InGRID
Supporting expertise in inclusive growth

SIG Workshop on non-probability samples

Experiments and projects on the computation of spatial consumer price level differences in Italy by using Big Data: issues related to the use of non-probability samples

Biggeri L., Giusti C., Marchetti S., Pratesi M.
ASESD, Camilo Dagum Centre

Trier University, 27-28 February, 2020

Motivations

- Why we present this kind of the paper on the computation of Sub-national Spatial Consumer Price Indexes (SCPIs) at this Workshop on non-probability surveys?
- In the Ingrid project, under the Monica Pratesi leadership, we are working on the methods to do adequate “real” comparisons of the disposal incomes, salaries, poverty and living conditions, at sub-national and local level
- To do it, it is necessary to compute sub-national SCPIs by using the data collected by the National Statistical I*nstitutes (NSIs)* for the computation of temporal Consumer Price Indexes (CPIs)
- This data are collected by using different sources: territorial surveys at the outlets by non-probability samples; use of administrative data, use of scanner data (Big data) and use of probability sampling data for the estimation of the weights to compute the aggregated indexes. Therefore, the SCPIs computation of the NSIs and our computation of SCPIs have to fight the issues which will be discussed during the Workshop.

Aims of the paper

- Present selected experiments carried out in Italy to compute different Spatial Consumer Price Indexes (SCPIs)
  - ❖ Conducted by the members of the Camilo Dagum Centre in collaboration with Istat's researchers

- Show data sources and methods of computation used and some results, also with reference to the estimation of SPIs for the poor

- Underline some specific issues to be solved by using data coming from different separate surveys, conducted with both non-probability sampling (administrative data and scanner data) and probability sampling

Overview of the presentation

1 Research conducted in Italy to compute Spatial Consumer Price Indexes, making use of:

1.1. Traditional CPI data

1.2. Scanner data and other sources of data

2 Research for the computation of spatial housing dwellings rents price indexes

3. Main Issues in using non-probability sampling data and big data for the computation of temporal and spatial consumer price indexes

1 Researches conducted in Italy to compute Spatial Price Indexes (SPIs)

- In 2004, Istat started the computation of official regional consumer SPIs (called PPPs), by **using CPI data and *ad hoc surveys*,** which published for the year 2006 and 2009 (Biggeri L., Laureti T., Polidoro F. , 2017; Biggeri L., Laureti T., 2018, 2019)
  - Results

The computations show **significant differences** in the level of consumer prices (especially for the house dwelling rents) **across the regional capitals**

  - Limits:

✔ **Labor-intensive preliminary analyses,** extensive data editing was necessary for using CPI data

✔ **High costs for carrying out *ad hoc* surveys for clothing and other products**

- Then, Istat, through a specific project included in the **National Statistical Plan, has planned** to regularly **produce spatial price consumer indexes** at regional level by using a **multi-sources approach** (**Data Warehouse** of microdata)

1.1 Researches conducted using traditional CPI - 1st Phase

We can distinguish different **phases of experiments**, with different aims, conducted by Istat Price Statistics staff in cooperation with Proff. Laureti and Biggeri (Laureti T., Rao, D.S.P. , 2018; Biggeri L., Laureti T., 2019)

- **First phase**, by **using only CPI data** (BiggeriL., Laureti T., Polidoro F., 2017)
  - To verify to what extent the type and characteristics of the data affect the estimates by comparing various **Country Product Dummy (CPD) models** based on data characterized by different levels of aggregation
  - Year 2014; Data set for food and beverages, 7 **Basic Headings (BHs)**
  - **218,228 monthly price quotes** from the 19 regional chief towns

- **Results**
  - The computations confirm that **methods and the characteristics of CPI data** for spatial comparisons **are reciprocally influenced** (ungrouped data is preferable)
  - Significant consumer **price level differences** across Italian region's chief town

1.2 Researches conducted by using scanner data: 2nd Phase

- **Second phase**
  - These researches started in 2014 with the **availability of the first set of scanner data** obtained from the retail trade chains of the modern distribution
  - At the beginning only a **limited part of the scanner data** were available in terms of territorial, BHs and products coverage
- **Two main researches**
  - By Laureti T., Ferrante C., Dramis B. (2017), with the following aims:
    - To explore the **potential advantages of the use of scanner data** for constructing sub-national SPIs (suitability of scanner data for making spatial comparisons)
    - To deal with the **methodological** and **empirical issues** deriving from the use of this new data source, also in combination with CPI data
    - ✔ Which method should be used at the lowest level of product aggregation (Basic Heading, BH)?
    - ✔ Are the BH suitable for estimating sub-national SPIs when they include heterogeneous group of items?

1.2 Researches conducted by using scanner data: 2nd Phase

Results

- The **stochastic approach** to spatial price indexes proved to be suitable for constructing sub-national PPPs in Italy using both scanner and CPI data
- Scanner data enables us to **use expenditure and quantity weights** thus increasing the reliability of the estimation results while ensuring both representativeness and comparability requirements
- It is **essential** to **estimate** sub-national PPPs **at lower levels** of the ECOICOP product classifications (consumption segments) when the subclass of products is composed of heterogeneous products
- There is a **great heterogeneity** across **consumption segments** within a subclass and across the **Italian regional chief towns** and the results depend on the product classification used

- By Laureti T., Polidoro F. (2017) with the following aims:
  - As for the previous analyses this research focuses on the **first stage of aggregation**, thus on obtaining estimates of sub-national SPIs at BH level

1.2 Researches conducted by using scanner data: 2nd Phase

- To conduct analyses at a very detailed level, both for Retail Chains, BH and Products, focusing on the Heterogeneous groups of products
- To **check the application and validity of the Weighted CPD Method**; If the products within a group are homogeneous, we expect weighted and unweighted CPD to produce similar SPIs

Results

- There is a **great heterogeneity** across consumption segments within a subclass and across the Italian regional chief towns, which affects the results
- Significant differences can be observed between the results obtained from the unweighted and weighted CPD
- Retail chains and type of the outlet show a significant influence in the estimation of SPIs, which reflects different characteristics of modern retail distribution in the Italian regions

1.2 Researches conducted using Scanner and CPI data: 3rd Phase

- **Third phase**
  - The main aims of this phase is to:
    - Compute the **general** household consumption sub-national Spatial Price Indexes (**SCPIs**)
    - ✔ By using only Scanner data
    - ✔ By using a combination of Scanner data with Traditional CPI data and other data sources
- **Two main researches**
  - By using **only** 2017 complete set of **Scanner data** (Laureti T., Polidoro F., 2018):
    - To assure the representativity of the price quotes used in the computation:
    - ✔ Stratified Random Sample of the 9,000 outlets by provinces have been used
    - ✔ Finest available classification of items
    - Computation of the SCPIs –of Food products and, separately, for non-Food products- for the provinces within each region and than for the regions

1.2 Researches conducted using Scanner and CPI data: 3rd Phase

- Among the various methods used for the computation of SCPIs, the Weighted Regional Product Dummy method (RCP) proved be the best taking into account the availability of the turnover data
- The differences in the regional SCPIs resulted rather small
- Several limitations of the scanner data must be noted:

✔ From territorial point of view there is the advantage that scanner data can cover all the cities across the whole country but it may be that the rural areas are not covered

✔ Coverage of scanner data is just for the purchased made in the outlets of the modern distribution chains, and scanner data cannot be used for perishables and seasonal products

✔ To attain a complete coverage of the products purchased by the households, other outlets and markets (as hard discount, small shops and open markets) must be used both for all the kind of grocery products and the other products and services

✔ It is surely important to evaluate how much is expenditures share cover by the scanner data

1.2 Researches conducted using Scanner and CPI data: 3rd Phase

- Ferrante C. , Laureti T., Polidoro F. (2019)
  - Explain what is necessary to do to overcome the mentioned limitations combining data coming from different sources of data for consumer price statistics available in Italy:

✔ Data coming from territorial data collection
✔ Data coming from centralized data collection
✔ Administrative data (for fuels and actual rents)
✔ Data coming from Household Expenditure sample survey (for the weights)
✔ Scanner data

  - Two set of Regional CSPIs have been obtained for the two retail trade channels (traditional and modern)
  - Istat is committed in the activity to achieve the objective of compiling reliable Regional CSPI to be produced regularly, solving the main issues emerged in the experimentations and by using all the different data sources covering the entire basket of products and services.

⇨ Reference framework: a **multistage stratified sampling design** should be used to select from the existing micro-data to obtain effective estimates of CPIs and SCPIs

1.2 Researches by using Scanner and CPI data: 4th Phase

- **Cooperation between Istat and Camilo Dagum Centre**

- In October 2018 Istat and Dagum Centre signed an agreement of on "Methodological advances in the field of Small Area estimation methods for poverty and living condition Indicators"
- Among the objectives:
- Integration between the different sources of elementary data (samples, administrative data, scanner data) also through probabilistic techniques
- Use of the *scanner data,* obtained from the retail trade chains of the modern distribution, and data coming from the CPI data collection to estimate the different level of prices at local level (CSPIs)

- The Dagum Centre obtained the availability to use the scanner data and the price data coming from the surveys to compute CPI

1.2 Researches by using Scanner and CPI data: 4th Phase

- Researchers of the Dagum Centre carry out very preliminary/limited analysis of data to verify the need to estimates **CSPIs for poor**
- Analyses for deciles of the price distribution for each product (**hypothesis** that the poor purchase the cheaper items of a product)
- Preliminary computation of the CSPIs for the different deciles of the distributions

Estimation of Regional CSPIs (2017)

	Mean	Dec1	Quart1	Median	Dec9
Piemonte	1,18	0,98	1,01	1,05	1,16
Valle d'Aosta	0,98	1,26	1,16	1,08	0,90
Lombardia	1,15	0,97	1,00	1,05	1,15
Trentino-Alto Adige	0,98	0,96	0,97	0,98	1,04
Veneto	1,00	1,02	1,01	1,02	1,05
Friuli-Venezia Giulia	1,00	1,04	1,02	1,03	1,02
Liguria	1,09	1,04	1,03	1,04	1,14
Emilia-Romagna	1,06	0,95	0,98	1,02	1,09
Toscana	1,07	0,94	0,95	1,00	1,10
Umbria	0,99	1,01	1,04	1,04	1,01
Marche	0,94	0,95	0,98	0,96	0,95
Lazio	1,05	0,94	0,96	1,03	1,09
Abruzzo	0,95	0,99	0,99	0,98	0,96
Molise	0,89	0,97	0,94	0,92	0,87
Campania	0,97	0,95	1,02	1,00	0,93
Puglia	0,99	1,01	0,99	0,98	0,99
Basilicata	0,81	0,95	0,91	0,86	0,76
Calabria	0,97	0,96	0,98	0,97	0,96
Sicilia	1,01	1,05	1,02	1,02	0,97
Sardegna	0,97	1,09	1,05	1,00	0,96

- Computation by using Scanner data
- Products food and non-alcoholic beverage
- Distribution of unit values for each product (within 102 groups of product of the ECOICOP 8 classification): millions of prices
- The Regional SPIS estimated by using unweighted RCPD
- Unfortunately the dummy of the regions are not statistically different (are not enough reliable), but

 Clear indication of some differences

2 The computation of spatial price indexes for housing dwellings rents

- **Spatial Price Indexes for Housing Rents** (SPIHR), measuring differences in rents across sub-areas within a country, are of **crucial importance** because may be used as **proxies** of spatial differences in cost of living since housing is the largest expenditure item in the household budget, **especially for poor households**.
- They are also important for:
  - ✔ Adjusting poverty thresholds
  - ✔ Comparing salaries and household disposable income;
  - ✔ **Designing housing policies at a local level**
- *In 2017* as yet no regular computation of them in Italy*, the Dagum Centre decided to carry out indipendently the estimation of* SPIHR:
  - By using the region-product-dummy-method (RPD) method, and then, in particular, the regression **hedonic price method (HPM)** that allows to take into account of the **structural characteristics** of the house, locality/neighborhood and environmental characteristics that indirectly affect the price/rent of an house

2 The computation of spatial price indexes for housing dwellings rents

- By using data coming from:

✔ **Archives** on purchases and rents recorded by Real Estate Market Observatory (**OMI**), which is part of the Italian Revenue and Tax Agency;Data provided, every semester for each Italian municipalities (more then 8,000) and homogeneous zones within the municipality territorial, for a total of **145,800 price rent observations**

✔ Rents and house information collected through the **Household Expenditure Survey**, conducted by Istat on a sample survey of about 23,000 units; The surveys collect a lot of information on the characteristic of the house occupied by the household

 The first estimates of the regional SPIHR have been conducted using 2017 OMI data (Biggeri L., Laureti T., 2018).

- Results show **significant rent level differences across various Italian regions** and support the notion that dwelling rents are higher in the Northern-Central regions than in the South. For example, considering as **base Italy=100**, SPIHR of economic and cheap dwelling-houses for **Liguria is 158, and for Sicily is 68!**

✔ The high variability of SPIHRs across regions, may depend on the presence of municipalities with a high propensity for tourism (i.e., sea, mountains, lake, etc.).

2 The computation of spatial price indexes for housing dwellings rents

- Subsequently, a more deep analysis as ben carried out, to **explore spatial heterogeneity in housing rents** within Italian regions we estimated RPD models by considering Italian provinces (Benedetti I., Laureti T. , Biggeri L., 2020, paper submitted for the publication)
- ✔ When moving from regional to provincial level, a higher level of heterogeneity emerged, especially within some Italian regions. Higher house rental prices are generally observed in regional capitals

- In 2019, the members of Dagum Centre (Biggeri, Giusti and Marchetti) implemented a research to compute Regional SPIHR, by **using the HES data**, which allows also to verify the effects of the different housing characteristics on the value of the rent
  - The estimations of the SPIHR have been done **for different typologies** of housing, of households, and occupancy status (i) all the occupancy status and only renting occupation: both for poor households and all the households; (ii) all the typologies of houses and separately for only the apartments: both for **poor households** and all the households
- ✔ The preliminary results, even **with statistical reliability problems** of the estimations for three regions, confirmed the **significant rent level differences across various Italian regions**.
- ✔ These estimation have been used in a paper presented at the 2019 ISI Congress, in in combination of the Grocery Index (taken by the international data base Numbeo), to estimate the sub-national Spatial CPIs (Marchetti S., et al, 2019)

3. Main Issues in using non-probability sampling data and big data

- **Coming to the topics** on which the **Workshop is focused**, we would like to remeber that Price index statisticians have **always faced the problems** of the use of non probability sample and of multiple data sources to compute CPIs (and consequently SPIs)

- It s long time go that I discussed these topics and the difficulties to obtain accurate estimates of CPIs and to measure the **sampling errors and bias** for CPIs (Biggeri and Giommi 1987: presented a classification of the errors and estimated the CPI variance by using Jacknife repeated Replication method (see also: ILO 2004, 2020; Manual of Eurostat 2015).

- In the last ten years **increasing use of scanner data**

  Many Issues: cut-off sampling, effects of sample selection bias, possible incomplete coverage and quality of the administrative data, and so on).

- Very **quick presentation**, as a bird flight, on the **experiments carried out in Italy** and on the **proposals by international experts**, just to give some ideas for the discussion during these days

3. Main Issues in using non-probability sampling data and big data

- **Experiments carried out in Italy** and proposals to improve the quality estimation of the CPIs and the Sub-National SPIs

1. Biggeri L., Falorsi P.D. (2006): some proposals and experiments on " A Probability Sample Strategy for improving the quality of the CPI survery using the **information on the Business Register**
2. De Gregorio, 2012: various proposals on the "Sample size for the estimate of consumer price sub-indices", under a **combination of alternative sample designs** and aggregation methods
3. Bernardini A. et al. (2016): deepen the work of De Gregorio, by considering the **use of scanner data**, also presenting some experiments by use a Montecarlo Simulation to evaluate bias and efficiency of the CPI
4. Brunetti A. et al. (2018): made proposal on the improvements in Italian CPI deriving from the use of **scanner data**, that now are corrently use by Istat
5. De Vitis C. et al. (2019): taking into account of the previous results, made proposals to **modernize the surveys for the collection of CPI data** (**Istat Workshop**)

3. Main Issues in using non-probability sampling data and big data

- New international proposals on making valid inferences from non-probability sampling and by integrating data from sample surveys and other sources of data (we cannot quotes all the very good papers that we found!)
  - Non probability sample, theory and practice: Mercer, A.W., Kreuter, F., Stuart, E.A. (2017); Kalton G. (2019)
  - Inference for non probability sample: Elliot, M.R., Valliant, R. (2017)
  - Limitation of big data and their uses: Kalton G. (2019), Eurostat (2015)
  - Combining surveys data with other data sources: Lohr, S.L., Raghunathan, T.E. (2017)
  - Cut-off, selection bias and paradoxes by using big data, Meng X.L. (2018); Yang, M., Ganesh, N., Mulrow, E., Pineau, V. (2018), Beresewicz, M., Lehtonen R., Reis, F., Di Consiglio, R., Karlberg, M. (Treating selectivity in big data (2018)
  - Big data and Small area estimation: Pfefferman, D., Sverchokov, M. (2007); but all of you are the experts!

Thank you for your kind attention

References -1-

Papers on researches conducted to compute Spatial Consumer Price Indexes by using traditional CPI data and scanner data

- Biggeri L., Laureti T., Polidoro F. (2017), Computing Sub-national PPPs with CPI Data: An Empirical Analysis on Italian Data Using Country Product Dummy Models. Special Issue, *Social Indicators Research*, 131, pp. 93-121
- Laureti T., Ferrante C., Dramis B. (2017), Using scanner and CPI data to estimate Italian sub-national PPPs, paper presented at the Scientific Meeting of the Italian Statistical Society (SIS), Fifrenzd, 28-30 June 2017
- Laureti T., Polidoro F. (2017), Testing the use of scanner data for computing sub-national Purchasing Power Parities in Italy, paper presented at *61st ISI World Statistics Congress 2017, Marrakech, Morocco*
- Biggeri L., Laureti T. (2018), Publications, experiments and projects on the computation of spatial price level differences in Italy, paper presented at the 3rd Meeting of the Country Operational Guidelines *Task Force of the ICP, World Bank*, Washington DC, September 27
- Laureti T., Polidoro F. (2018), Big Data and spatial comparisons of consumer prices, paper presented at the Scientific Meeting of the Italian Statistical Society (SIS), Palermo 20-22 June 2018
- Laureti T., Rao, D.S.P. (2018), Measuring Spatial Price Level Differences within a Country, Current Status and Future Developments, Estudios de Economia Aplicada, vol.36-1, pp.119-148
- Biggeri L., Laureti T. (2019), Short overview of sub-national consumer spatial price indexes in Italy, ppt presented at the paper presented at the 4th Meeting of the Country Operational Guidelines *Task Force of the ICP, OECD,* May 3, 2019, Paris
- Ferrante C., Laureti T., Polidoro F. (2019), Combining data coming from scanner, traditional CPI data collection and other sources to compile sub-national PPPs in Italy, poster presented at the Sixtheenth Meeting of the Ottawa Group, Rio de Janeiro, 8-10 May, 2019
- Marchetti S., Bertarelli G., Biggeri L., Giusti C., Pratesi M., Schirrippa-Spagnolo F. (2019), Small area poverty indicators adjusted using local Price indexes, ppt of the invited paper presented at *62th ISI World Statistics Congress 2019, 18-23 August, Kuala Lumpur, Malaysia*

References-2-

Papers on researches conducted to compute Spatial Price Indexes for housing dwellings rents

- Biggeri L., Laureti T. (2018), Publications, experiments and projects on the computation of spatial price level differences in Italy, paper presented at the 3rd Meeting of the Country Operational Guidelines *Task Force of the ICP, World Bank*, Washington DC, September 27
- Laureti T., Benedetti I., Biggeri L., Brandi M. (2018), Sub-national price indexes for housing rents: Methodological issue and empirical analyses for Italy, paper presented 11th International Conference of the ERCIM, 14 -16 December 2018, University of Pisa
- Benedetti I., Laureti T., Biggeri L. (2020), Sub-national price indexes for housing: Methodological issue and empirical analyses for Italy, Submitted for publication
- Biggeri L., Giusti C., Marchetti S. (2020), Estimation of price indexes for housing dwelling rents in Italy, by using the data come the Household Expenditure Survey, In preparation: internal note).

Papers on the main issues in using Big Data and non-probability sampling data

- Rao, J.N.K., Fuller, W.A. (2017), Sample survey theory and methods: Past, present and future directions, Survey Methodology, 43, 14-181.
- Pfeffermann, D., Sverchkov, M. (2007), Small-area estimation under informative probability sampling area and within the selected areas, Journal of the American statisticl association, 102, pp. 27-42
- Mercer, A.W., Kreuter, F., Stuart, E.A. (2017), Theory and practice in nonprobability surveys, Public Opinion Quarterly, 81, 250-279.
- Elliot, M.R., Valliant, R. (2017), Inference for nonprobability samples, Statistical Science, 32, 249-264
- Lohr, S.L., Raghunathan, T.E. (2017), Combining survey data with other data sources, Statistical Science, 32, 293-312.
- Kim, J.K., Wang, Z. (2018), Sampling techniques for big data analysis in finite population inference, Technical Report: arXiv: 1801.09728v1

References -3-

- Chen, Y., Li, P., Wu, C. (2018), Doubly robust inference with non-probability survey samples, Technical Report: arXiv: 1805.06432v1.
- Yang, M., Ganesh, N., Mulrow, E., Pineau, V. (2018), Estimation Methods for Nonprobability Samples with Companion Probability Sample, ASA Joint Statistical Meeting, Survey Research Methods Section.
- MengX.L. (2018),Statistical paradises and paradoxes in big data (I): Law of large population, big data paradox, and the 2016 US presidential election, Annal of Applied Statistics, 12, pp. 685-726
- Beresewicz, M., Lehtonen R., Reis, F., Di Consiglio, R., Karlberg, M. (2018), An overview of Methods for treating selectivity in Big data sources, Statistical Working Papers, Eurostat.
- Tam, S.M., Kim, J.K. (2018), Big data, selection bias and ethics – an official statistician's perspective, Statistical Journal of the IAOS, 34, 577-588.
- Yang, M., Ganesh, N., Mulrow, E., Pineau, V. (2018), , Developments in survey research over the past 60 years: A personal perspective, International Statistical Review, 87.
- Thompson, M.E. (2019), Combining data from new and traditional sources in population surveys, International Statistical Review, 87
-Kim, J.K., Park, S., Chen, Y., Wu, C. (2019), Combining non-probability and probability survey samples through mass imputation, Technical Report: ar Xiv: 181210694v2.
- Kalton G. (2019), Developments in survey research over the past 60 years: A personal perspective, International Statistical Review, 87, 510-530
- Yang S., Kim J., Song R. (2019)non-probability samples with high-dimensional data, Technical Report: arXiv: 1903.05212v1.
- European Statistical System(2015), Handbook for quality Reports, Eurostat, Luxembourg
- Eurostat (2018), Harmonized Index of Consumer Prices (HICP). Methodological Manual: Chapter 4, Sampling, pp. 51-68
- ILO (2020), Consumer Price Index Manual. Concepts and Methods: Chapter 4, Sampling, Draft of the Manual, pp 120-137

References -4-

Papers on some experiments carried out by researches of ISTAT to improve the quality estimation of CPIs and Sub-National SPIs, by using Big data (administrative and scanner data)

-Biggeri L., Falorsi P.D. (2006), A probability sample strategy for improving the quality of the consumer price index survey using the information of business registers, in: *Joint Unece/ILO Meeting on Consumer Price Indices*, 10-12 May, Ginevra

- De Gregorio C. (2012), Sample siz for the estimate of consumer price sub-indices with alternative statistical designs, Rivista di Statistica Ufficiale, N.1/2012, Istat, roma

- Bernardini, A., De Vitiis, C., Guandalini, A., Inglese, F., Terribili, M.D. (2016), Measuring inflation through different sampling designs implemented on scanner data, paper presented at Meeting of the Group of Experts on Consumer Price Indices, 2.4 May, Geneva

- Brunetti, A., Fatello, S., Polidoro, F., Simone, A. (2018), Improvements in Italian CPI/HICP deriving from the use of scanner data, paper presented at the Scientific Meeting of the Italian Statistical Society (SIS), 20-22 Giugno, Palermo.

- De Vitis, C., Guandalini, A., Inglese, F., Terribili, M. (2019), Sampling scheme using scanner data for the consumer price index, paper presente dat the Istat's Advisory Committee on Statistical Methods, Roma

- Biggeri L., Giommi A. (1987), On the accuracy and precision on the consumer price indices. Methods and applications to evaluate the influence of the sampling of households, Invited paper, *Proceedings of the 46th Session of International Statistical Institute*, Tokio, pp.1-18.

Appendices

CPD Method – in logarithmic form

$$\ln p_{knr} = \sum_{r=1}^{R} a_r D_r + \sum_{n=1}^{N} b_n D_n^* + v_{knr}$$

where, for each BH, p_{knr} denotes the annual price of product *n* in outlet *k* of area *r* (*n* = 1, 2,…*N*; *r* = 1, 2,…, *R*; *k* = 1,…, K_{nr});

D_r are dummies for the areas, D_n^* are dummies for the type of product.

The intra-national PPP for the area *r* is given by $PPP_r = e^{a_r}$ where a_r is the difference between the coefficient associated to area *r* and the coefficient corresponding to the reference area (in our case Rome)

NB: $PPP_r = e^{a_r}$ is a biased estimator however in our case given the small value of $\hat{\sigma}^2$ and the large value of **number of observations** the bias correction is negligible

CPD Method – in logarithmic form

$$\ln p_{knr} = \sum_{r=1}^{R} a_r D_r + \sum_{n=1}^{N} b_n D_n^* + v_{knr}$$

where, for each BH, p_{knr} denotes the annual price of product *n* in outlet *k* of area *r* (*n* = 1, 2,…*N*; *r* = 1, 2,…, *R*; *k* = 1,…, K_{nr});

D_r are dummies for the areas, D_n^* are dummies for the type of product.

The intra-national PPP for the area *r* is given by $PPP_r = e^{a_r}$ where a_r is the difference between the coefficient associated to area *r* and the coefficient corresponding to the reference area (in our case Rome)

NB: $PPP_r = e^{a_r}$ is a biased estimator however in our case given the small value of $\hat{\sigma}^2$ and the large value of **number of observations** the bias correction is negligible

Weitghed CPD Method – in logarithmic form

$$\sqrt{w_{ij}}\,\ln p_{knr} = \sum_{r=1}^{R} a_r \sqrt{w_{ij}} D_r + \sum_{n=1}^{N} b_n \sqrt{w_{ij}} D_n^* + \sum_{j=1}^{J} \lambda_j \sqrt{w_{ij}} Z_j + \sqrt{w_{ij}} v_{knr}$$

The weights are defined as:

$$w_{knr} = \frac{p_{knr} q_{knr}}{\sum_{n=1}^{N} p_{knr} q_{knr}} \qquad \sum_{n=1}^{N} w_{knr} = 1$$

- The use of weights is advisable in the context of index number literature in order to reflect the economic importance of each item (Silver, 2002, Diewert, 2002, Rao, 2005;2010)
- The expenditure share weights, w_{ij}, reflect the relative importance of different commodities as measured by turnover or sales (Rao, 2010; Aizcorbe and Aten, 2004)
- Quantity weights
- Rome=100

The Hedonic model to estimate Spatial Price indexes for House Rent (SPIHR)

The hedonic method suggests that rental price levels of an individual property *i* ($i = 1, \dots, N$) in *j-th* geographical area (j= $1, \dots, M$), r_{ij}, are estimated by regressing logarithms of rents on geographical areas and house characteristics (dummy variables). Considering a sample of n independent observations of houses' prices i (i = 1,2, …i,…,n) in the different areas (regions or provinces) j (j = 1,2, …, j,…,M), the model can be expressed in the semi-log formulation as follows:

$$lnr_{ij} = \sum_{j=1}^{M} \alpha_j A_j + \sum_{k=1}^{K} \sum_{h=1}^{H} \beta_{kh} C_{hk} + \varepsilon_{ij}$$

Where:

- A_j is a vector of geographical areas dummies;
- α_j is the vector of area prices;
- C_{hk} is the matrix of the characteristics (with k= 1,…,K) and their classifications (h=1,…,H).;
- β_{kh} is the matrix of hedonic regression coefficients called also characteristic shadow prices;
- ε_{ij} is the error terms, that satisfy the standard assumption of a multiple regression model.

➢ Once the α_j parameters are estimated, establishing the reference or base area, the SPIHRj of the area J with respect to the base area is given by

$$SPIHRj = \exp(\alpha_j)$$

Researches conducted by using CPI data

To refer at the paper by Biggeri L, Laureti T. and Polidoro F. (2017)

- **AIM**
  - Exploring the available **CPIs data** to understand whether and to what extent **data characteristics** affect the selection of the methods for computing intra-national PPPs which in turn influences the estimates obtained

- **DATA**
  - Year 2014
  - **Only 7 Basic Headings** (groups of products) have been selected weighing approximately 30.3% of the Food and non-alcoholic beverage group (16.5% on total)
  - The dataset used consists in **218,228 monthly price quotes** from the 19 regional chief towns

- **Methods (by using th five main different CDO methods)**
  - CPD model based on ungrouped data, using average prices: unweighted vs weighted
  - CDP models with averages prices vs individual price quotesf town

Researches conducted by using scanner data -1-

To refer at the paper by Laureti T., Ferrante C., Dramis B. (2017)

Aims

- To explore the **potential advantages of the use of scanner data** for constructing s SCPIs and the **methodological** and **empirical issues** deriving from their use

Data

By using a **limited part of scanner data** obtained from the retail trade chains of the modern distribution were available: data on turnover and quantities of items sold (unit value)

- **Year:** 2015
- **Product coverage:** Food products
- **Retailers**: selection based on available data
  - 931 outlets belonging to the 6 most important retail chains (Coop Italia, Conad, Selex, Esselunga, Auchan, Carrefour) covering 57% of the retail chains market
- **Territorial coverage**: 20 regional capitals (but only 37 among 109 Italian rovinces)
- **Turnover and quantity information**: 15,433 different products identified by GTIN codes (formerly known as EAN) code (the GTIN identifies a unique product and outlet type (Big data: **1,000,000 quotations every week**); 69 BHs or subclasses are considered in the scanner data set

Methods

- Weighted CPD by using Annual average price per outlet, and also Combining scanner data with CPI data

Researches conducted by using scanner data -2-

To refer at the paper by Laureti T., Polidoro F. (2017)

- **Aims**
  - To conduct analyses at a very detailed level, both for Retail Chains, BH and Products, focusing on the Heterogeneous groups of products
  - To **check the application and validity of the Weighted CPD Method**; If the products within a group are homogeneous, we expect weighted and unweighted CPD to produce similar SPIs
- **Data**
  - **Year:** 2017
  - **Product coverage:** grocery poducts for five divisions of the ECOICOP (01, 02, 05, 0,9, 12)
    - **Outlets**: Universe of 9,000 retailers belonging to the 16 most important retail chains (94% of modern retail chain distribution). Scanner data cover 55,4 of the total retail trade distribution for this category of products. Too many data: need to work on a sample of data.
      Outlets stratified, with probabilities proportional to the 2016 turnover, by province, distribution chains and kind of outlets (888 strata), 1,781 outlets (510 hypermarkets and 1,271 supermarkets
    - **Turnover and quantity information**: Items were selected with probabilities proportional to the 2016 turnover for each product aggregate (at 60% cut-off line)
  - **Territorial coverage**: 20 regional capitals and 109 Italian provinces)

- **Methods:** various methods for the computation of SCPIs at BH level and the aggregation above the BHs

Researches conducted by using scanner data -3-

To refer at the paper by Laureti T., Polidoro F. (2018)

- **Aims**
  - Compute the **general** household consumption sub-national Spatial Price Indexes (**SCPIs**)
  - ✔ By using only Scanner data
  - ✔ By using a combination of Scanner data with Traditional CPI data and other data sources
- **Data**
  - **Year:** 2017
  - **Product coverage:** grocery poducts for five divisions of the ECOICOP (01, 02, 05, 0,9, 12)
    - **Outlets**: Universe of 9,000 retailers belonging to the 16 most important retail chains (94% of modern retail chain distribution). Scanner data cover 55,4 of the total retail trade distribution for this category of products. Too many data: need to work on a sample of data.
      Outlets stratified, with probabilities proportional to the 2016 turnover, by province, distribution chains and kind of outlets (888 strata), 1,781 outlets (510 hypermarkets and 1,271 supermarkets
    - **Turnover and quantity information**: Items were selected with probabilities proportional to the 2016 turnover for each product aggregate (at 60% cut-off line)
  - **Territorial coverage**: 20 regional capitals and 109 Italian provinces)
- **Methods**
  - Hedonic CPD method