010 [a]					2008		2010	1q11		
	10	2010					2009		2010	1q11		
	11	2010					2009		2010	1q11	2q11	
	12	2010	1q11				2009		2010	1q11	2q11	

[a] Data available for the whole year

IV. ALGORITHM FOR SAMPLE COORDINATION

13. There are a number of methods for sample coordination which have been proposed in the literature, see e.g. Ohlsson (1995) and Nedyalkova et al. (2008a). The algorithm used at the FSO for coordinated sample selection is based on Poisson sampling and permanent random numbers. This method has been chosen for its theoretical simplicity, its ease of implementation and also for its ability to handle several rotating panels, negatively coordinated among them. The algorithm was developed, within the framework of a research convention between the FSO and the University of Neuchâtel, in collaboration with the Institute of statistics; see Nedyalkova et al. (2008b).

14. Each unit of the sampling frame receives a random number uniformly distributed between 0 and 1. This random number stays associated to the unit as long as it exists. For each survey, one defines for each unit a zone of selection. In the simplest case the zone of selection is an interval. In more complex situations the zone of selection can be the union of disjoint intervals.

The total length of the zone of selection corresponds to the inclusion probability for that unit. Finally, a unit is selected in the sample if its permanent random number falls within its zone of selection. Different types of coordination (negative, positive, rotation) can be achieved by an appropriate choice of the zones of selection.

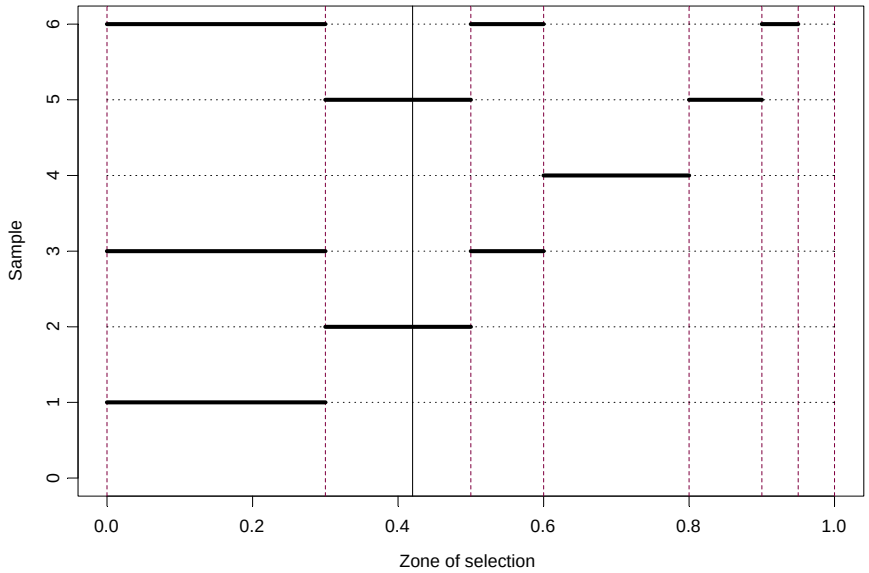
15. We consider as an example the selection of a unit in 6 samples. Table 2 below gives the sampling fractions (*f*) for that unit in the 6 samples, as well as the types of coordination desired. In this example we have two panels: the samples 1, 3 and 6 are three waves of the panel 1 and the samples 2 and 5 are two waves of the panel 2. The sample 4 is for a survey conducted only once.

Table 2
Sampling fractions and types of coordination

				Coordination with sample				
Sample	*f*	Panel	Wave	1	2	3	4	5
1	0.30	1	1					
2	0.20	2	1	N				
3	0.40	1	2	P	N			
4	0.20			N	N	N		
5	0.30	2	2	N	P	N	N	
6	0.45	1	3	P	N	P	N	N

16. Here we consider only negative (N) or positive (P) coordination. Globally the two panels and the sample 4 have to be coordinated negatively. Thus, for example, sample 4 has to be coordinated negatively with the samples 1, 2 and 3. For a panel, the current wave has to be positively coordinated with the earlier waves. Thus, for example, the second wave of the panel 2 (sample 5), is positively coordinated with the first wave of the panel (sample 2), and negatively coordinated with all the other samples (samples 1, 3 and 4). Figure 1 below shows the zones of selection associated to the 6 samples. Assuming that the permanent random number is equal to 0.42 for the unit under consideration, we see that the unit is selected in the waves 1 and 2 of the panel 2, and not selected in the samples 1, 3, 4 and 6.

Figure 1
Zones of selection for sample coordination



V. PERFORMANCE OF SAMPLE COORDINATION

17. The performance of a sample coordination algorithm can be measured by comparing the current practice of independent selections of samples with coordinated selections. We consider in this paper only the special case of global negative coordination in a population which is not changing.

18. It is important to note that there are characteristics of repeated sampling procedures which do not depend on the selection algorithm. These are the properties which depend on the first order inclusion probabilities only. One important example of such a property is the mean expected burden per unit. Thus, mean expected burden per unit stays the same, whatever method of coordination is chosen. This means that a sample coordination system cannot reduce the burden; it can only spread it as equitably as possible among the units in the population.

19. Another parameter of repeated sampling, essentially equivalent to mean expected burden per unit, is mean expected total time out of sample per unit. As for total burden, this parameter stays the same, whatever method of coordination is chosen.

20. The effect of coordinating the samples is on the distribution of burden. In particular the variance of burden does depend on the coordination algorithm. For negative coordination with Poisson sampling, burden can take only two consecutive values, which depend on the mean expected burden per unit. One can show that the variance of burden is then bounded by 0.25. On the other hand, for independent selections of T samples, burden can take any of the values from 0 to T, and the variance of burden in this case is unbounded. Thus, the effect of negative coordination is to drastically reduce the spread of the distribution of burden.

21. We illustrate these properties by an example, see Table 3. We assume that at each occasion the sampling fraction is the same for all units in the population (Bernoulli sampling). In this case, the mean expected burden per unit is simply the sum of the sampling fractions.

Table 3
Sampling fractions for negative coordination

Sample	Sampling fraction
1	0.4
2	0.3
3	0.5
4	0.5
5	0.6
Mean expected burden per unit	2.3

22. With a mean expected burden per unit of 2.3, burden under negative coordination can only take the values 2 or 3, whereas for independent selections the burden can take the values from 0 to 5. Table 4 compares the distribution of burden under independent and coordinated sampling. For both selection algorithms, the mean expected burden per unit is equal to 2.3, as was to be expected. The effect of spreading the burden is to concentrate the distribution of burden on the values 2 and 3. It follows that no unit has burden 4 or 5, but also that no unit has burden 0 or 1. Thus the mean expected burden of 2.3, which can be seen as an external parameter reflecting the selection of 5 samples with the indicated sampling fractions, is spread as equitably as possible. The concentration of burden distribution is also apparent in the variances: 1.19 for independent selection vs. 0.21 for negative coordination.

Table 4
Probability distribution of burden

	Probability of Burden	
Burden	Independent selection	Negative coordination
0	0.04	
1	0.19	
2	0.34	0.70
3	0.29	0.30
4	0.12	
5	0.02	
Mean	2.3	2.3
Variance	1.19	0.21

23. Total burden is still a rather coarse measure. One is often interested in the time between two selections in a sample. One has then to examine the different patterns of being in or out of sample. For the selection of 5 samples there are 32 such patterns, see Table 5. Pattern 00101 means that a unit is not selected in the samples 1, 2 and 4 and is selected in the samples 3 and 5. With the sampling fractions given in Table 3, the probability of obtaining this pattern under independent selections is 0.063, while under negative coordination the probability is 0.30. Here also, all patterns can occur under independent selections, while under negative coordination we have a concentration on a small number of patterns.

Table 5
Probability distribution of selection patterns

		Probability of Pattern				Probability of Pattern	
Pattern	Burden	Independent selection	Negative coordination	Pattern	Burden	Independent selection	Negative coordination
00000	0	0.042		11111	5	0.018	
00001	1	0.063		01111	4	0.027	
00010	1	0.042		10111	4	0.042	
00100	1	0.042		11011	4	0.018	
01000	1	0.018		11101	4	0.018	
10000	1	0.028		11110	4	0.012	
00011	2	0.063		00111	3	0.063	
00101	2	0.063	0.30	01011	3	0.027	
00110	2	0.042		01101	3	0.027	
01001	2	0.027		01110	3	0.018	
01010	2	0.018	0.30	10011	3	0.042	0.10
01100	2	0.018		10101	3	0.042	0.20
10001	2	0.042		10110	3	0.028	
10010	2	0.028	0.10	11001	3	0.018	
10100	2	0.028		11010	3	0.012	
11000	2	0.012		11100	3	0.012	

24. It would be possible to define characteristics summarizing some properties of the profiles and to compute their distributions under independent or coordinated selections. For examples one could compute the distribution of the lengths of time between two selections in a sample, considering that longer periods out of sample are desirable.

25. It is important to realize that a sample coordination system cannot guarantee that a unit will stay out of sample for a given length of time. If the sampling fractions are too high, the unit will have to be selected again, this being for example the case for the pattern 10011, occurring with probability 0.1 even under optimal negative coordination, see Table 5. All one can say is that small sampling fractions are desirable to make sample coordination worthwhile. With small sampling fractions one can have, for example, patterns with long periods out of sample, and the effect of coordination will be to concentrate the distribution on these good patterns.

VI. IMPACT ON SURVEYS

26. The introduction of a sample coordination system has consequences on the sampling designs of surveys. We argued in the preceding Section that small sampling fractions are desirable. This can be achieved by small sample sizes, i.e. by designing efficient sampling plans making optimal use of the available auxiliary information. Once the sample size is fixed, one can still try to make the sampling fractions as small as possible by defining large strata. Of course, small sampling fractions can only be used for the population of small and medium size enterprises. Since we use Poisson sampling, sample size is random and cannot be strictly controlled. It must also be noted that techniques for coordinating samples of different types of units, e.g. enterprises and establishment, are at present not available. For the sample coordination system being implemented at the FSO, it has been decided to spread the burden at the enterprise

level only. Thus, it appears that a sample coordination system imposes some constraints on the definition of the strata, the overall sample size and the sample sizes in the strata, the rotation rate for panels and the choice of the sampling unit.

27. One further aspect of the introduction of a sample coordination system is a shift in perspective from the planning of individual surveys to a global planning of several surveys. Thus one may attempt to adapt the calendars of the surveys to the updating of the sampling frame. Also one may consider parameters like mean expected burden per unit or mean expected time out of sample per unit as instruments for planning and designing an integrated system of enterprise surveys.

VII. REFERENCES

Ndeyalkova, D., Pea, J. and Tillé, Y. (2008a). Sampling Procedures for Coordinating Stratified Samples: Methods Based on Microstrata. *Int. Stat. Rev.*, 76, 368-386.

Ndeyalkova, D., Qualité, L. and Tillé, Y. (2008b). Tirages coordonnés d'échantillons à entropie maximale. Working paper. Institute of statistics. University of Neuchâtel.

Ohlsson, E. (1995). *Business Survey Methods*. Vol. 1, chap. 9, "Coordination of samples using permanent random numbers", pp. 153-169. New York: Wiley, Inc.

* * * * *