Measuring inflation through different sampling designs implemented on scanner data

A. Bernardini[1], C. De Vitiis[2], A. Guandalini[3], F. Inglese[4], M. D. Terribili[5],

Abstract

The Italian National Institute of Statistics (ISTAT) is planning a redesign of Consumer Price Survey (CPS). Register of local units and availability of scanner data (SD) from retail modern distribution, provided through an agreement with Nielsen, are the starting point of this transformation. Especially SD represents a big opportunity for introducing improvements in terms of both data collection and sampling perspective. This work aims to compare the properties of weighted and unweighted higher level elementary price indices when different sample schemes are implemented on SD (modern distribution for food and grocery data). Probabilistic and non-probabilistic samples of series have been drawn with different schemes. Furthermore, different criteria of sample allocation both for outlets and elementary items were also considered. The comparison among different sampling strategies (allocation, selection) has been carried out through a Montecarlo simulation evaluating bias and efficiency of the Consumer Price Index (CPI) derived using three classic aggregation formulas (such as Jevons, Fisher and Lowe). The simulation has been conducted on permanent series of some consumption segments (Italian COICOP6) for the Turin province.

Keywords: Consumer Price index, scanner data, sampling

1. Introduction

The Italian National Institute of Statistics (ISTAT) is planning a redesign of Consumer Price Survey (CPS). Scanner data (SD) represent a big opportunity for introducing improvements in the CPS in terms of both data collection and sampling perspective. Statistical sampling theory strongly recommends the use of probability sampling for all types of statistical surveys including the economic ones. Nevertheless several countries carry out CPS on the basis of a non-probability sampling scheme. Other national statistical offices, as well as the Italian National Institute of Statistics (ISTAT), produce estimates for CPS using a purposive multistage sampling selection scheme, in which the selection criterion at each stage is based on the concept of most representative units.

[1] Researcher (Istat), e-mail: bernardini@istat.it.
[2] Senior researcher (Istat), e-mail: deviitis@istat.it.
[3] Junior researcher (Istat), e-mail: alessio.guandalini@istat.it.
[4] Senior researcher (Istat), e-mail: fringles@istat.it.
[5] Junior researcher (Istat), e-mail: terribili@istat.it.

The CPS has recently undergone a review aiming to redesign different aspects of the data collection procedures and sampling methods. The goal of the project is introducing progressively in the CPS more rigorous sampling procedures (that is probabilistic) starting from the selection of outlets and elementary items for the sectors where this is feasible.

The register of local units and the availability of scanner data (SD) from the retail modern distribution are the starting point for the implementation of this transformation. Furthermore, the use of web-scraping for collecting prices referred to online market (tourism, mobile phone, etc.) is involved in the review of the survey.

Since 2014, ISTAT is collecting, thanks of a contract with Nielsen, SD referred to food and grocery market from the six main retail chains operating in Italy. Actually, SD are used for experimenting the computation of Consumer Price Index (CPI). Therefore, Italy places among countries using or testing the use of this source of data for compiling CPI.

SD from retail stores, containing both prices and quantities, and allows the use of different aggregation formulas at the elementary level, involving or not the use of quantities (weighted or unweighted formulas). In this paper different methods for selecting references are compared. The values of different elementary index formulas have been compared with the value of the corresponding price indices obtained from the whole set of references for the same elementary aggregate.

In the paper the project of redesigning the CPI based on use of scanner data are presented (par. 2) and the former procedure for treating these data (par. 3) are outlined; paragraph 4 focuses on the sampling scheme tested for the selection of references, while the results of these sampling experiments are illustrates in paragraphs 5 and 6. Finally, some conclusions and open issues are discussed in paragraph 7.

2. The use of scanner data in CPI and ISTAT redesign project

2.1 Use of scanner data for CPI

Scanner data files contain generally elementary information referred to single EAN codes (GTIN) for specific outlets consisting in turnover and quantities sold during a week. This information does not provide the “shelf price” of the product individuated by the EAN code (European Article Number) and outlet (*references* or *series*), but it allows to define a unit value or average weekly price. This feature introduces a crucial issue in the discussion among economists, which goes beyond the focus of this paper. Furthermore, usually SD do not include information about discounts or special sales. This is not a problem because the harmonized CPI considers the price actually paid, including discount.

Actually Switzerland, Norway, Netherlands and Sweden, are using SD in the compilation of the CPI, while Belgium and Denmark are going to use them from 2016. Scanner data can be exploited in different ways. The simplest way would be to use SD as an alternative source for price collection, which replaces collection in the stores, without changing the traditional principles of computing the price indices. This method is currently applied by the Swiss Federal Statistical Office (Vermeulen and Herren, 2006). Alternatively, as in Norway and Sweden, SD can be used as *universe* from which samples of references can be selected following different methods (Nygaard, 2010; Norberg, 2014). Finally, all (or almost all) SD can be used to compile price indices, without a strict selection

but with consequences on the theoretical definition of the index. In the Netherlands, the computation methods differ and data are used in a more extensive way to calculate price indices (Van der Grient and De Haan, 2010).

In a study perspective, moreover, SD from retail stores allows to evaluate how different aggregation formulas at the elementary level perform. In fact, official CPI are usually constructed in two broad steps. First, for narrowly defined, relatively homogeneous products, also known as elementary aggregates, elementary price indices are calculated. In a second step, these elementary indices are aggregated into a single consumer price index using expenditure weights. Elementary indices, named also *higher level elementary indices*, are therefore the building blocks of price index numbers. Their development over time measures the inflation of narrow product categories.

While the aggregation at higher level is carried out using generally *Laspeyres type* formulas with weights deriving from national account or expenditure survey data, official practices in elementary price index construction are still not uniform across countries, deserving further investigation in the consequences of different choices (Gábor and Vermeulen, 2014).

On the basis of SD recently obtained, ISTAT is planning an experimental phase of price index evaluation, starting from the comparison among different formulas for the elementary indices combined with different ways of using those data through sampling selection schemes. The choice, in fact, is between using all or the most of data or a sample of them. The first option is very appealing and seems to exploit all the potentialities of scanner data but leaves open relevant issues both theoretical, regarding the definition of the price index and operational, the managing of a huge quantity of elementary items (EAN codes) each month for the production of the CPI. In the former it is needed to face the choice of the more efficient sampling design. Moreover, it presents some risks due to the need of managing the shrinking during the year of a fixed representative sample of elementary items selected on the basis of the previous year universe of items.

2.2 ISTAT project

2.2.1 Sampling redesign

The current strategy actually adopted in CPS carried out at territorial level is based on three purposive sampling stages. The sampling units are respectively the municipalities, the outlets and the elementary items for which the prices are collected. The biggest municipalities are forced by law to participate to the survey. The Municipal Offices of Statistics select the outlets sample, where the prices of a fixed basket of products (including roughly 1,500 products) are collected. The outlets sample is chosen to be representative of consumer behavior in the municipality. For each product of the basket the most sold item is selected and the price of this item is collected throughout the year. At the end of each year, besides the sample of outlets and the elementary items, ISTAT refreshes the fixed basket of products.

The elementary price indices are currently computed at municipality level by unweighted geometric mean. The general price index is calculated by subsequent aggregation of elementary indices, using weights at different levels based on population and national account data on consumer expenditure.

A working group established at ISTAT is developing and testing different sampling

strategies (allocation, selection with respect to different aggregation formulas) to evaluate which combination leads to best results in terms of CPI estimates efficiency. The first step to make the sampling selection feasible, is the preparation of the sampling frame. The basis for this operation is the availability of an exhaustive list of outlets containing data for the unit identification (such as address, telephone, etc.) and data referring to the economic activity, number of employees, products sold. The Local Unit Archive of active enterprise in Italy (ASIA), yearly updated with administrative sources, represents a good starting list for the construction of the outlets list. In order to produce a suitable sampling frame, the information currently collected in this archive are appropriately enriched. Furthermore, the information available only referred to the total enterprise and not for each single local unit, such as turnover, has to be subdivided among its local units according to some rule or model.

For the elementary item level different approaches will coexist at the beginning. SD for food and grocery in the modern distribution allows the use of sampling methods and price index compilation using weights from quantities (or expenditures). For the traditional distribution (small shops) and the other retail sectors, data collection and price index compilation will continue unchanged at first. The outlets sample will be defined for all the economic activities for which the archive is available, independently on the availability of scanner data. In fact, also for the modern distribution with SD it is useful to follow an outlets sample in order to have the possibility to deal each month with a manageable quantity of data.

The selection of elementary items from SD 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. Further, the selection procedure can be based on Permanent Random Number (PRN) techniques and assuring a good overlap of the samples of different years.

2.2.2 Scanner data structure and uses at ISTAT

At the moment Nielsen provides to ISTAT weekly data of turnover and quantities at EAN code (elementary item) and outlet level for six modern retail distribution chains (Conad, Coop, Esselunga, Auchan, Carrefour, Selex) operating in the food and grocery market in 35 Italian provinces. In the experimental stage Nielsen provides backward data for at least one full year and the preceding month of December, starting from December 2013 or 2014 (depending on the starting point of delivery of each province). The coverage of the six chains with respect the entire modern distribution is quite high in all provinces, with some variability in the geographical areas of the country.

Furthermore, Nielsen provided the dictionary for the classification of EAN codes to GS1-ECR-Indicod product classification. ISTAT ensures internally the translation from ECR to COICOP, the classification of products used for the CPI. Consumption segments, not foreseen by the EU-COICOP, are the most detailed domain of estimate for Italian CPI, constitute groupings of homogeneous products; those defined for the food and grocery are 121 out of a total of 324.

For the aim of CPI computation, the reference for which the price are observed during a certain period are defined as *series*, individuated by an outlet and an EAN code. Furthermore, only the *relevant weeks*, defined as the first three full weeks (composed of seven days) in each month are considered. Also in the recommendation drafted by

EUROSTAT, the use of data referred to the first two full week for each month is advised.

3. The treatment of data

The use of SD for the production of the CPI in modern retail distribution implies a very important data processing stage. The main goal of the definition and implementation of quality checks is to achieve a "cleaned" data set.

The SD quality checks aims to ensure the completeness and correctness of the continuous flow of information weekly uploaded by Nielsen. First, formal checks have been defined to ensure the completeness of the data collected at provincial level, distribution chains, outlets and weeks.

Then, others quality checks on loaded data are made in the treatment stage introducing editing rules. At this stage missing or inadmissible values on the variables of interest (weekly turnover, quantities sold and unit prices per EAN code) are identified. The quality check implemented to identify and eliminate the problematic series are:

- quantity<1 not motivated by unit of measurement;
- decimal values on quantities>1 not motivated by unit of measurement;
- unit prices <= 0,01 €;
- missing turnover.

In this case the checks on the presence of the inadmissible values were carried out considering the data acquired for the relevant weeks in each month at provincial level.

To identify inadmissible unit prices several methods have been tested. For the sake of simplicity, also in terms of computational burden, the choice has been reduced between two methods. Both methods are based on the computation of the median unit prices considering the quantities sold for each single occurrence (EAN code, week, outlet) at provincial level. The first method consists in a fixed trimming method and the identified tolerance interval of prices is:

$$\left(\frac{Median_w}{K_1}, K_2 * Median_w \right),$$

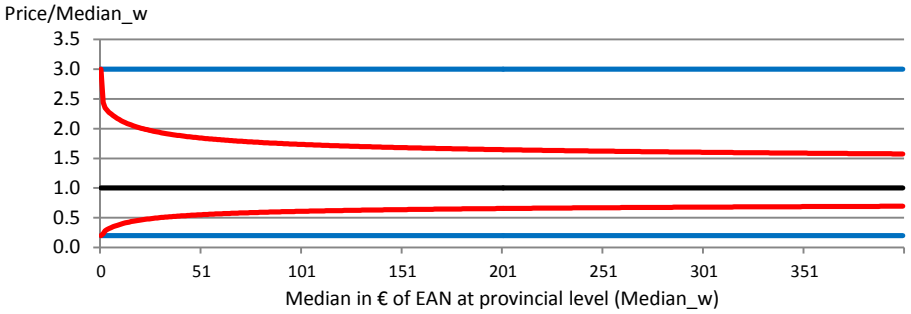
while the other it can be named mobile trimming and the related tolerance interval is

$$\left(Median_w - \frac{K_1 - 1}{K_1} \frac{Median_w}{log_{10}(Median_w + 10)}, Median_w + (K_2 - 1) \frac{Median_w}{log_{10}(Median_w + 10)} \right).$$

Trimming depends on the values assigned to K_1 and K_2. Assign values to K_1 and K_2 implies assumptions on maximum discount and maximum rise of prices with respect to the provincial median price of the product allowed. For food and grocery K_1=5 (it means that 80% of discount with respect to the median price of the product at most is allowed) and K_2=3 (it means that 3 times of rise with respect to the median price of the product at most is allowed) seem to be plausible values. However, while in fixed trimming the relative length of the tolerance intervals remain unchanged, with mobile trimming it becomes narrows as the median price of the product increases (see Figure 1). The mobile trimming is preferable

with respect to the fixed one because it narrows down, thanks to the log function, the extremes as the median price of the product increases.

Figure 1 – Tollerance interval limits of (prices/Medianw) with fixed and mobile trimming.



4. Sampling from Scanner Data

4.1 Outline

The first experiments on the SD have been aimed at the study of the performance of the weighted and unweighted higher level elementary price indices in different selection schemes of series. In this stage of the experiments, series life-cycle, seasonality problems and missing data substitutions haven't been taken into account.

For the first experiments reported here data cleaning is applied and a simplification is used: only *permanent* series are considered as universe for sampling and price index evaluation, so as not to have the necessity to deal with discontinued items. Permanent series are referred to those references with not-null turnover for at least one relevant week in each month of the considered year, starting from the December of previous year.

The experimental study has been developed in two phases. In the first phase probabilistic and non-probabilistic selection schemes of series are compared; selection schemes were analyzed separately for three consumption segments (coffee, pasta and mineral water) in Turin province (121 outlets). Several designs characterized by the use of different criteria of sample allocation, both for outlets and elementary items, and different selection methods of the sampling units have been considered. Herein, the market was assumed as a domain of interest and the sampling designs are referred to all markets (ECR groups) belonging to the six consumption segments (coffee, pasta, mineral water, olive oil, spumante and ice cream) in Turin province.

For each selection scheme considering, starting from the monthly price ratios with fixed base (December 2013) available for 2014, the higher level elementary indices were calculated using three classic aggregation formulas: Jevons (unweighted), Fisher (ideal) and Lowe (weights from quantities of previous year).

The choice of the price indices was made on the basis of theoretical and empirical considerations: Fisher ideal index is thus preferred by economic theory, it uses quantities in different times and allows for substitution effects; Lowe index is a modified Laspeyres in which a base month for the price (December 2013) and a base year (2013) for the quantities

are taken into account for the weight computation; Jevons index is, in the set of indices using only price information (as Carli and Dutot indices), the one that was found to show a more stable bias (measured with respect to the ideal Fisher index) among product categories (this result occurred from some empirical studies carried out on several consumption segments taken from the universe panel series SD).

Finally, in addition to Fisher index considered for its properties, Lowe and Jevons indices was dictated by attentions connected to the reference current context of CPI compilation.

Comparison between alternative selection schemes are made for each price index taking the corresponding universe (panel series SD) index value as benchmark. Indices performance were evaluated in terms of bias for all selection schemes of series. For probability selection schemes, accuracy (relative bias and sampling variance) of the price indices have been studied in a Montecarlo simulation scenario. In this context 500 samples have been selected, according to different sampling designs. Indices variability is measured considering the estimate of the relative sampling error, computed on the estimated indices in the replicated samples.

For the sample selection and weighting of price indices the total annual turnover was taken as reference.

4.2 Elementary index aggregation

The parameters of interest taken into account in this experimental stage are monthly Jevons, Fisher and Lowe indices defined at the consumption segment level.

Jevons index is an unweighted CPI that uses price information only (it assumes that expenditure shares remain constant), while Fisher and Lowe use also quantity information. These last indices consider expenditure shares at different times (current and reference period) as weights (Gábor and Vermeulen, 2014).

Indicating by the subscript t the current month (12 months in year 2014), t_0 the reference month (December 2013), m (m=1,…,M_i) the series and c (c=1,2,3) the consumption segment, Jevons, Fisher and Lowe formulas can be expressed as follows:

Jevons index - geometric average of the price ratios

$$JEVONS_{ct} = \prod_{m}^{M_c} \left(\frac{p_{cmt}}{p_{cmt_0}} \right)^{1/M_c} \quad ; \quad (4.1)$$

Fisher ideal index - geometric mean of the Laspeyres and Paasche

$$FISH_{ct} = \sqrt{LASP_{ct}} * PAAS_{ct} \qquad (4.2)$$

in which:

- Laspeyres index - arithmetic weighted average of the price ratios

$$LASP_{ct} = \sum_{m}^{M_c} \left(\frac{p_{cmt}}{p_{cmt_0}} \right) * \left(\frac{p_{cmt_0} * q_{cmt_0}}{\sum_{m}^{M_c} p_{cmt_0} * q_{cmt_0}} \right) \qquad (4.3)$$

- Paasche index - harmonic weighted average of the price ratios

$$PAAS_{ct} = \sum_{m}^{M_c}\left(\frac{p_{cmt_0}}{p_{cmt}}\right) * \left(\frac{p_{cmt} * q_{cmt}}{\sum_{m}^{M_c} p_{cmt} * q_{cmt}}\right); \qquad (4.4)$$

- Lowe index - arithmetic weighted average of the price ratios

$$LOWE_{ct} = \sum_{m}^{M_i}\left(\frac{p_{cmt}}{p_{cmt_0}}\right) * \left(\frac{p_{cmt_0} * q^{z}_{cml}}{\sum_{m}^{M_c} p_{cmt_0} * q^{z}_{cml}}\right) \qquad (4.5)$$

where

$$q^{z}_{cml} = \sum_{a=0}^{11} q_{cm(t_0-a)} \,.$$

The q^{z}_{cml} measure refers to the *m*-th quantity series of the *c*-th consumption segment in the previous year (2013).

4.3 Accuracy of price index estimates

The price indices (Lowe, Fisher and Jevons) are estimated in the sample using a plug-in estimator. Montecarlo simulations with 500 replicated samples have been performed for each sampling design.

Bias, with respect to the related index computed in the universe of SD and relative sampling error formulas shown below, are expressed for a generic parameter (price index) and with reference to simulation context.

For a generic estimated index in the *c*-th consumption segment (*c*=1,2,3), $\hat{\theta}_c$, the bias can be expressed as

$$RB\left(\hat{\theta}_c\right) = \frac{E\left[\hat{\theta}_c\right] - \theta_c}{\hat{\bar{\theta}}_c} \qquad (4.6)$$

in which $E\left[\hat{\theta}_c\right]$ is the expected value of the estimated index $\hat{\theta}_c$ in the consumption segment *c*, obtained from 500 samples, and θ_c is the corresponding index value computed on the reference universe (panel series SD).

The relative sampling error of a generic estimated index $\hat{\theta}_c$ in the consumption segment *c* can be expressed by

$$RE\left(\hat{\theta}_c\right) = \frac{\sqrt{Var\left(\hat{\theta}_c\right)}}{\hat{\bar{\theta}}_c} \qquad (4.7)$$

in which $\hat{\bar{\theta}}_c$ are is the mean of $\hat{\theta}_c$ and $Var\left(\hat{\theta}_c\right)$ variance of are calculated on the estimates generated from the selection of 500 samples in the consumption segment *c*.

5. First experimental studies

In the first experimental stage probabilistic and non-probabilistic selection schemes of series in each consumption segment are considered: probability proportional to size samples in the first approach; cut-off and representative elementary item samples in the second approach.

Probability sampling design considered in the experiments is a two-stages sampling with stratification of the first stage units. The primary stage units (PSU) are the outlets while the secondary stage units (SSU) are the EAN.

The outlets sample size has been fixed at a number of 30 out of 121 outlets available in SD. The sample size for the second stage is given by a sampling rate of 5 percent of the covered turnover by the elementary items in each sampled outlet to the first stage.

The outlets are stratified in the universe by chains (Conad, Coop, Esselunga, Auchan, Carrefour, Selex) and outlet types (hypermarket and supermarket). In each stratum, outlets sample has been allocated proportionally to the turnover of the strata.

The outlets selection is carried out on each stratum with a simple random sampling (SRS), while the elementary items selection is done with probability proportional to size (PPS) sampling. In the latest case, the inclusion probability of sampling unit is proportional to its turnover during the previous year, 2013. The elementary items selection was carried out by adopting two different methods for drawing PPS samples: Sampford (Sampford, 1967) and Pareto (Rosén, 1997a and 1997b) sampling.

As regards to non-probabilistic selection approach the selection of series is obtained on cut-off samples based on thresholds of covered turnover in previous year, 2013: two samples are formed with all elementary items covering respectively the 60 and 80 percent of all turnover in each selected outlet. Moreover, considering the current fixed basket approach, a second selection scheme was defined; in this case most sold elementary items for each representative product (of the fixed basket) in the selected outlets are considered.

Irrespective of the consumption segment, probability sampling always brings more efficient estimates than non-probability selection schemes. The cut-off at 80% is always better than the cut-off at 60% and both are closer to the real value of the index lower is the variability of the segments in terms of prices and turnovers of series (see De Vitiis et al., 2015 for further results).

6. Further experimental studies

6.1 Operational framework

In this stage of experiment, three different sampling designs have been considered: stratified sample of EAN, stratified one-stage sample of the outlets and two-stages sampling (outlets and EAN) with stratification of PSU and SSU. In each sampling design the size of EAN has been fixed in average at 7,400 to compare the different sampling strategies on

equal computational effort. Moreover, different criteria of sample allocation both for outlets and elementary items and different selection methods of the sampling units have been considered.

The stratified sampling design has been carried out stratifying the EAN by markets (ECR groups) in each six consumption segments (coffee, pasta, mineral water, olive oil, spumante and ice cream). In each market, sample size is defined through a Neyman[6] formula. Neyman allocation takes into account the prices relative variability of the elementary items in the markets observed in the reference year 2013. Two selection scheme for the selection of units have been considered: simple random sampling (SRS) and probability proportional to size sampling (PPS). In the former case, all the EAN in the same market have the same inclusion probability. In the latter case, they have inclusion probability proportional to the markets turnover during the previous year, 2013.

In Stratified one-stage sampling design a sample of outlets (14 out of 121 outlets) have been selected. The outlets have been stratified as described in paragraph 5.2 (by chains and types). In each stratum, two different allocation of outlets have been tested: proportionally to the strata turnover and optimal allocation by Neyman (based on the prices relative variability of the EAN observed in the strata during the previous year, 2013). Also in stratified one-stage sampling the units (in this case outlets) have been selected both with a simple random sampling (SRS) and with probability proportional to size sampling (PPS) with size equal to turnover (2013). All EAN in the selected outlets are included in the sample.

Two-stages sampling design is characterized by a stratification of both primary and secondary units. The adopted stratification of the PSU (outlets) and the SSU (EAN) are the same already described above. The outlets sample size has been fixed at a number of 30 out of 121 outlets present in the Turin province. For both outlets and elementary items, sample allocation in the strata is defined through the probability proportional to size sampling (PPS). The primary (outlets) and secondary (elementary items) sampling units are selected with a simple random sampling (SRS) and a probability proportional to size sampling (PPS). In both selection methods some sampling units are selected with certainty.

6.2 Indices evaluation under alternative sampling designs

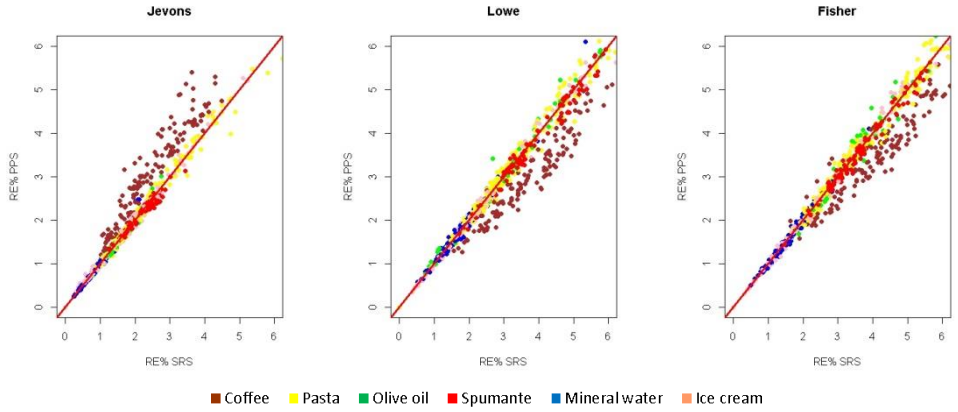
The comparison among behaviors in a sampling perspective of the three index aggregation formulas has been conducted with the aim of underline the differences among them under the different sampling strategies considered. The same sampling estimator (i.e. a plug-in estimator) has been considered for all the index aggregation formulas and the same overall sample size have been drawn under each sampling strategy. The CPI have been computed at market and consumption segment level for 13 months (from December 2013 to December 2014) for the province of Turin in the 88 markets related to the 6 consumption segments already listed above.

[6] See Neyman (1934) and Tschuprow (1923).

6.2.1 Stratified sampling

Under stratified sampling two different selection schemes in the strata (markets) have been implemented. The estimates, for all the aggregation formulas, are in both cases unbiased (RB approximately equal to 0). With respect to the variability, it possible to see a slight difference among the indices. Looking at Figure 2 the point cloud related to Jevons is usually below those related to Lowe and Fisher, that are almost at the same level.

Figure 2 – Relative percentage error (RE%) of estimates for Jevons, Lowe and Fisher indices in markets consumption under stratified sampling when simple random sampling (SRS) or probability proportional to size (PPS) selection methods are used. Turin, markets in coffee, pasta, olive oil, spumante, mineral water, ice cream. December 2013 – December 2014.



In the figure below, it is possible to notice also a different behavior of the estimators of the indices with respect to the consumption segments. In particular for coffee segment the estimator of Jevons index is more efficient when EAN are selected under SRS, whilst those for Lowe and Fisher are more efficient when selection of units is based on PPS. For the other segments there is no significant differences between the selection methods.

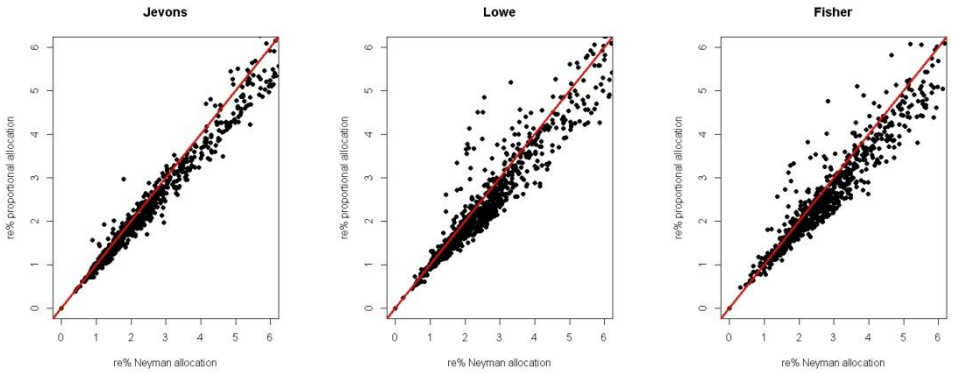
6.2.2 Stratified one stage sampling

In stratified sampling two different allocation methods for outlets in the strata – defined as chain and type (ipermarket and supermarket) - have been considered: optimal Neyman allocation based on the prices relative variability of the EAN observed in the strata during 2013 and proportional allocation based on the turnover of the strata in 2013. For each allocation method two selection schemes of outlets, SRS and PPS methods, have been considered. To keep in average around 7,400 EAN the outlets sample has been fixed at 14.

In all the scenarios, also in this case, the estimator of the indices are unbiased. However, when the outlets are selected through a PPS the estimates are more efficient, under both the allocation methods. Instead, between them the optimal Neyman allocation seems to be less efficient for the outlets. Then the proportional allocation based on turnover of previous year is preferable. This advantage is more remarkable when the interest parameter are weighted

price indices, such as Lowe and Fisher (see Figure 3).
In this case there are no significant difference among consumption segments and among the level of RE% of the estimator of the indices.

Figure 3 – Relative percentage error (RE%) of estimates for Jevons, Lowe and Fisher indices in market consumption under stratified one stage sampling when proportional allocation for the outlets is based on the turnover of the previous year is adopted. Comparison between simple random sampling (SRS) and probability proportional to size (PPS) selection of outlets. Turin, December 2013 – December 2014.



6.2.3 Stratified two stages sampling

In this case two stages of selection, outlets (PSU) and EAN (SSU) has been considered. The PSU have been stratified by chain and type, allocated through a proportional allocation and selected with PPS based on the turnover of previous year. The SSU in the selected outlets have been allocated proportionally with the Neyman allocation defined for the stratified sampling and selected with SRS and PPS. To keep in average around 7,400 EAN in this case the sample of outlets has been fixed at 30.

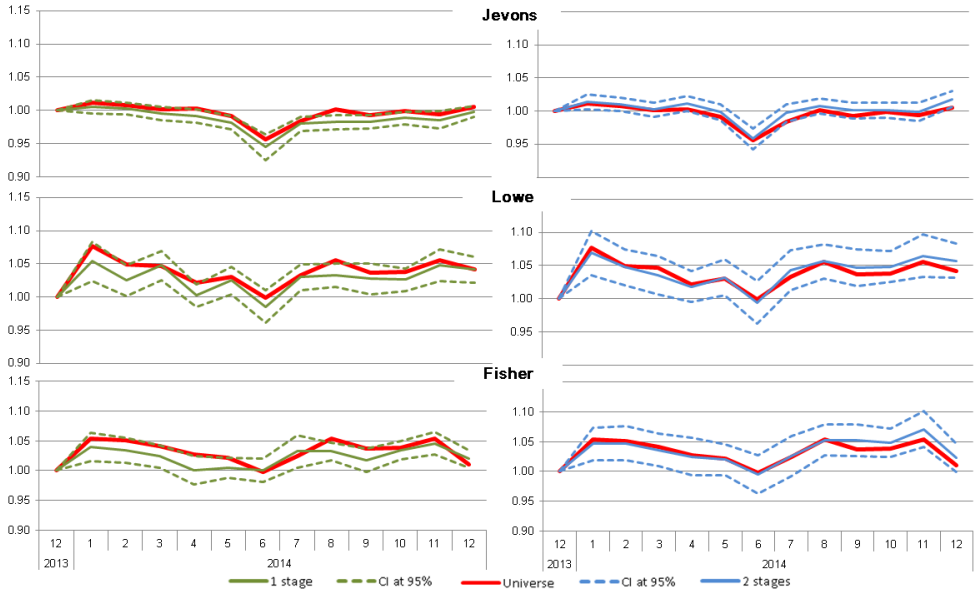
No significant differences arise using SRS or PPS, probably because the most of the variability is at outlet level. This result points the attention on the importance on the allocation and selection method to be used at PSU level and on the size of the PSU sample. In some markets this could bring an advantage in terms of RE% even if usually two stages sampling implies a higher design effect.

Looking at Figure 4 is possible to notice the difference among the indices estimated under the two different sampling strategies. The two sampling strategies compared are stratified one stage, proportional allocation of outlets and PPS selection versus stratified two stages, proportional allocation and PPS selection of outlets and furthermore ri-proportionated Neyman allocation of EAN in the selected outlets in which EAN have been selected with SRS. The results of estimates on a single sample for the indices in coffee segment in Turin obtained under the two sampling strategies has compared with the real value (computed on the universe of SD of coffee segment in Turin). All the estimates seems to catch properly the level and the trend of the related real index. The estimator of Jevons index has in both cases more narrow confidence intervals (CI) with respect to the other two, it means that its RE% is lower. In general the length of CIs are wider under two stages

sampling than under stratified one stage, even if the difference does not seem so large.

In terms of bias with respect to stratified sampling, in one stage and two stages sampling designs the RB increases slightly, but the estimators can be still considered unbiased.

Figure 4 – Jevons, Lowe and Fisher indices for coffee segment estimated on one sample, confidence interval (CI) of estimates at 95% and real value (computed on the universe of SD). Turin, December 2013 – December 2014.



7. Open issues and conclusions

This work is an overview of the first analyses conducted in ISTAT for testing the sampling strategies for computing CPI from SD. The results presented show the advantages in using probability sampling designs with respect to the non-probability ones. Among the probability sampling designs several sampling strategies have been tested for estimating Jevons, Lowe and Fisher indices. The estimator of the indices are always unbiased and in terms of sampling variability the Jevons index has a slightly advantage with respects to the other two that are more similar between each other.

Besides the stratified sampling that represents an ideal but hard to realize benchmark, due to the huge quantity of data to manage for the monthly estimates for all the Italian territory, the sampling efficiency of the estimators in stratified one stage and the stratified to stages are not so different. The choice between these two strategies is not simple and it is not only a decision from a sampling perspective, that is considering just the efficiency of estimates, but also the computational burden has to be taken into account. Moreover also impact of attrition and possible substitution of elementary items during the yearly data collection in these sampling design must be considered and evaluated.

For this reason the studies conducted will be extended to all consumption segments and other provinces. The experiments should also be expand by introducing alternative price index formulas (as monthly chained matched-item index), here not considered and evaluating the robustness of price index formula.

Moreover an open issue to deal with is the integration between scanner data and traditional data for CPI compilation, as SD do not cover all outlet typology. A further problem to be addressed concerns the combination of price indices obtained with different approaches.

References

Anderberg M. R. (1973). Cluster Analysis for Applications, Academic Press Inc., New York.

De Vitiis C., Casciano M. C., Guandalini A., Inglese F., Seri G., Terribili M.D., Tiero F. (2015). Sampling design issues in the first Italian experience on scanner data. Under review on Rivista di Statistica Ufficiale – ISTAT.

Feldmann B. (2015). Scanner-data-current-practice, http://www.istat.it/en/files/2015/09/5-WS-Scanner-data-Rome-1-2-Oct_Feldmann-Scanner-data-current-pratice.pdf.

Gábor E. and P. Vermeulen (2014). New evidence in elementary index bias.

ILO, IMF, OECD, Eurostat, United Nations, World Bank (2004). Consumer Price Index Manual: Theory and Practice, Geneva: ILO Publications.

Nygaard R. (2010). Chain drift in a monthly chained superlative price index. Workshop on scanner data, Geneva, 10 may 2010.

Norberg A. (2014). Sampling of scanner data products offers in the Swedish CPI. *Draft version 8 – Statistic Sweden*.

Neyman, J. (1934). On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection, *Journal of the Royal Statistical Society*, 97: 558-606.

Rosén B. (1997a). Asymptotic theory for order sampling. *Journal of Statistical Planning and Inference*. 62(2):135–158.

Rosén B. 1997b. On sampling with probability proportional to size. *Journal of Statistical Planning and Inference*. 62(2):159–191.

Sampford M.R. 1967. On sampling without replacement with unequal probabilities of selection. *Biometrika*. 54(3-4):499–513.

Tschuprow, A. A. (1923). On the Mathematical Expectation of the Moments of Frequency Distributions in the Case of Correlated Observations, *Metron*, 2:461-493.

SAS Institute (1988). SAS ISTAT User's Guide. Release 6.03 Edition. SAS Inst. Cary, NC.

Van der Grient, H. A. and J. de Haan (2010). The Use of Supermarket Scanner Data in the Dutch CPI, Paper presented at the Joint ECE/ILO Workshop on Scanner Data, Eurostat (2015).

Vermeulen B. C. and H. M. Herren (2006). Rents in Switzerland: sampling and quality adjustment. 11th Meeting - Ottawa Group - Neuchâtel 27-29 May.