Measuring geographical and population coverage in CPI internet price collection

An application with groceries web scraping in Italy

Tiziana Laureti[a] and Luigi Palumbo[ab*]
[a]Università degli Studi della Tuscia, Viterbo, Italy
[b]Banca d'Italia, Roma, Italy [1]

Meeting of the Group of Experts on Consumer Price Indices

[1] The views expressed herein are those of the authors and do not necessarily represent the views of the Bank of Italy and/or the Eurosystem.

Agenda

Coverage and CPI

- Modernization of CPI process to reduce biases and errors (Smith, 2021)
- Geographical dimension is key to assess CPI soundness (Berry et al., 2019)
- Prices may vary widely across space (Aten, 1996; Biggeri et al., 2017; Montero et al., 2020; Rao, 2001)

However:

- No consensus on the optimal degree of spatial disaggregation (Diewert, 2021)
- Limited attention devoted so far to CPI coverage in the literature (Diewert, 2021; Guerreiro et al., 2022; Hawkes and Piotrowski, 2003)

Objective

Our aim: propose a coverage metric for CPI data collection

- Represent geographical and population coverage
- Enable comparisons between areas of different extensions and population
- Enable weighted aggregations for the coverage metric

In addition:

- Demonstrate impact on CPI soundness when coverage experiences an abrupt change

Assumptions

The growth of e-shopping is generating both competing and complementing dynamics, with an impact on shopping travel pattern, e.g. duration, distance, (Shi et al., 2019; Le et al., 2022) and in-store landscape (Maat and Konings, 2018; Shah et al., 2021)

- Consumers are willing to travel for a limited time when doing purchases.
- Maximum acceptable travel time may vary according to frequency and magnitude of purchases.
- E-commerce shops have limited delivery range on certain categories of goods, in particular groceries.
- Consumer prices differ across geographical areas

→ Validity of price information in a certain location decays with travel distance

Fuzzy Coverage Index - Membership function

Geospatial fuzzy index (Zadeh, 1977; Zimmermann, 2011) to represent CPI data collection coverage. The Fuzzy set theory does not require the identification of a clear cut-off line thus taking into account the problem of imprecision.

Linear specification

$$lc(x) = max(1 - \frac{x}{D}, 0) \tag{1}$$

Inverse sigmoid specification

$$c(x) = 1 - \frac{1}{1 + e^{-k(x - \frac{D}{2})}} \tag{2}$$

Fuzzy Coverage Index - Aggregation

Unweighted aggregation

$$C_{mun} = \frac{\sum_{i=1}^{n} c_i}{n} \tag{3}$$

Population-weighted aggregation

$$C_{pop} = \frac{\sum_{i=1}^{n} c_i * pop_i}{\sum_{i=1}^{n} pop_i} \tag{4}$$

Spatio-temporal Price Indices

Time-interaction-Region Product Dummy (TiRPD) model (Aizcorbe and Aten, 2004)
➔ Ability to derive spatial and temporal parities in a single equation.

$$lnP_{ijt} = \sum_{i=1}^{N} \beta_i D_{ijt} + \sum_{t=1}^{T} \sum_{j=1}^{M} \delta_{jt} R_{ij} T_{jt} + \eta_{ijt} \tag{5}$$

The TiCPD provides the same answers as separate CPD or TPD models, with the advantage that it normalizes the relationships on a single region and time period.

Abrupt change indicator

Bayesian Estimator of Abrupt change, Seasonal change, and Trend (BEAST) (Zhao et al., 2019)

- Decomposition of time series into multiple trend and season signals
- Probability for each of the time series point to be an abrupt change point

Beast model:

$$y_i = T(t_i; \Theta_t) + \varepsilon_i \tag{6}$$

Trend change points are implicitly encoded in Θ_t.
Trend component in each segment:

$$T(t) = a_j + b_j t \text{ for } \tau_j \leq t < \tau_{j+1}, j = 0, ..., m \tag{7}$$

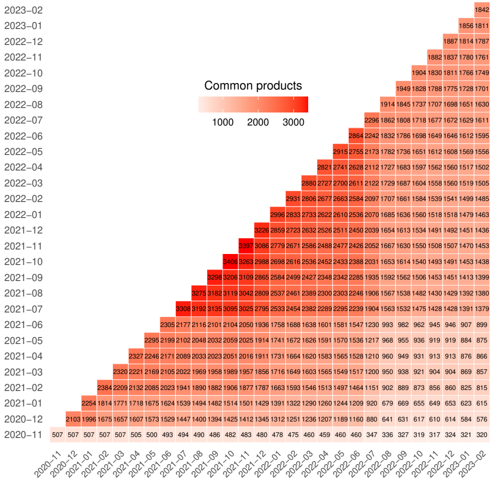
Grocery prices web scraping

- November 2020 to February 2023
- 23 online supermarket chains
- 616 outlets with GPS coordinates
- 19 Italian regions

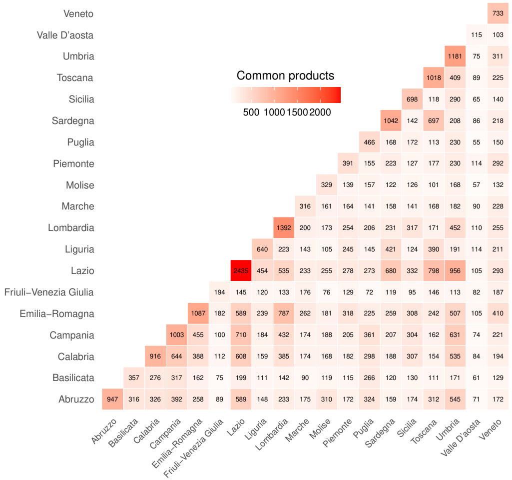
Coffee category (ECOICOP code 01.2.1.1)

- Average Italian household monthly expenditure: 11.91 EUR
- HICP weight: from 0.38% in 2020 to 0.43% in 2023
- 5338 unique products (2056 identified with GTINs)
- 1221755 total observations

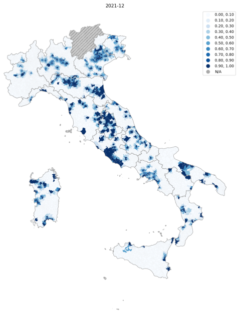
Common products across time

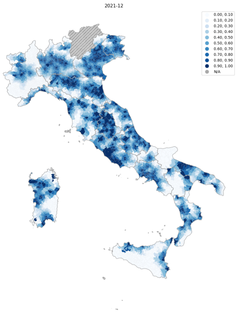


Common products across regions

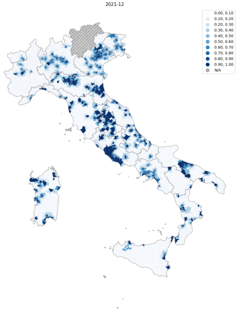


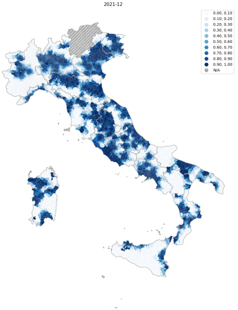
Coverage - Linear function at selected distance values





Coverage - Inverse sigmoid function at selected distance values





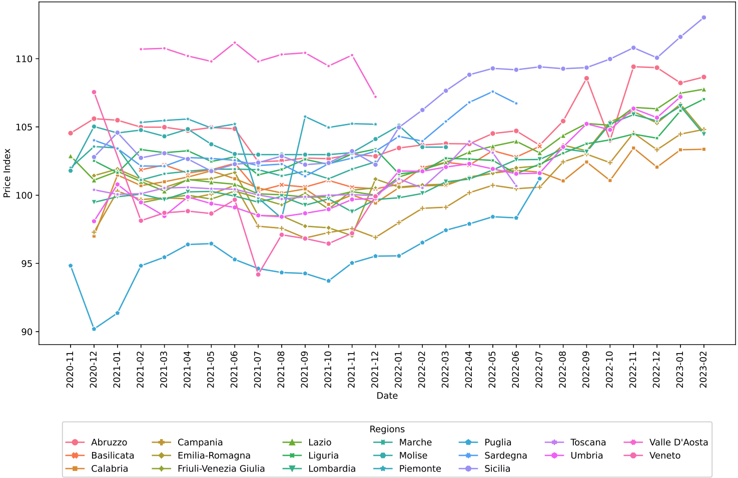
Coverage index stability

Spearman Rank correlation between pairs of indices with different D parameters and weighting methods:

- D parameters: 20 mins, 30 mins, 40 mins, 50 mins
- Correlation always positive and significant
- Correlation always > 0.86 within the same weighting method (municipalities or population)

→ **Proposed coverage index is stable regardless of the parameters or specification chosen.**

TiRPD indices



Coverage and TiRPD abrupt changes correlation

Region	Correlation	p-value
Abruzzo	0.359	*(0.061)*
Basilicata	0.399	*(0.101)*
Calabria	0.142	*(0.481)*
Campania	0.735	*(0.000)*
Emilia-Romagna	0.010	*(0.961)*
Friuli-Venezia Giulia	1.000	*(0.000)*
Lazio	-0.078	*(0.695)*
Liguria	0.410	*(0.034)*
Lombardia	-0.048	*(0.813)*
Marche	0.815	*(0.000)*
Molise	0.993	*(0.000)*
Piemonte	0.994	*(0.000)*
Puglia	0.300	*(0.186)*
Sardegna	-0.047	*(0.848)*
Sicilia	-0.033	*(0.868)*
Toscana	0.954	*(0.000)*
Umbria	0.554	*(0.003)*
Valle d'Aosta	0.999	*(0.000)*
Veneto	0.919	*(0.000)*

→ Significant correlation between abrupt changes in coverage and TiPRD indices.

Conclusions

- Abrupt changes in price data collection coverage can have significant impact on the CPI
- Coverage can provide relevant insights on the CPI quality
- National Statistical Institutes should consider calculating and publishing coverage metrics as complementary information for their CPI statistics.

Thank you for your attention

laureti@unitus.it
luigi.palumbo@bancaditalia.it

References I

Aizcorbe, A. and Aten, B. (2004). An approach to pooled time and space comparisons. SSHRC Conference on Index Number Theory and the Measurement of Prices and Productivity, Vancouver, Canada.

Aten, B. (1996). Evidence of spatial autocorreilation in international prices. *Review of Income and Wealth*, 42(2):149–163.

Berry, F., Graf, B., Stanger, M. M., and Ylä-Jarkko, M. (2019). Price statistics compilation in 196 economies: The relevance for policy analysis. *International Monetary Fund Working Papers*, 2019.

Biggeri, L., Laureti, T., and Polidoro, F. (2017). Computing sub-national PPPs with CPI data: an empirical analysis on Italian data using country product dummy models. *Social Indicators Research*, 131(1):93–121.

Diewert, W. (2021). Elementary indexes. *Consumer Price Index Theory*.

Guerreiro, V., Baer, M. A., and Silungwe, A. (2022). *The Availability, Methodological Soundness, and Scope of Consumer Price Statistics in 2020*. International Monetary Fund.

Hawkes, W. J. and Piotrowski, F. W. (2003). Using scanner data to improve the quality of measurement in the consumer price index. In *Scanner data and price indexes*, pages 17–38. University of Chicago Press.

References II

Le, H. T., Carrel, A. L., and Shah, H. (2022). Impacts of online shopping on travel demand: a systematic review. *Transport Reviews*, 42(3):273–295.

Maat, K. and Konings, R. (2018). Accessibility or innovation? store shopping trips versus online shopping. *Transportation Research Record*, 2672(50):1–10.

Montero, J.-M., Laureti, T., Mínguez, R., and Fernández-Avilés, G. (2020). A stochastic model with penalized coefficients for spatial price comparisons: An application to regional price indexes in Italy. *Review of Income and Wealth*, 66(3):512–533.

Rao, D. S. P. (2001). Weighted EKS and generalised CPD methods for aggregation at basic heading level and above basic heading level. In *Joint World Bank-OECD seminar on Purchasing Power Parities, Recent Advances in Methods and Applications*, Washington DC.

Shah, H., Carrel, A. L., and Le, H. T. (2021). What is your shopping travel style? heterogeneity in us households' online shopping and travel. *Transportation Research Part A: Policy and Practice*, 153:83–98.

Shi, K., De Vos, J., Yang, Y., and Witlox, F. (2019). Does e-shopping replace shopping trips? empirical evidence from chengdu, china. *Transportation Research Part A: Policy and Practice*, 122:21–33.

References III

Smith, P. A. (2021). Estimating sampling errors in consumer price indices. *International Statistical Review*, 89(3):481–504.

Zadeh, L. A. (1977). Fuzzy sets and their application to pattern classification and clustering analysis. In *Classification and clustering*, pages 251–299. Elsevier.

Zhao, K., Wulder, M. A., Hu, T., Bright, R., Wu, Q., Qin, H., Li, Y., Toman, E., Mallick, B., Zhang, X., and Brown, M. (2019). Detecting change-point, trend, and seasonality in satellite time series data to track abrupt changes and nonlinear dynamics: A Bayesian ensemble algorithm. *Remote Sensing of Environment*, 232:111181.

Zimmermann, H.-J. (2011). *Fuzzy set theory—and its applications*. Springer Science & Business Media.