

Sampling solutions to the problem of undercoverage in CATI household surveys due to the use of fixed telephone list

Claudia De Vitiis, Paolo Righi¹

Abstract: The undercoverage of the fixed line telephone number list, used as sampling frame for survey on households, is becoming a pressing problem. Some methodological studies conducted by ISTAT have pointed out the risk of biased estimates for particular phenomena of interest due to this type of frame imperfections. To deal with the bias problem a research group in ISTAT has been set up. The research group has highlighted for the Holidays and Trips Survey, for which data are collected by means of CATI technique on a sample selected from the phone directories, a risk of estimation bias due to the undercoverage effect. In the paper two possible approaches to achieve unbiased estimates are described. The first one is based on the enhancement of the estimator aiming to take into account the frame imperfections. An experiment is shown, consisting in choosing the calibration variables more carefully and producing a good improvement on the accuracy of the estimates. The second approach defines a new sampling design based on a list not suffering of undercoverage, the municipal population register, a mixed data collection mode and a two stage sampling design.

Keywords: Frame Undercoverage, CATI Surveys, Sample Design and Estimation, Official Statistics

1. Introduction

In recent years the fixed line telephony has lost increasing quotas of coverage with respect to the Italian population, caused on one hand by the spreading of exclusive mobile phone and on the other hand by the recourse to reserved fixed lines numbers. Moreover, a part of the population persists in not having a telephone at all. The overall percentage of households not covered by a fixed telephone line was in 2008 about 40% (ISTAT, 2009a). These circumstances have produced relevant implications for the household sampling surveys carried out by means of the CATI (Computer Assisted Telephone Interview) technique, which are commonly based on a sample selected from the list of fixed line telephony. The estimates produced by these surveys could be affected by a certain degree of bias if referred to the whole population, because of the undercoverage of the list, if the population units belonging to the list are different from the units excluded by the list with respect to the target variables. The degree of bias depends, in general, on the combination (Särdal and Lundström 2005) of the undercoverage rate and the difference between the two subpopulations, to be evaluated.

¹Italian National Statistical Institute (ISTAT) (e-mail: devitiis@istat.it, parighi@istat.it)

To deal with undercoverage bias a research group has been established at the Italian National Statistical Institute (ISTAT). In order to verify and to assess this suspected bias, several analysis have been carried out on the data of the Multi Purpose Everyday Life Aspect Survey (in the following indicated as MPS), a survey based on a sample of households selected from the municipal registers. For this reason, MPS can be considered not affected by frame undercoverage and, therefore, producing unbiased estimates. This survey collects some variables which are collected also by the main CATI household surveys and, moreover, collects the presence of a fixed phone line in the sample households; this fact allows to individuate in the overall sample the subsample of the households covered by fixed telephone, which can be considered similar to the sample contactable through the CATI surveys selected from the telephone list.

The results of the research group highlighted significant differences between the estimates of some interest variables obtained on the whole MPS sample and the same estimates obtained on the subsample of the phone-owner households, consequently reachable through a survey using the telephone list. In particular, the analysis carried out by the group has confirmed the suspected bias for one of the main target variable of the CATI Holidays and Trips Survey (HTS), while for another important survey, the Citizen Victimization Survey, the findings of the group did not highlight a risk of bias.

In this paper the attention is concentrated on the analysis of the possible solutions to the problem of the bias caused by frame undercoverage. Two types of solution are discussed: the first one improves the estimation procedure without modifying the frame of the survey; the second one consists in a complete redesign of the sampling design, based on the use of a different population list and a mixed data collection technique. It is relevant to stress the fact that the CATI technique still remains the preferable way of conducting interviews for many important ISTAT surveys, for the timeliness and the quality of data and for cost reasons as well.

The paper is articulated in the following sections: section 2 describes the driving steps for defining a new estimator taking more carefully into account the undercoverage bias of one of the main target variable of the HTS. The study is supported by an empirical analysis of the property of the new estimator in terms of bias. The experiment is based on MPS data, in which such variable is observed in the same terms as in HTS. Section 3 is devoted to a definition of a two stage sampling design for the HTS, using an exhaustive selection list, guaranteeing to preserve the estimates from undercoverage bias. Section 3.1 gives an empirical evaluation of the design effect of the estimates based on MPS data, as the current HTS is based on a one stage sampling design and the proposed design can possibly produce a negative impact on the sampling errors. In Section 4 some conclusions and a future research path are given.

2. Reweighting approach

The first approach to deal with the bias deriving from the undercoverage of the telephone list leaves unchanged the sampling design and the selection frame and aims to improve the estimation procedure. This is done by introducing in the estimator, which is generally a calibration estimator, additional or different auxiliary variables, more correlated with the survey variables or with the undercoverage phenomenon. The reweighting approach tries to reduce the estimation bias due to the undercoverage of the

sampling list by means of a suitable estimator. We suppose the existence of a relationship among the target variables and some auxiliary variables observed on the sample and for which are known the totals at population level. If not, we assume that unbiased estimates with small sampling variance of these totals are available. Then, we use these variables in the calibration procedure, modifying the current calibration estimator. Adding these benchmarking variables, the final estimates of the totals of the new auxiliary variables will be equal to the known totals. The use of the calibration approach is one of the possible ways to deal with undercoverage at the estimation phase (Särndal and Lundström 2005, ch. 14). The approach is based on the following assumption: if the new estimator corrects the biased sampling distribution of the explaining (auxiliary) variables of the target variable, then the estimator corrects, at least partially, the bias of the estimate of the target variable as well.

We made an unpretentious experiment based on the ISTAT MPS. The sample of households is drawn from an exhaustive demographic list (the municipal registers) according to a two stage sampling design. Data are collected by means of PAPI (paper and pencil personal interview) technique. MPS uses currently a calibration estimator benchmarking the number of persons by sex and age groups at regional level. Some of the variables observed in the MPS are also collected in the HTS, such as: structural variables (sex, age, educational level, professional position, etc.), the “yes/no variable about at least one Overnight Stay lasting Four nights or more (OSF) in a collective accommodation establishment for holiday, during the last twelve months”. As in the MPS, the variable “yes/no Phone Owner, except mobile phone” is recorded, it is possible to distinguish the subsample of MPS reachable by a survey using CATI techniques like the HTS. On the subsample of MPS composed by the phone owner households, the estimation of the frequency of “yes” to the OSF variable has been carried out by means of the HTS estimator. This estimate simulates the estimation process performed currently in HTS, even though there are some substantial differences such as: different sample design, different sample size and different survey techniques. Nevertheless, assuming that MPS produces unbiased distribution estimates (hereinafter denoted by reference distributions), we may suppose that the differences among the reference distributions and the distributions estimated through the above procedure based on the HTS estimator² (table 1), depend essentially on the undercoverage problem of the phone owner population with respect to the overall household population.

Table 1: Estimates of relative (%) distributions of the number of person per at least one Overnight Stay lasting Four nights or more (OSF). Estimates obtained by the current MPS estimator on the overall sample of the MPS (estimate type a) and by the current HTS estimator on the subsample of the phone owner households (estimate type b)

	Estimates type a and confidence interval		Estimates type b and confidence interval	
No	48.7	(47.8-49.5)	44.4	(43.4-45.5)
Yes	50.3	(49.5-51.2)	54.6	(53.5-55.7)
No-response	1.0	(0.9-1.2)	1.0	(0.8-1.2)

The main result shown in table 1 is the not overlapping of the confidence intervals computed by means of the two estimation procedures. Starting from these findings, the experiment on the MPS referred to 2008 data has concerned the following steps:

² The HTS current estimator is calibrated with respect to the following known totals: population by sex and region; population by municipal type; population by age groups; households by size.

- a logistic regression model, fitted on MPS data, identified the significant explaining structural variables of the probability to have yes for the variable OSF (ISTAT 2009a);
- the estimation of the sampling distribution of the significant structural variables based on the subsample of phone owner households were computed using the current HTS estimator;
- the comparison among the estimates of these structural variables computed with the current MPS estimator based on the overall household sample and the estimates obtained with the HTS estimator on the phone owner households was performed;
- finally, some structural variables presenting different estimated distributions when obtained with MPS and HTS estimators are put in the system of constraints of the HTS calibration estimator if they are not already used in the current estimator.

Table 2 and 3 show the estimated distributions of Educational Level (EL) and Professional Position (PP) obtained by means of the MPS and the HTS estimator, being EL and PP resulted significant explaining variables for OSF variable. The estimated distributions obtained through the two estimators for each variable appear quite different.

Table 2: *Estimates of relative (%) distributions of Educational Level. Estimates obtained by the current MPS estimator on the overall sample of the MPS (estimate type a) and by the current HTS estimator on the sub sample of the phone owner households (estimates type b)*

Level	Estimate type a and confidence interval		Estimate type b and confidence interval	
Doctorate, degree	8.9	(8.5-9.2)	10.8	(10.1-11.5)
Upper secondary school certificate	26.0	(25.4-26.6)	30.1	(29.0-31.2)
Lower secondary school certificate	35.3	(34.7-35.9)	32.7	(31.6-33.8)
Primary school certificate, no education	29.9	(29.4-30.4)	26.4	(25.4-27.4)

Table 3: *Estimates of relative (%) distributions of Professional Position. Estimates obtained by the current MPS estimator on the overall sample of the MPS (estimates type a) and by the current HTS estimator on the sub sample of the phone owner households (estimates type b)*

Position	Estimates type a and confidence interval		Estimates type b and confidence interval	
Manager	1.0	(0.9-1.2)	0.7	(0.5-0.9)
Executive / Clerk	14.2	(13.8-14.6)	23.5	(22.5-24.5)
Workman / workwoman	12.6	(12.2-13.0)	15.4	(14.6-16.2)
Entrepreneur, Professional	3.3	(3.1-3.5)	2.9	(2.5-3.3)
Coordinated free-lance worker	1.3	(1.2-1.4)	-	-
Self-employed, Collaborator in the family Business	5.7	(5.4-6.0)	5.1	(4.6-5.6)
Not employed	61.8	(61.3-62.2)	52.3	(50.9-53.7)

Hence, the EL and PP variables have been introduced in the set of calibration variables of the HTS estimator. Table 4 shows that the estimated distribution of the OSF variable with EL variable in the calibration system, based on the subsample of the phone owner households (estimates type c), is more similar to the reference distribution with respect to the type b estimated distribution shown in table 1. Table 4 also shows the results of the calibration process on the subsample of phone owner households when EL and PP are in the set of calibration variables (estimate type d). We may observe a further reduction of the bias with confidence intervals almost overlapping the confidence intervals of the reference distribution.

This results show that it is possible to obtain significant improvements in reducing the bias produced by the undercoverage of the telephone list by introducing different

variables in the calibration procedure. What remains still to be verified is the possibility to utilize such an estimator in the current HTS, which is based on a relatively small sample, on which the convergence of the calibration procedure is not guaranteed.

Table 4: *Estimates of relative (%) distributions of the number of person per at least one Overnight Stay lasting Four nights or more (OSF), obtained on the subsample of the phone owner sample of the MPS calibrating on the auxiliary variables currently used in the HTS estimator plus EL variable (estimates type c) and plus EL and PP variables (estimates type d)*

	Estimates type c and confidence interval		Estimates type d and confidence interval	
No	46.0	(45.0-47.0)	46.4	(45.4-47.4)
Yes	53.0	(52.0-54.0)	52.6	(51.6-53.6)
Not response	1.0	(0.8-1.2)	1.0	(0.8-1.2)

3. Redesigning a CATI survey: two stage sampling selection from municipal registers and mixed data collection technique

The second approach here proposed is based on the use of different selection frames, which can substitute or be added to the current fixed telephone list, and requires planning a completely new sampling strategy. The strong hypothesis at the basis of this redesign is that the new lists guarantee a very high level of coverage of the target population. At present, the only complete list of the Italian population at ISTAT's disposal is the one currently used for the household surveys carried out through face to face interview, constituted by the set of municipal registers available at each Italian municipality. Resorting to such a different frame will imply the definition of a different sampling scheme and the use of different interview technique to be added to the usual CATI method for those units not reachable through a fixed phone. The CATI technique remains, nevertheless, the main interviewing mode.

This solution is based on the assumption that the municipal registers are not affected by undercoverage as it covers the whole population of households. As a unique list is not available and each municipal register is obtainable in each municipality, the sampling selection scheme will be similar to those ones commonly used for household survey with selection from municipal register: a two stage sample scheme, in which the municipalities are the PSUs (primary sampling units) and the households the SSUs (secondary sampling units). After the selection of the sample households, a delicate phase of linking of the sample units to the list of fixed telephone numbers is performed. The linkage will not be possible for all the sample households, both for errors of linkage and for real absence of a fixed line. Therefore, while for the sub-sample of units linked to a fixed line number the CATI method will be applied, for the remaining ones the interview will be carried out through a different technique: by CAPI (Computer assisted Personal Interview), which it is the most expensive one, or, more conveniently, the CAWI (Computer Assisted Web Interview). The CAPI technique will be used also for those households not reachable through the linked telephone numbers, in order to avoid the substitution of such units, which is a not recommended practice if the aim is to reduce the bias of the estimates.

Another relevant issue under discussion is the possibility to conduct CATI interviews through mobile telephone numbers. Although this interviewing mode has been already experimented giving good performances, at present a complete list of

mobile telephone numbers to be linked to a sample of individuals or households is not available (mostly for legal reasons). Therefore, this possibility cannot be considered in our context and mobile interviewing can be considered only for taking appointments after a first contact through a fixed line.

With regards to the sampling aspects, a relevant element to be taken into account is the fact that a two stage sampling scheme produces in general an inefficient effect due to the level of association of elementary units within the selected clusters, with respect to the survey variables: the more similar to each other the units belonging to the clusters, the higher the increase of inefficiency of the estimates. This effect can be limited, in general, by selecting a large number of PSUs and therefore a small number of final units in each PSU. In the case under study, the only reason to limit the number of sample municipalities is due to the necessity to get access to each municipal register. The use of CATI technique, eliminating the need for the interviewer to go personally to the houses of the sample households on one hand allows to widen the sample of PSUs obtaining an improvement of the efficiency of the estimate, on the other hand enables to reduce the cost of the survey, being the CATI much cheaper than a face to face interview. In order to obtain an adjustment of the cost of the redesigned survey, it is possible to operate on three alternatives: selecting a very large number of municipalities keeping the number of sample final units similar to the current one, increasing the number of final units in the sample keeping the number of municipalities small, or balancing both numbers at the same time. In any case, to fix the parameters of the two stage selection scheme (the size of first and second stage sample and the number of sample households for each sample municipality) it is necessary to carry out evaluations on the efficiency of the estimates to be obtained, which varies from survey to survey depending on the intracluster correlation coefficient of the specific variables.

To summarize, it is useful to point out that this sampling approach presents very good advantages together with some drawbacks. The advantages are: it mostly resolves the problem of undercoverage; it is based on the use of well consolidated sampling, estimation and data collection techniques; it allows to limit the increase of the costs of the survey as a high quota of CATI interviews is preserved. On the other hand, the disadvantages are: the problems of linking between the list of sample households and the telephone list; the need to handle with different no-response and measure errors, due to the use of a mixed data collection technique; the difficulty to know in advance the quota of interviews to be conducted by CATI and CAPI, while the knowledge of this subdivision of the sample is fundamental to fix the parameters to plan the two stage sample design (number of PSUs, SSUs per PSU and total sample size).

3.1. First evaluations on a redesign of the Holidays and Trips Survey

The survey on Holidays and Trips has resulted to be exposed to the risk of bias deriving from the undercoverage of the selection list, as shown in (ISTAT 2009a) and in section 2. To follow the redesign approach, the use of an exhaustive population archive is required. As stated in previous section, the most natural population archive is the municipal registers, giving rise to a two stage sampling scheme and a multi-technique data collection, CATI for the part of the sample to which it is possible to link a fixed phone number, CAPI or CAWI for the other part. It is important to underline that for

this survey the use of a computer assisted interview is crucial, essentially for timeliness reasons, being the survey obliged to send quarterly data to EUROSTAT; in this context the CATI and the CAWI technique would be better because they would allow also to limit the survey costs.

By adopting such a sample strategy it should be possible to reduce consistently the part of the total error of the estimates due to the undercoverage of the list, maintaining at the same time the sampling error at the same level as it is currently, by increasing if necessary the whole sample size.

In this context some analysis have been carried out to obtain some evaluation about the required sample size to realize a two stage sample scheme, with the aim not to worsen the precision of the estimates of the survey. The analysis has been made in term of design effect of the estimate of the parameter related to the yes/no OFS variable, described in the previous section 2, the only variable collected both in MPS and in HTS. The design effect is a measure of the impact on the sampling variance of an estimate, deriving from the use of a complex sample design, in comparison with a simple random sample with the same sample size. For a two stage design this quantity is generally greater than unity, owing to the *intracluster correlation coefficient* ρ , which expresses the similarity among units belonging to the same cluster with respect to a given survey variable.

The expected necessity to increase the number of sample units derives from the fact that at present the survey is based on a single stage stratified sample selection, directly from the telephone register, being such a design a very efficient one. For example, in general, a design effect equal to 2 would require an increase of the sample size equal to the square root of 2, passing from a simple random sample to a two stage design.

From the empirical analysis on the MPS data, which derive from a two stage sample design municipality-household, the first finding is that the ρ shows very high values within the "household" clusters, while the effect of grouping of the sample in sample municipalities is weak. Besides, surprisingly, for the analogous OSF variable collected in the HTS (based on a single stage design) it can be observed a design effect higher than in MPS: 1.83 for the latter and 1.90 for the HTS, both at national level for the last available survey year 2008. Going into details of MPS data separately for the self representative (SR) and non-self representative (NSR) part of the sampled municipalities, the data show that the design effect is around 1.6 for the SR part and around 2 for the NSR part. It is useful to underline the meaning of the comparison between the design effect evaluated on the SR part of the MPS sample and the HTS sample. In fact, the two sample scheme are similar: HTS sample is selected from the telephone list by means of a stratified sample scheme and the SR part of the MPS sample is the part of the sample selected through a single stage design, in which municipalities are one stratum each and selected in the sample with certainty. The result of the evaluation of the design effect states that, with a similar selection scheme, MPS design effect is much lower than the HTS one. This result allows to suppose that the impact of undercoverage of the selection list produces a bias (an overestimation) of the design effect. In other words, it seems that in the part of the population covered by fixed phone the intracluster correlation coefficient is higher than in the whole population, observed through the MPS. This evidence is confirmed by the analysis of MPS data limited to the subsample covered by fixed phone: the design effect estimated on this subset of households is higher than the design effect estimated on the whole sample. This result emerges both in the SR and in the NSR part of the sample. Although this analysis is valid with respect to the considered variable OSF, the result is very appealing and would deserve a more in depth investigation. The consequence of

this result on the definition of the sample size for a two stage sample design is relevant. In fact, it seems that to reduce the bias produced by the use of the telephone list affected by undercoverage, it is not required an increase in sample size to keep unchanged the sampling error of the estimates. The reason of this fact is that the current HTS estimates are already affected by a high design effect due to the strong similarity of the individuals belonging to same household and it is reasonable to expect that, with a two stage design with selection from municipal registers, this design effect will not rise. Therefore, by utilizing a population frame not affected by undercoverage, it would be possible to obtain a remarkable reduction of the total error of the estimates, deriving from the reduction of the bias, together with the maintaining of the level of the current sampling error.

It would be necessary, however, an experimental phase, through a pilot survey, in order to get estimates of all the unknown parameters about variability and design effect for the other variables of the HTS (number of trips, number of nights spent in the trips) not considered in the reported analysis because they are not collected in the MPS, which is at the moment the only source of reference unbiased information.

4. Conclusions

ISTAT CATI surveys may suffer from the undercoverage of the fixed telephone list, with the consequence of producing biased estimates. In the paper two possible approaches to achieve unbiased estimates are described. The first approach, maintaining unchanged the sampling design, is based on the enhancement of the estimator to take into account some more auxiliary variables, related to the coverage phenomenon. The second approach defines, instead, a new sampling design based on a list not suffering from undercoverage (the municipal register) and a different sampling scheme. Some experiments exploiting the data of the Multi Purpose Survey have been carried out. This survey is not affected by the undercoverage problems and it collects two useful information: the phone owner households and the yes/no variable about at least Four Overnight Stays, one of the most relevant interest variable of the CATI Holidays and Trips Survey. This survey has resulted to be exposed to the risk of bias deriving from the undercoverage of the selection list. In this empirical context we have assessed the undercoverage estimation bias in the subsample of phone owner households, in comparison to the estimate obtained on the overall MPS sample. Hence, the two proposed approaches have been applied. The results have been encouraging on both sides, even though further analysis must be arranged on the real data of the Holidays and Trips Survey. These are the further aims of the research group which has been established at ISTAT to study the undercoverage bias of CATI surveys.

References

1. ISTAT: Documento di sintesi sullo stato delle indagini CATI presso le famiglie e sull'analisi delle criticità legate alla copertura delle indagini basate su liste di telefoni fissi. Technical report. (2009a)
2. ISTAT: Documento di sintesi sulla prospettiva di ampliamento delle indagini CATI sulle famiglie e sulle possibili soluzioni da adottare per risolvere i problemi di copertura e qualità delle indagini telefoniche. Technical report. (2009b)
3. Särndal, C.-E., Lundström, S.: Estimation in Surveys with Nonresponse. Wiley (2005)