

A Method of Probability Selection of Points of Purchase for the Consumer Price Survey

*Un metodo di selezione probabilistica di punti vendita per
l'indagine sui prezzi al consumo*

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

Riassunto: L'indagine italiana sull'indice dei prezzi al consumo è oggetto di alcuni studi volti a definire un rigoroso processo di revisione. Tra gli altri aspetti, è stata esaminata l'ipotesi di un disegno di campionamento che prevede la selezione probabilistica dei comuni, dei punti vendita e delle referenze. Questo lavoro, in particolare, propone uno schema per la selezione probabilistica dei punti di vendita, definito a partire da una lista di unità locali coinvolte nella vendita di beni e servizi e basato su una procedura di coordinamento dei campioni.

Keywords: Consumer Price Survey, Coordinated Samples.

1. Introduction

The Italian Survey on Consumer Price for the production of the Consumer Price Index has recently undergone a rigorous review aiming at redesigning different aspects of the data collection procedures and estimation methods. Among other suggestions, it has emerged the need to modify the sample design in order to implement a fully probabilistic scheme having the goal of producing estimates of price index for each cell defined by the cross-classification of a geographic area level and the *type of product* (goods or services, in the following TP). The type of product defines an element of the partition of the population of good and services, at a lower level with respect to the COICOP classification. The recent availability of a Local Unit Archive offers the main support to the sampling redesign, as it can be utilized to construct a list of Points of Purchase (POP in the following) 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 TPs sold by a given POP with a good approximation. The proposed sampling scheme may be summarised as follows: the municipalities are the primary sampling units, the POPs are the secondary sampling units and the *items* are the ultimate sampling units. The inclusion probabilities at each sampling stage are proportional to the turnover value (known by the register or suitably imputed), used as a proxy of sales. The sampling strategy is described in detail in Istat (2005).

The current survey design consists, on the contrary, at each sampling stage, of a purposive sample: the municipalities involved in the survey are fixed by law, as the most populated ones; the points of purchase are chosen as the *most representative* ones; the items on which the prices are collected are chosen as the *most sold* ones; such a design is founded on strong hypothesis, that if not satisfied can cause bias in the estimates.

In this paper the focus is on the description of the proposed procedure for the selection of the POPs: for each municipality involved in the survey, a sample of POPs is selected for each type of product; if these samples were selected independently of each other, the number of sample POPs could result very large. On the contrary, the proposed sample scheme aims to coordinate the samples in such a way to guarantee a maximum overlap of the selected POPs and so to minimise the number of sample POPs. A given POP is included in the sample for one or more specific types of product. The

proposed sample scheme could be utilised for the selection of the POPs apart from the method of selection of municipalities, either probability or purposive.

2. Coordination of POPs probability sample

In this paper a probability method for the selection of POPs is proposed, on the basis of a POP frame with the characteristics discussed in the next section. This method implies the selection of multiple samples from the same frame, being each selection carried out for each TP. The increase of the overlap between samples is desirable in order to contain survey costs. A way to handle with this kind of issue is using a coordinating procedure (Ohlsson, 1995).

As outlined in the previous paragraph, the aim of the project is to define a *pps* sample scheme based on probabilities proportional to the turnover of the sampling units. The main estimation goal is producing estimates of price index for each cell defined by the cross-classification of a geographic area level (i.e. administrative region) and the TP. For operational and administrative reasons, the geographical areas are organised in local districts (i.e. municipalities, but they could be defined in a more apposite way regarding the survey purposes).

Therefore, to define the inclusion probability of the POP v that retails the TP d , let ${}_d m_{(a)}$ denote the number of POPs to sample, for the type of product d within the geographical area a . The number ${}_d m_{(a)}$ could be fixed both on the basis of the minimum variance criterion, and on cost consideration. Within the geographical area, a number $n_{(a)}$ of local districts can be selected on the basis of law indication, as it currently happens, or on the basis of a probability sample, being π_c the inclusion probability of the local district c . So the number ${}_d m_c$ of POPs to sample in the local district c has to be determined so that the inclusion probability of the POP will be proportional to the turnover of the POP within the local district

$${}_d m_c = {}_d m_{(a)} \frac{\tilde{F}_c}{(\pi_c \tilde{F}_{(a)})} \quad (1)$$

where $\tilde{F}_c$ and $\tilde{F}_{(a)}$ are, respectively, the total turnover of the POPs selling the TP d within the local district c and in the region a . Formula (1) assures the following condition

$$\sum_{c=1}^{N_{(a)}} \pi_c {}_d m_c \left[\frac{\tilde{F}_c}{(\pi_c \tilde{F}_{(a)})} \right] = {}_d m_{(a)}.$$

Moreover, formula (1) guarantees some optimal proprieties in the context of super-population approach (see Särndal et al., 1992 and Vaillant, 1999) when π_c is defined proportional to total turnover of all the POPs belonging to c . The resulting final probability for the POP v for the selection relative to the TP d is defined as

$${}_d \pi_{c,v} = {}_d m_{(a)} \frac{\tilde{F}_{c,v}}{\tilde{F}_{(a)}} \quad (2)$$

where $\tilde{F}_{c,v}$ is the turnover of the POP v relative to the TP d . In order to coordinate the POP samples relative to all the TPs sold by the POPs in the local district c , a Sequential Poisson sampling technique is implemented through the use of Permanent Random Number (Ohlsson, 1995). This selection scheme has been chosen as it guarantees a fixed sample size, even if it is not a strict *pps* procedure.

3. The Archive of Points-of-Purchase

To implement the proposed sampling scheme, the availability of a list of POPs is required. For each unit the frame has to report the whole turnover subdivided among the different TPs sold by the POP itself. 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 all information related to 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 codes and the classification of products has to be established. It should be stressed that the current classification of the TPs is not always appropriate to be linked to the ATECO codes. In some cases it is not even possible to define an exact correspondence between the TPs and the ATECO codes. This difficulty suggests the need to review the classification of product, sub-classifying the COICOP codes taking into account the compatibility with the ATECO codes.

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. Currently, some studies are carried out on this topic, like, for example, analyses conducted on data collected for taxation purpose.

However, in this paper an application of the proposed selection method is presented, based on an training archive. This archive has been constructed starting from the list of enterprises collected with the 2001 Italian Census of Industry and Services, taking into account only the sector of retail, i.e. outlets selling goods. As the census archive does not contain information about the turnover, it has been reconstructed using the Archive ASIA, where the turnover is registered at enterprise level, through imputation methods taking into account the geographical location, the ATECO code and the number of employees. The operation of assigning to each POP the turnover related to each TP has been made by means of the national weight system currently used for the production of the Consumer Price Index.

4. Experimentation of the methodology

In this work, the proposed method is applied to select POPs in three municipalities: Milano, Roma and Napoli. The choice of these municipalities is due to the fact that they are involved in the current consumer price survey, they are collocated in different geographical areas and they would be included with probability equal to 1 in a probability proportional to turnover sample of municipality. The aim of this experimentation is essentially to test the coordination procedure and to compare the current POP sample with a probability and coordinated one.

In order to assign the number of POPs to be selected and the inclusion probability to the POPs in each municipality, as defined in section 2, the monthly number of items currently pointed out in each municipality are considered. The sample sizes are from a month of the 2002 consumer price index survey. The comparison between the current procedure and the coordination scheme is conducted under the assumption that, when a POP is selected for the TP d , a number of prices equal to the number of elements currently in the *basket*¹ classified in the type of product d has to be collected.

The results are reported in tables 1, where some statistics are shown about the frame of POPs and about the current and the coordinated sample of POPs, in terms of size of the samples, average and maximum number of type of product collected in each selected POP

¹ The current survey is based on the use of a fixed basket of sample products, classified in TPs

Table 1. *Results of the coordination in the sample of POPs*

Municipality	Milan	Rome	Naples
Population	1,247,052	2,540,829	1,008,419
POPs in the frame	14,861	31,821	13,293
Number of items currently pointed out	15,582	7,873	5,118
POPs in the current sample (CuS)	1,110	432	680
Average number of TPs for the CuS	4.9	7.1	4.2
Maximum number of TPs for the CuS	92	67	90
POPs in the coordinated sample (CoS)	591	380	354
Average number of TPs for the CoS	3.7	3.8	3.3
Maximum number of TPs for the CoS	63	55	34

The municipality of Milan is a special case in the current survey, as Milan is in the forefront in data collection and even it collects a larger than required number of prices. So the number of POPs and items currently pointed out in Milan are rather larger than the number of POP and items currently pointed out in Rome and Naples.

The main outcome of the experimentation is the evidence that with the coordinated sample the number of POPs required to collect the same number of items is always lower than that in the current sample, even if in Rome the difference is not so stressed. Anyway, the reduction of the sample size occurred with the coordinated selection is a result of the coordination as well as of the fact that the current sample size is the result of a real collection procedure, whereas the coordinated sample size is not tested on field.

References

- Istat (2005) *Proposta di un piano di campionamento probabilistico per la rilevazione territoriale dei prezzi al consumo*, Technical Report.
- Ohlsson E. (1995). *Coordination of Samples Using Permanent Random Number*. In Business Survey Methods. A Cura di Cox B. G., Binder D. A., Chinnappa B. N., Christianson A., Colledge M. J., Kott P. S., 153-170. New York, Wiley.
- Särndal 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.