

Comparing Probability Sample Designs for the Selection of Municipalities in the Consumer Price Survey

*Confronto tra disegni campionari probabilistici per la selezione
dei Comuni nell'indagine sui prezzi al consumo*

N. Cibella, C. De Vitiis, E. Scavalli, P. Righi, T. Tuoto
ISTAT, tuoto@istat.it

Riassunto: Nel lavoro si illustrano i primi risultati di una sperimentazione per la selezione probabilistica di un campione di comuni per l'indagine per la stima della variazione dei prezzi al consumo condotta dall'Istituto Nazionale di Statistica. Tale sperimentazione rientra in un più vasto progetto che prevede la ridefinizione di tutte le fasi dell'indagine compresa quella del disegno del campione in tutti gli stadi di selezione. L'indagine attuale si basa su un disegno campionario di tipo ragionato, fortemente vincolato da un modello teorico implicito, che può condurre a stime distorte, se non rappresenta correttamente la dinamica del fenomeno. Per evitare elementi distorsivi sulle stime si propone un disegno campionario probabilistico bilanciato.

Keywords: Consumer Price Survey, Balanced Sampling

1. Introduction

The current Consumer Price Survey is based on a multistage purposive sampling design in which the selection criterion at each stage is based on the concept of *most representative* units. Sampling units are respectively the municipalities, the point of purchases (POPs) and the items on which the price are collected. Such a design is founded on strong hypothesis, that if not satisfied can cause bias in the estimates. Therefore, it appears interesting to implement a fully probabilistic scheme. The recent availability of a Local Unit Archive offers the main support to the sampling redesign, as it can be utilised to construct a list of POPs from which the sample can be drawn. This archive is yearly updated and has a good coverage of the population of POPs; the economic activity classification - 5 digit NACE - allows to define the *types of product* (goods or services, in the following TP) sold by each POP with a good approximation. The types of product, currently in number of nearly 200, define the elements of the partition of the population of good and services, at a lower level with respect to the COICOP classification.

The general sampling scheme of the proposed redesign is similar to the current survey, but the selection at each stage is up to the inclusion probabilities proportional to the turnover value (known by the register or suitably imputed), aiming at producing estimates of price index for each cell defined by the cross-classification of a geographic area level and *type of product*. The proposed sampling strategy is described in detail in ISTAT (2005). In the present paper the focus is on the comparison of different schemes for the selection of municipalities, based on *pps* and *balanced sampling* design, having

the aim of making the most of the auxiliary variables known from the Outlet Register, in order to obtain unbiased and accurate estimates and a good coverage of the types of product at geographical area level.

2. The balanced sampling

In this paper two different types of sampling design are compared. The first one is a *probability proportional to size (pps)* sampling design using for each municipality the overall turnover of the points-of-purchase belonging to. The balanced sampling is the second type of design considered. In this paper, a balanced sample is defined as a sample in which the Horvitz-Thompson estimates for a set of auxiliary variables are equal to their known population totals (Deville and Tillé, 2004). A formal description of balanced sampling introduces the general definition of sampling design. Let S be the set of all the samples s of the population U , and let S be the random variable representing S that takes the value $\mathbf{s} = (s_1, \dots, s_k, \dots, s_N)'$, with $s_k = 1$ if the unit k ($k \in U$) belongs to s and $s_k = 0$ otherwise. Given $\boldsymbol{\pi} = (\pi_1, \dots, \pi_k, \dots, \pi_N)'$ the vector of inclusion probabilities, we consider the sampling designs $p(\cdot)$ consisting in assigning a probability $p(s)$ to each sample s such that $E(\mathbf{s}) = \sum_{s \in S} p(s) \mathbf{s} = \boldsymbol{\pi}$. Furthermore, given a vector $\mathcal{X} = (X_1, \dots, X_h, \dots, X_H)'$ of auxiliary variables available at the sampling stage, the sampling design $p(s)$ is said to be balanced with respect to $\mathcal{X}$, if and only if it satisfies the balancing equations given by :

$$\sum_{k \in U} (\mathbf{x}_k / \pi_k) s_k = \sum_{k \in U} \mathbf{x}_k, \quad (1)$$

for all $s \in S$ such that $p(s) > 0$, where $\mathbf{x}_k = (x_{k1}, \dots, x_{kh}, \dots, x_{kH})'$ is the observed vector of $\mathcal{X}$ on the unit k . The right and left hand side of formula (1) are respectively the Horvitz-Thompson estimators and the known population totals.

One of the main problem of balanced sampling has always been the implementation of a procedure giving a multivariate balanced random sample. Recently, Deville and Tillé (2004) proposed a method, called *cube method*, that allows the selection of balanced samples (or approximately balanced) for a large set of auxiliary variables.

For our purposes, the inclusion probability π_k of each municipality is taken for each unit as proportional to the total turnover at municipality level over the total turnover at the level of the geographical area chosen as domain of estimate, i.e. the region.

It is important to underline that the choice of the turnover as auxiliary variable and the definition of such inclusion probabilities implies some important optimal properties in the context of super-population approach (see Särndal et al., 1992, Vaillant, 1999, Istat 2005). An other reason to introduce the methodology of balance sample is to obtain a good level of coverage of the TPs in the selected municipalities of each geographic area.

3. The frame of municipalities

The selection of sample municipalities requires the availability of a frame listing the sampling units and containing the auxiliary information needed for the definition of the inclusion probabilities and the implementation of the sample design.

This frame can be obtained starting from a list of point of purchase in which for each unit the whole turnover, subdivided among the different TPs sold by the POP, is reported. The Local Unit-Italian Active Enterprise Archive (ASIA-UL) is the more suitable archive for this purpose, when taking into consideration the sections regarding enterprises selling goods and services. This archive will be soon available containing the total turnover of the enterprises classified on the basis of the Economic Activity Code (ATECO, or NACE). In order to identify the TPs sold by each POP, a correspondence between the ATECO code and the classification of products has to be established. Once constructed the list of POPs with the indication of the TPs sold by each one, it is also necessary to subdivide the whole turnover of the POPs among each type of product. For this purpose it is necessary that a source of information about the structure of sales is available. Presently, some studies are carried out on this topic, like for example analyses conducted on data collected for taxation purpose.

In this paper a Monte Carlo simulation study of alternative selection methods is presented, based on a *training archive*. This archive has been constructed starting from the list of enterprises collected by the 2001 Italian Census of Industry and Services, taking into account only the sector of retail, i.e. outlet selling goods, covering 127 types of product out of the about 200 ones currently considered. As the census archive does not contain information about the turnover, it has been imputed using the Archive ASIA, where the turnover is registered at enterprise level. The operation of assigning the POP turnover to each TP has been made through the national weight system currently used for the production of the Consumer Price Index.

Municipalities in which no point of purchase is resident do not participate, obviously, to the sample selection. In this way the final reference universe is composed by 7,900 municipalities, out of the 8100 Italian municipalities.

4. Simulation and results

The experimentation for the choice of the sample design focused on nine alternative schemes of selection of the municipalities, on the basis of the list obtained from the *training archive* of POPs.

We implemented, firstly, six different procedures of balanced sampling, considering two alternative auxiliary variables and three level of their known population total. The chosen **X** variables are: (i) the turnover of the POPs belonging to each municipality for each type of product, 127 variables ; (ii) the number of POPs in each municipality selling each type of product, 127 variables. In both of the two kinds of **X** variables the balancing equation (1) is defined to obtain the known population totals, alternatively, at regional, main geographic area and national level. Another balanced procedure is performed including with certainty the 20 regional chief towns (Self Representative municipality, SR) and balancing at regional level with respect to the known turnover total. This choice allows to have in the sample at least a large municipality for each region. We have also considered two *pps* sampling designs without balancing but using the same inclusion probabilities, one fixing the 20 SR municipalities and one without. It is important to point out that all the schemes with SR municipalities do not respect the optimality properties of the sample design because of the removal of the proportionality to the total turnover.

Five hundred replications for each design were performed, on the basis of the same sample size as the current survey, nearly 80 sample municipalities, distributed among the regions partly uniformly and partly proportionally to the total turnover.

The selection procedures worked well, even if the algorithm for the balanced sampling has been converged on an only approximate solution.

The alternative schemes have been compared with respect to two different indicators: the mean (over the 500 replications) of the relative errors of the estimates of the known totals with respect to the known population totals themselves, reported in table 1; the coverage of the types of product in the drawn municipalities, (not reported for sake of brevity) resulted very close to 100% in all regions but the smallest ones.

The main results of the simulation are shown in table 1, where we report the errors on the estimation of the *turnover*, which are lower than the ones obtained balancing on the variable *number of POPs*. Notice that the estimation errors become balancing errors if we consider the errors for the variable on which the sample is balanced. We put in bold the balancing errors. The numbers show that the best outcome in term of balancing error is obtained balancing at national level (3.2% against 6.3% of the non-balanced sample), even if in this situation the percentage of non-coverage of the types of product (not shown for sake of brevity) becomes slightly higher than in the other cases. The two non-balanced designs produced, as expected, worse results.

Table 1: Mean relative errors (%) for the estimation of turnover totals for alternative sampling schemes .

Sample scheme		Turnover Estimate at level:		
		<i>Italy</i>	<i>Main geographic area</i>	<i>Region</i>
		Mean relative error	Mean relative error	Mean relative error
Not balanced	With SR	6,2	12,0	18,8
	Without SR	6,3	12,4	19,5
Balanced with respect to turnover at level:	Italia	3,2	10,0	17,2
	Main geographic area	4,6	8,7	16,5
	Region	5,2	10,2	16,0
	Region with SR	6,7	12,0	18,5

In conclusion, the outcomes of the simulation show that the sampling procedure based on balanced sampling yields very good results when the domain of estimate is the national one, but when the domains of estimate are more disaggregate (i.e. the regions), the gain of the balanced sampling decrease and, furthermore, optimality conditions are met less strictly.

References

- Deville J. C., Tillé Y. (2004). *Efficient balanced sampling: the cube method*, Biometrika, 91, 893-912.
- Istat (2005) *Proposta di un piano di campionamento probabilistico per la rilevazione territoriale dei prezzi al consumo*, Technical Report.
- Sämdal C. E., Swensson B., Wretman J. (1992). *Model Assisted Survey Sampling*. New York, Springer.
- Valliant R. (1999). *Uses of Models in the Estimation of Price Indexes: a Review*. In Proceedings of the Survey Research Methods Section, American Statistical Association.