Big data and spatial price comparisons of consumer prices

Big data e confronti spaziali dei prezzi al consumo

Tiziana Laureti and Federico Polidoro

Abstract An accurate measurement of price level differences across regions within a country is essential for assessing inequality in the distribution of real incomes and consumption expenditures. However, systematic attempts to compile sub-national Purchasing Power Parities (PPPs) on a regular basis have been hampered by the labour-intensive analyses required in processing traditional price data. The increasing availability of big data may change the current approach for estimating sub-national PPPs, although only for household consumption expenditures. This paper estimates spatial price indexes using a scanner dataset set up for experimental CPI computations in 2017 and includes information on prices, quantities and quality characteristics of products at barcode level. Our dataset refers to grocery products sold in a random sample of approximately 1,800 outlets across Italy belonging to the most important retail chains (95% of modern retail trade distribution), covering 55.4% of total retail trade distribution for this product category. We use various weighted index formulas for calculating consumption sub-national PPPs at detailed territorial level and at the lowest aggregate level at which no quantity or expenditure weights are usually available. Finally we report preliminary estimates of between-regional spatial price indexes for specific product groups and for Food and Non-Food consumption aggregates.

Riassunto*: Misurare i differenziali di prezzo tra le regioni di un Paese è fondamentale per analizzare la disuguaglianza nella distribuzione dei redditi e dei consumi. Tuttavia, la produzione regolare di Parità del Potere di Acquisto (PPA) a livello sub-nazionale è stata ostacolata dalle difficoltà di utilizzare dati da fonti tradizionali. La crescente disponibilità di big data potrebbe cambiare l'attuale approccio per la stima delle PPA infra-nazionali sebbene solo in riferimento ai*

[1] Tiziana Laureti, Department of Economics, Engineering, Society and Business Organization University of Tuscia, Italy; email:laureti@unitus.it

Federico Polidoro, Living conditions and Consumer prices Unit, Istat; email: polidoro@istat.it

consumi delle famiglie. Il lavoro presenta stime preliminari di indici spaziali regionali utilizzando un dataset di dati scanner costruito per la stima sperimentale degli IPC e include informazioni su prezzi, quantità e caratteristiche dei prodotti a livello di codice a barre per il 2017. Il dataset si riferisce ai prodotti grocery venduti in un campione di circa 1,800 negozi in Italia rappresentativo delle più importanti catene di vendita (95% del fatturato totale della distribuzione moderna e 55,4% del commercio al dettaglio totale di questi prodotti). Si utilizzano diversi metodi con strutture di ponderazione per il calcolo delle PPA infra-nazionali al livello più disaggregato, dove solitamente non sono disponibili informazioni su quantità e spese. Si riportano stime preliminari degli indici spaziali regionali per specifici gruppi di prodotti e per gli aggregati di consumo alimentari e non alimentari.

Key words: regional price levels, spatial price indexes, sub-national PPPs

1 Introduction

An accurate measurement of price level differences across regions within a country is essential in order to better assess regional disparities, thus enabling policy makers to adequately identify and address areas of intervention. Regional values of economic indicators such as Gross Domestic Product (GDP), income and poverty levels, should be adjusted for regional price differences, following the same logic according to which the economic well-being of different countries is compared by taking into account *international purchasing power parities* (PPPs).

At international level, PPPs for countries are compiled by the International Comparison Program (ICP), administered by the World Bank with the collaboration of the OECD, EUROSTAT and other organizations (World Bank, 2013).

Though not as widespread as international comparisons of prices through ICP, there have been research projects and studies on the compilation of spatial price indexes, also referred to as *sub-national purchasing power parities*, carried out by NSOs and individual researchers in various countries including USA, Brazil, India, Indonesia, China, Italy, Australia, New Zealand and the United Kingdom (Biggeri et al, 2017; Laureti and Rao, 2018).

The Italian National Institute of Statistics (Istat) is one of the few National Statistical Offices (NSOs) that carried out official experimental sub-national PPPs computations by using price data from Consumer Price Indexes (CPIs) and ad-hoc surveys and focusing on comparing consumer prices across the 20 Italian regions. Significant price differences emerged in 2010 and in 2008, which encouraged Istat to confirm the project for producing sub-national PPPs on a regular basis (Biggeri et al., 2017). However, systematic attempts to compile sub-national PPPs on a regular basis have been hampered by the labour-intensive analyses required for processing traditional price data, i.e. data used for CPIs, and by the cost involved for carrying out ad-hoc surveys for collecting price data. In this context, the use of big data is

both a challenge and a possible solution to more than one difficulty NSOs face when constructing spatial price comparisons worldwide. A range of alternative sources, including scanner data, information obtained from administrative sources or by adopting web-scraping techniques must be investigated for making both temporal and spatial price comparisons. Since the primary use of scanner data is not to measure temporal and spatial price differences, methodological and empirical issues regarding scanner data quality, such as under-coverage for missing outlet types (i.e. hard discount) and products (i.e. perishables and seasonal products) and over-coverage (i.e. business expenditure) should be solved by price statisticians. It is worth noting that, in order to produce sub-national PPPs for the entire household consumption expenditure, other sources should be considered in addition to scanner data.

By focusing on grocery products (for which retail scanner data are currently available), it may be fruitful to use the itemized information contained in scanner data (turnover and quantities for each well-specified item code) in order to compile weighted spatial price indexes at detailed territorial level. When reviewing international practices, it is important to note that even if a fifth of EU countries have been using scanner data for compiling CPIs using different methods, as yet few studies have explored the use of new data sources for compiling spatial indexes (Laureti and Polidoro, 2017). Within the European Multipurpose Price Statistics project, Istat has recently introduced scanner data in the official CPI computation and has been exploring the possibility of using them for compiling sub-national PPPs.

The aim of this paper is to estimate household consumption sub-national PPPs for Italy in 2017 by using a scanner dataset constructed for experimental CPI computation after having carried out data cleaning and trimming outliers processes. This data set refers to a random sample of approximately 1,800 hypermarkets (more than 500) and supermarkets (almost 1,300), concerning the grocery products sold in the most important retail chains (95% of modern retail chain distribution that covers 55.4% of total retail trade distribution for this category of products). Since data are available for the 107 Italian provinces we estimate within-regional and between-regional spatial price indexes for the food and non-food consumption aggregates included in the scanner dataset.

2 Scanner data and spatial price comparisons in Italy

The first and a fairly critical step when compiling household consumption sub-national PPPs is to prepare a long list of goods and services that will be priced in the regions involved in spatial price comparisons. The list used in price comparisons needs to meet and balance the requirements of comparability and representativity. Comparability means that identical products of the same or similar quality should be priced across all the regions so that the PPPs based on these data solely reflect price level differences. Representativity is a concept associated with the relative “importance” of individual products within a group of similar products (called Basic

Heading, BH[1]): ideally, a product's importance should be determined using information on expenditure.

In this paper we refer to a list of products derived from the dataset described above, that specifically covers 54 grocery product aggregates, belonging to five divisions of the ECOICOP (01, 02, 05, 09, 12). Annual provincial average prices for each item were used which were obtained by aggregating the weekly prices of each GTIN code sold in the supermarkets and hypermarkets of 16 modern distribution chains located in the 107 Italian provinces using turnover weights, thus obtaining a total of 487,094 different products (GTINs).

The identification of the items is based on barcodes (GTINs), which univocally classify the products across the entire national territory. In each outlet (approximately 1,800), items were selected in order to cover up to 60% of the total turnover of the product aggregate. It is worth noting that perishables and seasonal products such as vegetables, fruit and meat were excluded from the scanner data because these products are sold at price per quantity and are not pre-packaged with GTIN codes.

Although this dataset was constructed for CPI compilation, it can also be used for making spatial price comparison among Italian regions bearing in mind that only the best selling products, which are typically consumed in each Italian province and region, may have be included according to the CPI selection procedure. Therefore these products may not be strictly comparable across different provinces and regions. It is useful to note that not all of the listed products must be priced in all of the regions. However, reliable regional price comparisons can be made as long as there is reasonable overlap in the items priced in different regions. We checked this requirement by verifying if product overlaps exhibit a chain structure.

3 Methods

In order to estimate regional spatial indices for products sold in modern distribution chains by using data for the 107 Italian provinces, a two-step procedure similar to the one used in the ICP was adopted whereby provinces are grouped into regions (World Bank, 2013). In the first step, within-regional PPPs are computed by comparing price and quantity data referring to products sold in the various provinces within each region while in the second step, between-regional PPPs are obtained for each region by using deflated price data for each province.

Moreover, as in international practice, sub-national PPP compilation is undertaken at two levels, viz., at BH level and at a more aggregated level (Food and Non-Food products). The methods selected for making multilateral comparisons is based on several axiomatic properties, including two basic properties: transitivity and base region invariance. Transitivity simply means that the PPP between any two regions should be the same whether it is computed directly or indirectly through a

[1] The smallest level of aggregation at which expenditure data are available are known in ICP parlance as basic headings. Although scanner data include expenditure data at item level, we still use the term "basic heading" to indicate a group of similar products which corresponds to a sub-class in the COICOP.

third region. The second requirement is that the PPPs be base region–invariant, which means that the PPPs between any two regions should be the same regardless of the choice of base region.

Aggregation methods at BH level

First step: within-regional PPPs

Let us assume that we are attempting to make a spatial comparison of prices between *R* regions, *r=1,…,R,* with *Mr* provinces in each region *r*. In the first stage of aggregation of price data at item level, which leads to price comparisons at BH level, p_{ijr} and q_{ijr} represent price and quantity of *i-th* item in *j-th* province and in *r-th* region $(i=1,2,...,N;\ j=1,2,...,M_r;\ r=1,\ldots,R)$.

In order to compute within-regional PPPs, we explored using different methods[1], however, due to lack of space, we only illustrate the *Region-product-dummy (RPD) method* which was also used to compute between-regional PPPs. All methods are implemented using R. If product overlaps exhibit a chain structure thus the RPD method exhibits some aspects of spatial chaining.

The RPD is the regional version of the country-product-dummy (CPD) method used in international comparisons. This method suggests that price levels are estimated by regressing logarithms of prices on provinces for each *province* and product dummy variables; the model is given for each BH by:

$$\begin{aligned}\ln p_{ijr} &= \ln PPP_j + \ln P_i + \ln u_{ijr} \\ &= \pi_j + \eta_i + v_{ijr} \\ &= \sum_{j=1}^{M_r} \pi_j D^j + \sum_{i=1}^{N} \eta_i D^i + v_{ijr}\end{aligned} \qquad (1)$$

where D^j is a provincial-dummy variable that takes value equal to 1 if the price observation is from *j-th* province in the *r-th* region; and D^i is a i-dummy variable that takes value equal to 1 if the price observation is for *i-th* commodity. The random disturbance is assumed to satisfy the standard assumptions of a multiple regression model. In order to estimate parameters of this model we impose normalization $\sum_{j=1}^{M_r} \pi_j = 0$ thus treating all regions in a symmetric manner. If $\hat{\pi}_j\ (j=1,2,...,M_r)$ are estimated parameters, the within-regional PPP for the province *j* in region *r* is given by $WR\_PPP_j = e^{\hat{\pi}_j}$. The RPD method based price comparisons are transitive and base-invariant. With the aim of taking into account the economic importance (representativity) of each product expressed by expenditure weights w_{ijr} based on turnover we used a weighted RPD model:

$$\sqrt{w_{ijr}} \ln p_{ijr} = \sum_{j=1}^{M_r} \pi_j \sqrt{w_{ijr}} D^j + \sum_{i=1}^{N} \eta_i \sqrt{w_{ijr}} D^i + \sqrt{w_{ijr}} v_{ijr} \qquad (2)$$

Second step: between-regional PPPs

[1] We used various spatial index formulae, including Fisher based GEKS, Geary-Khamis and CPD (World Bank, 2013; Laureti and Rao, 2018). We found interesting results which suggest a high variability of prices within various regions. However, they cannot be reported here due to lack of space.

In order to use provincial prices adjusted for differences among provinces within the *r-th* region item prices in all the provinces of region *r* are converted by using:

$$\hat{p}_{ijr} = \frac{p_{ijr}}{WR\_PPP_{jr}} \tag{3}$$

The deflated prices (in log form) were used for estimating a weighted RPD model with regional dummies and weights defined by deflated expenditure for each item in the *r-th* region.

$$\sqrt{w_{ijr}} \ln p_{ijr} = \sum_{k=1}^{R} \pi_k \sqrt{w_{ijr}} D^k + \sum_{i=1}^{N} \eta_i \sqrt{w_{ijr}} D^i + \sqrt{w_{ijr}} v_{ijr} \tag{4}$$

The between-regional PPP for the region *r* is given by $R\_PPP_r = e^{\hat{\pi}_r}$ and transitive price comparisons based on RPD method are given by:

$$P_{rk}^{RPD} = \frac{\exp(\hat{\pi}_k)}{\exp(\hat{\pi}_r)} \quad for\ all\ r,k = 1,2,...,R \tag{5}$$

Aggregation method for aggregation above basic heading level

The next and final step for compiling regional price comparisons is to aggregate the results from BH level comparisons to higher level aggregates. Let us assume that there are *L* basic headings (*l=1,…,L*) and e_i^r expenditure for *i-th* BH in region *r*. We decided to use the Fisher price index since it has a range of axiomatic and economic theoretic properties. The Fisher index is given by:

$$P_{rk}^{Fisher} = \sqrt{P_{rk}^{Laspeyres} \cdot P_{rk}^{Paasche}} \tag{6}$$

where

$$P_{rk}^{Laspeyres} = \frac{\sum_{l=1}^{L} p_l^k q_l^r}{\sum_{l=1}^{L} p_l^r q_l^r} = \sum s_i^r \left(\frac{p_l^k}{p_l^r} \right), \qquad P_{rk}^{Paasche} = \frac{\sum_{i=1}^{N} p_l^k q_i^k}{\sum_{i=1}^{N} p_l^r q_i^k} = \left[\sum_l s_l^k \left(\frac{p_l^k}{p_l^r} \right)^{-1} \right]^{-1}$$

with $s_i^r = \frac{e_i^r}{\sum_{l=1}^{L} e_l^r} = \frac{p_l^r \cdot q_l^r}{\sum_{l=1}^{L} p_l^r \cdot q_l^r}$.

As the Fisher binary index in (6) is not transitive, it is possible to use the procedure suggested by Gini (1931), Elteto and Koves (1964) and Szulc (1964) referred to as the GEKS index to generate transitive multilateral price comparisons across different regions. The resulting index is given by:

$$P_{rk}^{GEKS-Fisher} = \prod_{r=1}^{R} \left[P_{rs}^{Fisher} \cdot P_{sk}^{Fisher} \right]^{1/R} \tag{7}$$

The GEKS-Fisher based formula is used in cross-country comparisons made within the ICP at the World Bank (2015) and the OECD-Eurostat comparisons. In order to obtain a set of R_PPPs that refer to the group of regions (Italy) we standardized the GEKS-Fisher based PPPs (S-GEKS).

4 Results

In order to compute within and between regional spatial price indexes, we ran weighted RPD for all available BHs in scanner data (L=54) using expenditure weights defined by turnover. We then aggregated the results from BH level comparisons to higher level aggregates, i.e. food and non-food products both for within and between-regional PPPs. All results obtained cannot be reported, therefore Table1 illustrates RPPP for 2 BHs while Figure 1 shows aggregated RPPPs.

Table 1: WRPD estimation results for "Pasta products" and "Non-electrical appliances" Lazio=100

	Pasta Products (BH1)				**Non-electrical appliances (BH2)**			
Region	*Coef.*	*std.error*	*p-value*	*RPPPs*	*Coef.*	*std.error*	*p-value*	*RPPPs*
North-Center								
PIEMONTE	0.0187	0.0039	0.000	101.89	-0.0806	0.0079	0.000	92.26
VALLE D'AOSTA	0.0526	0.0039	0.000	105.41	0.0305	0.0081	0.000	103.10
LIGURIA	0.0482	0.0044	0.000	104.94	-0.0269	0.0079	0.001	97.35
LOMBARDIA	0.0264	0.0038	0.000	102.67	-0.0509	0.0079	0.000	95.04
TRENTINO A.A.	0.0716	0.0039	0.000	107.42	0.0051	0.0080	0.523	100.51
VENETO	0.0347	0.0038	0.000	103.53	-0.0309	0.0079	0.000	96.96
FRIULI V.G.	0.0435	0.0038	0.000	104.45	-0.0285	0.0079	0.000	97.19
EMILIA-ROMAGNA	0.0227	0.0041	0.000	102.30	-0.0580	0.0079	0.000	94.37
TOSCANA	-0.0050	0.0039	0.201	99.50	-0.1294	0.0079	0.000	87.86
UMBRIA	-0.0094	0.0039	0.015	99.06	-0.0185	0.0079	0.019	98.17
MARCHE	0.0557	0.0041	0.000	105.73	0.0077	0.0079	0.327	100.77
South and Islands								
ABRUZZO	0.0561	0.0040	0.000	105.77	-0.0163	0.0079	0.039	98.38
MOLISE	0.0471	0.0041	0.000	104.82	0.0142	0.0080	0.076	101.43
CAMPANIA	-0.0097	0.0040	0.014	99.04	0.0167	0.0079	0.035	101.69
PUGLIA	-0.0388	0.0040	0.000	96.20	-0.0267	0.0079	0.001	97.37
BASILICATA	-0.0410	0.0040	0.000	95.98	0.0021	0.0080	0.791	100.21
CALABRIA	-0.0286	0.0040	0.000	97.18	0.0087	0.0080	0.275	100.87
SICILIA	-0.0598	0.0044	0.000	94.19	0.0667	0.0080	0.000	106.89
SARDEGNA	0.0336	0.0046	0.000	103.41	-0.0266	0.0079	0.001	97.37
Obs.	18,007				3,453			
Root MSE	0.09538				0.10105			
AIC	-19261.88				-4447.601			

Regional spatial price indexes for two specific groups of products, that is "Pasta products" (BH1), which belongs to the aggregate Food products, and "Non-electrical appliances" (BH2, e.g. razors, scissors, hairbrushes, toothbrushes, etc.) included in the Non-Food aggregate, confirm large differences in price levels among Italian regions even if BH2 shows a higher territorial heterogeneity than BH1 (range is equal to 19.03 and 13.23 respectively). In the case of BH1, 5 regions located in the South and Islands (out of 8) and 2 Northern- Central regions (out of 11) show lower prices than Lazio while for BH2 higher price indexes are observed in 3 Southern regions and 8 regions in Northern and Central Italy. This different territorial pattern of consumption spatial price indexes is not confirmed when aggregated regional PPPs are computed for Food and Non-food products (Italy=100).

As shown in Figure 1, Southern regions appear to have price levels that are below the national average both for Food and Non-Food products, with the exception of Abruzzo (101.90 and 101.33, respectively), Molise (102.90 and 101.24) and Sardegna (101.93 and 101.57). However, it is worth noting that some

Northern regions also show lower price levels than the national average, such as Emilia-Romagna (98.31 and 98.40), Veneto (99.09 and 98.48) and Piemonte for Food products (99.80). On average, Toscana proved to be the less expensive region for both product aggregates (96.24 and 95.17). These results seem to suggest that when considering the retail modern distribution, the expected relationship in terms of price levels between the North and South of Italy partially changes (for "Pasta products" locally made goods could play a key role for maintaining the traditional price differences) and propose an interesting line for future research on the influence of the various distribution channels when defining sub-national PPPs. Caution is required when interpreting these results since: a) they may be influenced by the characteristics of the modern retail trade which is not uniformly distributed across Italian territory in terms of types of retail chains and market share; b) we excluded two groups of products "Whole Milk" and "Low-Fat Milk" since there were no reliable overlaps among regions enablig spatial price comparisons; c) these results are based on data selected for CPI compilation and hard discounts are excluded.

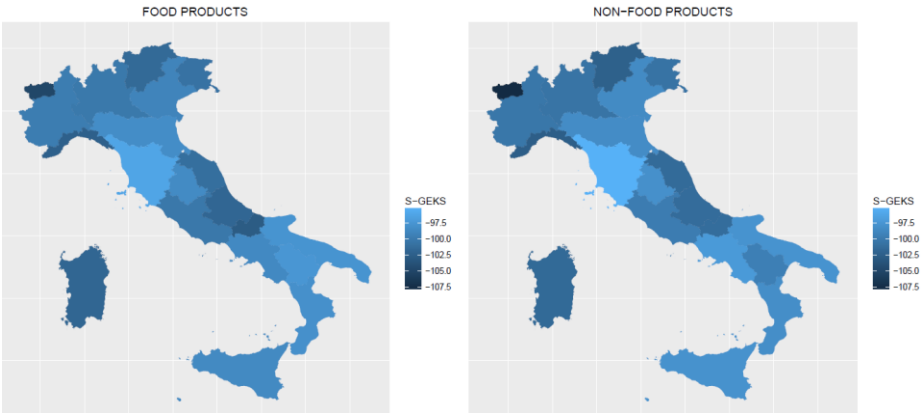

Figure 1: Between-regional PPPs for Food and Non Food products (Italy=100)

References

Biggeri, L., Laureti, T., and Polidoro, F.: Computing sub-national PPPs with CPI data: an empirical analysis on Italian data using country product dummy models. Soc Indic Res, 131(1), pp. 93-121, (2017).
Laureti, T., and Polidoro, F. Testing the use of scanner data for computing sub-national Purchasing Power Parities in Italy, Proceeding of 61st ISI World Statistics Congress, Marrakech, (2017)
Laureti, T., and Rao, D. P.: Measuring Spatial Price Level Differences within a Country: Current Status and Future Developments. Estudios de economía aplicada, 36(1), pp.119-148, (2018).
World Bank: Purchasing Power Parities and the Real Size of the World Economies-A Comprehensive Report of the 2011 International Comparison Program, Washington, DC(2015)