Changes in the selection of dwellings in the labor force survey of Argentina: A Simulation

Authors: Claudio Comari[1] , Augusto E. Hoszowski[2]

Abstract

This paper summarizes the main considerations and results of the studies conducted in the *Instituto Nacional de Estadística y Censos* of Argentina, to evaluate the impact that the change in the methodology of selection of households within each area (Secondary Sampling Unit) could have in the variances of certain estimators due to the passage from the method of selection of dwellings by *systematic sampling* to the method of selection of compact segments, in the household surveys. The results of a simulation exercise with microdata files from the 2010 Population Census are presented, comparing the different results obtained in relation to estimators for several population totals and their corresponding variances.

It has been proven that going from systematic sampling to cluster sampling, during the last stage of selection, does not conduct to a significant increase in sampling errors, having therefore cluster sampling evident benefits in the organization of fieldwork.

Keywords: Sampling, variance estimation.

[1] Master in Methodology in Social Research from Universidad Nacional de Tres de Febrero and Università Degli Studi di Bologna - National Directorate of Living Conditions in Instituto Nacional de Estadística y Censos of Argentina - Director of Master in Design and Analysis of Surveys in Universidad Nacional de La Matanza

[2] Lic. in Mathematics from Universidad de Buenos Aires - Coordinator of Sampling Methodology in Instituto Nacional de Estadística y Censos of Argentina - Coordinator of Master in Design and Analysis of Surveys in Universidad Nacional de La Matanza

Historical background

Sampling in the Household Surveys System in Argentina
General characteristics

In Argentina, household surveys carried out by the National Statistical System (*Sistema Estadístico Nacional)* are mainly based on the sample design of the Permanent Household Survey (EPH), the labor force survey that measures unemployment. This is not surprising, given the fact that it is the most important households' survey of Argentina due to the importance of the subjects involved, the time extension of the data collected, and the large territorial area covered.

- EPH started to be carried out in 1974, almost continuously providing the unemployment rate of the main urban centers of Argentina. At the beginning it was carried out twice a year (generally on March and October) and since 2003 it has been carried out on a continuous basis, with 4 annual estimations.

- EPH is carried out in 31 *urban agglomerations*, most of them being capitals of the provinces. They are virtually all of the urban agglomerations with more than 100,000 inhabitants. The Survey covers more than 62% of Argentina's population.

For some of the other surveys, such as the Household Expenditure Survey or the Living Conditions Survey, the National Institute of Statistics and Censuses (INDEC) extended the sample of the EPH so as to obtain estimations of the *urban total* of each province or of the population as a whole in areas with more than 5,000 inhabitants.

2. Sample design of the Permanent Household Survey in Argentina

EPH samples are based on a probabilistic sample of 'areas'. The design is similar to the one used for most of the labour force surveys:

Estimation Domains
Every 'urban agglomeration' with more than 100,000 inhabitants ('*agglomerations*'). *Independent* samples in each of these 'agglomerations'.

Sample design in each agglomeration
Two-stage and stratified design.

Primary sampling units (PSU)
Areas that in general constitute enumeration districts (geographical delimitations designed for the Population Census). On average, an enumeration district covers an area with 350 dwellings; it is a geographical delimitation designed for the census proceedings. In some cases the areas are divisions of enumeration districts or groups of enumeration districts, to guarantee a minimum number of dwellings inside the areas.

Number of areas selected in each agglomeration
The number varies between 60 and 120 areas, except for Greater Buenos Aires (~12,692,000 inhabitants) where approximately 520 PSU are selected.
PSU selection method
Stratified sampling, *pps* inside each stratum.

Secondary sampling units (SSU)
Private dwellings.

Number of dwellings selected in each SSU
It varies between 8 and 12 dwellings.

Dwelling selection method inside each selected PSU
Selection based on systematic sampling. The list of dwellings is geographically sorted before the selection takes place.

3. Advantages and disadvantages of the systematic sampling in the selection of households inside each PSU

In order to guarantee that the sample of dwellings inside each PSU remains *spread* around the whole area, the list of households is organized geographically, as it is usually done. The result is a decrease of the variance of the second stage of most of the estimators, assuming that the dwellings which are geographically close to each other present a positive association or correlation with most of the characteristics measured in household surveys.

The other method used to select households is the following: considering that surveys in Argentina are carried out in person and that the same interviewer generally surveys all the dwellings that are situated inside the same area, the interviewer must move from one selected dwelling to another one that may be far away from the previous one. This procedure has a cost in terms of time and may affect the response rate.

Although there are approximately 350 households in each area, the areas vary: from *one hectare* in the City of Buenos Aires, to *hundreds of hectares* in some less populated areas.

Synthetically, a dispersed sample provides:

- Lower variance of estimators
- Higher cost of the data-gathering process, in terms of working-time for the survey-taker.
- Reduction in the time available for the recovery of ‘non-responses’.

4. Compact segment sampling
Segments and compact segments

Although it is a widespread technique, it has been rarely used in official Argentine statistics.

This alternative results obvious if the ‘areas’ are very extensive (this is the case of the PSU in the CPS of the United States for example), but it can also be applied to any PSU. A disadvantage would be a potential increase of the variance of estimators, depending on what we want to estimate and on the corresponding ‘intra-class correlation’. The countries that use this strategy usually increase the final number of dwellings that are going to be selected in order to counteract the increase of the variance (which can be done up to certain limit) without having an important increase of the costs thanks to the proximity between dwellings.

The literature on this selection strategy results too profuse and it may receive different names: compact segments[3], compact groups of households or conglomerates[4], X size compact conglomerates[5], ultimate sampling unit clusters[6], etc.

[3] Anthony G. Turner (2003).
Paul Milligan, Alpha Njie and Steve Bennett (2004).

The two most popular versions of the selection of groups of adjacent dwellings are the following:

A) A sample of k adjacent addresses according to the available list is selected. All the dwellings of the selected addresses are surveyed. If a new dwelling appears between two selected addresses, it won't be surveyed.

B) A 'compact segment' sample is selected. The beginning and end of the segments (their ends) are marked and all the dwellings situated between these two addresses are surveyed.

The advantage of the A) alternative is that it can follow a fixed size group of selected dwellings, even though they may stop being 'adjacent' eventually.

The B) alternative results appropriate in the cases where the list has *omissions*: they are automatically detected, as all the dwellings situated between both selected ends must be surveyed. It results an interesting alternative for specific surveys that are carried out based on lists that have not been updated. In any case, this method detects the changes that the group of dwellings in the area may have suffered. In the case of omissions, estimators are equally unbiased.

The disadvantage is that the sample of dwellings does not have a fixed size and, occasionally, much more work may be assigned to the interviewer, what would require subsequent solutions from the regional office.

Synthetically, to survey all of the dwellings situated between two selected limits provides:

Advantages

- It provides possible solutions to any imperfection in the list (omissions)
- It requires less time for the data collection (proximity between the dwellings to be surveyed)

Disadvantages

- Possible excessive growth of segments that require solutions from the central offices
- Variability in the sample size
- Variance increase

It follows some of the final sizes of the dwellings segments (or conglomerates) in some of the household surveys of the official statistical offices.

Country	**Survey**	**Final sizes in segments of households**
EEUU	Current Population Survey	4
	American Community Survey	1
Argentina	Encuesta Permanente de Hogares	1

[4] Fernando Riveros Siguiura, Metodología del diseño de muestra de la encuesta continua de hogares y condiciones de vida de Bolivia, available at <http://www.eclac.cl/deype/mecovi/docs/TALLER10/14.pdf>

[5] INE Uruguay (2005).

[6] Bureau of Labor Statistics (2003).

	Encuesta de Gastos de los Hogares 2012	5
España	EPA (Encuesta de Población Activa)	1
Francia	Enquête Emploi (1968 - 2001)	40 in rural area 20 in urban area
	Enquête Emploi en Continu (2001 -)	~ 20
Guatemala	Encuesta de Empleo e Ingresos	6
Bolivia	Encuesta Trimestral de Empleo (ETE)	1
India	Employment and Unemployment Survey	1
China	Urban Labour Force Sampling Survey	50-100
Uruguay	Encuesta Continua de Hogares	3-4

5. Workload and non-response

The non-response is one of the most serious problems faced by household surveys. It is well known that there is not an imputation method that could compensate for the lack of response. This is why the final quality of a survey strongly depends on the measures taken to guarantee the maximum response rate and the quality of the information collected.

The non-response rate in a survey depends on a number of factors, among which we can mention the following.

- General willingness among the population to cooperate with the survey.
- Advertising or public awareness campaigns about the importance of responding official surveys.
- Specific characteristics of the areas to be surveyed (area of apartments or area of houses, etc.).
- Training of the group of interviewers.
- Interviewers experience.
- Workload demanded to households.
- Workload demanded to interviewers.
- Etc.

The non-response may be a source of bias (as well as generating an increase of the variance of the estimators due to the reduction of the sample size) and may be influenced by the workload demanded to interviewers (that comprises both the extension and the complexity of the questionnaire as well as the difficulty to locate the dwellings to be surveyed and achieve the interview).

In order to face these difficulties and improve the final quality of the products, it was decided, among other adopted measures, to investigate the effects of making changes in the sample design using the selection of compact segments of dwellings that contribute to the decrease of the non-response, contracting the workload of interviewers and seeking the usufruct of the “contagion effect[7]”.

Due to the fact that INDEC does not register precedents in sampling plans with selection of compact segments of dwellings in household surveys, a simulation was used to evaluate the increase of the variance when applying this methodology instead of selecting dwellings using systematic sampling.

The main objective of the simulation was then to evaluate the changes in the variance of certain estimators when modifying the sampling plan inside each PSU, contrasting

[7] Several experiences show that households are more likely to answer when they perceive that other neighboring dwellings are being surveyed too.

the results obtained through systematic selection of dwellings, simple random selection of dwellings, and selection of dwellings by segments.

6. Simulation: description of the experiment carried out

The simulation was carried out based on micro-data provided by the National Population, Household, and Housing Census 2010[8] of the District of La Matanza in the Province of Buenos Aires.

The District of La Matanza has a population of more than 1,700,000 inhabitants. The database of households includes more than 450,000 records and the database of enumeration districts (PSU) includes 1,341 records.

The variance of the estimators of various totals was studied, on variables that represent different degrees of 'intra-class correlation'.

The totals to estimate are:

- Total of unemployed people
- Total of active population
- Unemployment rate
- Total of Heads of Households with complete university education
- Total of households in dwellings with drain from the bathroom to a sewer

From these 5 totals to estimate, the fifth one is the one that, in principle, comes from a variable with high 'intra-class correlation':

$Y_i = 1$, if the dwelling ***i*** has drain from the bathroom to a sewer

$Y_i = 0$, if the dwelling ***i*** does not have drain from the bathroom to a sewer

The variable is strongly associated to a geographical area.

Four sampling designs have been compared. In all of them, PSUs are enumeration districts.

The primary sampling frame was thus the group of enumeration districts of the District of La Matanza in the Province of Buenos Aires, which gave rise to a primary frame with 1,341 PSU.

16,000 independent samples were selected from this sampling frame:

Method 1

4,000 samples according to the following design:

- First stage: Selection of 200 PSU, selected according to *pps* (Madow, proportional to the number of dwellings)
- Second stage: Selection of 12 dwellings (approximately) in each PSU selected by means of *Systematic Sampling.*

Method 2

4,000 samples according to the following design:

- First stage: Selection of 200 PSU, selected according to *pps* (Madow, proportional to the number of dwellings)
- Second stage: Selection of 12 dwellings in each PSU selected by means of *Simple Random Sampling.*

[8] The experiment was also carried out in the provinces of Salta, Santiago del Estero, and Mendoza.

Method 3
4,000 samples according to the following design:

- First stage: Selection of 200 PSU, selected according to *pps* (Madow, proportional to the number of dwellings)
- Second stage: Selection of 2 *compact segments* of 6 dwellings, by means of *Simple Random Sampling.*

Method 4
4,000 samples according to the following design.

- First stage: Selection of 200 PSU, selected according to *pps* (Madow, proportional to the number of dwellings)
- Second stage: Selection of 1 *compact segment* of 12 dwellings, by means of *Simple Random Sampling.*

The segments were built sorting the dwellings that belong to each enumeration district according to the geographical order appearing in the micro-data bases to ensure that the segments consist of geographically adjacent dwellings.
The exercise was programmed with the statistical software Stata 9.2.

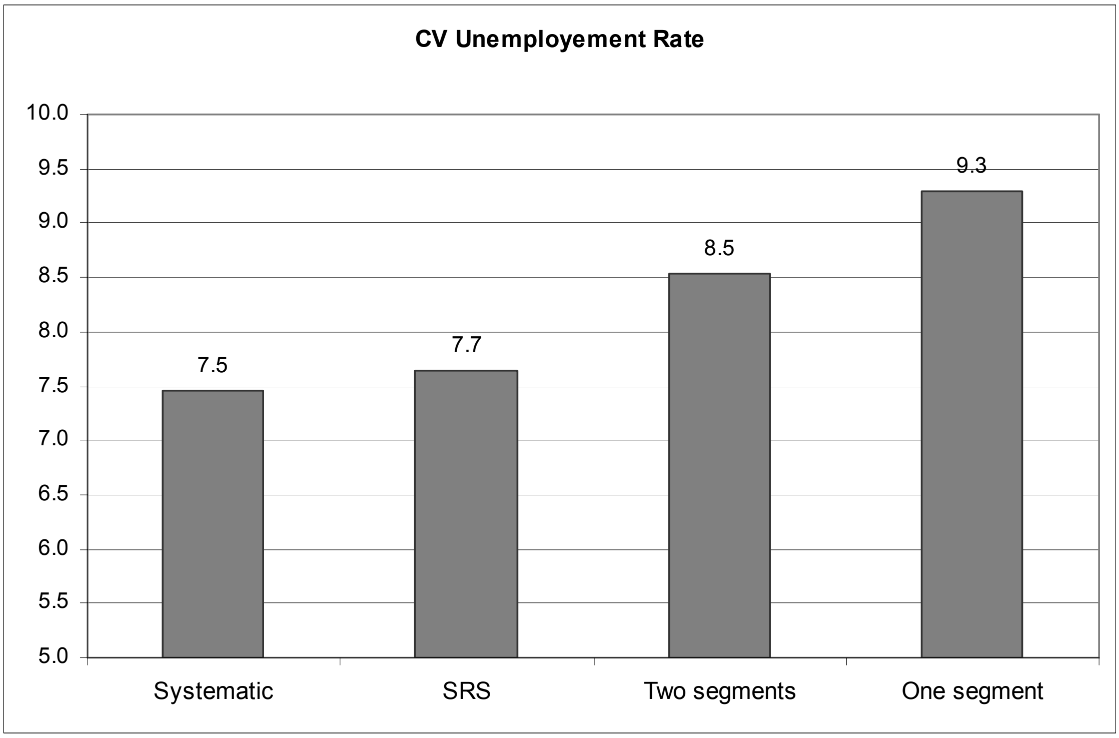
Results of the simulation in the District of La Matanza, Province of Buenos Aires

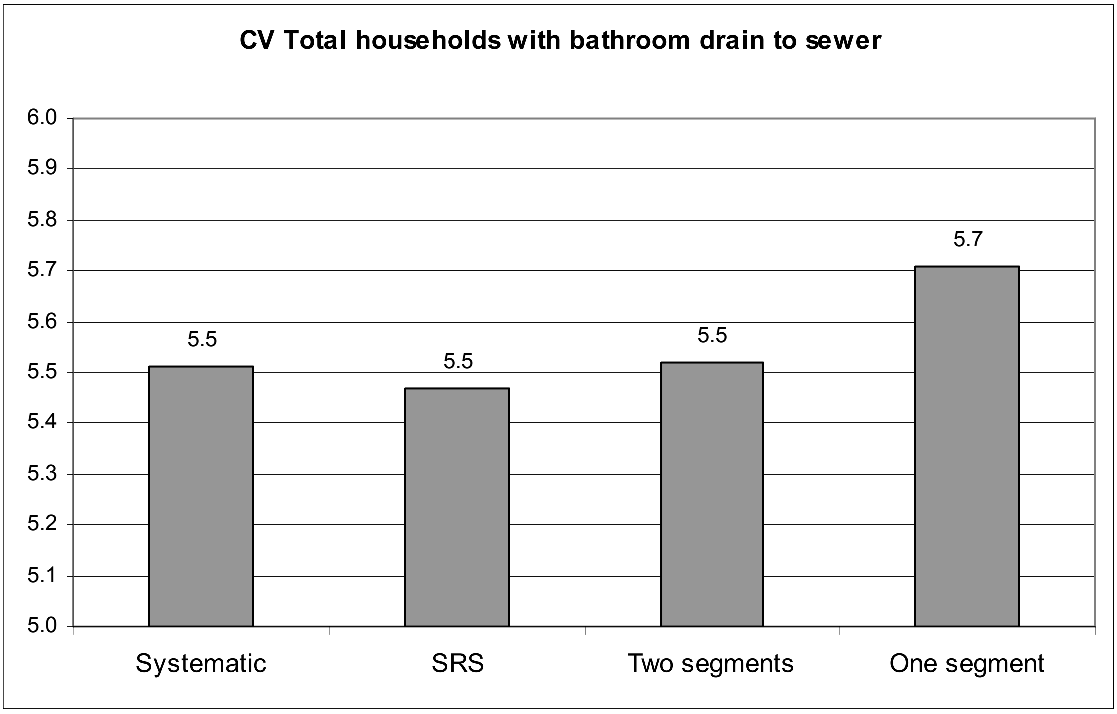
Method of selection within PSU	CV Total unemployed (%)	CV total in EAP (%)	DS unemployment rate (%)	CV unemployment rate (%)
Systematic (n=12)	7.49	2.01	0.43	7.45
SRS (n = 12)	7.70	2.00	0.44	7.65
2 segments with 6 dwellings	8.64	2.18	0.49	8.54
1 segment with 12 dwellings	9.44	2.33	0.54	9.29

Method of selection within PSU	CV total of head household without instruction (%)	CV total of head household with university degree (%)	CV Total households with bathroom drain to sewer (%)
Systematic (n=12)	4.88	11.57	5.51
SRS (n = 12)	4.95	11.30	5.47
2 segments with 6 dwellings	5.29	11.95	5.52
1 segment with 12 dwellings	5.66	12.23	5.71

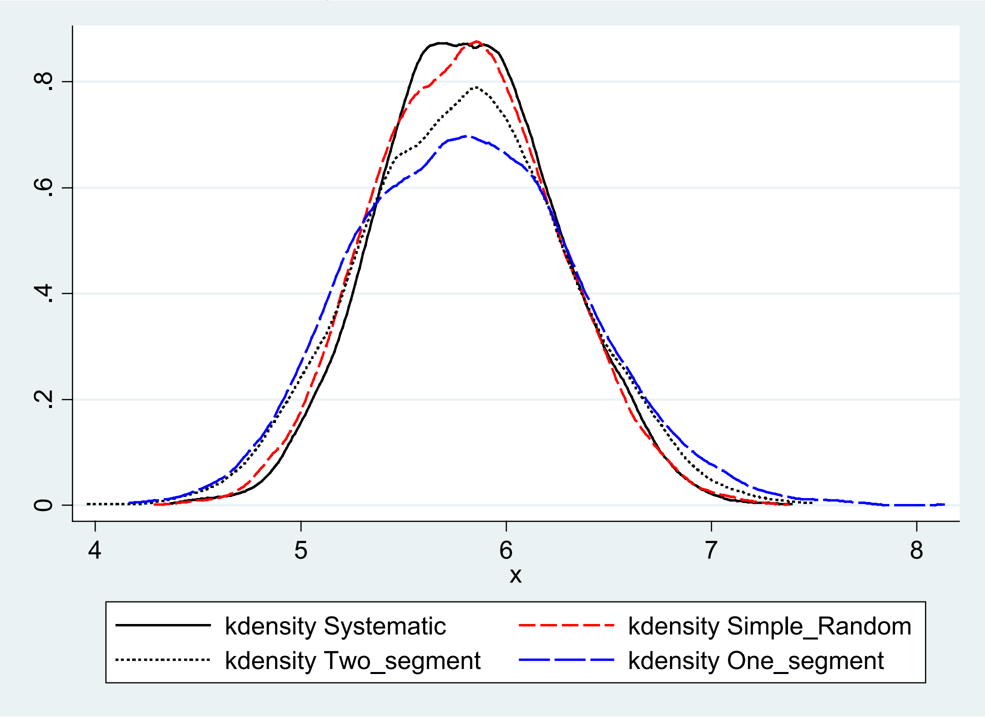
Method of selection within PSU	DS unemployment rate * (%)	CV unemployment rate (%)
Systematic (n=12)	0.43	7.45
SRS (n = 12)	0.44	7.65
2 segments with 6 dwellings	0.49	8.54
1 segment with 12 dwellings	0.54	9.29

(*) Unemployment rate in Census: 5.8%


CV Unemployement Rate
10.0
9.5
9.0
8.5
8.0
7.5
7.0
6.5
6.0
5.5
5.0
7.5
7.7
8.5
9.3
Systematic
SRS
Two segments
One segment

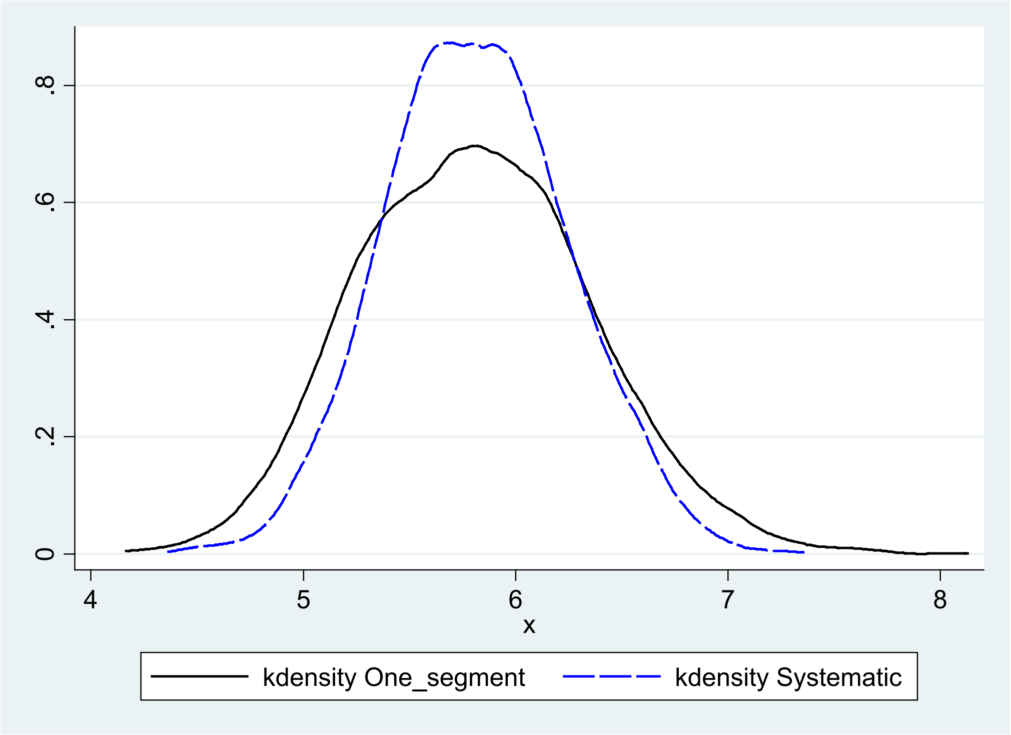

CV Total households with bathroom drain to sewer
6.0
5.9
5.8
5.7
5.6
5.5
5.4
5.3
5.2
5.1
5.0
5.5
5.5
5.5
5.7
Systematic
SRS
Two segments
One segment

Graphic of the density of the estimators of the census unemployment rate according to the four selection methods used.



The observation of the densities allows us to confirm that the systematic selection has the highest "concentration" or "positioning", while the density that corresponds to the selection of only one segment of 12 dwellings shows the curve with less 'positioning'.

The four densities are symmetric and 'centered' in the population parameter.
In order to better appreciate the difference between the distribution arising from the systematic sampling of 12 dwellings and a cluster sampling of 12 consecutive dwellings, the following chart shows only these two densities:



This chart shows the lower variance and the higher concentration of the estimator obtained by means of a plan of systematic sampling (inside the areas) and the normal behavior that both estimators show, which so far confirms the theory.

But when deciding whether to implement one strategy or the other, and from the point of view of the 'total quality' that must prevail in the design and administration of a household survey, we must focus not only on the variance, the concentration, or the mean squared error of an estimator, but also on other equally important factors, such as *non-sampling errors*. As mentioned previously, cluster sampling has major advantages for fieldwork, mainly for the reduction of the non-response rate.

The previous table shows that the coefficient of variation (CV) of the estimator of the unemployment rate changes from 7.45% to 8.54%, going from a systematic sampling to a plan where two segments of 6 dwellings are selected.

But in these simulations there is not an important factor, such as the *non-response*. Not only does this phenomenon introduce biases into the estimators (less probability of selection of single households, certain population categories, etc.), but it also causes the increase of the variance when reducing the sample size.

We assume that the gain in variance caused by the decrease of the non-response, plus the consequent minimization of the biases, widely compensates the low increase in the variance of the estimators if we move to a plan of selection of two conglomerates of 6 dwellings, instead of a systematic sampling of 12.

7. Conclusions

The simulation carried out in the District of La Matanza shows that the way of selecting dwellings in the last sampling stage in the exercise does not have a *significant* effect in the variance of the estimators, even in variables with high intra-class correlation. A correct interpretation is the assumption that most of the variance comes from the first stage of the selection. Another correct interpretation is the assumption that enumeration districts are, on average, homogeneous enough to go from a selection of dwellings by means of systematic sampling to a selection by compact segment without having an important influence on the variances. Therefore, the diagram that guarantees a better fieldwork and a reduced non-response must be given priority.

Bibliography

BLS. *"Handbook of Methods Chapter 1. Labor Force Data Derived from the Current Population Survey"*. Available at: <http://www.bls.gov/opub/hom/>

Bregger, John E. "*The Current Population Survey: a historical perspective and BLS's role"* Monthly Labor Review, June 1984, pp. 8-14.

Cochran, G., "*Sampling Techniques*". Wiley, 1977.

INE Uruguay. "*Diseño de la Muestra para una Encuesta de Hogares Ampliada*". 2005. Available at: <*http://www.ine.gub.uy/biblioteca/metodologias/ENHA_2006_DISE%D1O%20DE%20LA%20MUESTRA.pdf*>

United Nations: "*Household Sample Surveys in Developing and Transition Countries*", Studies in Methods, Series F No. 96, Department of Economic and Social Affairs, Statistics Division, 2005

Paul Milligan, Alpha Njie and Steve Bennett. "*Comparison of two cluster sampling methods for health surveys in developing countries*". International Journal of Epidemiology Vol.33 no.3, International Epidemiological Association, 2004
Riveros Siguiura, Fernando. "*Metodología del diseño de muestra de la encuesta continua de hogares y condiciones de vida de Bolivia*". Available at: <*http://www.eclac.cl/deype/mecovi/docs/TALLER10/14.pdf*>

Turner, Anthony G. "*Sampling frames and master samples* ". Expert Group Meeting to Review the Draft Handbook on Designing of Household Sample Surveys, 2003.

Verma, Vijay. "*Sampling for Household-Based Surveys of Child Labour*", ILO, International Programme on the Elimination of Child Labour (IPEC), 2008.