SOME EMPIRICAL EVIDENCE FROM THE USE OF SCANNER DATA TO ESTIMATE PRICES OF FOOD PRODUCTS INCLUDED IN THE ABSOLUTE POVERTY BASKET[1]

Francesco Altarocca, Cristina Dormi, Stefania Fatello, Carlo Matta

Abstract. This work was developed within the activities of the Inter-Institutional Scientific Commission (IISC) established by ISTAT in order to review the methodology for estimating absolute poverty. The paper describes the main aspects of the new methodology used to estimate the prices of food products included in the absolute poverty basket, implemented thanks to the availability of scanner data, the new data source used by Istat for survey on consumer prices. The large availability of data allowed a selection of references (identified by barcodes/GTINs[2]) in order to calculate the annual minimum average prices of products necessary to satisfy the food needs of poor households. The paper shows some empirical evidences obtained from the analysis of scanner database. The causes of the high differences that emerged at a territorial level are also investigated.

1. Introduction

The Istat approach for measuring absolute poverty involves the identification of primary needs and the calculation of the cost of the basket of goods and services necessary to satisfy these needs (Istat, 2023c). Regarding the food component, the basket of products was identified through a nutritional model defined by Istat and sector experts. The monetary value of the absolute poverty basket is estimated on the basis of the prices acquired with the Istat survey on consumer prices (Istat, 2023b).

In 2009, in order to measure absolute poverty, Istat developed a methodology for calculating the household's "minimum acceptable expenditure"[3] (Istat, 2009). In 2023, Istat revised the methodology in order to take into account both changes in households' primary needs and the availability of new data sources to use. All the details on the revision of the main components of the absolute poverty basket are

[1] The work is the result of the overall contribution of the authors. However, Sections 1, 2 and 4 are attributable to Stefania Fatello, Section 3.1 is attributable to Cristina Dormi, Sections 3.2 and 3.3 are attributable to Carlo Matta and the appendix, related to the Information Technology architecture, is attributable to Francesco Altarocca.

[2] The Global Trade Item Number (GTIN) is a unique product identifier that is recognized internationally. If available, the GTIN code is indicated next to the barcode on product packaging.

[3] This is defined by the monetary value at current prices that constitutes the minimum amount of expenditure that the family must support in order not to find themselves in a condition of absolute poverty.

described in related papers published on the Special Issue of RIEDS "*New approaches for measuring poverty: studies and perspectives*".

This paper shows the results obtained using the new scanner data source relating to food products included in the absolute poverty basket. Section 2 describes the main aspects of the new methodology implemented using scanner data to estimate the annual minimum average price for the products considered. In Section 3 the results of the empirical analyses are shown. In particular, the main evidences relating two of the products included in the basket are highlighted, through the analysis of prices and quantities sold in the province of Rome and the comparison between the variability of minimum average prices at regional level. In the last section there are some concluding remarks. Finally, in the appendix some IT aspects related to the processing of big data are illustrated.

2. Main aspects of the new methodology

This section describes the main aspects of the new methodology implemented to valorize the poverty basket starting from the scanner data that Istat receives to estimate consumer price indices. To know in depth the details of the new methodology see Brunetti *et al.* (2024).

The methodological choices made were based on the analysis of the large database available. First of all, to estimate the minimum average price, a mapping was made which traces each food products included in the poverty basket to one or more products belonging in the consumer price basket. In total 61 food products, referring to packaged products, were associated with products that coming from monetary transaction data.

There are two important aspects of the new methodology implemented by looking at the empirical evidence on the data which shown in the following paragraphs: the first is represented by the selection of references, identified by barcodes/GTINs, to be included, while the second concerns the estimate of the minimum average prices for all food products considered at different levels of territorial aggregation.

2.1. The selection of references

For food products associated with scanner data collection it was necessary to identify the amount of GTINs sold for each product. Thus each product of poverty basket has been associated with the corresponding ECR[4] markets selected to

[4] ECR markets are the lowest level of the ECR classification (classification shared by industrial and distribution companies) and they have been linked to the aggregates of product of ECOICOP classification.

calculate consumer price indices and specifically connected to 171 markets of the ECR classification.

In this way all the GTINs relating to these markets are selected for subsequent analyses. After identifying the products to include in the poverty basket, it was essential to identify the references to consider. The choices made were supported by analyzes of the universe of GTINs available for each product in the year 2022. For example, regarding the product "Rice" all GTINs sold during the year in all available outlets were considered.

Scanner data provide prices and quantities of all the items actually purchased by household including all types of formats and packaging of the products. Therefore, for each selected market, it was possible to divide the products based on their characteristics and in particular by type of packaging and format.

Among these homogeneous subgroups of products, the ones most purchased by households were identified, based on annual data of quantities sold; a high turnover share for each group of products was included. Looking again at the product "Rice", the analysis show that the packages consisting of "1 pack, 1 kilogram" and "1 pack, 1/2 kilogram" represent 87% of the turnover in 2022. By selecting these two packages we were able to greatly simplify the processing without losing much information.

Obviously, to calculate the average prices of products with different packaging it was necessary to use a standardized unit of measurement for each different product.

Therefore, strata have been defined for each product considering: province, retail trade channel[5], combination of packaging type and most common format (example of stratum for rice is "*Rome, supermarket, 1 pack, 1 kilogram*") and GTINs were selected monthly within these strata. In fact, with scanner data it is possible to have data separated by retail trade channels: this enabled to evaluate the differences in sales prices between the different channels; the analysis of price distributions did not show evidence to exclude some retail trade channels. Furthermore, there is no information available on where the poorest households buy the products. So all retail trade channels were included in the calculations, not only Discounts.

In relation to each stratum thus defined, all GTINs sold in each month of 2022 were considered in order to study their price distribution. The aim was in fact to identify the GTINs that are presumably mostly purchased by the poorest households and therefore at relatively lower prices. Indeed, the new data source provides us with a large amount of data. This allowed us to make a selection of the GTINs sold taking into account the distribution of all prices paid for each product and selecting only the cheapest ones.

[5] For food product we have four retail trade channels: hypermarkets, supermarkets, discounts, small outlets with surface between 100 and 400 s.q.m.

Then the selection of GTINs, based on the price distribution in each stratum, occurs considering all GTINs belonging to the lower tail of the distribution (first quintile). The assumption underlying this choice is that having many GTINs available that satisfy the same food needs, the choice of the poorest households is oriented towards those with lower prices for the same quantity purchased.

Considering as an example the province of Rome, on which the analyses described below were carried out, the monthly average number of GTINs of all products included in the absolute poverty basket goes down from 10,064 available to 2,823 considered after the selection of the references. Similarly, if we consider the price quotations recorded for the same GTINs, the total number goes down from 465,114 available to 59,458 considered after the selection of the GTINs.

2.2. The process of estimating minimum average prices

As previously described, the methodology implemented involved a selection of GTINs based on the price distribution considering only the products belonging to the first quintile of the price distribution. This made it possible to estimate the minimum expenditure necessary to ensure the consumption of the recommended quantities of food products by households.

For each identified stratum, given by province, retail trade channels, packaging and format of the products, all the GTINs belonging to the first quintile of the price distribution were selected for each month of 2022. Then, for each stratum, the monthly average price was calculated as the weighted arithmetic average of the GTINs of the lower tail of the distribution, with weights proportional to the quantities sold. Formerly, for each stratum, the monthly minimum average price, calculated with the procedure just described, was reported to the specific unit of measurement.

The aggregation of the monthly minimum average prices then follows the steps described below for each food product:

1. the monthly minimum average price by province and retail trade channel is calculated as the weighted average of the monthly minimum average prices of the different combinations of packaging type and format. The weights are proportional to the quantities sold in terms of units of measurement;
2. the monthly minimum average price by province is obtained as the weighted arithmetic average of the monthly minimum prices of the retail trade channels. The weights are proportional to the importance of the channels in terms of provincial turnover[6];

[6] The provincial turnover is calculated as the sum of the turnover of all the GTINs sold in the outlets of a certain province.

3. the annual minimum average price by province is given by the simple arithmetic mean of the monthly minimum average prices;
4. the annual minimum average price by region is calculated as the weighted arithmetic average of the provincial average prices referred to point 3. The weights are proportional to the population resident in the provinces.

The greater availability of data at a more disaggregated territorial level allowed the use of data estimated at regional level without further territorial aggregations as in the past. The results of the processing showed important differences between the annual minimum average prices of each product in the different regions.

3. Analysis of the results: evidences on data

3.1. Analysis on prices and quantity sold

The objective of this section is to propose an exploratory analysis of the implications of the application of some criteria for the selection of references from the scanner database to be used for the calculation of minimum average prices, aimed at valorising the food basket of the poverty.

For the analysis, referring is made to two widely consumed products, rice and olive oil, focusing on the data relating to the province of Rome for the year 2022. Table 1 shows the composition by retail trade channels of the outlet sample for the province of Rome and for the entire national territory, relating to the year 2022 (Istat, 2023a).

Table 1 – *Number of outlets sampled by retail trade channels, Province of Rome and Italy - Year 2022.*

Retail trade channels	**Rome**	**Italy**
Hypermarkets	9	471
Supermarkets	57	1,453
Discounts	22	567
Small outlets (with surface between 100 and 400 s.m.)	26	1,000
Total sample	**114**	**3,491**

Source: Elaborations on scanner data

The selection of GTINs to include, computed at national level, is based first of all on the analysis of the most frequently purchased types of packaging. The aspects considered concern the characteristics of the package (single, multiple) and the quantity of product contained.

Table 2 shows the type of packaging that were selected for rice and olive oil: in particular, we consider the number of GTINs and the average price (respectively for kilogram and for liter). The last two columns of each table indicate, in decreasing order, the percentage of quantities sold and turnover relating to each packaging type. For each product there is only a subset of the different typologies of packaging type of all references sold. Most of them are not included in the table since the percentage of GTINs contained within is very small. The selected packaging types are highlighted in blue.

Table 2 – *Selection of packaging type for rice and olive oil, Italy - Year 2022.*

Rice						*Olive oil*					
Packaging type	N° GTINs	Average price per kilogram	% GTINs	% quantity	% turnover	Packaging type	N° GTINs	Average price per liter	% GTINs	% quantity	% turnover
1000GR 1 PACK	973	2,3	62,5	71,8	71,4	1000ML 1 PACK	805	4,8	34,4	76,5	71,1
500GR 1 PACK	324	2,1	20,8	17,1	15,6	750ML 1 PACK	765	5,6	32,7	17,0	18,7
2000GR 1 PACK	90	3,3	5,8	3,5	4,9	500ML 1 PACK	429	4,8	18,4	2,9	2,7
250GR 1 PACK	31	1,7	2,0	2,8	2,0	3000ML 1 PACK	111	13,9	4,7	1,0	2,7
800GR 1 PACK	13	2,3	0,8	1,5	1,5	250ML 1 PACK	91	3,2	3,9	0,7	0,5
850GR 1 PACK	17	2,3	1,1	1,1	1,1	5000ML 1 PACK	62	24,0	2,7	0,7	3,3
5000GR 1 PACK	62	5,5	4,0	0,9	2,1	450ML 1 PACK	1	3,9	0,0	0,4	0,3

Source: Elaborations on scanner data

Note that also the number of units in the package are specify because for some products included in the poverty basket like eggs, tuna, yogurt many GTINs are sold in multiple packages. For rice there are two main packaging type (single 1000 gr and single 500 gr), that have a coverage in terms of turnover of about 87% (about 89% in terms of quantity sold). Regarding olive oil the selected packaging types are two (single 1000 ml and single 750 ml), and the percentage of coverage is about 90% in terms of turnover and 93.5% in terms of quantity sold.

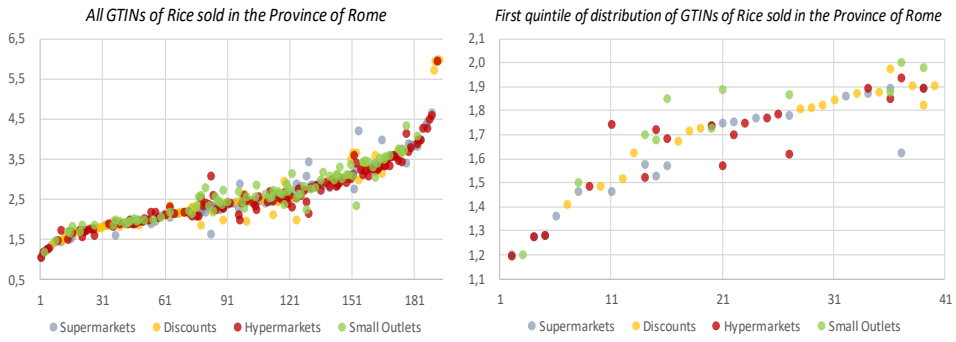
After identifying the types of packaging to include, some analysis were carried out on the distribution of prices and quantities sold. Figures 1 and 2 show, respectively, the distribution of the annual average price of rice and of olive oil relatively to all GTINs sold (on the left side) and to GTINs belonging to the first quintile of distribution (on the right side). The different colors indicate the types of retail channel where they are sold.

In general, it is customary think that Discounts could be the type of retail channel that has the lowest prices compared to other channels but this evidence does not emerge from the analysis on scanner data. In fact, from the figures it is clear that GTINs that occupy the lower positions in the ranking do not belong exclusively to the Discounts but to all type of retail channel. Furthermore, looking the distribution of the prices of GTINs selected into the first quintile (graphs on the right in Figure 1 and Figure 2), we can see that they belonging to all type of retail channel;

furthermore, GTINs prices appear very close to each other for olive oil while for rice they have an increasing trend.

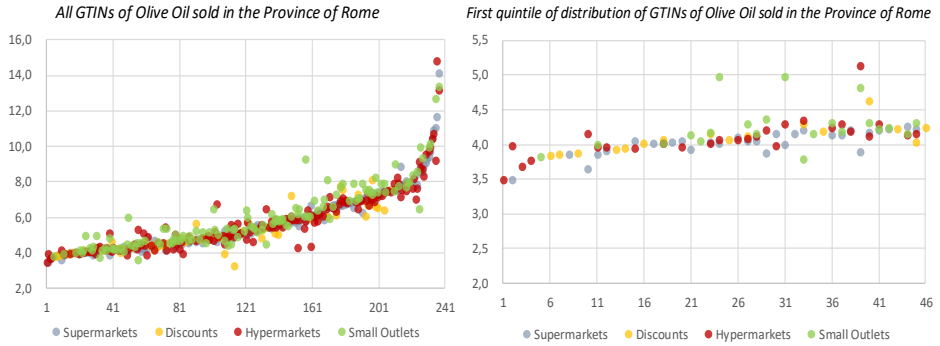
In particular, the first quintile of GTINs of rice are sold more by Hypermarkets (20 out of 40 GTINs) and Supermarkets (20/40) respect to the Small outlets (17/40) and Discounts (11/40). The same scenario looms for the first quintile of GTINs of olive oil: the chains where there are more products sold, referring to the lowest prices of distribution, are Supermarkets (32 out of 48 GTINs) and Hypermarkets (27/48) while there are fewer GTINs in the Small Outlets (22/48) and Discounts (16/48).

Figure 1 – *Distribution of the annual average prices of Rice by retail channels - Year 2022.*



Source: Elaborations on scanner data

Figure 2 – *Distribution of the annual average prices of Olive oil by retail channels - Year 2022.*

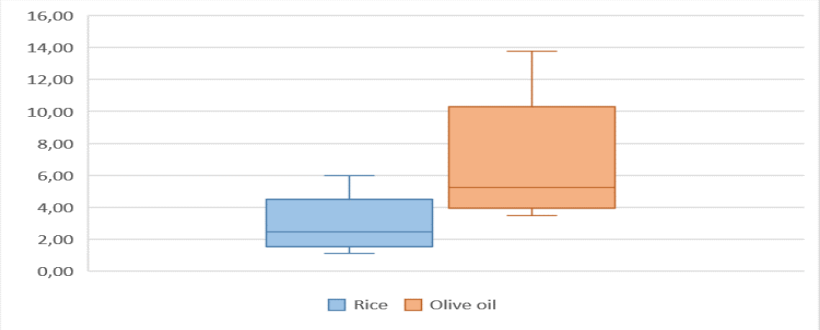


Source: Elaborations on scanner data

Figure 3 shows the boxplot of the two products considered for the analysis: this graph helps to describe the characteristics of a distribution and to identify the presence of asymmetry and outliers. The average annual prices of the 193 GTINs

selected for the rice, in ascending order point out a minimum value equal to 1.09 and a maximum equal to 5.99. The median is equal to 2.45 and it shows that the distribution of the minimum average prices is slightly asymmetrical to the left (positive asymmetry of the distribution). This means that the mean of the distribution of the minimum average prices (equal to 2.60) is higher than the median and the values are grouped in the low values part, with a long tail towards the higher values.

Figure 3 – *Boxplot of the annual average prices of Rice (single, 1000 gr) and Olive oil (single, 1000 ml) in the province of Rome - Year 2022.*

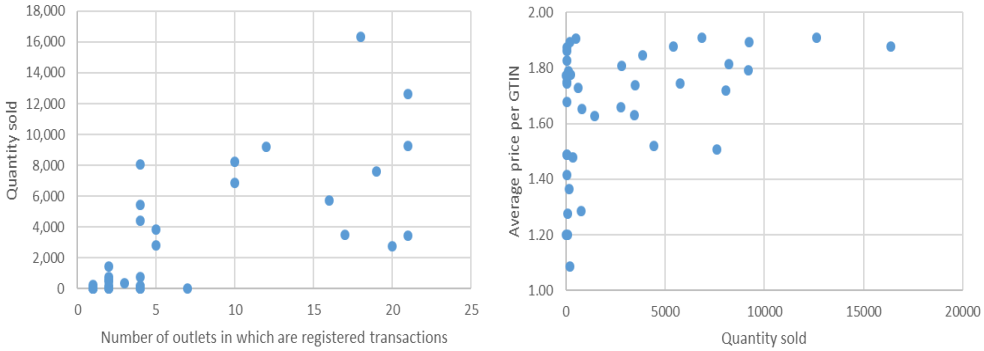


Source: Elaborations on scanner data

Regarding olive oil, the distribution of the average annual prices of the 237 considered GTINs have the same situation: the range goes from 3.49 to 13.77 and the median that is equal to 5.24 is below the mean (5.69). In fact, as the graph shows, the median is closer to the lower quartile Q1 than to the upper quartile Q3.

Figure 4 and 5 show the results of the analysis carried out considering the quantities sold for each GTINs and relating them to the number of outlet in which they are sold and the relative average prices. GTINs considered in these analyses belong to the first quintile of the distribution prices.

Figure 4 – *Quantity sold, number of outlets and average price per GTIN for Rice (single, 1000 gr) in the Province of Rome, first quintile of distribution - Year 2022.*

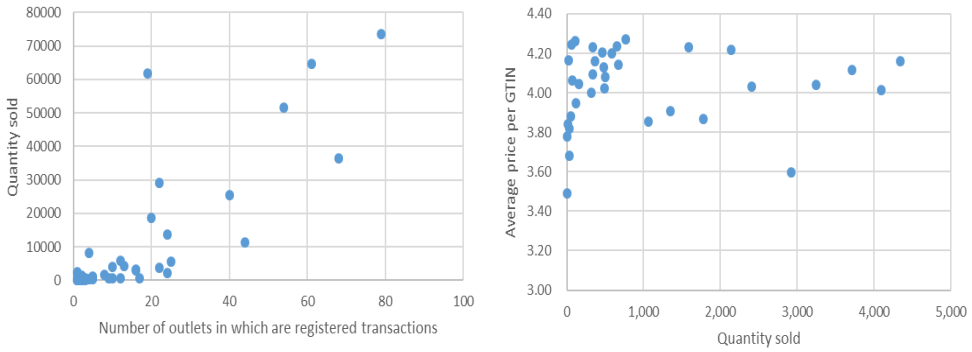


Source: Elaborations on scanner data

Considering GTINs of rice, the variability of the quantities (number of packages) sold during the year is very large: the minimum value is equal to 1, the maximum is equal to 16,357 and the average is equal to 2,907.

Looking at the distribution of the olive oil, the variability range widens even further with a minimum value of 1, a maximum of 73,659 and an average equal to 9,186. Furthermore, Figure 4 and 5 highlight that there is a certain level correlation between quantities sold of the single GTIN and the number of outlets in which sales were recorded (more in the case of olive oil respect to rice). Looking at the number of outlets in which are sold the GTINs of the first quintile of the distribution, on average, references are sold in 6.7 outlets for rice and in 14.2 outlets for olive oil.

Figure 5 - *Quantity sold, number of outlets and average price per GTIN for Olive oil (single, 1000 ml) in the Province of Rome - Year 2022.*



Source: Elaborations on scanner data

Finally, concerning the GTINs of the first quintile, there is not a clear correlation between average prices for GTIN and quantities sold. Regarding the trend of quantity sold respect to the average price per GTIN, considering rice we note that sales values are more variable in terms of price levels than in the case of olive oil.

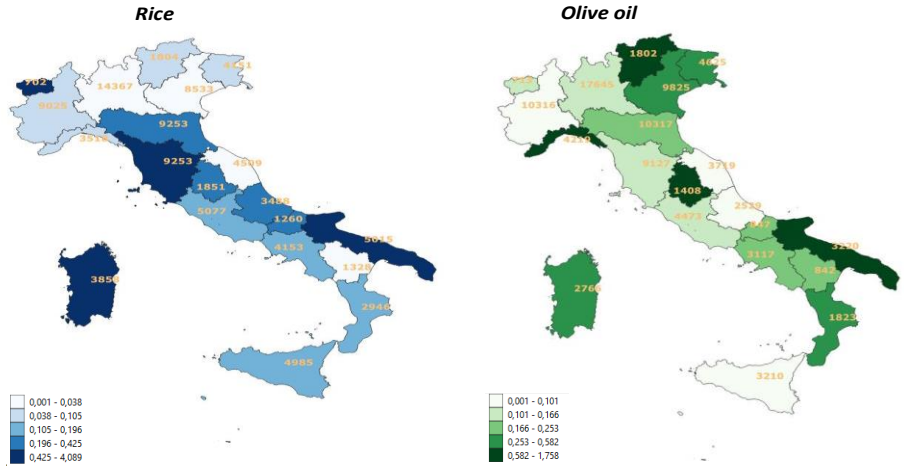
At the same time, quantity sold are relatively low for GTINs whose average prices are at the extreme ends of the range examined. Although, in few cases for both products, we find some GTINs with high prices that have rather high quantities sold.

3.2. Analysis of the variability of the annual average prices: a regional comparison

This subsection describes the results of the variability analysis of the annual average prices by making a regional comparison.

Figure 6 shows the territorial distribution of price variability among the Italian regions: the differences between the coefficients of variation (CV) of prices at territorial level are evident in the maps. The graphs also show the number of quotations used for the calculation of the minimum average price in each region for the two products considered. Note that the coefficient of variation express by how much the regional minimum average price considered varies compared to the national minimum average price.

Figure 6 – *Distribution of the coefficient of variation of prices and number of quotations used for Rice (single, 1000 gr) and Olive oil (single, 1000 ml) - Year 2022.*



Source: Elaborations on scanner data

From the figure we notice that some regions show a high level of variability respect to the minimum average prices: for the rice (map on the left) Valle d'Aosta, Toscana, Puglia and Sardegna show high values of the variation coefficient respect to the national minimum average prices.

In particular, Valle d'Aosta and Sardegna have a higher value of CV associated to higher prices respect to the Italian figure while Toscana and Puglia have them lower. This means that a high CV can be related to a higher or a lower level of regional minimum average price.

Regarding olive oil (map on the right), the regions with high values of CV are Liguria, Trentino Alto Adige, Umbria and Puglia. Specifically, Trentino Alto Adige and Liguria have higher prices while Umbria and Puglia have them lower.

The number of quotations highlighted in Figure 6 do not show a particular relationship with the CV values in the different Italian regions. A deeper analysis of the number of quotations recorded for the different retail chains could produce more evidences regarding this aspect.

The differences across regions from the CV point of view could be due to two different situations: one relating to the assortment present on the shelves of the outlets and the other relating to the number of quotations available. The analysis of the variability related to all the products within the poverty basket shows that higher values of CV are not directly related to lower values of number of quotations.

The question relating to the assortment of products in the outlets and therefore the variety of products offered by a specific chain derives from a differentiation in distribution and localization in the territory. At the same time, the number of quotations actually present for a given outlets can also influence the level of the regional minimum average price (in particular, if we consider the different retail trade channels). This is evident, especially comparing two regions, within which there is a different composition in terms of types of outlet. The strong presence of some types of retail trade channels with products at lower prices and/or on offer can significantly influence the level of the regional minimum average price.

3.3. A case study: focus on products with high variability

Looking at the variability of the minimum average prices of the different group of products, we can focus on the territorial distribution of the coefficient of variation (CV). In the Table 3, there are the products within the poverty basket with higher CV: these products show more volatile prices for some regions. The selection was made by considering the products that point out a regional CV higher than the Italian CV for a certain number of regions.

Table 3 – *Distribution of the coefficient of variation of some selected products for the Italian regions (with minimum value, mean and maximum) - Year 2022.*

Product	Measure unit	Piemonte	Valle d'Aosta	Lombardia	Trentino-Alto Adige	Veneto	Friuli-Venezia Giulia	Liguria	Emilia-Romagna	Toscana	Umbria	Marche	Lazio	Abruzzo	Molise	Campania	Puglia	Basilicata	Calabria	Sicilia	Sardegna	Min	Mean	Max
COOKED HAM	KG	3.06	21.65	8.60	24.99	8.79	3.29	2.82	0.59	5.91	2.61	0.33	2.60	3.78	2.58	6.15	3.87	9.37	4.83	4.28	15.71	0.33	6.79	24.99
RAW HAM	KG	3.00	75.26	7.10	2.44	5.37	0.33	6.52	3.14	0.07	2.49	1.53	0.85	0.60	0.02	24.82	0.35	33.01	24.02	12.15	10.95	0.02	10.70	75.26
CHOCOLATE	KG	1.55	12.51	0.49	0.47	9.10	13.14	1.46	1.68	13.83	0.27	5.17	1.18	0.88	1.33	1.81	2.81	4.45	13.85	0.80	1.25	0.27	4.40	13.85
DRIED FRUIT	KG	0.03	76.73	0.58	8.02	0.47	1.33	20.01	7.01	0.80	1.20	0.21	4.19	4.21	22.75	3.20	14.49	24.50	6.36	5.68	0.07	0.03	10.09	76.73
DEPARTURE CHILD MILK	LT	1.24	11.47	1.78	10.58	4.75	6.17	5.89	0.33	2.45	0.25	5.46	0.71	6.97	5.69	0.02	3.64	54.13	1.48	96.64	35.11	0.02	12.74	96.64
FROZEN FISH	KG	1.47	43.55	0.11	0.96	2.70	5.36	3.64	0.26	0.01	0.58	0.19	0.03	0.27	0.36	5.34	14.44	12.10	0.11	19.21	0.50	0.01	5.56	43.55
SALMON	KG	0.15	72.25	0.96	0.98	28.71	24.79	0.69	7.93	1.88	0.43	1.42	0.32	3.73	2.20	0.08	0.04	1.95	1.78	0.35	18.03	0.04	8.43	72.25

Source: Elaborations on scanner data

In particular, the regions that have the highest values of the coefficients of variation are Valle d'Aosta, Trentino Alto Adige, Friuli Venezia Giulia, Abruzzo, Puglia, Sicilia and Sardegna.

Specifically, a detailed analysis of the trend of the "*Departure child milk*" product was conducted in order to understand the dynamics of the average prices[7] based on the quantities sold by every GTIN in the regions considered. Table 4 shows, for the regions that assumes the highest values of coefficient of variation for this product, the selection of the main GTINs sold in the different geographical areas.

Table 4 – *Average price* and quantity sold for the GTINs of product "Departure child milk" in 4 Italian regions - Year 2022.*

Product	Trentino Alto Adige		Veneto		Abruzzo		Sicilia	
	Average price	Quantity sold	Average price	Quantity sold	Average price	Quantity sold	Average price	Quantity sold
GTIN 1					5,78	21		
GTIN 2					6,11	29		
GTIN 3	5,09	974	4,84	5.182	5,25	155	5,89	412
GTIN 4					5,57	464		
GTIN 5	5,78	444	5,34	9.616	5,38	220	6,50	6
GTIN 6			4,61	357	5,44	1.024		
GTIN 7	**6,62**	**8.523**	**6,67**	**63.089**	**7,31**	**6.437**	7,61	217
GTIN 8			5,38	285	5,71	709	6,07	10
GTIN 9			3,32	592	3,11	802	**3,25**	**17.000**
Total		9.941		79.121		9.859		17.645

Source: Elaborations on scanner data

** The average prices are referred to the quantity sold of 1000ml.*

[7] The average prices of the different products was brought back to the quantity sold of 1000ml since the GTINs have different formats with different quantities inside.

Looking at the composition of the "*Departure child milk*" in terms of GTINs we can see how in each region there is a prevalence of sales of just one product: indeed the GTIN 7 is the best seller for 3 of the 4 regions observed (Trentino Alto Adige, Veneto and Abruzzo) despite having the highest minimum average price in all three.

Sicilia has a different most selled product (GTIN 9) which is, at the same time, a product with a very low minimum average price (3.25 euro for 1000ml); this also significantly decreases the average regional minimum price. This situation is due, on the one hand, to a different behaviour of consumers towards the purchase of the GTINs in question and, on the other, to the different assortment within the outlets in the different regions.

The evidence shows that the variety of GTINs offered for some products is not homogeneous across the national territory. The offer of retail distribution chains between the Italian regions is also quite different and this affects the minimum average prices.

Since the product "*Departure child milk*" is composed by a relatively small number of GTINs inside, the analysis was made also for other products of the poverty basket with an amount of GTINs that are greater than the previous observed. In particular, regarding the product "*Cooked ham*" there is a greater assortment of products sold in the regions considered. Table 5 shows only the first 12 GTINs in terms of quantity sold not all the product sold.

Table 5 – *Average price* and quantity sold for GTINs within the product "Cooked Ham" in 4 Italian regions - Year 2022.*

	Trentino Alto Adige		**Toscana**		**Calabria**		**Sardegna**	
Product	**Average price**	**Quantity sold**	**Average price**	**Quantity sold**	**Average price**	**Quantity sold**	**Average price**	**Quantity sold**
GTIN 1	16,22	24.958	13,63	179.242				
GTIN 2	13,31	121.468	**13,53**	**2.045.273**	**13,66**	**313.236**	**13,55**	**510.892**
GTIN 3	**18,53**	**167.211**	17,34	792.602	18,31	42.163	24,94	1.738
GTIN 4	18,73	106.653	14,51	1.406.422	19,46	32.204	20,85	40.843
GTIN 5	12,56	26.215	12,55	15.473	12,56	48.000	12,58	100.322
GTIN 6	7,55	37.245	6,95	1.478	7,89	37.185	7,20	97.674
GTIN 7	10,99	62.263	10,92	14.033	11,07	96.515	11,01	174.450
GTIN 8			10,89	98.919	10,26	65.089		
GTIN 9			8,26	674.220			10,60	4.396
GTIN 10			9,42	537.374	9,51	24.823		
GTIN 11	11,91	39.054	10,33	16.366	12,06	134.902	11,59	232.599
GTIN 12	15,10	1.894	14,42	15.528	12,06	25.553		
Total		586.961		5.796.928		819.670		1.162.914

Source: Elaborations on scanner data

** The average prices are referred to the quantity sold of 1000gr*

The GTIN 2 is the best seller in three of four considered regions, Toscana, Calabria and Sardegna with very similar prices each other, whereas GTIN 3 is the best seller in Trentino with an higher price than GTIN 2. The evidence again shows that the minimum average price is influenced by the different variety of GTINs purchased by households and/or offered by retail chains.

4. Concluding remarks

This paper shows some empirical evidence on the basis of which the new methodology was implemented to valorise the absolute poverty basket using scanner data. First of all, thanks to such a rich data source, it was possible to associate each food product with a very high number of GTINs sold characterized by very heterogeneous packaging (both in terms of quantity and number of pieces included in the package). This allowed us to make a selection to identify the packaging most purchased by households.

Second, the distribution of prices for individual products highlighted that the GTINs with lower prices are not concentrated in specific type of retail channel but are present in all the types of outlets. Therefore no type of retail channel can be excluded a priori. Third, thanks to the high number of GTINs sold for each product it was possible to select those belonging to the first quintile of the price distribution. In fact, it is assumed that poorer households can satisfy their food needs by purchasing GTINs with lower prices.

Finally, for each product, it was possible to calculate the annual minimum average prices at more disaggregated territorial level and the results show a high variability of prices at regional level. A detailed analysis of some products with high coefficients of variation between regions has shown that this variability is explained, not only by different pricing policies, but above all by the different assortment of products sold. This is also due to the different distribution of large-scale retail trade chains across Italy. Therefore the difference between the GTINs most purchased by consumers significantly influences the minimum average prices estimated at territorial level.

Further analyzes on scanner data may be carried out in the future to investigate the price distribution of food products. In the analysis carried out so far, we have chosen to study the distribution of prices relating to the packaging most purchased by consumers. A possible development is the study of the scales of savings that can be achieved by increasing the contents of the packages. These are the least sold GTINs but they could be those purchased by larger poor households to save money.

Appendix: Information technology architecture, methodologies and paradigms

In this appendix, the main features of the technological infrastructure will be presented as well as the important aspects that had led to choices regarding technologies, paradigms and implementation. Because of the growth of retail data, it was necessary to abandon RDBMS[8] systems and adopt a big data platform solution. In fact, RDBMS systems had begun to show some critical issues such as the difficulty of scaling over a larger amount of data, the inability to easily store and process multiple annuals of the same survey, the increasing time for processing data.

The big data management platform, combined with modern data management techniques, overcome most of the problems mentioned above. In particular, the use of a data-lake built on HDFS (Apache Hadoop Distributed File System) has enabled the efficient processing of more than 700 million records per year for the poverty analysis and 1.3 billion records per year for the consumer price survey.

Specifically, data of the consumer price survey are weekly acquired through a centralized supply system ARCAM[9] and then are ingested into the data-lake. Thus, the data used for the valorisation of poverty basket were already available in our systems. It was sufficient to define a logical link between the data contained in the CSVs[10] files and the logical data model whereas the technical details and parallelism were handled directly by the platform engines.

From a technical point of view, thanks to the design and implementation of the new scanner data information system, including data structures, procedures and pipelines, the effort to deal with new challenges in new research activities and scenarios are significantly reduced.

Most of the modules of the production system were implemented in Scala programming language and using Spark framework. That permits us, out of the box, to take advantage of the distributed processing with high levels of performance and shorter production time. In addition, since Scala is a strongly typed language, it made the system extremely robust and less prone to errors.

The architecture of the application was realized by dividing it into modules: each module implements a defined part of the process. The guiding principles that inspired the design of the architecture are separation of responsibilities, maximum cohesion and minimum coupling of functionalities, as well as abstraction in design to make the modules more extensible and reusable (Gamma *et al.*, 1994), (Fowler *et al.*, 2002).

[8] Relational DataBase Management System.
[9] ARCAM is the portal for securely acquiring administrative archives from public and private organizations.
[10] The supplier stores the data in simple comma separated values (CSV) files whose structure is defined by the following fields: week and year, product id, outlet id, quantity of the product sold and related turnover.

Starting from the elementary data in the data repository dedicated to the production of consumer price indices, many packages, procedures and workflows are reused without major changes. The creation of the new environment was straightforward thanks to the abilities of big data management systems to define data structures regardless of the underlying data format, partitions and locations.

References

BRUNETTI A., FATELLO S., PATACCHIA O., RICCI R. 2024. New data sources for the valorization of the absolute poverty thresholds, *Rivista Italiana di Economia Demografia e Statistica*, Special Issue "New approaches for measuring poverty: studies and perspectives" ,Vol. LXXVIII, No. 4, pp. 111- 122.

FOWLER M., FOEMMEL M., HIEATT E., MEE R., RICE D., STAFFORD R. 2002. *Patterns of Enterprise Application Architecture*. Addison-Wesley Professional.

GAMMA E., HELM R., JOHNSON R., VLISSIDES J. M. 1994. *Design Patterns: Elements of Reusable Object-Oriented Software.* Addison-Wesley Professional.

ISTAT 2009. La misura della povertà assoluta. *Metodi e Norme n. 39*. https://www.istat.it/wp-content/uploads/2021/01/misura_della_poverta_assoluta.pdf

ISTAT 2023a. Gli indici dei prezzi al consumo. Aggiornamenti del paniere, della struttura di ponderazione e dell'indagine - Anno 2023. *Nota informativa*. https://www.istat.it/wp-content/uploads/2023/02/NOTA-INFORMATIVA-PANIERE-2023.pdf

ISTAT 2023b. Indici dei prezzi al consumo. Aspetti generali e metodologia di rilevazione - Edizione 2022. *Letture Statistiche - Metodi.* https://www.istat.it/it/files/2013/04/Indice-dei-prezzi-al-consumo.pdf

ISTAT 2023c. Le statistiche dell'Istat sulla povertà - Anno 2022. *Statistiche report*. https://www.istat.it/wp-content/uploads/2023/10/REPORT-POVERTA-2022.pdf

---

Francesco ALTAROCCA, Istat, fraltaro@istat.it
Cristina DORMI, Istat, dormi@istat.it
Stefania FATELLO, Istat, fatello@istat.it
Carlo MATTA, Istat, matta@istat.it