Working Papers

06/2010

Sampling coordination of business surveys in the Spanish National Statistics Institute

Dolores Lorca

M. Concepción Molina

Gonzalo Parada

Ana Revilla

First draft: July 2011

This draft: July 2011

Sampling coordination of business surveys in the Spanish National Statistics Institute

Abstract

The Spanish NSI works in several alternatives in order to reduce the statistical burden in business surveys. One of them is the use of sampling coordination techniques to reduce the overlap between samples of different surveys.

The use of the same sampling frame for business surveys (the Central Business Register) has allowed to obtain coordinated samples using the Permanent Random Number (PRN) technique. A statistical burden function is defined and used to coordinate the samples obtained each year. .

Keywords

Statistical burden, sampling coordination, business survey

Authors and Affiliations

Dolores Lorca, M. Concepción Molina, Gonzalo Parada and Ana Revilla

S. G. of Sampling and Data Collection

I. INTRODUCTION

The most common sample design used in business surveys is the stratified simple random sampling. The definition of strata is made by region, main economic activity and size, according to the number of employees.
In business populations, the study variable distributions are usually very skewed. Few large businesses account for a large portion of the totals for several survey variables. This produces high sampling fractions (more than 30%) in medium and large business strata, which implies an important statistical burden.
To reduce this statistical burden, in a way that it is evenly spread among the population units over repeated selections, the business samples are drawn in a coordinated way applying the Permanent Random Number (PRN) technique together with a statistical burden function.
Since 1992 the use of the same sampling frame for business surveys, the Central Business Register (CBR), it has allowed to apply sampling coordination techniques. Initially, we drew positive coordination samples maximizing the overlap between them. Our goal was to achieve coherence in the results. For example, the samples of the Retail Trade and Services Sector Activity Indices with the sample of Annual Trade and Services Survey or the sample of the Technological Innovation in Companies Survey with the sample of the Industrial Companies Survey were positively coordinated.

The increase in the demand of statistical information during the years 1999 to 2005, due to new European regulations and to the requirement of disaggregate data for the autonomous regions, originated the necessity of carrying out new surveys or to rise the sample size of the current surveys. This has caused that one of our main concerns is to reduce the statistical burden and for this reason, nowadays the sampling coordination techniques are usually applied to obtain negative coordinated samples, minimizing the overlap among them.

Spanish NSI keeps on working to reduce the statistical burden. There exist several ways to reach this objective. We point out some of the strategies:

-Using administrative data reducing sampling sizes without loss in quality statistic. The SIMFAES project is developing to reach this objective.

-To collect data through WEB of all economic surveys.

-To collect data using business's information system. Sampling businesses can send us XML files obtaining from their own information system.

-To provide free of charge tailored data to sampling businesses in most economic survey.

-To coordinate the data colleted of structural economic surveys trough the specialized units of data collection.

In the following section we introduce the statistical burden function that use together the PRN technique for the negative sampling coordination. We distinguish between negative sampling coordination of different surveys in a period and over time. In section

3 the effectiveness of the coordination sampling is calculated. Finally, some conclusions are provided.

II. SAMPLING COORDINATION: A FUNCTION OF STATISTICAL BURDEN

The goal of sampling coordination is to reduce the statistical burden spreading evenly the burden among the population units over repeated selections. For that matter, we apply the PRN technique.
This technique consists in assigning to each unit in the sampling frame a random number drawn independently from the uniform distribution on the interval [0,1]. This random number is associated to the frame unit on a permanent basis. To draw a simple random sample without replacement of size n, firstly, the frame units are sorted in ascending (descending) order by the random number. Secondly, a starting point or origin in the interval [0,1] and a sample direction (right o left) is selected. Then, the first n units in the ordered list give rise to a simple random sample without replacement.
If the starting points and the sample directions are chosen properly, the samples for the different surveys will be positively coordinated (maximum overlap) or negatively coordinated (minimum overlap).
When the surveys have different stratifications, the coordination is not perfect but the election of an origin and a sample direction properly help to obtain samples, which tend to be positively or negatively coordinated, as we desired.
We must take care with those surveys whose sampling strata are formed by the merge of sampling and take all strata of other surveys. In those cases, to avoid biased samples, it is necessary to divide the sampling stratum of the survey in take all and sampling stratum of the other survey allocating the sample size in each part before obtaining the coordinated sample.

II. 1. A FUNCTION OF STATISTICAL BURDEN

In order to coordinate negatively the sampling of different surveys in a year, a statistical burden function is created. This function is formed by two components: the time needed to fill the questionnaires and the number of surveys in which the business participate during that year.
Firstly, we generate a random number from an uniform distribution U[0,1] to all the frame units and in all of them we assign the value (0,0) to the statistical burden function. The first sample is drawn independently using a simple random sampling in each stratum. The statistical burden function value of the selected units in the sample is now equal to time needed to fill the survey questionnaires and 1.

First survey: estimated time to fill the questionnaire= 60’

f: Sampling frame $\rightarrow R^2$

$u_i \rightarrow f(u_i)=(x, y)=$(time, survey number)$=(60',1)$ if $u_i \in s$

The second and next samples are drawn coordinated with the previous ones using the same random numbers and the statistical burden function. Before obtaining the next sample, in each stratum, the frame units are sorted by ascending order according to first component of the statistical burden function (time), the second component (survey

number) and the random number. For units selected in this sample the statistical burden function is updated accumulating the time and the survey number.

Second survey: estimated time to fill the questionnaire= 120'

f: Sampling frame $\rightarrow R^2$

$u_i \rightarrow f(u_i)=(x, y)=$(accumulated time, survey number)$=(180',2)$ if $u_i \in s_2 \cap s_1$

II. 2. COORDINATION OVER TIME

Ours goal and the reason of applying negative sampling coordination over time is to avoid that a business is selected more than *x* consecutive years in the same survey. The value of *x* varies depending on each survey. For example, in the Annual Trade and Services Survey *x* is equal to 2 and in the Annual Industrial Companies Survey *x* is equal to 5.

In this case, the statistical burden function has a third component that takes value 1 if the business is in the same sample more than consecutive x years and 0 otherwise. Before obtaining the sample, the frame units are sorted by ascending order according to the third component of the statistical burden function, the first component, the second component and the random number.

The negative sampling coordination over time avoids the 'bad random chance' in the case of small businesses, which belong to large strata with small sampling fractions. But in the case of medium and large businesses, which belong to small strata with large sampling fractions, we may be disturbing the selection probabilities of all units of stratum. The selection bias would become important if the number of units with more than *x* consecutive years excluded from the sample were large and had a different behavior to the rest of population units.

In Table 1 we show the first quartile Q_1, the median M and the third quartile Q_3 of the distribution of the percentage of units with more than x=2 consecutive years with regard to the population total into each stratum according with the type of activity and the stratum size measured by employee number.

Table1: Percentage of units with more than x=2 consecutive years

Activity: Retail Trade

Quartiles	From 0 to 2	From 3 to 9	From 10 to 19	From 20 to 49	More than 49
Q3	3,2	20,2	66,7	100	100
Medium	0,7	6,7	33,3	50	100
Q1	0,2	1,8	11,1	25	66,7

Activity: Wholesale Trade

Quartiles	From 0 to 2	From 3 to 9	From 10 to 19	From 20 to 49	More than 49
Q3	10,5	20	50	100	100
Medium	2,8	7,7	22,2	50	100
Q1	0,7	3	11,1	30,8	77,8

Activity: Services Sector

Quartiles	From 0 to 2	From 3 to 9	From 10 to 19	From 20 to 49	More than 49
Q 3	4,3	13,3	35,4	80	100
Medium	1,6	4,8	16,7	50	92,3
Q1	0,3	1,6	7,8	25	70,1

In strata with small businesses the number of units with more than *x* consecutive years in the same sample is usually low. In those strata we avoid the bad random chance so the selection bias is negligible. That is not the case in strata with medium or large businesses with large sampling fractions. Here the percentage of units with more than *x* consecutive years in the same sample with regard to the population total into each stratum can be large.

III. EFFECTIVENESS OF SAMPLING COORDINATION

Following to McKenzie and Gross [7] we measure the effectiveness of the negative sampling coordination carried out in year 2008 with the following five surveys: Annual Trade and Services Survey, Annual Industrial Companies Survey, Technological Innovation in Companies Survey, The Use of ICTs and e-Commerce Survey Sector and Environment Surveys, calculating the expected load, the actual load and the avoidable load for frame units.

The expected load for a unit is defined as the sum of its selection probabilities in the surveys. The actual load represents the number of surveys in which a unit is selected. Finally, the avoidable load is defined as the number of selections in multiple surveys in excess of the optimal value that is determined by the expected load.

Let p_i be an integer and d_i a number between [0,1). Then, the expected load is defined as

$$\exp ected\ load(i) = \sum_{s=1}^{S} \pi_s = p_i + d_i$$

where S is the number of surveys, π_s represents to selection probability of unit i to the sample s. The optimal value of the number of times that the unit i is selected in some survey is p_i+1.

Let m_i be the actual load of i, number of times that unit i is selected in some survey. The avoidable load of the unit i is determined as

$$avoidable\ load(i) = \begin{cases} 0 & si\ m_i \le p_i + 1 \\ m_i - (p_i + 1) & si\ m_i > p_i + 1 \end{cases}$$

The sum over the population of avoidable load provides a measure of how effective a selection method is in controlling overlap among several surveys.
To measure this effectiveness we obtain the samples of the five surveys considered in two different ways: independently and negative coordinated using the statistical burden function and the PRN technique. In both cases, we calculate the actual and avoidable load.
Table 2 and Table 3 show the expected load against the actual load and the expected load against the avoidable load for the independent samples and for the negative coordinated samples. In Table 2 all frame units except units from take all strata are considered and the Table 3 only frame units between 10 to 49 employees are considered.

Table 2: Expected, actual and avoilable load for frame units not in take all strata

independent samples

expected	actual load					
load	0	1	2	3	4	total
[0,1]	2917983	144131	5311	279	4	3067708
(1, 2]	432	24049	5716	517	3	30717
(2, 3]		14	199	75		288
>3				10		10
total	2918415	168194	11226	881	7	3098723

negative coordinated samples

expected	actual load				
load	0	1	2	3	total
[0,1]	2914659	149678	3355	16	3067708
(1, 2]	213	24822	5674	8	30717
(2, 3]		3	235	50	288
>3				10	10
total	2914872	174503	9264	84	3098723

independent samples

expected	avoidable load					
load	0	1	2	3	total(1-3)	total
[0,1]	3062114	5311	279	4	5594	3067708
(1, 2]	30206	508	3		511	30717
(2, 3]	288				0	287
>3	10				0	10
total	3092618	5819	282	4	6105	3098723

negative coordinated samples

expected	avoidable load				
load	0	1	2	total(1-3)	total
[0,1]	3064337	3355	16	3371	3067708
(1, 2]	30709	8		8	30717
(2, 3]	288			0	288
>3	10			0	10
total	3095344	3363	16	3379	3098723

Table 3: Expected, actual and avoilable load for frame units between 10 to 49 employees

independent samples

expected	actual load					
load	0	1	2	3	4	total
[0,1]	114389	34171	4456	279	4	153299
(1, 2]	432	18299	5612	517	3	24863
(2, 3]		14	199	74		287
>3				10		10
total	114821	52484	10267	880	7	178459

negative coordinated samples

expected	actual load				
load	0	1	2	3	total
[0,1]	111199	39599	2485	16	153299
(1, 2]	213	19020	5622	8	24863
(2, 3]		3	234	50	287
>3				10	10
total	111412	58622	8341	84	178459

independent samples

expected	avoidable load					
load	0	1	2	3	total(1-3)	total
[0,1]	148560	4456	279	4	4739	153299
(1, 2]	24343	517	3		520	24863
(2, 3]	287				0	287
>3	10				0	10
total	173200	4973	282	4	5259	178459

negative coordinated samples

expected	avoidable load				
load	0	1	2	total(1-3)	total
[0,1]	150798	2485	16	2501	153299
(1, 2]	24855	8		8	24863
(2, 3]	287			0	287
>3	10			0	10
total	175950	2493	16	2509	178459

The frame units considered in Table 3 belong to medium stratum, businesses between 10 to 49 employees. In those strata the sampling fractions are large and so the sampling coordination can be useful. Units included in take all strata are not shown because in them the coordination is useless. The comparison between both tables shows us that in strata with small businesses where the sampling fractions are small there is not large difference between independent and coordinated samples

In medium strata, the negative sampling coordination is effective. The avoidable load is reduced more than 50% in the case of negative coordinated samples against the independent samples. The avoidable load total in the case of independent samples is 5,259 and in the case of coordinated samples is then 2,509. With the coordinated samples none of the respondents carry out more than 3 surveys and with the independent samples there are 7 ones that participate in 4 surveys. The number of businesses in 3 surveys decreases from 880 to 84, so it is reduced by 90 per cent and those that carry out 2 surveys decrease from 10,267 to 8,341, a 19% less with coordinates samples against independent samples. Finally, the businesses that carry out only one survey are increased by a 10% in the case of coordinated samples.

IV. CONCLUSIONS

In sampling strata with medium or large businesses, sampling coordination is quite effective achieving a more even distribution of the statistical burden. For the small businesses belonging to strata with large population and small sampling fractions, improvement with sampling coordination is negligible.

Using sampling coordination some control is achieved within surveys and among them. This control allows us to reduce the statistical burden on businesses participating in more than a survey. To find a balance between random sampling and the control on statistical burden, we only take into account statistical burden features as are the number of time that a business participates in the samples, in a year or over time, independently of its respond status.

V. REFERENCES

1) Azorín, F. y Sánchez-Crespo, J. (1986). Métodos y Aplicaciones del Muestreo. Alianza Editorial, Madrid.
2) Brewer, K.R.W. (1999). PRN Sampling: The Australian Experience. Proceedings Acts of Institute International of Statistics, 1999.
3) Cotton, F. and Hesse C. (1992). Coordinated Selection of Stratified Samples. Proceedings of Statistics Canada Symposium, 1992.
4) De Ree, J. (1999). Coordination of Business Samples using Measured Response Burden. Proceedings Actes of Institut Internationale de Statistique. Helsinki, 1999.
5) Ernst, L.R. (1999). The Maximization and Minimization of Sample Overlap Problems: A Half-Century of Results. Proceedings Actes of Institut Internationale de Statistique. Helsinki, 1999.
6) Hesse, C. (1998). Sampling Coordination: Theory and Practices. Project SUPCOM 96 lot 7. Eurostat.
7) McKenzie, R and Gross, B (2000). Synchronized Sampling. Proceedings of the Second International Conference on Establishment Surveys, American Statistical Association
8) Ohlsson, E. (1995). Coordination of Samples using Permanent Random Numbers. Chapter 9 in Business Survey Methods, edited by Cox, Binder, Chinnappa, Colledge and Kott, New York: Wiley, pp 153-169
9) Rainer, N (2004). Measuring Response Burden: The Response Burden Barometer of statistic Austria. European Conference on Quality and Methodology in Official Statistics. Mainz, Germany.
10) Rivière, P. and Laflamme, G (1997). OCEAN: Toward a General System for Samples Selection and Coordination. Work Session on Statistical Data Editing, Prague, Czech Republic
11) Riviére,P (2001). Coordination Sample using the Microstrata Methodology. Proceedings of Statistics Canada Symposium.
12) Sigman, R.S. and Monsour N.J (1995). Selecting Samples from List Frames of Businesses. Chapter 8 of Business Survey Methods, edited by Cox, Binder, Chinnappa, Colledge, Kott. John Wiley & Sons, Inc.
13) Srinath, K.P and Carpenter R.M. (1995). Sampling Method for Repeated Business Surveys. Chapter 10 of Business Survey Methods, edited by Cox, Binder, Chinnappa, Colledge, Kott. John Wiley & Sons, Inc.
14) Sunter, A.B. (1977). Response Burden, Sample Rotation, and Classification Renewal in Economics Surveys. International Statistical Review, 45.
15) Sunter, A.B. (1986). Implicit Longitudinal Sampling from Administrative Files: A Useful Technique. Journal of Official Statistics, Vol 2, Nº 2,1986. Statistics Sweden.