Improvements in Italian CPI/HICP deriving from the use of scanner data

Alessandro Brunetti, Stefania Fatello, Federico Polidoro, Antonella Simone[1]

Abstract Scanner data are a crucial innovative "big data" source to estimate inflation, providing several advantages that derive from the detailed information available about sales and quantities at weekly frequency, GTIN by GTIN, outlet by outlet throughout the entire national territory. In paragraph 1 the paper makes the point about the state of play of the use of scanner data (introduced for grocery products in Italian CPI/HICP in 2018) and evaluates the benefits in terms of accuracy of inflation estimation coming from the improvement of the territorial coverage. The data of a sample of more than 1,700 hyper and supermarkets, for years 2017 and 2018, have been processed in order to calculate price indices differentiated according to outlet location (inside and outside municipal borders of the provincial chief towns) and price indices for the 80 provincial chief towns previously involved, to be compared with the indices calculated for the entire Italian territory (paragraph 2). The results in terms of improved accuracy are analyzed at national and geographical area level (paragraph 3). Perspectives of Italian scanner data project (brought forward by ISTAT) are finally sketched in paragraph 4.

Key words: Scanner data, inflation, accuracy, territorial coverage

1. The use of scanner data to estimate inflation in Italy: the state of play

Starting from January 2018 ISTAT introduced scanner data of grocery products (thus excluding fresh food) in the production process of estimation of inflation. This

[1] Alessandro Brunetti, ISTAT, albrunet@istat.it
Stefania Fatello, ISTAT, fatello@istat.it
Federico Polidoro, ISTAT, polidoro@istat.it
Antonella Simone, ISTAT, ansimone@istat.it

innovation concerns 79 aggregates of product belonging to 5 ECOICOP Divisions (01, 02, 05, 09, 12).

Since the end of 2013 a stable cooperation was established among ISTAT, Association of modern distribution, retail trade chains (RTCs) and Nielsen. Scanner data of grocery products have been collected by ISTAT through Nielsen for years 2014, 2015 and 2016 for about 1400 outlets of the main six RTCs for 37 provinces.

Afterwards, in view of the inclusion of scanner data into price indices calculations, a probabilistic design has been implemented for the selection of the sample of outlets, for which Nielsen provided ISTAT from December 2016. Scanner data for 1.781 outlets (510 hypermarkets and 1.271 supermarkets) of the main 16 RTCs covering the entire national territory are monthly collected by ISTAT on a weekly basis at item code level. Outlets have been stratified according to provinces (107), chains (16) and outlet-types (hypermarket, supermarket) for a total of 867 strata, taking into account only the strata with at least one outlet. Probabilities of selection were assigned to each outlet based on the corresponding turnover value. Table 1 shows the number of the strata, the number of the outlets and the coverage in terms of turnover, at regional and national levels for years 2018. The coverage for the year 2017 is slightly lower because a small RTC has been excluded from the analysis.

Concerning the selection of the sample of items, a static approach that mimics traditional price collection method has been adopted[1]. Specifically, a cut off sample of barcodes (GTINs) has been selected within each outlet/aggregate of products (covering 40% of turnover but selecting no more than the first 30 GTINs in terms of turnover). The products selected in December are kept fixed during the following year. A "thank" of potentially replacing outlets (258) and GTINs (until a coverage of 60% of turnover within each outlet/aggregate) has been detected in order to better manage the possible replacements during 2018.

About 1.370.000 price quotes are collected each week to estimate inflation. For each GTIN, prices are calculated taking into account turnover and quantities (weekly price=weekly turnover/weekly quantities). Monthly prices are calculated with arithmetic mean of weekly prices weighted with quantities.

Scanner data indices of aggregate of products are calculated at outlet level as unweighted Jevons index (geometric mean) of GTINs elementary indices. Provincial scanner data indices of aggregate of products are calculated with weighted arithmetic mean of outlet indices using sampling weights. Finally, for each aggregate of products, scanner data indices and indices referred to other channels of retail trade distribution are aggregated with weighted arithmetic mean using expenditure weights.

To calculate weights for the integration of regional indices of modern and traditional distribution at regional level, data are broken down using regional estimates from National Account (at ECOICOP sub-class level), regional

[1] The static approach to sampling is discussed in EUROSTAT [2017].

expenditure by type of distribution from Ministry of Economic Development and qualitative information on the shopping habits of consumers coming from HBS.

Table 1: Sample size: number of strata, number of outlets and coverage in terms of turnover -Year 2018

Region	North								Centre				South								ITALY
	Piemonte	Valle d'aosta	Liguria	Lombardia	Trentino Alto Adige	Veneto	Friuli Venezia Giulia	Emilia Romagna	Toscana	Umbria	Marche	Lazio	Abruzzo	Molise	Campania	Puglia	Basilicata	Calabria	Sicilia	Sardegna	
Num. of strata	78	4	31	148	12	85	45	84	65	16	43	38	34	11	35	34	6	25	44	29	**867**
Num. of outlets	152	6	62	286	35	161	77	163	142	32	83	106	58	18	88	85	10	51	111	55	**1781**
% market shares (hyper + super)	95,9	76,0	99,8	94,8	99,1	87,0	94,3	98,5	99,9	92,2	97,4	93,2	92,9	96,6	77,5	91,3	67,9	87,2	86,4	98,0	**93,7**

2. The improvements of the territorial coverage of indices and its effect on the accuracy of inflation estimates

Scanner data allow calculating the indexes for the entire national territory using data from outlets of all Italian provinces and located both in the municipal area and outside. With the aim of evaluating the benefits in terms of accuracy of inflation estimation coming from the improvement of the coverage in territorial terms, price indices are calculated by taking into account the outlet location (i.e. inside municipal area of the provincial chief towns: HICP SD MA) and by distinguishing the 80 provincial chief towns previously involved in the consumer price survey from the rest of the provinces whose data now are made available by scanner data (HICP SD 80P).

Table 2 shows the number of outlets used for the calculation of the different indices at national and macro-regional level for years 2018.

Table 2: Number of outlets at national and macro-regional level – Year 2018

Macroregion	All outlets		Outlets in 80 provinces		Outlets in municipal area	
	N° outlets	*% outlets*	*N° outlets*	*% outlets*	*N° outlets*	*% outlets*
North	942	52,9	857	58,8	304	50,6
Centre	363	20,4	286	19,6	133	22,1
South	476	26,7	315	21,6	164	27,3
Italy	**1781**	**100,0**	**1458**	**100,0**	**601**	**100,0**

In order to point out the methodology used for this analysis, it is necessary to start with a short description of the procedure for the aggregation of indices[1].

[1] For a detailed description of the procedures adopted by Istat for the calculation of the consumer price indices, see ISTAT [2012].

Let us introduce the following symbols[1]:

- *n* denotes the *n*-th product aggregate[2] (*n*=1,…,*N*);
- *g* denotes the g-th region (g=1,…,G=20);
- j denotes the j-th province (j=1,…,J(r));
- h denotes the *h*-th outlet (*h*=1,…,*H*(*j*)).

Let:

$P_{ng,j} = \sum_{h \in j} w_{ngj,h} \cdot P_{ngj,h}$ be the provincial index of the product aggregate *n*;

$P_{n,g} = \sum_{j \in g} w_{ng,j} \cdot P_{ng,j}$ be the regional index of the product aggregate *n*;

$P_n = \sum_{g} w_{n,g} \cdot P_{n,g}$ be the national index of the product aggregate *n*;

$P = \sum_{n} w_n \cdot P_n$ be the general index at the national level.

where:

$$w_{ngj,h} = \frac{e_{ngj,h}}{\sum_{h \in j} e_{ngj,h}} \; ; \; w_{ng,j} = \frac{e_{ng,j}}{\sum_{j \in g} e_{ng,j}} \; ; \; w_{n,g} = \frac{e_{n,g}}{\sum_{r} e_{n,g}} \; ; \; w_n = \frac{e_n}{\sum_n e_n}$$

and $e_{ngj,h}$ is the expenditure estimate for product aggregate *n* in the outlet *h* of province *j* in region *g*[3].

For the scope of the present analysis, the general index *P* is to be compared with the index $\hat{P}$ which is calculated using:

- Transaction prices of the outlets (*h*') situated inside municipal borders of the provincial chief towns (HICP SD MA);
- Transaction prices of the outlets of the 80 provincial chief towns previously involved in the consumer price survey (HICP SD 80P).

Concerning the first case, the general index *P* can be usefully expressed as the weighted arithmetic mean of provincial product aggregate indices:

$$P = \sum_{nj} \frac{e_{nj}}{\sum_n e_n} \cdot P_{nj} = \sum_{nj} \pi_{nj} \cdot P_{nj}$$

[1] The notation used is adapted from that one suggested in Biggeri L., Giommi A. [1987].

[2] Product aggregates indices are indices calculated at the lower level of aggregation of product-offers.

[3] $e_{ngj,h}$ incorporates the sampling coefficient attached to the outlet *h*.

Accordingly, the impact of the improvement of territorial coverage is calculated as follows:

$$P - \hat{P} = \sum_{nj} \pi_{nj} \cdot \left(P_{nj} - \hat{P}_{nj}\right)$$

where:

$$\hat{P}_{n,j} = \sum_{h' \in j} \frac{e_{nj,h'}}{\sum_{h' \in j} e_{nj,h'}} \cdot P_{nj,h'}$$

The impact can also be decomposed as suggested in Biggeri L., Brunetti A. e Laureti T. [2008]. By indicating with *k* the product $N \cdot J$, with δ_{nj} the difference between sub-indices $\left(P_{nj} - \hat{P}_{nj}\right)$, with $s_{\pi_{nj}}$ and $s_{\delta_{nj}}$ the standard deviations of π_{nj} and δ_{nj}, with $R_{\pi_{nj},\delta_{nj}}$ the linear correlation coefficient between π_{nj} and δ_{nj}, with $\bar{\delta}_{nj}$ the arithmetic mean of δ_{nj}, we have:

$$P - \hat{P} = k \cdot s_{\pi_{nj}} \cdot s_{\delta_{nj}} \cdot R_{\pi_{nj},\delta_{nj}} + \bar{\delta}_{nj}$$

As for the second case, it is convenient to express the general index *P* as the weighted arithmetic mean of regional product aggregate indices:

$$P = \sum_{ng} \frac{e_{ng}}{\sum_{n} e_{n}} \cdot P_{ng} = \sum_{ng} \pi_{ng} \cdot P_{ng}$$

Consequently, it is possible to write:

$$P - \hat{P} = \sum_{ng} \pi_{ng} \cdot (P_{ng} - \hat{P}_{ng})$$

where:

$$\hat{P}_{n,g} = \sum_{j' \in g} \frac{e_{ng,j'}}{\sum_{j' \in g} e_{ng,j'}} \cdot P_{ng,j'}$$

and, with similar notation:

$$P - \hat{P} = k \cdot s_{\pi_{ng}} \cdot s_{\delta_{ng}} \cdot R_{\pi_{ng},\delta_{ng}} + \bar{\delta}_{ng}$$

3. Results

By comparing indices calculated on the whole national territory and the corresponding indicators compiled taking into account only the outlets in the

municipal area of the provinces (figure 1) moderate differences emerge in the first months of 2017 and 2018 and in the middle of the first year. However, when the geographical breakdown is considered, the divergences tend to be relatively larger and persistent, especially in the South of Italy (islands included) (table 3). For example, the difference of the indices, calculated on March 2018, shows that the HICP SD MA index of the South is about 0.3 percentage points below the corresponding HICP SD index, while it is 0.24 in the North and 0.15 in the Centre). The main factors explaining this divergence seem to be the relatively higher value of the standard deviation of the differences of sub-indices and the relatively high value of the linear correlation coefficient $R_{\pi_{nj},\delta_{nj}}$ (higher differences in the level of sub-indices tend to have higher weights).

Figure 1: Comparison between HICP SD and HICP SD MA - Years 2017-2018

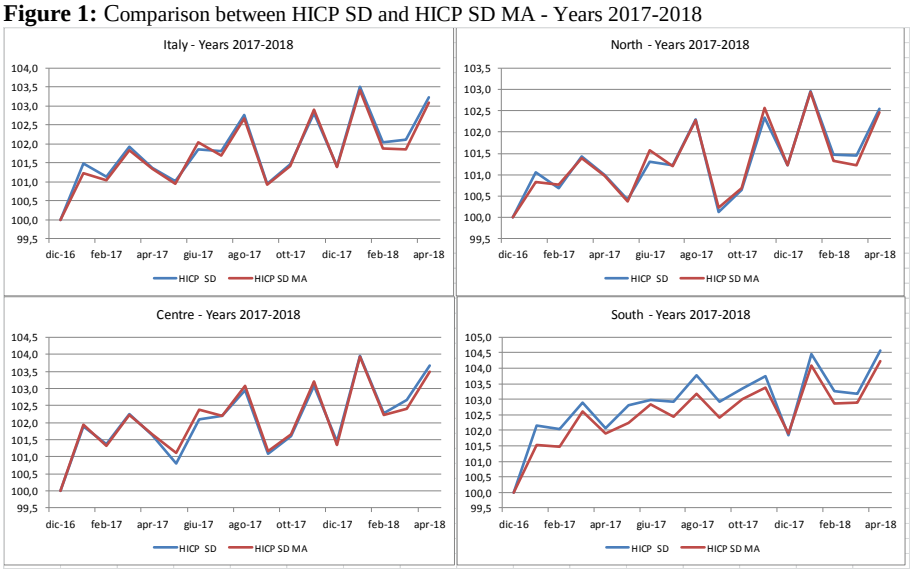
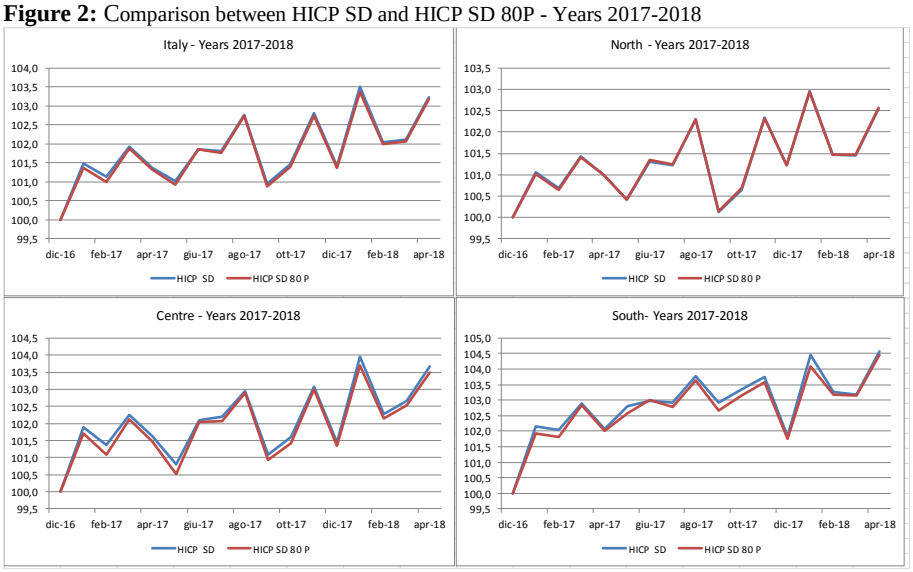


Table 3: Decomposition of the difference between HICP SD and HICP SD MA. March 2018.

	Italy	**North**	**Centre**	**South**
K	8.368	3.713	1.738	2.917
$s_{\pi_{nj}}$	0,0002	0,0004	0,0012	0,0005
$s_{\delta_{nj}}$	3,3349	2,9171	2,8525	4,0235
$R_{\pi_{nj},\delta_{nj}}$	0,0139	-0,0088	0,0082	0,0460
$\bar{\delta}_{nj}$	0,1522	0,2732	0,1046	0,0267
HICP SD	100,6830	100,2363	101,2078	101,2760
HICP SD MA	100,4517	100,0011	101,0555	100,9741
HICP SD - HICP SD MA	0,2314	0,2352	0,1523	0,3019

Regarding the comparison between HICP SD and HICP SD 80P, major divergences tend to be concentrated in the South of Italy (figure 2) as well. This result reflects the fact that the share of provinces participating to the survey before the introduction of scanner data, in this part of the country, was relatively low as compare to the North and the Centre.

Figure 2: Comparison between HICP SD and HICP SD 80P - Years 2017-2018



Generally speaking, for what concerns modern retail trade distribution and comparing indices compiled from scanner data source in a time span of 16 months, the improvement in terms of accuracy coming from the coverage of the entire provincial territories are limited to three months at national level with a maximum of three decimal points of differences between grocery index calculated inside the municipal borders and that one compiled with the outlets of the entire municipal areas. For the South the differences between the two indices are more frequent and wider along the time span considered. In all the cases, the level of the indices referred to the entire provincial areas are slightly higher than those ones compiled just within the municipal borders.

If we consider the grocery index compiled taking into consideration the outlets of the 80 provinces that were involved in 2017 in the territorial data collection, the comparison with the grocery index calculated from the data of all the 107 Italian provinces, shows just some local differences (in the South and in the Centre of Italy) but without important consequences in the estimation of the grocery index at national level.

4. Next steps in the development of the ISTAT project on scanner data

Scanner data project (brought forward by ISTAT) is still on the way and it is possible to sketch some further steps.

The first one is the adoption of the so called dynamic approach (abandoning the static one) to the selection of the elementary items (GTINs) to be considered for the compilation of indices. It should be implemented starting from January 2019 and it means the use of all the elementary price quotes of all the GTINs sold monthly in the sample of outlets selected. Next months of 2018 will be dedicated to solve the main crucial issues in sight of this next step, starting from that regarding relaunches and IT environment and procedures. Dynamic approach is the choice towards other National Statistical issues at European level are converging and could represent a further improvement in the accuracy of Italian CPI/HICP.

The following steps concern the extension of the use of scanner data to other retail trade channels (discount, outlets with surface between 100 and 400 square meters) and other goods such as fresh products with variable weight and no grocery products.

Toward these aims the role of the collaboration with the modern distribution and the representative association (Association of Modern Distribution, ADM) keeps its crucial importance.

References

Biggeri L., Brunetti A. e Laureti T. (2008), "*The interpretation of the divergences between CPIs at territorial level: Evidence from Italy*", Invited paper presented at the Joint UNECE/ILO meeting on Consumer Price Indices, May 8-9, Geneva., Proceedings UNECE, Geneva.

Biggeri L., Giommi A. [1987], "*On the accuracy and precision of the consumer price indices. Methods and applications to evaluate the influence of the sampling of households*", Proceedings of the 46th Session of the ISI (International Statistical Institute). Book 2, Tokyo, 134-157.

EUROSTAT [2017], Pratical Guide for Processing Supermarket Scanner Data, available on EUROSTAT website (https://circabc.europa.eu).

ISTAT [2012], Indici dei prezzi al consumo: aspetti generali e metodologia di rilevazione, available on ISTAT website (https://www.istat.it/it/files/2013/04/Indice-dei-prezzi-al-consumo.pdf).