

A Methodology for the Selection of Items in Points of Purchase for the Consumer Price Survey

Una metodologia per la selezione delle referenze nei punti di vendita per la rilevazione territoriale dei prezzi al consumo

Claudia De Vitiis, Piero Demetrio Falorsi, Gianni Marliani

ISTAT, {;devitiis;falorsi}@istat.it;
Università di Firenze, marliani@ds.unifi.it

Riassunto: L'indagine italiana sui prezzi al consumo, su cui è basata la produzione dell'indice dei prezzi al consumo, sta subendo un rigoroso processo di revisione. Tra gli altri aspetti, è in corso di revisione il disegno di campionamento con lo scopo di definire uno schema di selezione probabilistica. Oggetto di questo lavoro è una proposta di procedura per la selezione probabilistica delle referenze, basata sulla costruzione di una classificazione gerarchica dei prodotti e sulla definizione di un disegno campionario multi fase.

Keywords: Consumer Price Survey, hierarchical classification, multiphase sampling design.

1. Introduction

The Italian Consumer Price Survey (ISCP) for the production of the Consumer Price Index (CPI) has recently undergone a rigorous review aiming at redesigning different aspects of the data collection procedures and estimation methods. A scientific committee, composed by university professors and researchers of various institutions (e.g. Ministry of Finance and Economy), has been established at National Statistical Institute for steering and monitoring the ISCP innovation process. Among other suggestions, the committee has stressed 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 *type of product* (goods or services). The type of product defines an element of the partition of the population of goods and services. The recent availability of an Outlet Register (OR) offers the main support to the sampling redesign. OR is yearly updated and has a good coverage of the population of outlets; the economic activity classification - 5 digit NACE - allows to define the types of product sold by a given outlet with a good approximation. The general sampling scheme may be summarised as follows: the municipalities are the Primary Sampling Units, the outlets are the Secondary Sampling Units and the *items* are the ultimate sampling units. For each selected municipality, a sample of outlets is selected for each type of product; these samples are coordinated in such a way to guarantee a maximum overlap of the selected outlets. In this design a given outlet is included in the sample for one or more specific types of product. The inclusion probabilities at each sampling stage are proportional to the turnover value (known by the register or suitably imputed). The sampling strategy is described in detail in Istat (2005). In this paper the focus is on the description of a feasible probability item selection procedure, based on the construction of a hierarchical classification of the universe of products within each type of product and on a multi-phase sampling scheme. The proposed scheme allows to overcome the drawbacks of the solution adopted

for the current survey, based on the use a fixed *basket of products*, easy to implement, but which may introduce an unknown bias in estimates if the purposive sample is not representative enough. A similar approach has been experimented for the British CPI (Ball et al., 2004), based on the use of scanner data.

2. General Scheme

The phase of item selection from selected outlets is the most delicate and complex issue of the entire sampling plan. The current plan is based on a practically easy solution consisting in the definition of a fixed basket of products, selected independently from the first two sampling stages (municipalities and outlets). For each product in the basket, the most sold item is chosen in the outlet at the beginning of each year and the prices of this item are collected for the twelve month of the year. As the most important scope of the proposed design is introducing a probability sample at every stage, there is the need for a sampling scheme giving rise to an item selection procedure which the interviewer could easily put in practice at every outlet. The main operational difficulty is the construction of the list of all the items sold in the outlet with regard to the types of products for which the outlet is included in the sample. The list construction is not practically feasible, above all for complex types of products or for large outlets which have a very large list of items for each type of products.

A useful strategy to overcome the above difficulties is proposed in the following. All the products of each type of products are a priori classified using *hierarchical tree structure* with an assigned number of levels. Each node of the tree represents a specific *Aggregate of Products* (AP). Each parent node is split at a lower level into a maximum tractable (e.g. 10) number of child nodes. An AP of the *final level* identifies a very small subset of homogeneous items. A probabilistic *selection tree procedure* is defined. Starting from the selection of one AP of initial level, in the subsequent step, one AP is selected among the child nodes of the selected AP of the previous level. The procedure is repeated until one AP of *final level* is selected. This procedure is implemented in proper software for the interviewers' laptops and performed just once at the sample outlet at the beginning of each year. The item selection procedure is a two step multiple phase sampling. In the *first step*, if $\bar{j}$ items must be selected in a given AP for a given type of products, the selection tree procedure is repeated $\bar{j}$ times independently. Each time a multiple phase sampling is applied. In this way it is possible to individuate $\bar{j}$ different sample final level APs or, at the extreme opposite situation, the same final level AP may be selected $\bar{j}$ times. The number of items to be selected in a given sample final level AP is defined by the number of times the AP is selected in the $\bar{j}$ replications of the selection tree procedure. In the *second step*, the interviewer constructs the list of the items belonging to the sample final level APs selected in the first step and then selects on or more items (depending from the first step outcomes) from the list defined for each final level sample AP with a sequential Poisson procedure (Ohlsson, 1995).

3. Selection probabilities

In order to describe formally the item selection procedure, let's introduce the following notation. Let's consider the d -th type of products and let's define with: λ ($\lambda = 1, \dots, dA$) the level of the *hierarchical tree structure*, being dA the *final level*; i_1 ($i_1 = 1, \dots, dI$) the identification label of the initial level AP; $i_2 | i_1$ the label of the AP of level 2, that is a child node of the AP i_1 ; $i_\lambda | i_1, \dots, i_{\lambda-1}$ ($i_\lambda = 1, \dots, dI_{i_1, \dots, i_{\lambda-1}}$) the label of the AP of level λ , that is a child node of AP $i_1, \dots, i_{\lambda-1}$. With

reference to the sample outlet v selected in the municipality c , let's define with: $d, i_\lambda | i_1, \dots, i_{\lambda-1} \Psi_{c,v}$ a probability value for the AP $i_\lambda | i_1, \dots, i_{\lambda-1}$ - being $d, i_\lambda | i_1, \dots, i_{\lambda-1} \Psi_{c,v} = 0$ for the items of the AP

$i_\lambda | i_1, \dots, i_{\lambda-1}$ not sold in the outlet (c,v) , being $\sum_{i_\lambda=1}^{d, i_\lambda | i_1, \dots, i_{\lambda-1}} d, i_\lambda | i_1, \dots, i_{\lambda-1} \Psi_{c,v} = 1$.

Let $j | i_1, \dots, i_{A_d}$ be the label ($j | i_1, \dots, i_{A_d} = 1, \dots, d, j | i_1, \dots, i_{A_d} \Psi_{c,v}$) identifying a specific item in the list (constructed by the interviewer) of the items belonging to the sample final level AP $i_{A_d} | i_1, \dots, i_{A_d-1}$; let $d, j | i_1, \dots, i_{A_d} \Psi_{c,v}$ be a probability value for the item $j | i_1, \dots, i_{A_d}$ - being $d, j | i_1, \dots, i_{A_d} \Psi_{c,v}$
 $\sum_{j=1}^{d, j | i_1, \dots, i_{A_d}} d, j | i_1, \dots, i_{A_d} \Psi_{c,v} = 1$.

In order to describe the proposed AP selection procedure, let's suppose that in the sample outlet (c,v) $d, \bar{j}(c,v)$ items of the product type d must be selected. $d, \bar{j}(c,v)$ different Permanent Random Numbers (PRNs) (drawn from an uniform $[0,1]$ distribution) are assigned to each AP for the yearly selection, to be used for the $d, \bar{j}(c,v)$ random selections where the AP is involved. Let's add that: (i) the $d, \bar{j}(c,v)$ PRNs are selected independently; (ii) they PRNs remain unchanged for the consecutive years in which the outlet is included in the sample. Condition (i) guarantees that the final level APs are selected independently while condition (ii) assures the maximum overlapping between the samples of consecutive years. Let's then denote with $d, i_\lambda | i_1, \dots, i_{\lambda-1} X_{c,v}(u)$ the PRN assigned to the AP $i_\lambda | i_1, \dots, i_{\lambda-1}$, for the u -th ($u=1, \dots, d, \bar{j}(c,v)$) sample selection.

For the u -th selection, the AP $i_\lambda | i_1, \dots, i_{\lambda-1}$ is drawn conditionally by the selection of the previous level parent APs and it is chosen among the $d, i_{\lambda-1} | i_1, \dots, i_{\lambda-2}$ AP child nodes of the parent node $i_{\lambda-1} | i_1, \dots, i_{\lambda-2}$. The selection is carried out through a *pps* sampling using the probabilities $d, i_\lambda | i_1, \dots, i_{\lambda-1} \Psi_{c,v}$ and the PRNs $d, i_\lambda | i_1, \dots, i_{\lambda-1} X_{c,v}(u)$. Through this procedure the APs of subsequent levels are selected until a final level AP A_d is chosen.

Having selected a final level AP, the interviewer will make (out) the list of all the items sold in the outlet and draw $d, i_{A_d} | i_1, \dots, i_{A_d-1} \gamma_{c,v}$ items from it, denoting with $1 \leq d, i_{A_d} | i_1, \dots, i_{A_d-1} \gamma_{c,v} \leq d, \bar{j}(c,v)$ the number of times the AP $i_{A_d} | i_1, \dots, i_{A_d-1}$ has been chosen randomly in the first step sampling procedure. The selection of the $d, i_{A_d} | i_1, \dots, i_{A_d-1} \gamma_{c,v}$ items is done using *pps* sampling with inclusion probabilities equal to $d, i_{A_d} | i_1, \dots, i_{A_d-1} \gamma_{c,v} \times d, j | i_1, \dots, i_{A_d} \Psi_{c,v}$. If $d, i_{A_d} | i_1, \dots, i_{A_d-1} \gamma_{c,v} = 1$, a selection procedure similar to that used for the selection of the APs is adopted. Whereas, if $d, i_{A_d} | i_1, \dots, i_{A_d-1} \gamma_{c,v} > 1$ a sequential Poisson sampling is implemented. Both techniques use PRNs assigned to the items for different year selection, assuring the maximum overlap in consecutive year sample.

The described sampling procedure makes use, at every selection step, of probabilities ($d, i_\lambda | i_1, \dots, i_{\lambda-1} \Psi_{c,v}$ for the AP and $d, j | i_1, \dots, i_{A_d} \Psi_{c,v}$ for the items), whose values have to be defined in an appropriate way on the basis of the information available at the sample outlet. The optimal situation would occur if these probabilities were equal to the *proportion* of the turn over relative to the unit ($i_\lambda | i_1, \dots, i_{\lambda-1}$ -th AP or

$j | i_1, \dots, i_{A_d}$ -th item) to the total turn over of the outlet for the set of units among which the selection has to be carried out at the specific level. In practice, it is possible to define these probabilities in different ways: (i) if the outlet has an informative system that records all the sales of each item (identified by a specific bar code), the probabilities may be directly calculated; (ii) in small outlets the proportions may be directly collected by a proper interview to the person in charge of the outlet; (iii) the probabilities may be determined by proxy variables, as for example the surface of the shelves occupied by the units; (iv) if no information is available at the outlet level, external information, previously collected by means of proper surveys (or from other sources) and available in the interviewer's laptop, can be used. The last option can be adopted only for the selection of the APs, since the item list is constructed only at the moment of data collection.

In the procedure for the estimation of CPI, the survey data must be weighted by a sample weight expressed as product of factors each referred to a different sampling stage. The factor at *outlet level* may be obtained using the theory of the π^* estimator for multiphase sampling (Sämdal, 1992, sec. 9.3) as

$$d, j \omega_{c,v} = (d, i_1 \Psi_{c,v} \times \dots \times d, i_{\lambda} | i_1, \dots, i_{\lambda-1} \Psi_{c,v} \times \dots \times d, i_{A_d} | i_1, \dots, i_{A_d-1} \Psi_{c,v} \times d, j | i_1, \dots, i_{A_d} \Psi_{c,v} d \bar{j}_{(c,v)})^{-1}.$$

It is easy to demonstrate that if the probabilities are proportional to the turnover, the $d, j \omega_{c,v}$ factor is expressed as $d, j \omega_{c,v} = d, F_{c,v} / (d, j F_{c,v} d \bar{j}_{(c,v)})$ being $d, F_{c,v}$ and $d, j F_{c,v}$ respectively the total turnover of the type of products d and the turnover of item j in the outlet (c,v) .

4. Final remarks

The method for the selection of the items shown in this paper allows to implement a sufficiently feasible field procedure and to overcome the potential source of bias of the procedure adopted in the current survey, based on subjective choices; while the sampling strategy here proposed allows obtaining unbiased estimates of the interest parameters. Further, the selection procedure is based on PRN techniques and assures a good overlap of the samples of different years. The main drawbacks are the implementation costs of the classification of the *aggregates of products* and the collection of the information (from different sources) needed for the definition of the probabilities described in point (iv) of section 3.

References

- Ball A., Waldron K., _Smith K, Hughes J. (2004) *Changes to methodology employed in the CPI and RPI from February 2004*, Economic Trends, No. 604, pp 64-69.
- 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*, Springer Verlag, New York.
- 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.