Computation of consumer spatial price indexes over time using Natural Language Processing and web scraping techniques

Laureti T.[1], Benedetti I.[1], Palumbo L.[1], and Rose B.[2]

[1] University of Tuscia, [2] Starsift LLC

Outline

Background and aims

Web-scraped data and price statistics

- Current status
- Developments for Spatial Price Index computation

Methodological approach and data set

- NLP for product categorization
- Web-scraped data from US multichannel supermarkets
- Adding a time dimension: TiRPD models
- Estimation results

Conclusion

Background and aims

- The share of **electronic sales** to consumers in total turnover has greatly increased in most EU countries.

- The COVID-19 crisis has accelerated an expansion of e-commerce towards new firms, customers and types of products (OECD, 2020).

- A recent international survey showed that, following the pandemic, more than half of the respondents shop online more frequently than before (UNCTAD, 2020). Some of these changes in purchasing patterns will likely be of a long-term nature.

- The official statistics providers should not ignore this rich data source especially in the price statistics domain

Background and aims

Several EU countries have started to collect data from online retailers and use them for official CPI calculation.

As yet, to my knowledge, only few studies has been carried out on using web-scraped data for compiling consumer spatial prices indexes at international level (Cavallo, 2017, Cavallo et al, 2018)

- In this context web-scraped data may enable countries to construct regional or sub-national spatial prices indexes (SPIs)

The aims of this paper are to:

- Exploring the potential advantages of using web-scraped data for constructing sub-national SPIs
- Presenting first results obtained using TiRPD model and US multichannel supermarkets .

Web-scraped data in official price statistics

Statistics Finland	Istat – Istituto Nazionale di Statistica	STATBEL – Belgium in figures	Insee – Institut national de la statistique et des études économiques – Mesurer pour comprendre
Python API interfaces (Amadeus)	Web Browser Automation (IMacros)	R packages (Rvest, Rselenium)	Python Selenium Browser automation
Air tickets	Air Tickets, Train tickets, Consumer Electronics	Over 60 categories: Clothing, Air tickets, Hotels, Supermarkets,…	Laptops, Train tickets

Other NSIs leveraging web-scraping techniques for price statistics:
Statistics Netherlands, Statistics Austria , Statistics Slovenia, Statistics Luxemburg and Statistics Norway

- ***Eurostat 2020, Practical guidelines on web scraping for the HICP***

Using Web Scraped Data for computing Spatial Price Indexes

- The importance of spatial price comparisons within a country has been acknowledged in literature over the last decades (Laureti and Rao, 2018)

Advantages of using web-scraped price data

- ✓ Feasible alternative to scanner data (Chessa and Griffioen,2019)
- ✓ Daily frequency
- ✓ Collecting additional metadata (e.g. product characteristics)
- ✓ Aligned with offline prices (Cavallo, 2017)

Drawbacks:

- ❖ Product matching (ID codes may not be available)
- ❖ No quantities sold (lack of weights)
- ❖ Anti-scraping measures implemented by webmasters

Methodological Approach and data set

Product classification and matching

- **Natural Language Processing**

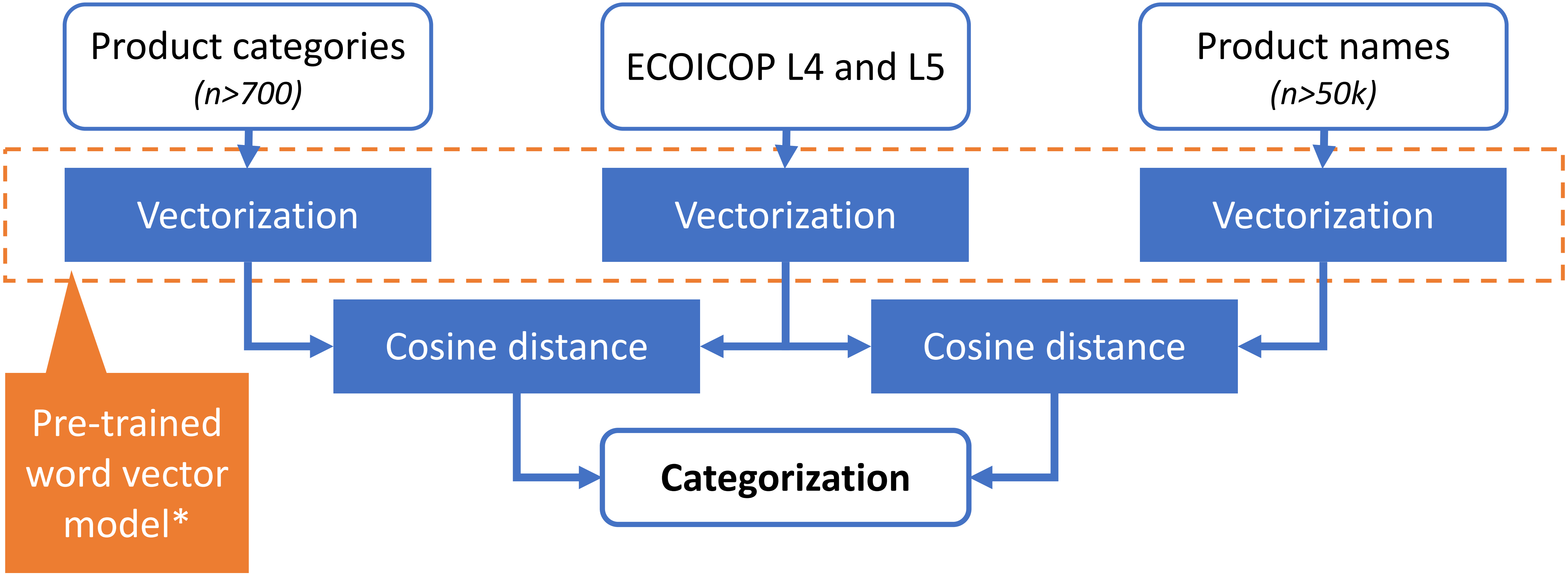
  Word and sentences vectorization for product classification and product matching

Adding a time dimension

- **Time-interaction-Region Product Dummy model (TiRPD)**

  Reconciling consumer price indexes across space and time (Aizcorbe and Aten, 2004; Dikanov, 2010) and using all the information available

NLP for product classification



* Vector model details at https://spacy.io/models/en-starters#en_vectors_web_lg

Methodological Approach and data set

US web-scraping dataset provided by Starsift LLC

~120 Mn data points

10.93% CPI-U Coverage

- Food at home
- Alcoholic beverages (at home)
- Housekeeping supplies
- Medicinal drugs (non-prescription)
- Cigarettes & Tobacco products
- Personal care products

4 retailer chains

11 US cities

Daily prices from January 2017 to May 2018

Monthly average prices using unweighted arithmetic mean

Methodology: Time-interaction-Region Product Dummy model

$$\ln p_{nrt} = \sum_{r=1}^{R}\sum_{t=1}^{T} \delta_{rt} D_{nr} T_{nt} + \sum_{n=1}^{N} b_n^{'} D_{rnt}^{*} + v_{nrt}$$

where, for each BH, p_{nrt} denotes the price of product *n* in area *r* at time t (*n* = 1, 2,…*N*; *r* = 1, 2,…, *R*; t=1,…,T)

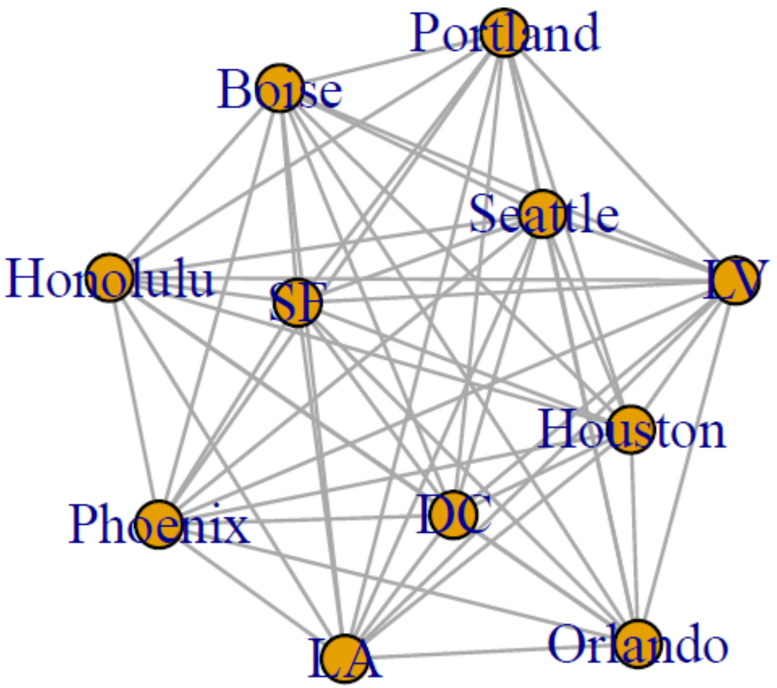
D_{rnt}^{*} are dummies for product n in area r at time t.

$D_{nr}T_{nt}$ are dummy variables for each combination of area and time period with *n=1,…,N; r=1,…,R* and *t=1,…,T.*

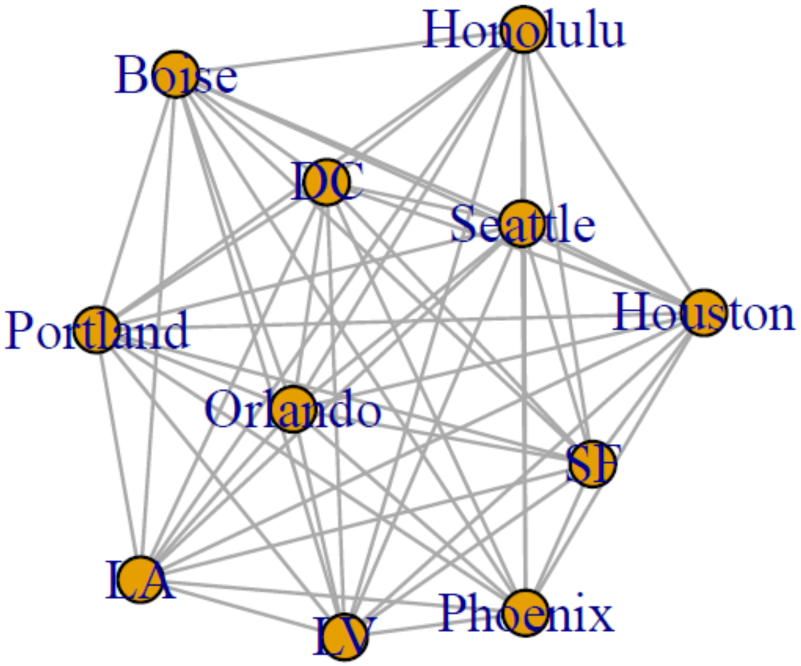
The intra-national SPI for the city *r* at time t is given by $\boxed{exp\left(\delta_{rt} - \delta_{Orlando,t=1}\right)}$

RESULTS: Product overlap across cities

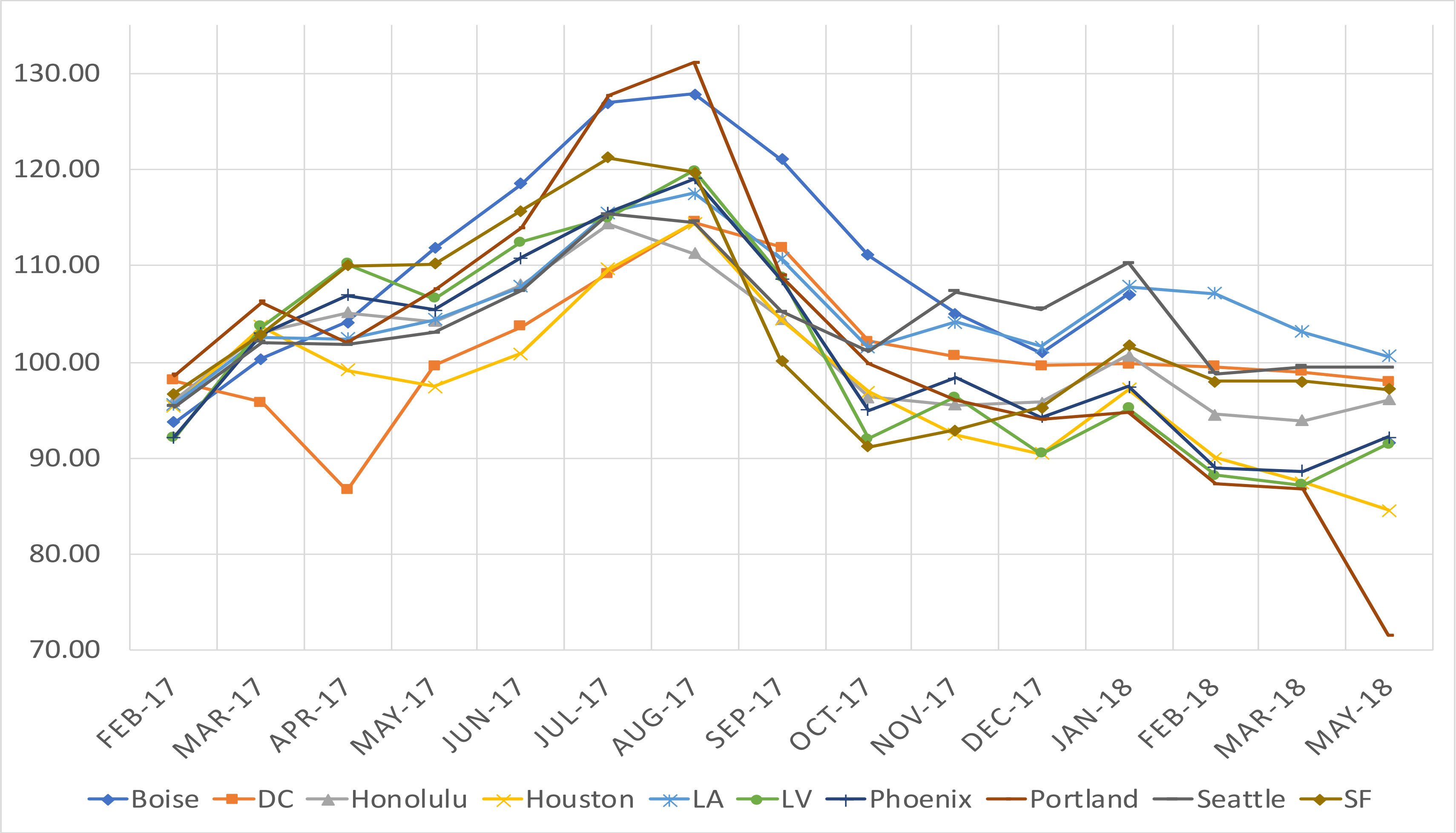
Chocolate



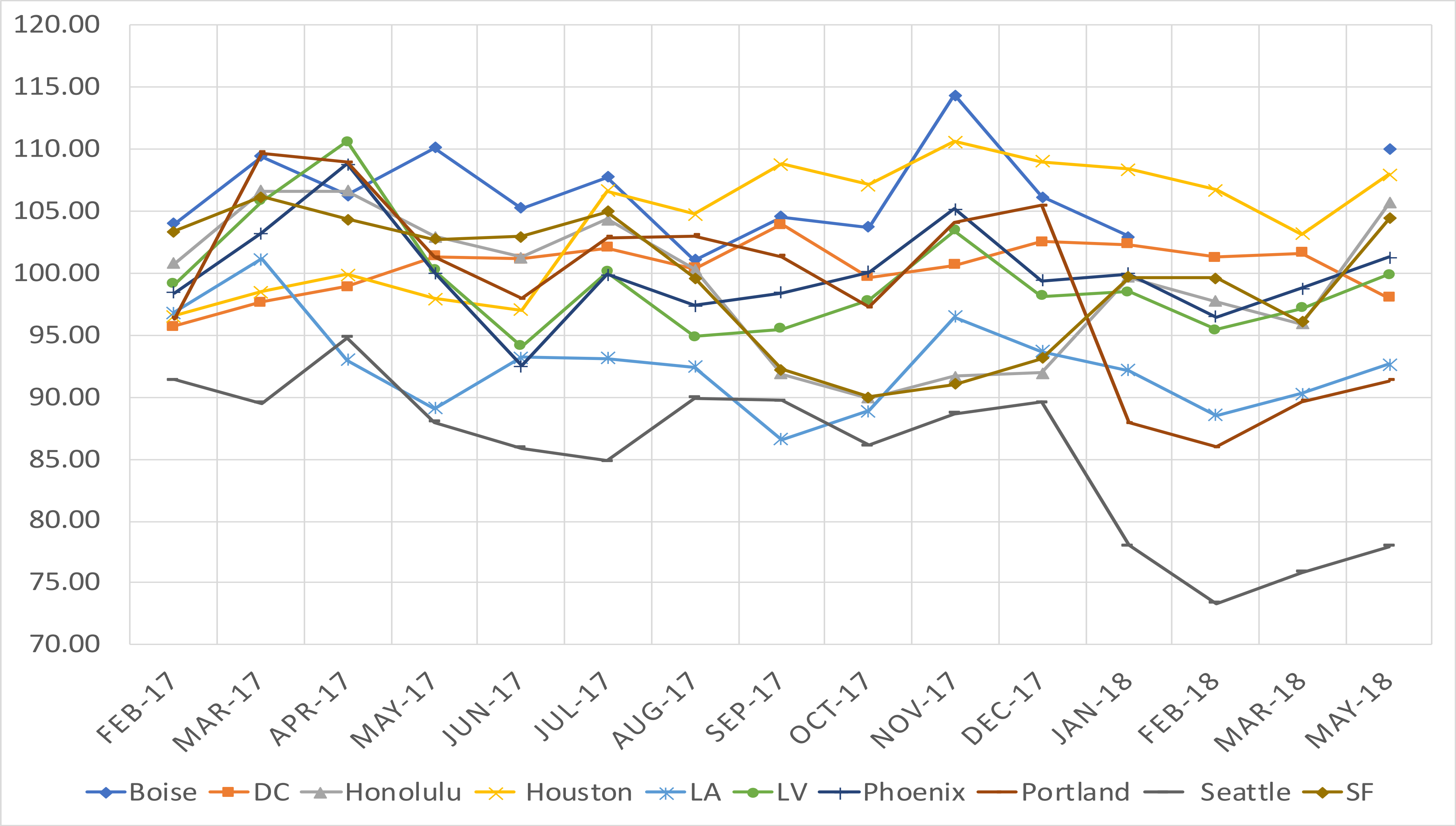
Apples



SPI estimation results - Apples



SPI estimation results - Chocolate



Concluding remarks and ongoing research

- ❑ On line data together with the use of TiRPD methods could allow the computation of sub-national SPI
- ❑Web-scraped data may reduce data collection burden and supplement information on items.

- ➢More IT resources are required due to the huge amount of data obtained
- ➢Not all countries may have access to this type of data
- ➢On line data my cover a limited number of product categories

Further research is underway for

- ❑ Web-scrabing prices for other product aggregates and in other countries
- ❑ Testing the use of other multilateral methods

Thank you for your attention

Prof. Tiziana Laureti (laureti@unitus.it)

Luigi Palumbo (luigi.palumbo@unitus.it)

Dr. Ilaria Benedetti (i.benedetti@unitus.it)

Brandon Rose(brandon@starsift.com)

References

Virgillito, A. Polidoro, F. (2019) “Big Data Techniques for Supporting Official Statistics: The Use of Web Scraping for Collecting Price Data”. In Web Services: Concepts, Methodologies, Tools, and Applications (pp. 728-744). IGI Global.

Eurostat (2020). Practical guidelines on web scraping for the HICP

Cavallo, A., Diewert, W. E., Feenstra, R. C., Inklaar, R., & Timmer, M. P. (2018). Using online prices for measuring real consumption across countries. In AEA Papers and Proceedings (Vol. 108, pp. 483-87).

Cavallo, A. (2017) “Are Online and Offline Prices Similar? Evidence from Multi- Channel Retailers,” American Economic Review 107, 283–303.

Gábor, K. Zargayouna, H. Tellier, I. Buscaldi, D. Charnois, T.(2017) “Exploring Vector Spaces for Semantic Relations,” Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 1814–1823. Copenhagen, Denmark, September 7–11.

Laureti, T. D.S. Prasada. Rao, (2019) “Measuring spatial price level differences within a country: Current status and future developments”. Studies of Applied Economics, 36(1), 119-148.

Aizcorbe, A., and Aten, B. (2004). An Approach to Pooled Time and Space Comparisons. In SSHRC Conference on Index Number Theory and the Measurement of Prices and Productivity, Vancouver, Canada.

Methodological Approach and data set

- BEA publishes RPPs yearly using several data sources
- **Base City for our study**: Orlando