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

Keywords: consumer spatial price indexes, data scraping, spatial index, time comparison, big data, NLP

1. INTRODUCTION

During the last years, the use of Big Data in official statistics has become a major topic of several initiatives both at national and international level [1]. Among all the possible types of Big Data sources, the "Internet as a data source" is a highly popular one due to the increased share of consumers electronic sales in total turnover. In this framework, the development of "web scraping" techniques as tools to capture big amount of price data proved to be useful for compiling Consumer Price Indices (CPIs) in several EU countries. However, National Statistical Institutes (NSIs) are becoming aware of the potential issues that may be encountered when using web-scraped data, which is aided by the recent Eurostat guidelines on web scraping for the HICP [2]. Contrastingly, to the authors' knowledge, only Cavallo et al [3] carried out a research study on using web-scraped data for constructing spatial consumer price indexes (SPIs) at international level. However, the wide availability of online information on product prices on a daily basis represents a great opportunity for computing consumer SPIs at sub-national level and analyse their evolution over time. Web-scraped data may be considered as a feasible alternative to scanner data which are collected by retailers as products are scanned through the till. Price and quantity information can then be derived from this transactional data. However, scanner data may not be available for smaller retailers and may not contained metadata information needed for constructing price statistics. Web scraping collects price information directly from the retailer's websites. A web scraper is a tool that reads the HTMLs on the website and extracts the data needed. Recent literature showed how online prices can be used as a close proxy of general prices [4], even if online transactions represent a small share of overall purchases and they are close to scanner data. In this paper we explore the use of web scraped data for compiling sub-national SPIs by using monthly average prices (arithmetic mean) for 93 categories according to COICOP level 4, covering 10.9% of CPI-U index, and track their evolution over time in 11 US cities reported in our dataset. Given space limitation, we only report results for 2 of those categories: Apples and Chocolate.

1.1. Dataset

This study leverages a dataset composed by scraped product prices[1], with over 120 million data points recorded between January 2017 and May 2018 across 11 US cities. The dataset presents a daily average of around 270k data points, representing around 50k unique items in about 750 commercial categories. There is a high variability of data points across days, as scraping routines were not always able to access and collect data. Data has been scraped by 4 online shops which belong to the same retail group. Categories are reported as assigned by each shop in its product classification, and mostly cover grocery products.

[1] The dataset has been gently provided by Starsift LLC.

2. METHODS

2.1. Data reclassification according to COICOP categories

We leverage a set of Natural Language Processing (NLP) techniques to rearrange the dataset in order to perform our study and calculate, in this first step of our study, Time-interaction-Region Product Dummy (TiRPD) for computing spatial price differences among cities and time in the US has been used. Each product has a category classification based on the online retailer category organization. In order to calculate price indexes for each category, the first step is to rearrange the retailer classification according to COICOP categories.

To this aim, we computed the word/sentence similarity between the retailer classification and labels for COICOP classification at level 4 and 5 [5], assigning the retailer category to the COICOP category with the highest similarity. The basis for the semantic similarity calculation is a pre-trained word vector model composed by 1.1 million unique vectors with over 300 dimensions.[2]

The first step is to assign a vector representation to each retailer category and COICOP category. Similarity between words/sentences pairs is calculated as the cosine similarity between their vector representation, returning a value between 0 and 1. After assigning the retailer category to the COICOP category (level 4 or level 5) with the highest similarity, we reassigned all COICOP level 5 categories to their parent COICOP level 4 category in order to obtain a homogeneous categorization of our data.

2.2. Consumer spatial price indexes calculation

Various methods can be used to produce spatial price indices [6]. In the case of web scraped data, the fact that there is no expenditure information limits the methods that can be used. We considered data on monthly basis, due to dataset size and limited availability of computational resources. As a first exploratory analysis of this large dataset, the price for each product in each city has been calculated as the arithmetic mean of prices observed across different retailers. We used the Time-interaction-Region Product Dummy (TiRPD) suggested by Aizcorbe and Aton [7]. The TiRPD is an extension of the CPD, which is an implementation of the hedonic approach accounting for the quality variations in price data. It provides a regression analysis-based econometric methodology for constructing multilateral price index numbers that accounts for the quality variations in the cross-area price data and time.

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

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 PPP for the area *r* at time t is given by $exp\left(\delta_{rt} - \delta_{Orlando,\, t=1}\right)$.

We applied this method by considering 15 months (January 2017 to May 2018) in order to monitor the evolution of spatial price differences. The results of this first analysis

[2] More details on the vector model at https://spacy.io/models/en-starters#en_vectors_web_lg

provide a basis for further statistical developments in the estimation of spatial price indexes using web scraped data including the possibility of measuring consumer price differential across area and over time using an econometric approach.

3. RESULTS

Using NLP techniques as described above, we reclassified all retailer provided categories in 93 COICOP level 4 categories. We used Orlando as reference city to calculate relative price indexes across cities and time periods for all 93 categories. In Figure 1 and Figure 2 we report results for Apples and Chocolate categories. Price differences show different spatial patterns for different consumption categories. We can note a marked trend in the price for apples, with price indexes moving mostly in the same direction. This may be due to seasonal price variation for this fresh product.

Figure 1. Spatial price index by Apple category – January 2017-May 2018

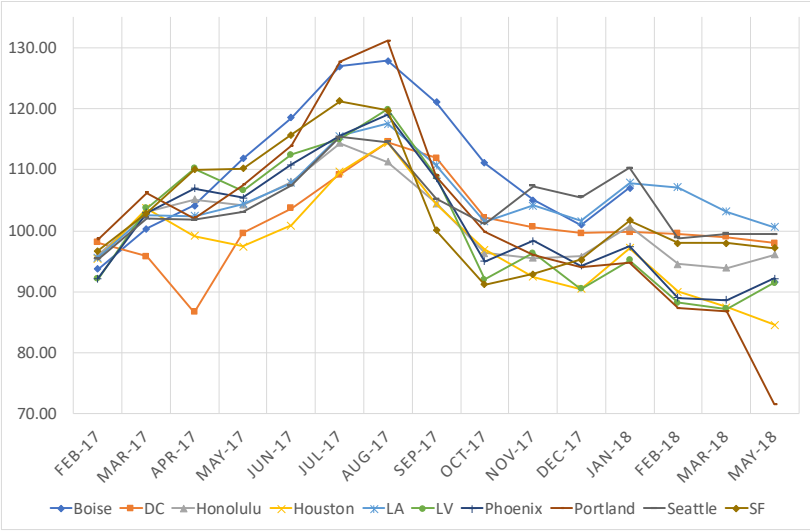
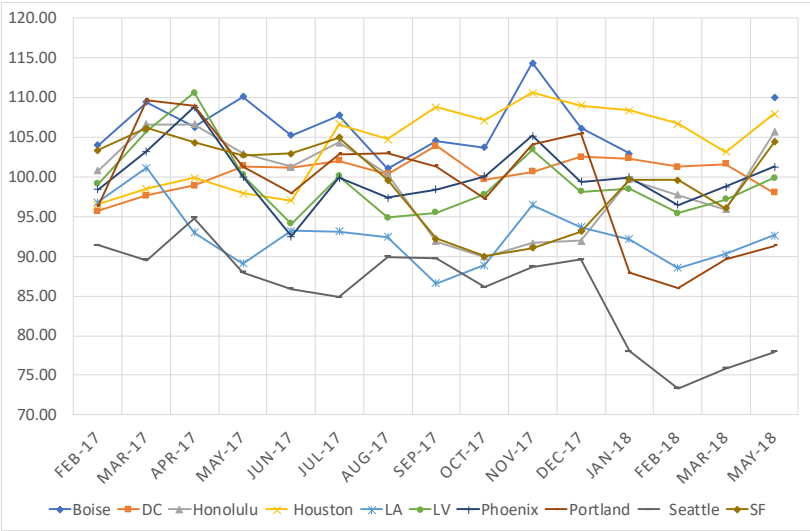


Figure 2. Spatial price index by Chocolate category January 2017-May 2018



4. CONCLUSIONS

Web scraping offers the potential to improve greatly the quality and efficiency of consumer price indices. However, there remain a number of limitations to using this data to construct price indices, including problems with processing and cleaning large datasets. Questions remain around whether all web scraped data should be used (that may

reduce the representativeness of the products included within the analysis) or whether a sample should be taken.

This work demonstrates how scraped data, with the support of NLP and data analysis techniques, can be a useful tool to monitor the evolution of spatial price indexes over time, providing nearly real-time information for policy and business decisions.

4.1. Future research

As with many scraped datasets and big data in general, data quality is a point of concern. Given the unfeasibility to manually check every single record, it may be appropriate to develop automatic routines to identify and correct anomalies which may affect results. Also, NLP techniques to extract brand information from product description may prove useful to analyse a brand spatial pricing strategy.

Moreover, different methods may be applied to calculate spatial and temporal price indexes on our dataset, leading to additional insights and offering the chance to perform comparisons. For instance, the computation of price indexes using a shared base for all period and weights derived from external sources (such as CEX) will ensure the comparability of price indexes evolution over time for all cities and categories. The larger dataset of scraped product prices we have available would enable us to add further data points on this line of research.

REFERENCES

[1] A. Virgillito, F. Polidoro, (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.

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

[3] 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).

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

[5] K. Gábor, H. Zargayouna, I. Tellier, D. Buscaldi, T. Charnois, (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.

[6] T. Laureti, D.P. Rao, (2019) "Measuring spatial price level differences within a country: Current status and future developments". Studies of Applied Economics, 36(1), 119-148.

[7] 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.