A geostatistical fuzzy index to measure geographical and population coverage of consumer prices collection: The case of groceries webscraping in Italy

Luigi Palumbo[a*], Tiziana Laureti[a], Gianni Betti[b], Ilaria Benedetti[a]

[a]Università degli Studi della Tuscia, Viterbo, Italy

[b]Università degli Studi di Siena, Siena, Italy

[*]Corresponding author: Luigi Palumbo; luigi.palumbo@unitus.it

Abstract

Prices collection for consumer price indices (CPI) compilation has come a long way in the past 20 years. While ideally, the index should include expenditure made by all households, urban and rural, throughout the country, usually CPIs in various countries had limited geographic coverage both for price collection and consumption expenditures. The introduction of new data sources, such as webscraping and scanner data, have contributed to reduce price collection costs and increase the reach across national territories, thus allowing to enhance the accuracy and quality of the CPI. The geographical dimension, which is related to the scope of the index, becomes more important the smaller the region to which the index relates. While classifications of CPI geographical coverage usually refer to administrative areas (regional, capital city, urban-rural) a finer measurement would provide price statisticians with better insights on the actual reach of data collection. This would be particularly useful in cases where prices vary substantially across space, as it is proven that consumers only travel within limited extents for their purchases and a sparse network of outlets may lead to biased measurements. We propose a geostatistical fuzzy index to measure the reach of data collection in terms of geographical and population coverage of outlets where prices are collected. This index uses a fuzzy membership function to calculate the coverage value for each municipality, inversely proportional the driving distance in minutes from the closest outlet where prices have been collected. Total coverage value for a given territory is calculated as mean of each municipality coverage value, either simple or weighted by each municipality population. Using a dataset deriving from geo-localized groceries webscraping in Italy, we provide a practical application calculating coverage at a regional level and comparing results from two different functional forms – linear and non-linear – as well as a set of different parameters for spatial decay of coverage. Our findings corroborate the robustness of the index, as rank correlations amongst different parameters or functional values are close to 1 and statistically significant. We believe this index may be used to evaluate the degree of coverage for price data collection, both in the context of probabilistic and non-probabilistic outlet selection.

Keywords: geographical coverage; geostatistics; fuzzy logic; prices; webscraping.

1 Introduction

Prices collection for consumer price indices (CPI) compilation has come a long way in the past 20 years. While ideally, the index should include expenditure made by all households - urban and rural - throughout the country, usually CPIs in various countries had limited geographic coverage both for price collection and consumption expenditures, consequently CPI is a sample statistic that represents the change in prices over the target universe in the two periods (International Monetary Fund, 2020).

Due to the fact that it is impossible to regularly record all the prices of the universe, sampling techniques are used to select a subset of prices that enter the CPI compilation. Consequently, a CPI compilation is based on probabilistic and non probabilistic samples. The sampling process occurs on geographical location, outlet type, products and time dimensions. Within each of the different sampling levels, the sampling approach can differ from country to country, reflecting different administrative arrangements. Either probability or non-probability sampling methods can be considered in each dimension.

Unless it is possible to sample outlets directly from a national sampling frame such as a business register (which often cannot identify small outlets or the precise range of products available in them), the sampling of outlets generally needs to be done in two stages. In the first stage, a sample of locations such as cities or shopping areas is drawn/selected in each region of the country, and in the second stage outlets are sampled. When sampling locations, two major factors must be borne in mind: representativeness and cost effectiveness. Areas where the

bulk of consumer purchases take place need to be covered with certainty or by a probability sample to make the sample representative. In small countries it is common to select a few of the larger cities for price collection. This leaves out smaller towns and rural areas, but as consumers living in areas close to city will go there for some of their shopping the effect of their exclusion will be smaller than might be inferred from population numbers, and a sufficient coverage may still be achieved

Geographical coverage, which refers to either the coverage of expenditure or the coverage of price collection of the CPI (International Monetary Fund, 2020), is a key component in assessing the methodological soundness (Berry, Graf, Stanger, & Ylä-Jarkko, 2019). Many countries have CPIs with limited geographic coverage — capital city, including few of the largest areas (such as large and medium-sized cities) and prices are collected in urban areas only because their movements are considered to be representative of the price movements in rural areas.

The geographical dimension, which is related to the scope of the index, becomes more important the smaller the region to which the index relates. While classifications of CPI geographical coverage usually refer to administrative areas (regional, capital city, urban-rural) a finer measurement would provide price statisticians with better insights on the actual reach of data collection. Since CPI compilation is becomes more important in economic planning and inflation monitoring, efforts should be made to expand the CPI to cover more geographic areas including all urban and rural areas. This would be particularly useful in cases where prices vary substantially across space, as it is proven that consumers only travel within limited extents for their purchases and a sparse network of outlets may lead to biased measurements.

The popularity and availability of new data sources for the compilation of the CPI, such as web-scraping and scanner data, has increased over the past twenty years and have contributed to reduce price collection costs and increased the reach across national territories, thus allowing to enhance the accuracy and quality of the CPI (Brunetti, Fatello, Polidoro, & Simone, 2018). Although the compilation of price indices from such large datasets is not straightforward, these new sources of data have proved to be of benefit to CPIs thanks to the detailed information available for individual products (product characteristics, quantity sold, etc.), the wide coverage both in terms of product groups and territorial areas, the opportunity to implement superlative index and greater precision or lower variance.

Using a dataset deriving from geo-localized groceries webscraping in Italy, we provide a practical application calculating coverage at a regional level and comparing results from two different functional forms – linear and non-linear – as well as a set of different parameters for spatial decay of coverage. Our findings corroborate the robustness of the index, as rank correlations amongst different parameters or functional values are close to 1 and statistically significant. We believe this index may be used to evaluate the degree of coverage for price data collection, both in the context of probabilistic and non-probabilistic outlet selection. Concerning the methodology, we adopted a geostatistical fuzzy index (Zadeh, 1977; Zimmermann, 2011) to measure the reach of data collection in terms of geographical and population coverage of outlets where prices are collected. This index uses a fuzzy membership function to calculate the coverage value for each municipality, inversely proportional the driving distance in minutes from the closest outlet where prices have been collected. Total coverage value for a given territory is calculated as mean of each municipality coverage value, either simple or weighted by each municipality population. A properly designed membership functions may enable us to achieve a better classification of the data, smoothing distortions caused by outliers while still including them into the analysis. Another advantage of the fuzzy set theory approach is to overcome the limits of discrete classifications of data, preserving a higher degree of information for analysis.

2 Methodology

The fundamental concept behind our proposed measure of coverage for price collection is that price information decays with space and travel time. Consumers may travel for certain distances and time to make purchases, thus providing an incentive for sellers to maintain competitive prices in different municipalities. However, consumers' inclination to commit time and money for purchasing trips is directly connected to the expected economic benefit in terms of savings.

Given the average basket value for groceries shopping, it is reasonable to affirm that there are limits to shopping trips distances, even if those may vary between consumers because of different travel costs, cost-opportunity of travel time and other individual characteristics.

Empirical evidence of spatial effects underlying consumer price differences among geographical areas have been observed both at country and sub-national level (Aten, 1996; Rao, 2001; Biggeri, Laureti, & Polidoro, 2017; Montero, Laureti, Mínguez, & Fernández-Avilés, 2020)

Therefore, we need to conclude that prices may be different between municipalities situated at a certain distance, and the information value of collected prices in a certain location will decay with space and travel time.

In order to appropriately model this decay we resort to Fuzzy Set theory, as it seems inappropriate to specify hard boundaries regarding the validity of price information in binary terms. We then propose two different membership functions to calculate the coverage value for each municipality. The first one is a simple linear function, where coverage is inversely proportional to travel distance.

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

Where x is the travel time by car in minutes between a municipality and the closest municipality where prices have been collected, and D is a parameter indicating at which travel time level the price information is considered no longer valid.

The second type of membership function is based on an inverse sigmoid modeling of price information decay. In fact, it is reasonable to assume that consumer willingness to travel for purchases is not linear, therefore price persistence in space is relatively stronger at short distances and weaker at longer ones. We can then propose a different membership function as follow:

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

Where x is again the travel time by car in minutes between a municipality and the closest municipality where prices have been collected, D is a parameter indicating at which travel time level the price information is considered no longer valid (and $\frac{D}{2}$ is the midpoint of the inverse sigmoid), and k is a parameter indicating the steepness of the inverse sigmoid curve.

Once we calculate coverage values for all municipalities in a region, we need to synthesize a metric to indicate the overall coverage for the region. In order to do so, we can aggregate individual municipalities as units or by weighting them according to their population.

If we chose to treat municipalities as individual units, the coverage for a given region could be expressed as a simple arithmetic mean as in 3.

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

On the other side, if we chose to weight coverage in each municipality by its population the overall Region coverage would be:

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

Formulas for coverage are applied at a regional level, since the main purpose is to provide a coverage metric for sub-national Consumer Price Indexes (CPIs) over space and time.

3 Data

Data used for the empirical validation of our proposed methodology has been scraped from 464 online supermarkets belonging to 14 different chains in 19 Italian Regions during September 2021[1]. Each online supermarket has been located with GPS coordinates and placed in a specific municipality using geographical merging functions. We collected prices for each supermarket using the "pick up" option for purchase delivery. Therefore, validity of price information is considered linked to its geographical position.

Driving distances between municipalities have been obtained from a distance matrix published by the Italian National Institute of Statistics (Istat). Istat calculated the driving distance between centroids for all Italian municipalities in 2013 using a commercial road graph (Istat, 2019). We performed a basic elaboration in order to adjust for merging between small municipalities in the 2013-2021 period, also excluding minor islands and municipalities disconnected from the road graph[2]. The total amount of population living in excluding municipalities is marginal when compared to the relative Region population.

4 Results

In tables 1 to 4 we report the results for coverage by Region calculated according to the different membership functions presented in (1) and (2), using both individual municipalities and their population in order to calculate the overall regional coverage. Figures 1 to 4 are graphical representations of municipalities' coverage values according to the above mentioned membership functions at selected values of D.

[1] No data has been collected for the Trentino-Alto Adige region.

[2] Municipalities of Monte Isola (BS) and Campione d'Italia (CO) do not have any connection with the road graph used for distance calculation. Istat only provides distance from the closest municipality for them. For minor islands Istat provides a travel time by ferry to the closest port.

Table 1: Coverage value by Region and different values of D
Linear Membership Function - Individual municipalities.

Region	50 min	40 min	30 min	20 min
Piemonte	0.29218	0.21942	0.14842	0.08021
Valle d'Aosta	0.69496	0.61870	0.49571	0.31731
Lombardia	0.55029	0.46168	0.34909	0.21715
Veneto	0.65666	0.58342	0.48814	0.34745
Friuli-Venezia Giulia	0.68383	0.60966	0.50085	0.33339
Liguria	0.36113	0.26454	0.16302	0.07881
Emilia-Romagna	0.60585	0.52594	0.41919	0.28257
Toscana	0.19177	0.13226	0.07506	0.03210
Umbria	0.50411	0.40526	0.28376	0.16569
Marche	0.57025	0.46779	0.33907	0.19591
Lazio	0.47404	0.38090	0.26861	0.15694
Abruzzo	0.59122	0.49945	0.38058	0.23928
Molise	0.34682	0.25933	0.15944	0.07382
Campania	0.51522	0.43984	0.35138	0.25011
Puglia	0.38707	0.28523	0.19086	0.11255
Basilicata	0.31025	0.22515	0.13734	0.06901
Calabria	0.46548	0.37478	0.25849	0.14435
Sicilia	0.31868	0.24579	0.16448	0.08847
Sardegna	0.46519	0.37716	0.28565	0.19339

Table 2: Coverage value by Region and different values of D
Sigmoid Membership Function - Individual municipalities.

Region	50 min	40 min	30 min	20 min
Piemonte	0.2865	0.21476	0.14635	0.08458
Valle d'Aosta	0.78896	0.65572	0.49880	0.32831
Lombardia	0.59117	0.47832	0.35344	0.22481
Veneto	0.72609	0.62674	0.49970	0.35611
Friuli-Venezia Giulia	0.77372	0.65564	0.50814	0.34375
Liguria	0.34134	0.24260	0.15592	0.08573
Emilia-Romagna	0.65795	0.54581	0.42036	0.29072
Toscana	0.17094	0.11422	0.06918	0.03500
Umbria	0.51390	0.39232	0.27774	0.17647
Marche	0.60164	0.46826	0.33379	0.20597
Lazio	0.48450	0.37217	0.26301	0.16553
Abruzzo	0.63666	0.51543	0.38206	0.24961
Molise	0.33771	0.24074	0.15277	0.07979
Campania	0.54536	0.46004	0.36172	0.25215
Puglia	0.36103	0.26716	0.18470	0.11779
Basilicata	0.28836	0.20405	0.13136	0.07547
Calabria	0.47939	0.36370	0.25287	0.15206
Sicilia	0.31465	0.23172	0.15801	0.09342
Sardegna	0.47015	0.38046	0.28763	0.19860

Table 3: Coverage value by Region and different values of D
Linear Membership Function - Municipalities weighted by population.

Region	50 min	40 min	30 min	20 min
Piemonte	0.52764	0.47135	0.40994	0.34002
Valle d'Aosta	0.72720	0.65900	0.54587	0.34486
Lombardia	0.73997	0.68081	0.59267	0.45858
Veneto	0.70509	0.63459	0.53309	0.38641
Friuli-Venezia Giulia	0.65340	0.57173	0.49491	0.36215
Liguria	0.61105	0.55418	0.49100	0.42899
Emilia-Romagna	0.79169	0.74292	0.67086	0.56743
Toscana	0.33048	0.23505	0.13823	0.08521
Umbria	0.73736	0.67749	0.60297	0.51917
Marche	0.72002	0.65089	0.54590	0.39817
Lazio	0.81844	0.78639	0.74789	0.70363
Abruzzo	0.79657	0.74859	0.68223	0.59085
Molise	0.47338	0.38694	0.30057	0.23043
Campania	0.78030	0.73413	0.66684	0.56630
Puglia	0.59796	0.51552	0.41808	0.33844
Basilicata	0.51367	0.44390	0.36988	0.30448
Calabria	0.61839	0.55468	0.47422	0.39424
Sicilia	0.56974	0.52419	0.46586	0.39793
Sardegna	0.65002	0.59322	0.52793	0.45692

Table 4: Coverage value by Region and different values of D
Sigmoid Membership Function - Municipalities weighted by population.

Region	50 min	40 min	30 min	20 min
Piemonte	0.54207	0.47999	0.41248	0.34295
Valle d'Aosta	0.86241	0.73612	0.55706	0.35872
Lombardia	0.81625	0.72396	0.60372	0.46554
Veneto	0.77906	0.67199	0.54185	0.39419
Friuli-Venezia Giulia	0.71412	0.62374	0.50939	0.36775
Liguria	0.60669	0.53942	0.48138	0.43459
Emilia-Romagna	0.84696	0.77104	0.67590	0.57318
Toscana	0.28481	0.19753	0.13354	0.08785
Umbria	0.75718	0.68064	0.60112	0.52594
Marche	0.79560	0.68429	0.55061	0.40839
Lazio	0.83069	0.79220	0.74970	0.70570
Abruzzo	0.84162	0.77516	0.6888	0.59763
Molise	0.45239	0.36579	0.29405	0.23271
Campania	0.84495	0.78111	0.68819	0.56559
Puglia	0.58973	0.49690	0.41446	0.34306
Basilicata	0.50259	0.43069	0.36471	0.31123
Calabria	0.62960	0.54898	0.47093	0.39917
Sicilia	0.58741	0.52710	0.46402	0.40337
Sardegna	0.66540	0.60167	0.53221	0.46096

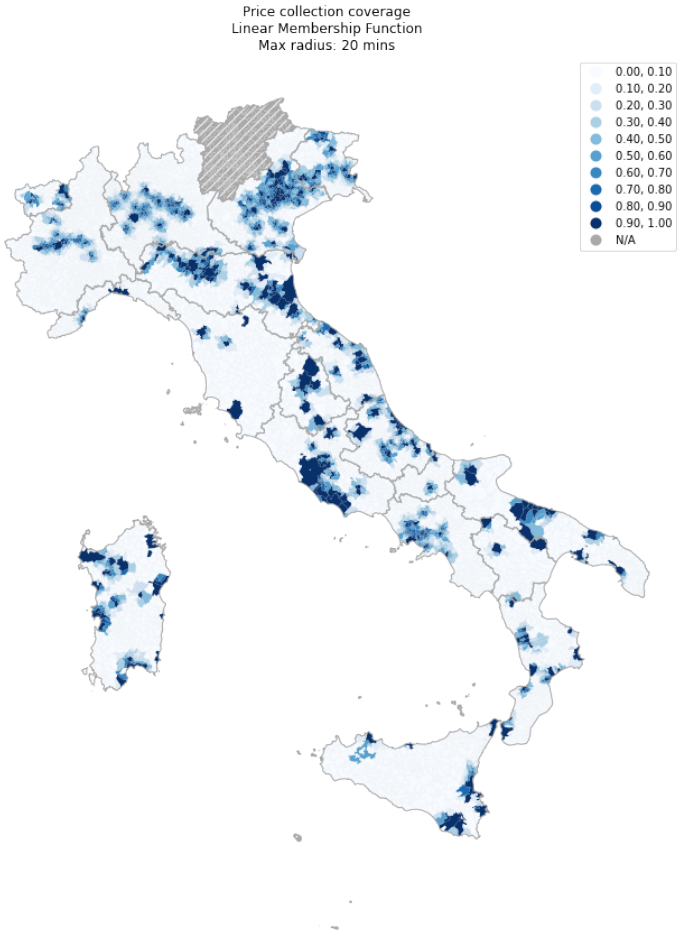

Figure 1: Coverage representation - Linear membership function - D: 20 min.

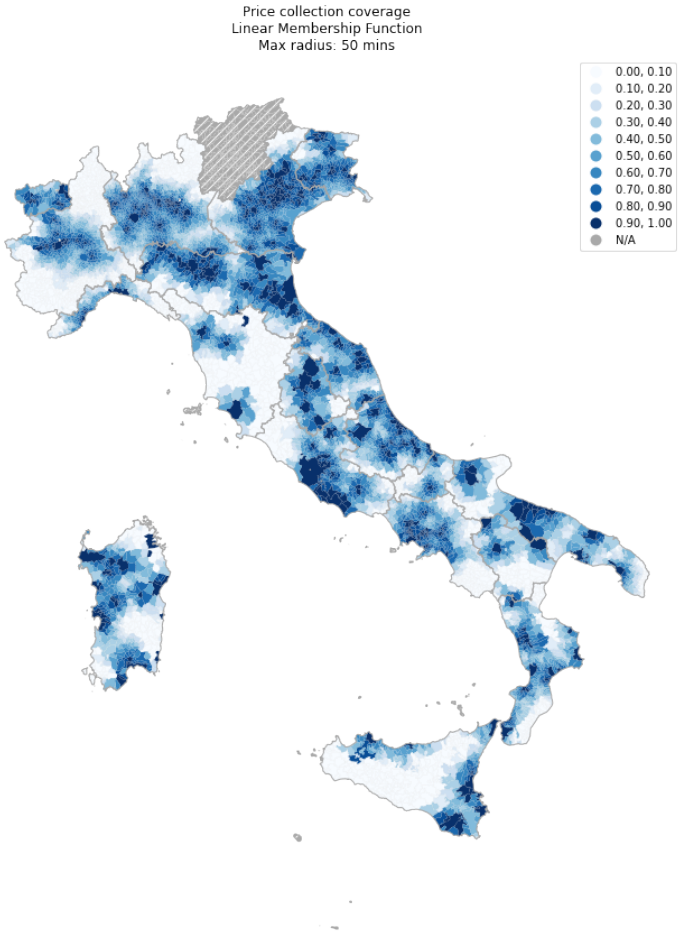

Figure 2: Coverage representation - Linear membership function - D: 50 min.

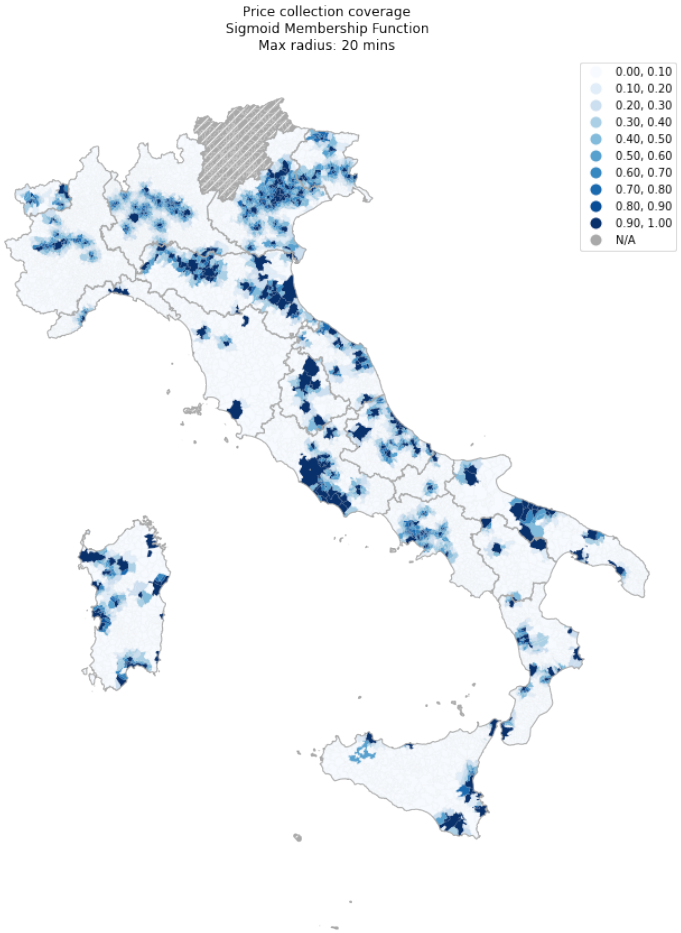

Figure 3: Coverage representation - Sigmoid membership function - D: 20 min.

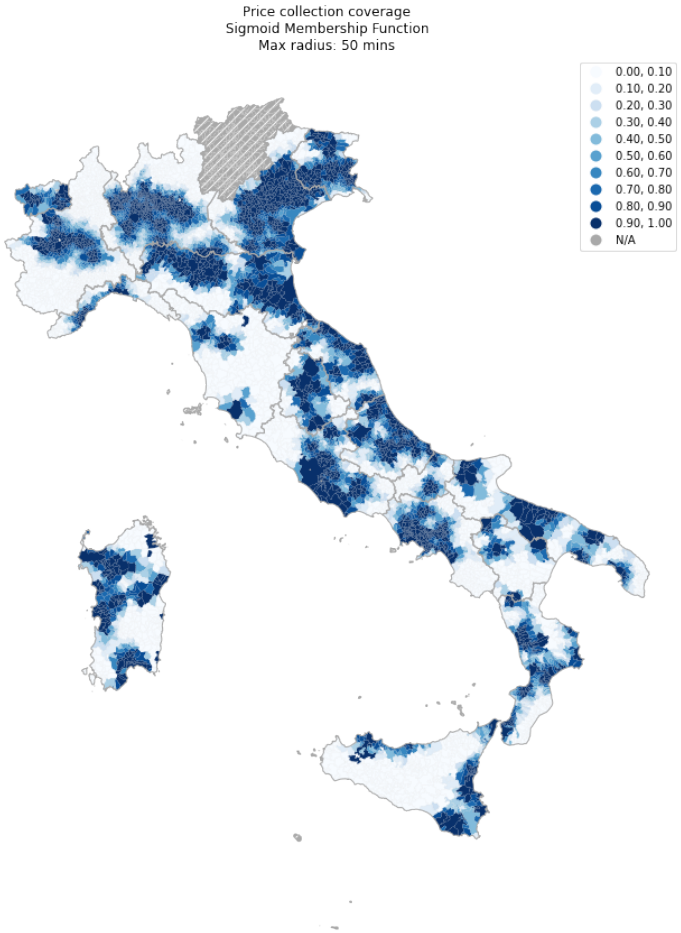

Figure 4: Coverage representation - Sigmoid membership function - D: 50 min.

In order to evaluate the stability and consistence of our coverage metrics, we performed a series of measurement leveraging the Spearman Rank Correlation non-parametric test on the coverage values calculated for each region, as presented in Tables 1 to 4. The first set of tests was focused on rank correlation between consecutive pairs of D parameters when using the same membership function and aggregation method. Results are presented in Table 5. The second set of test was focused on rank correlation between same values of D using the same aggregation method and different membership functions. Results are presented in Table 6.

Table 5: Spearman Rank Correlation Test between consecutive pairs of D

Membership Function	Aggregation	50-40 min	40-30 min	30-20 min
Linear	Municipalities	0.99825	0.97895	0.98772
		(0.000)	*(0.000)*	*(0.000)*
Linear	Population	0.99649	0.99474	0.91228
		(0.000)	*(0.000)*	*(0.000)*
Sigmoid	Municipalities	0.99123	0.98246	0.99474
		(0.000)	*(0.000)*	*(0.000)*
Sigmoid	Population	0.96316	0.98596	0.90175
		(0.000)	*(0.000)*	*(0.000)*

Table 6: Spearman Rank Test Correlation between different Membership Functions

Aggregation	D	Linear-Sigmoid
Municipalities	20 min	0.99825
		(0.000)
	30 min	0.99825
		(0.000)
	40 min	0.99474
		(0.000)
	50 min	1.00000
		(0.000)
Population	20 min	0.99825
		(0.000)
	30 min	0.99474
		(0.000)
	40 min	0.98421
		(0.000)
	50 min	0.93860
		(0.000)

All results report a strong and significant positive correlation between rankings, indicating that our proposed indicator can deliver robust and consistent results irrespective of the parameters chosen.

5 Conclusions and future research

We believe coverage information is a relevant metric for price statistics. Modelling accurately where price collection happens and embedding this information in CPIs can provide tremendous insights at several levels in the price statistics compilation and utilization process.

During selection and sampling of outlets for price collection it would be important to have an accurate view of geographical and population coverage in order to make sure that no dark spot is left systematically in price surveys and there is continuity and consistent overlap over time for the covered area.

When using price statistics this coverage view would be equally important. Local dynamics in economic and social measures are object of a growing number of studies, and granular coverage information could help to better integrate price statistics in this stream of research.

We plan to apply our methodology to a larger dataset of geolocalized prices from webscraping, from December 2019 onward, and provide the coverage information as complement to our monthly CPIs.

Other services may be explored for obtaining updated travel time calculations in the future, such as Google Distance Matrix API or TravelTime API. Furthermore, we foresee additional application for a measure of information decay over space or travel time as presented in this work.

Funding Statement

This research received no external funding.

Conflicts of Interest

Authors declare no conflict of interest.

References

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., & Ylä-Jarkko, M. (2019). Price statistics compilation in 196 economies: The relevance for policy analysis. *International Monetary Fund Working Papers*, *2019*. DOI: hiips://doi.org/10.5089/9781513508313.001

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. *Social Indicators Research*, *131*(1), 93–121.

Brunetti, A., Fatello, S., Polidoro, F., & Simone, A. (2018). Improvements in Italian CPI/HICP deriving from the use of scanner data. In *50th scientific meeting of the italian statistical society.* Retrieved from hiip://meetings3.sis-statistica.org/index.php/sis2018/50th/paper/viewFile/1484/32

International Monetary Fund. (2020). *Consumer price index manual.* London, England: International Monetary Fund.

Istat. (2019). *Matrici di contiguità, distanza e pendolarismo.* Retrieved 2022-04-20, from hiips://wwwistatit/it/archivio/157423

Montero, J.-M., Laureti, T., Mínguez, R., & 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.

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

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