doi.org/10.26398/asaproc.0015

# Compiling sub-national spatial price indices using scanner data and traditional data sources

Alessandro Brunetti[a], Anna Di Franco[a], Barbara Dramis[a], Tiziana Laureti[b], Federico Polidoro[c], Agostina Zanoli[a]

[a] National Institute of Statistics (ISTAT), Rome, Italy

[b] Department of Economics, Engineering, Society and Business Organization, University of Tuscia, Viterbo, Italy

[c] World Bank, Rome, Italy

## 1. Introduction

Sub-national spatial price indices (SN-SPIs) or regional purchasing power parities (RPPPs) measure the differences in price levels across regions within a country at a given point of time. They express how many currency units a given quantity of goods or services costs in different areas. SN-SPIs are essential for comparing real income, standards of living and consumer expenditure patterns. In countries characterized by large territorial differences in consumer preferences as well as in household characteristics, the calculation of SN-SPIs acquires considerable importance.

The Italian National Institute of Statistics (Istat) is one of the few national statistical offices that carried out official experimental sub-national spatial price indices computations. The analyses, based on price data coming from consumer price indices (CPIs) and *ad hoc* surveys, focused on the comparison of consumer prices across the 20 Italian regions (Istat, 2010). The availability of scanner data, used since 2018 in the CPIs production process, provided new impulse to the research (Laureti, Ferrante, Dramis, 2017; Laureti, Polidoro, 2022).

Moreover, the increasing request for sub-national indicators both at the national and European levels has pushed Istat to develop a project to estimate SPI at the regional level on a regular basis. Due to the wide information amount necessary, a multi-source approach has been used to select for each product the best available information. Data sources used in the analyses are scanner data, traditional CPI data and *ad hoc* surveys.

The methods adopted to properly consider the information provided by each source at the basic heading (BH) level are weighted and unweighted regional product dummy (RPD) methods. BH-SPIs are obtained and aggregated at higher levels of classification by Gini-Elteto-Koves-Szulc (GEKS) technique based on Fisher indexes (in accordance with international practices).

Experimental results (Istat, 2023), referred to the first three expenditure divisions of the European classification of individual consumption according to purpose (ECOICOP) (food, including tobacco, beverages, clothing and footwear), are presented and discussed in the rest of this paper: data sources are illustrated in Section 2; the methodological aspects are addressed in Sections 3; Section 4 provides a picture of the results. Finally, in Section 5 some concluding remarks are sketched.

## 2. Data sources

SN-SPIs computation needs many detailed data at a local level that satisfies two main requirements: comparability and representativeness. The first one entails that items being compared should be identical or at least equivalent, to ensure that differences in prices are not influenced by differences in quality. As for the second requirement, products that enter into the calculation of the SN-SPIs should be those that are frequently purchased and widely available at the regional level.

To this aim, a multi-source approach has been used, coherently with the approach adopted for the compilation of the Italian CPIs.

The first source consists of scanner data coming from a sample of outlets of twenty-one retail trade chains [1] covering 60% of the total turnover for the hard-discount channel and over 90% for each other distribution shop types at national level. Scanner data are referred to processed food, including non-alcoholic and alcoholic beverages.

The outlets sample[2], widespread in all the municipalities within the 107 Italian provinces, is the same sample extracted for the inflation estimates. The sampling plan is probabilistic, stratified by province and store type, with the probability of selection of the outlets proportional to their turnover.

The detailed information contained in these data (turnover and quantities sold for each global trade item number-GTIN in each outlet) can be used to compile weighted SPIs at a highly detailed territorial level and allows to fulfil both the requirements for spatial comparison of prices: comparability and representativeness.

Data for turnover and quantities sold in the first three weeks of each month (weekly data) are used to calculate the annual unit value price for each GTIN, at the outlet level, as the ratio between annual turnover and annual quantities sold (in the outlet). These annual unit values are then aggregated across outlets of the same province to obtain the average price for GTIN at the provincial level. For the aggregation, a weighted average is used, with outlet and outlet-type weights. Only GTINs sold in at least two regions are considered: 140,197 for Food and non-alcoholic beverages and 25,121 for Alcoholic beverages.

In addition, CPI data coming from the traditional price collection has been used for fresh products such as fish, fruits and vegetables for which CPI definition is enough accurate to guarantee comparability. Moreover, products included in CPI basket are widespread in almost all 80 provinces. The data have been cleaned for outliers and only the prices of the months in which these products are "in season" have been included. The annual provincial average prices are the unweighted mean by product at the provincial level of all the prices since the weights relating to the quantities sold are not available. Therefore, the annual price quotations included in the analysis were 87,784. The regional average price per product was then calculated as a weighted mean of average provincial prices, using population as weights, in line with what occurs in inflation calculating.

Thirdly, *ad hoc* surveys have been carried out by product category for which the use of other sources is prevented by comparability issues. A specific basket of comparable products is defined as subset of the one used for International Comparison Program (ICP) at European level[3].

The twenty-one municipalities involved in the survey are the 19 regional and the two autonomous provincial capitals (with the exception of L'Aquila, replaced by Pescara). The outlet sampling design in each municipality is the one defined for the CPI survey. The *ad hoc* price surveys are based on a cyclical system of surveys coinciding with those of international comparison program for purchasing power parity (PPP) computations. Overall, there are six surveys, each on a specific basket of goods and services, which take place over a three-year cycle.

The basket of food products included in the index calculation is made up of 39 items, while 79 items are considered for clothing and footwear. The number of quotations entered in the indices calculation for food products was 8,945, while for clothing and footwear 7,885. The trimming of invalid quotations had an impact above all on clothing and footwear (16% vs. 5.5% in food division).

In general, *ad hoc* surveys involve a considerable effort by municipalities in terms of organization, time and human resources. The number of collected and validated quotations for the municipalities of Napoli, Pescara and Potenza, in clothing and footwear, is small for several products. Therefore, the results for the Campania, Abruzzo and Basilicata regions have to be read taking into consideration this limit and require significant further study. Despite these difficulties, the BHs with the greatest weight are overall well represented.

Given that the prices collected through *ad hoc* surveys refer to the two months in which each

[1] Scanner data are provided by Nielsen to Istat, thanks to an agreement with the retail trade chains.

[2] The sample of outlets includes about 4,000 shops of different type (supermarkets, hypermarkets, minimarkets and discounts).

[3] Eurostat is a partner in the ICP. EUROSTAT-OECD Methodological manual on purchasing power parities (PPPs)

survey was conducted, the monthly data were made representative of the year, by considering the inflation recorded in the Municipality in the other months of the year. Temporal adjustment factors (TAFs) were calculated based on harmonized index of consumer price (HICP) data.

Finally, for tobacco products, the administrative source has been used (Customs and Monopolies Agency).

The first three ECOICOP expenditure divisions represent about 32% of household expenditure based on the 2021 HICP basket. The first division (20.5%) is affected by the all three data sources, even if with different importance: 53% scanner data, 27% *ad hoc* survey and 20% CPI traditional price collection. The second division represents about 3.8% of Italian consumption and the drinks are entirely covered by scanner data. The third division data (approximately 7.5% of the basket) comes entirely from *ad hoc* surveys. The tobacco products whose prices have no variability across the Italian territory (but have different weights in the regions) have been included in the calculation with parity equal to 100.

## 3. Methodological approach

PPP indicators are calculated using recognized methodological tools and used in numerous international studies and empirical experiences conducted in various countries. The ICP published a guideline in 2021 (ICP, 2021), to which Italy contributed, for the calculation of sub-national PPP at different territorial or geographical area (frequently regions within a country). In accordance with these guidelines, the methods used are:

- RPD model at BH level
- GEKS (Gini - Èltetö-Köves-Szulc) method for aggregations above BH level.

### *3.1 Regional Product Dummy*

The RPD method is the regional version of the country-product-dummy (CPD) method used in international comparisons. The idea behind this method (Rao and Hajargasht, 2016; World Bank, 2013) is that the price $p_{nr}$ of an item n (n=1…N) belonging to the BH in a region r (r=1...R) is function of a specific regional factor $RPPP_r$ (parity or general price level of the area considered with respect to other areas), of the average price of the *n*-th commodity/item $P_n$ and of a random error $u_{nr}$:

$$p_{nr} = P_n * RPPP_r * u_{nr} \quad (1)$$

Considering the logarithms, equation (1) can be written as:

$$ln\, p_{nr} = \sum_{r=1}^{R} a_r D_r + \sum_{n=1}^{N} b_n D_n^* + v_{nr} \quad (2)$$

where: $D_r$ is the dummy variable that takes value equal to 1 if the price quotation is from region *r* and *0* otherwise, $D_n^*$ is the dummy variable for commodity n which takes value equal to 1 when item considered is *n* and *0* otherwise, $a_r$e $b_n$ are the differences in the effects associated with the regions and the commodity type, respectively, $v_{nr}$ are random error normally distributed with a zero mean and variance $\sigma^2$.

Parameters of this model can be estimated using ordinary least squares, imposing a restriction that a coefficient corresponding to a specific area or region is set equal to zero ($a_i$=0) or equivalently $PPP_i$ =1 thus considering it as reference area to which the coefficient estimates are referred.

The SN-PPP between an area r and the reference area is $\widehat{RPPP}_r$=exp($\widehat{a_r}$). The parities thus estimated satisfy the property of transitivity and invariance of the basis.

Having weights in terms of value or quantity for each commodity, the weighted RPD model can be written as:

$$\sqrt{w_{nr}}\, ln\, p_{nr} = \sum_{r=1}^{R} a_r \sqrt{w_{nr}} D_r + \sum_{n=1}^{N} b_n \sqrt{w_{nr}} D_n^* + \sqrt{w_{nr}} v_{nr} \quad (3)$$

where $w_{nr}$ are the weights in terms of value or quantity share that reflect the economic importance of the different commodities consumed in the area.

The model was estimated for each BH, making the most of the information available depending on the data source used. Within each BH, only one of the three considered sources was used.

In the case of scanner data, they provided information on turnover and quantities sold for each GTIN, in all 107 Italian provinces. The revenue-weighted model was then used to obtain the estimates. Furthermore, a two-step procedure was adopted (Laureti, Polidoro, 2022) for each BH:

- Step 1: In each region, a model (equation 3) was estimated to obtain provincial PPP. These indicators have been used to 'deflate' the initial prices and the turnover of each GTIN within each region.
- Step 2: Model (equation 3) was estimated to obtain RPPPs, using the 'deflated' prices and turnover obtained in step1.

In the case of the data from the local survey of consumer prices and the *ad hoc* surveys, weighted model cannot be used (the quantities sold are not available) and the unweighted RPD (equation 2) has been used to estimate RPPPs at the BH level.

### *3.2 GEKS (Gini - Èltetö-Köves-Szulc)*

Parities at the upper level of the BH are calculated using the GEKS method aggregating RPPP estimated at the BH level. The weights, based on household expenditure, are the same used in HICP. The index compiled by the GEKS method satisfies the properties of transitivity and invariance of the basis. Furthermore, it is as close as possible to the corresponding binary indices (Diewert, 2013). Aggregate RPPPs are obtained as follows:

$$RPPP_{jk}^{GEKS} = \prod_{l=1}^{R} (F_{jl} \cdot F_{lk})^{1/R}$$

The parity for region *k* with reference to region *j* chosen as the base is given by the geometric mean of the Fisher indices of all direct comparisons between region *j* and region *k*, and indirect across all possible links between the *R* regions ($l, k, j \in R$). Fisher type ($F_{jl}$) indices are obtained as the geometric mean of the corresponding Laspeyres type and Paasche type indices, calculated based on the parities for BH weighted respectively by the expenses of the base region (Laspeyres) and of the partner region (Paasche).

These first results lack data on some BHs, because of difficulties in data collection. The weights of these BHs were distributed among the BHs of the same consumption segment or at a higher level if it was the only BH of the segment.

To express the parities referring to the national average, each RPPP is divided by the geometric mean of the price level indices of the participating regions, and conventionally, multiplied by 100.

## 4. Results

The analyses show significant differences in consumer price levels between the Italian regions, in 2021. In general, considering the aggregation of the first three ECOICOP expenditure divisions, the prices recorded in the northern regions are higher than those of the central ones, except for Tuscany, and of the south and islands, except for Sardinia (Table 1).

The most expensive regions are Alto Adige with prices 5.3 % higher than the national average (Italy=100), Lombardy (+ 5%) and Liguria (+ 4.7%). The least expensive regions, compared to the national average, are Campania (prices 9.5% lower than the average), Abruzzo (-6.2 %) and Basilicata (- 5.2%). Thus, the difference in price levels between Campania (least expensive) and Alto Adige (most expensive) is almost 15 percentage points. The estimated RPPPs for Food Products show that the regions with price levels above the national average are both in Northern Italy and in the Centre, while those with price levels systematically lower than the national average are in the South (except for, also in this case, Sardinia). Lombardy and Alto Adige have price levels

above the Italian average by more than 6 %, while Campania and Basilicata have price levels below the average by 11% and 9 % respectively. In this case, the difference in price levels between Campania (least expensive) and Lombardy (most expensive) is almost 17 percentage points.

Soft drinks and Alcoholic beverages show the least heterogeneity in consumer price levels between regions and low differences in price levels compared to the Italian average. There is not territorial trend that leads to characterize groups of regions as in the other product categories. It should be noted, in fact, that in both cases a northern region is the least expensive (Veneto for soft drinks, Lombardy for alcoholic ones). This seems to highlight how for products with a longer supply chain, greater efficiency in logistics and infrastructure tends to change the traditional comparative geography of price levels between Italian regions.

**Table 1. Consumer spatial price indices by ECOICOP divisions, groups of product and regions, year 2021- Italy=100**

| Region | Food and non-alcoholic beverages | | | Alcoholic beverages and tobacco | | | Clothing and footwear | | | First three divisions |
|---|---|---|---|---|---|---|---|---|---|---|
| | Food | Non-alcoholic beverages | All items | Alcoholic beverages | Tobacco | All items | Clothing | Footware | All items | |
| **Piedmont** | 103.27 | 100.06 | 103.00 | 98.04 | 100.00 | 99.29 | 99.10 | 95.23 | 98.32 | **101.49** |
| **Valle d'Aosta** | 102.66 | 103.23 | 102.70 | 101.90 | 100.00 | 100.71 | 103.00 | 97.10 | 101.82 | **102.29** |
| **Lombardy** | 106.59 | 98.50 | 105.90 | 96.71 | 100.00 | 98.79 | 106.08 | 104.92 | 105.82 | **105.01** |
| **Trentino** | 100.96 | 99.63 | 100.85 | 99.18 | 100.00 | 99.70 | 101.31 | 117.89 | 104.40 | **101.57** |
| **Alto Adige** | 106.08 | 102.14 | 105.77 | 102.27 | 100.00 | 100.87 | 107.05 | 103.81 | 106.37 | **105.33** |
| **Veneto** | 103.86 | 95.20 | 103.16 | 98.03 | 100.00 | 99.26 | 101.42 | 117.93 | 104.50 | **103.00** |
| **Friuli-Venezia Giulia** | 103.91 | 97.64 | 103.40 | 100.06 | 100.00 | 100.04 | 99.24 | 113.76 | 101.94 | **102.67** |
| **Liguria** | 104.47 | 102.21 | 104.28 | 101.25 | 100.00 | 100.47 | 111.55 | 98.59 | 108.92 | **104.77** |
| **Emilia-Romagna** | 104.78 | 98.39 | 104.27 | 98.35 | 100.00 | 99.38 | 104.48 | 106.56 | 104.87 | **103.80** |
| **Tuscany** | 103.94 | 97.45 | 103.40 | 97.91 | 100.00 | 99.26 | 102.97 | 95.24 | 101.38 | **102.46** |
| **Umbria** | 100.65 | 98.28 | 100.45 | 99.13 | 100.00 | 99.71 | 96.37 | 84.14 | 93.81 | **98.85** |
| **Marche** | 103.07 | 99.96 | 102.81 | 100.70 | 100.00 | 100.27 | 88.64 | 88.35 | 88.55 | **98.95** |
| **Lazio** | 101.12 | 102.72 | 101.25 | 100.88 | 100.00 | 100.33 | 93.84 | 97.48 | 94.53 | **99.59** |
| **Abruzzo (*)** | 94.54 | 101.63 | 95.08 | 101.62 | 100.00 | 100.53 | 87.25 | 85.19 | 86.77 | **93.75** |
| **Molise** | 94.94 | 99.42 | 95.28 | 100.34 | 100.00 | 100.12 | 93.40 | 101.86 | 95.21 | **95.80** |
| **Campania (*)** | 88.99 | 98.22 | 89.69 | 99.79 | 100.00 | 99.95 | 84.28 | 100.41 | 87.61 | **90.47** |
| **Puglia** | 93.31 | 98.66 | 93.72 | 99.21 | 100.00 | 99.76 | 109.88 | 103.51 | 108.41 | **97.52** |
| **Basilicata (*)** | 91.03 | 99.99 | 91.71 | 100.49 | 100.00 | 100.18 | 101.44 | 99.28 | 100.93 | **94.72** |
| **Calabria** | 94.42 | 100.66 | 94.90 | 100.38 | 100.00 | 100.14 | 107.19 | 108.26 | 107.42 | **98.25** |
| **Sicily** | 97.54 | 101.14 | 97.85 | 101.55 | 100.00 | 100.50 | 103.54 | 98.55 | 102.47 | **99.16** |
| **Sardinia** | 102.67 | 105.41 | 102.91 | 102.45 | 100.00 | 100.77 | 103.52 | 90.62 | 100.66 | **102.15** |
| *Min* | *88.99* | *95.20* | *89.69* | *96.71* | *100.00* | *98.79* | *84.28* | *84.14* | *86.77* | *90.47* |
| *Max* | *106.59* | *105.41* | *105.90* | *102.45* | *100.00* | *100.87* | *111.55* | *117.93* | *108.92* | *105.33* |
| *Coefficient of Variation* | *5.09* | *2.28* | *4.70* | *1.56* | *0.00* | *0.56* | *7.12* | *9.10* | *6.58* | *3.87* |

(*) Due to the low number of observations for clothing and footwear division data of Basilicata, Campania and Abruzzo regions are not reliable.
Source: Istat

For Clothing and footwear, five southern regions are more expensive than the national average, almost in line with the northern ones. Instead, the central regions show price levels on average lower than the national average (except for Tuscany, which ranks slightly above). There are more than 20 percentage points of difference between the least expensive region (Abruzzo) and the most expensive one (Liguria). The most expensive regions are Puglia and Liguria, with values more than 8 points above the average, while Campania and Abruzzo are the least expensive. Likely, this more accentuated dispersion of the results, related to clothing and footwear, is due to the greater volatility present in the elementary data, attributable to the limited number of price observations recorded

(especially in Campania, Basilicata and Abruzzo) and deserving of further insights.

## 4. Concluding remarks

Our multi-source approach provides the opportunity to achieve the important objective of compiling sub-national spatial price indices for the first three expenditure divisions of the ECOICOP classification.

The preliminary results provide very interesting insights in the comparison of prices between the Italian regions (it is not obvious that are the southern ones those with the lower prices).

The project uses mainly data sources already acquired by Istat for the inflation estimates (scanner data and traditional price collection), thus optimizing the available resources, in terms of data, software used, personnel skills. Specifically, scanner data bring such an important value added, given the availability of information on quantities sold that allow a more precise consideration of the requirement of representativeness.

The data source with major issues is the *ad hoc* survey, which has a low territorial coverage and the least number of quotations used in the process. In several cases, it was not possible to collect a number of price quotations sufficient to guarantee robust results. Possible solution to this problem should be to extend the data collection to other municipalities and/or alternative sources of data.

The results will be updated and progressively extended to other divisions. The next release will cover ECOICOP division 04 (Housing, water, electricity, gas and other fuels) and 11 (Restaurant and accommodation services). SN-SPIs or RPPPs produced on a regular basis will improve our knowledge of the real economic territorial differentiation in Italy and in particular, the measurement of relative poverty by taking into account the price dimension, which is now not considered, given that a uniform threshold at the national level has been adopted.

## References

Diewert, W.E. (2013). Methods of aggregation above the basic heading level within regions, in *Measuring the Real Size of the World Economy,* World Bank, Washington D.C., pp. 121-167

ICP (2021). *A Guide to the Compilation of Subnational Purchasing Power Parities (PPPs).* World Bank.

Istat (2010). La differenza nel livello dei prezzi al consumo tra i capoluoghi delle regioni italiane. *Comunicato Stampa.* Roma, Italy.

Istat (2023). Indici spaziali dei prezzi al consumo – anno 2021, *Statistica Sperimentale*, 3 agosto 2023.

Laureti, T., Ferrante, C., Dramis, B. (2017). Using scanner and CPI data to estimate Italian sub-national PPPs, in *Proceedings of 49th Scientific Meeting of the Italian Statistical Society – SIS*, Palermo, Italy, pp. 581–588.

Laureti, T., Polidoro, F. (2022). Using scanner data for computing consumer spatial price indexes at regional level: An empirical application for grocery products in Italy. *Journal of Official Statistics,* **38**(1), pp. 23-56.

Rao, D.S.P, Hajargasht, G. (2016). Stochastic approach to computation of purchasing power parities in the International Comparison Program (ICP). *Journal of Econometrics,* **191**(2), pp. 414-425.

World Bank (2013). *Measuring the Real Size of the World Economy: The Framework, Methodology, and Results of the International Comparison Program (ICP).* World Bank.