University of Cyprus
Economics Research Centre

Economic Policy Papers

The 'Billion Prices Project' in Cyprus: Measuring the Consumer Price Index Inflation of Cyprus using Big Data

Elena Andreou

Economics Research Centre, University of Cyprus

Andriani Panagi

Economics Research Centre, University of Cyprus

Maria Papageorgiou

Economics Research Centre, University of Cyprus, and University of Leeds

George Syrichas

Economics Research Centre, University of Cyprus

Kyriacos Vitalis

Economics Research Centre, University of Cyprus

No. 03-24

May 2024

Publication Editor: Vasiliki Bozani

ERC Sponsors (in alphabetical order)

Central Bank of Cyprus

Directorate General for European Programmes, Coordination and Development

Economics Department, University of Cyprus

Ministry of Finance

University of Cyprus

Εφαρμογή του ‘Billion Prices Project’ στην Κυπριακή Οικονομία: Μέτρηση του Πληθωρισμού της Κύπρου με βάση τον Δείκτη Τιμών Καταναλωτή χρησιμοποιώντας Μεγάλα Δεδομένα

Έλενα Ανδρέου, Ανδριανή Παναγή, Μαρία Παπαγεωργίου, Γιώργος Συρίχας και Κυριάκος Βιτάλης

ΠΕΡΙΛΗΨΗ

Η μεθοδολογία του ‘Billion Prices Project’ εφαρμόζεται για πρώτη φορά στην κυπριακή οικονομία και χρησιμοποιεί προηγμένες τεχνικές συλλογής/απόξεσης δεδομένων από το διαδίκτυο (web data scraping) για τον δυναμικό υπολογισμό των ποσοστών πληθωρισμού με βάση τον Δείκτη Τιμών Καταναλωτή (ΔΤΚ), προσφέροντας κατ’ αυτόν τον τρόπο μια προοπτική σχετικά με τις οικονομικές τάσεις σε πραγματικό χρόνο. Μέσω της αυτόματης και συχνής εξαγωγής δεδομένων για προϊόντα και υπηρεσίες – που περιλαμβάνονται σε ένα σταθερό και αντιπροσωπευτικό καλάθι ΔΤΚ– από διάφορες ψηφιακές (online) πηγές, συμπεριλαμβανομένων πλατφορμών ηλεκτρονικού εμπορίου και ψηφιακών αγορών, το έργο στοχεύει στην παροχή μιας πιο συνολικής και ενημερωμένης κατανόησης της δυναμικής του πληθωρισμού. Αξιοποιώντας τεχνικές εξόρυξης και ανάλυσης ‘Μεγάλων Δεδομένων’ για να ενισχύσει την ακρίβεια, η κλιμακούμενη και προσαρμοστική φύση της μεθοδολογίας εξασφαλίζει μια ολιστική αναπαράσταση των προτύπων κατανάλωσης του μέσου Κύπριου καταναλωτή. Η συγκριτική ανάλυση με παραδοσιακούς δείκτες πληθωρισμού υπογραμμίζει την δυναμική του έργου να προσφέρει έγκαιρες πληροφορίες, καθιστώντας το ένα πολύτιμο εργαλείο για οικονομολόγους, επιχειρήσεις και υπεύθυνους χάραξης πολιτικής που πλοηγούνται στην πολυπλοκότητα των σύγχρονων ψηφιακών οικονομιών.

CONTENTS

The 'Billion Prices Project' in Cyprus: Measuring the Consumer Price Index Inflation of Cyprus using Big Data

Elena Andreou, Andriani Panagi, Maria Papageorgiou, George Syrichas and Kyriacos Vitalis*

Abstract

The 'Billion Prices Project' methodology is for the first time in the Cypriot economy applied, and it employs advanced web-scraping techniques to dynamically calculate inflation rates, offering a real-time perspective on economic trends. Through automated and frequent extraction of price data from diverse online sources, including e-commerce platforms and digital marketplaces, the project aims to provide a more comprehensive and up-to-date understanding of inflation dynamics. Leveraging data analysis techniques to enhance accuracy, the methodology's scalability and adaptability ensure a holistic representation of consumer spending patterns. Comparative analysis with traditional inflation indices highlights the project's potential to offer timely insights, making it a valuable tool for economists, businesses, and policy makers navigating the complexities of modern digital economies.

Keywords: Web Scraping, Online Prices, Big Data, Inflation, Macroeconomy

* Corresponding Author. Address: Economics Research Centre, University of Cyprus, P.O. Box 20537, CY-1678, Nicosia, Cyprus. Email: vitalis.kyriacos@ucy.ac.cy

1. Introduction

The *Billion Prices Project* [1] (Cavallo and Rigobon, 2016) stands at the forefront of a change in thinking in economic measurement, leveraging innovative web-scraping[2] technologies to redefine how we understand and analyse inflation rates in contemporary economies. In a world increasingly interconnected and digitally driven, traditional methods of gauging economic indicators face challenges in capturing the rapid and nuanced dynamics of online commerce. This work delves into the pioneering efforts of the Billion Prices Project (BPP), an academic initiative at MIT and Harvard, which was designed to harness the power of web scraping to collect and analyse vast amounts of real-time price data from diverse online sources.

In particular, the BPP shows how online prices can be used to construct daily price indices in multiple countries/economies and to avoid measurement biases, emphasizing how *'Big Data'* technologies are providing macro and international economists with opportunities to stop treating the data as 'given' and to get directly/personally involved with the data collection process. The BPP data comprises high-frequency daily price indices, monthly, and annual inflation rates for Argentina and the US, as well as monthly data with annual inflation rates for Argentina, Brazil, China, Germany, Japan, South Africa, UK, US, 3 US sectors, and global aggregates (including Eurozone).

Founded in 2008 by Alberto Cavallo and Roberto Rigobon, the BPP remained active until 2016. The research continues through related projects such as PriceStats, a private company founded in 2011 that uses web-scraping technologies to monitor online prices day by day (raw micro data) to provide, among others, daily inflation series for 25 countries[3]. The objective of the series (most start in 2007–2008) is to anticipate major shifts in inflation trends, and this could be achieved because the online real-time prices tend to react to shocks faster than offline prices.

As global economies continue to undergo digital transformations, the landscape of consumer transactions has evolved dramatically. Traditional measure of inflation, reliant on periodic surveys and data collection methods, often lag the pace of real-world economic changes. The BPP addresses this discrepancy by embracing the agility and responsiveness afforded by web scraping. By automating the extraction of

[1] See The Billion Prices Project.
[2] Web scraping is the process of automatically collecting information from the web by converting unstructured data (typically in HTML format) into structured datasets that can be stored and analyzed.
[3] These are Argentina, Australia, Brazil, Canada, Chile, China, Colombia, France, Germany, Greece, Ireland, Italy, Japan, Mexico, Netherlands, New Zealand, Poland, Russia, South Africa, South Korea, Spain, Turkey, United Kingdom, Uruguay, and the United States.

price information from hundreds of online retailers' platforms, ranging from e-commerce websites to digital marketplaces, the project aims to provide a more accurate and timely reflection of inflationary pressures within a given economy.

The methodology employed in BPP is rooted in the principle of inclusivity. Unlike traditional indices that may focus on a limited basket of goods and services, web scraping allows for a comprehensive examination of a wide array of products, capturing the full spectrum of consumer spending patterns. This inclusivity, coupled with the scalability and adaptability of the web-scraping approach, presents an opportunity to revolutionize how we conceptualize inflation rates consequently make informed decisions in a rapidly changing economic landscape.

In the context of a global economy increasingly shaped by online interactions, the BPP serves as beacon of innovation, illustrating the possibilities that arise when we embrace modern technologies to enhance our understanding of economic phenomena. As we navigate the complexities of the digital age, the insights gleaned from this project contribute to the ongoing discourse on refining economic measurement techniques, paving the way for a more accurate and responsive approach to analysing rates in the 21st century.

1.1 Objective

The objective of this project is to compute the inflation rate in Cyprus by analysing changes over time in the *Consumer Price Index (CPI)*. The CPI serves as a metric for gauging fluctuations in the cost and price levels of a predetermined basket of goods and services ('shopping/market basket') that is representative of consumer spending patterns. In this project, the CPI is estimated using online prices for a carefully chosen and representative basket of goods and services.

It is worth noting that the official CPI measurement is traditionally carried out by the Cyprus Statistical Service (CYSTAT). This process involves deploying a substantial workforce to physically visit hundreds of stores monthly for price data collection in the offline domain. In contrast, this methodology leverages web-scraping algorithms to traverse the websites and pages of major retailers in Cyprus, extracting prices for a *fixed and representative* set of goods and services available online. To elaborate, a web-scraping algorithm is employed to download specific public webpages. Subsequently, it scrutinizes the HTML code, extracts, or scrapes relevant price data, and organizes this information into a database. Finally, we apply standard techniques and proprietary methodologies to calculate price statistics and indices, ensuring a

robust and accurate representation of inflation dynamics in Cyprus based on the online pricing landscape.

1.2 The Importance of the 'Billion Prices Project'

Emerging data sources are already reshaping the landscape of inflation and price statistics measurement. Traditional methods of collecting raw price data have faced longstanding challenges, characterized by the constraints of manual data collection processes. National statistical offices employ a considerable workforce to visit numerous stores monthly, compiling prices for a predetermined basket of goods and services. This method, however, proves to be costly, intricate, and often sluggish, posing difficulties in adapting to changing market dynamics.

The limitations of infrequent sampling and delayed updates to these baskets further complicate the accurate adjustment for quality changes and the introduction of new products. In the face of economic crises, the demand for faster and more precise data has escalated among policymakers and other users of these statistics. Unfortunately, the resources available to national statistical offices are dwindling, placing additional strain on their essential work.

A promising solution to these challenges lies in the realm of *'Big Data'*, particularly in the realm of online prices, as outlined by Cavallo and Rigobon (2016). The appeal of online prices stems from their widespread distribution across hundreds of websites and thousands of webpages. Recent advancements in automated scraping software have democratized large-scale data collection on the web, offering detailed information on each product. This method proves to be cost-effective, swift, and accurate, making it an ideal complement to traditional data collection methods, especially for goods and services well-represented in online markets.

Cavallo's and Rigobon's (2016) approach stands out from other attempts to leverage big data in economics as it prioritizes *measurement* over prediction. Recognizing that an abundance of data does not necessarily equate to better information, they emphasize applying traditional data collection principles, such as ensuring a *representative* sample. Their focus on *large multi-channel* retailers, catering to both online and offline consumers, rather than online-only retailers, reflects a commitment to capturing a more comprehensive picture of retail transactions. Additionally, their preference for collecting prices of goods within official CPI baskets, for which consumer expenditure weights are available, underscores the importance of aligning with established methodologies. In their research, they present compelling examples of

how online prices can serve as an effective *alternative* source for constructing price indices that mirror the patterns observed in official CPIs.

1.3 Advantages and Disadvantages of Online as opposed to Offline Price Data

Online prices serve as an alternative and increasingly valuable data source, each with its distinct advantages and drawbacks outlined in Table 1, as expounded upon in the comprehensive study by Cavallo and Rigobon (2016). This categorization sheds light on the strengths and weaknesses of online prices, offering valuable insights for both measurement and research applications. A parallel examination in Table 2 presents a comparative analysis of the advantages and disadvantages associated with their offline counterparts.

In their seminal 2016 paper, Cavallo and Rigobon assert that online prices harbor the transformative potential to significantly augment the quantity and quality of micro price data accessible for academic research. The unique characteristics of online data don't only facilitate the reevaluation of longstanding empirical puzzles but also empower researchers to address questions that were previously insurmountable. Beyond its academic utility, online data emerges as a reliable resource for inflation measurement, marked by its alignment with 'big data' characteristics, such as the automatic inclusion of 'uncensored price spells' upon the introduction of a new product and comprehensive coverage of all product varieties and models.

The key takeaway from their findings emphasizes the optimal use of online prices as an alternative data source rather than treating it as a separate sector requiring special treatment. Notably, Cavallo and Rigobon (2016) highlight online prices' ability to forecast changes in inflation trends, demonstrating a unique real-time information advantage, surpassing the mere ability to expedite data collection and publication. This anticipatory capability typically manifests with a lead time of 2-3 months. Online prices exhibit quicker reactions and adjustments to economic shocks due to lower adjustment costs, coupled with the diminished sensitivity of online shoppers to price changes. Furthermore, the heightened and transparent competition in online markets contributes to this accelerated responsiveness.

Importantly, Cavallo and Rigobon (2016) predict the widespread accessibility of online data in the coming years, underscoring the growing significance of this data source in shaping future research and measurement methodologies.

TABLE 1
Advantages and Disadvantages of Online Price Data

Advantages	Disadvantages

• Cheap to collect (low cost per observation)
• High frequency of collection (daily)
• Can be collected remotely
• Granularity:
  • All product details (brands size, anything shown online)
  • All goods and varieties available for sale (census) on display
  • New products automatically sampled (*uncensored* price spells that ease quality adjustments and control for new-product biases)
• Easier to compare internationally (useful in research applications that use cross-country comparisons)
• Available in real time (particularly useful for policymakers)
	• Cover fewer retailers and locations than official CPI data
	• Fewer product categories (not all categories of goods and services are online, at least not yet)
	• Short time series
	• Lack information on quantities sold (must be combined with weights from official consumer expenditure surveys)
	• Online and Offline prices *may* behave differently (see next section)

TABLE 2
Advantages and Disadvantages of Offline Price Data

Advantages	Disadvantages

• Representative sample:
  • carefully chosen products
  • many products categories
  • many retailers and locations
• Information on quantities sold/ expenditure weights
• Long time series
• Collection of posted prices in stores
	• Very costly to collect and access
	• Low frequency of collection (monthly)
	• Limited number of goods and varieties
	• New product only when the previous disappears from the stores (*censored* prices)
	• Some unit values and imputed prices
	• Difficult international comparisons
	• Not available in real time

1.4 Offline vs. Online Prices – Are they Different?

A natural question which arises is whether online prices are similar to the prices collected offline in physical stores where most retail transactions take place. Therefore, the question of whether online prices are representative of the economy is of

fundamental importance since online sales/purchases are still only about 10% of retail sales in developed countries.

The first attempt to answer this question was carried out by Cavallo (2017) who simultaneously collected prices on the online (websites) and offline (physical) stores for over 24,000 products in 56 of the largest retailers in 10 countries worldwide for over a year (December 2014 – March 2016). This large-scale comparison was possible thanks to the combination of a smartphone app, crowdsourced workers, and web-scraping techniques. More than 300 freelance workers used their phones to scan barcodes in about 500 physical stores, manually enter prices, take photos of the price tags, and upload the information to the BPP servers. Then, the barcodes were used in the offline data to collect the prices for those exact same goods at the website of the same retailer within a seven-day time window. This direct comparison between online and offline prices revealed a high degree of similarity in price levels, as well as in both the frequency and size of price changes. *On average, about 72% of price levels were identical in the offline and online samples*. The similarity was highest in clothing and electronics retailers, and lowest in drugstores and office-supply retailers that also tend to price differently across offline stores. While price changes do not have the exact same timing online and offline (only about 20% of weekly price changes occur simultaneously), they tend to have similar frequency and average sizes. This suggests that the price spells for individual goods may not be synchronized online and offline, consistent with evidence discussed previously that online prices may adjust faster to shocks and anticipate later price changes.

Despite the general similarity between online and offline pricing, the study results also revealed a great deal of heterogeneity among pricing behaviors, suggesting that researchers using relatively few sources of data should be cautious to understand particular pricing patterns and control for any sampling biases. For national statistical offices, these results imply that the web can be effectively used as an alternative data-collection technology to obtain the same prices found offline. Prices collected through the web are very similar to those that can be obtained at a much higher cost by physically walking into a store. However, many challenges to the use of online data in CPIs remain, such as the more limited sectoral coverage or the lack of quantity data mentioned above. Moreover, the results have implications for people interested in the effects of the Internet on retail prices. The fact that online prices are the same for all locations and similar to offline prices collected from many different zip codes implies there is little within-retailer price dispersion. In practice, most retailers seem to adopt a uniform pricing behavior by selling at a single price for the majority of products,

regardless of the location of the buyer and whether the product is sold online or at a particular offline store.

2. Background

The genesis of the BPP can be traced back to 2008 when a collaborative effort between MIT Sloan School of Management and the Harvard Business School sought to revolutionize the way economists and policymakers measure inflation. At the time of its inception, the traditional methods of tracking inflation were grappling with the challenges posed by an increasingly digitized global economy. The limitations of periodic surveys and data collection methods became apparent as the pace of economic transactions accelerated, particularly in the realm of online commerce.

Motivated by the need for a more agile and real-time approach to economic measurement, the visionaries behind the BPP embarked on a pioneering journey to leverage the vast potential of web-scraping technologies. Recognizing the transformative impact of the digital age on consumer behavior and market dynamics, the project aimed to create a dynamic system capable of capturing and analyzing a billion prices observations from diverse online sources every day.

The initial collaboration between MIT Sloan and Harvard Business School brought together interdisciplinary expertise, blending insights from economics, computer science, and data analytics. The project's architects understood that traditional indices, reliant on a fixed/pre-selected and representative basket of goods and services and periodic data collection, struggle to keep pace with the dynamic nature of online pricing. By automating the collection of price data from a myriad of online platforms, including e-commerce websites, the BPP sought to provide a more comprehensive, timely and accurate representation of inflationary pressures within different economies.

The following sections of the paper delve deeper into the technical intricacies, outcomes, and implications of this groundbreaking initiative, which are applied to the Cyprus' economy for the first time to calculate the national inflation rate using big-data technologies.

3. Analysis

3.1 Methodology

The 'Cyprus BPP' employs a methodology centred around web-scraping technologies to capture, analyse, and interpret real-time pricing data for the calculation of the inflation rates. The web-scraping techniques were implemented in Python and the

automated process can be found in GitHub[4]. The following outline the key steps and components of the methodology:

1. Matching between the items/commodities (goods and services) included in the official CPI basket and those provided from online retailers in Cyprus:
    - The analysis focuses on aligning items within the official CPI basket with those offered by large online retailers in Cyprus. Employing the ECOICOP hierarchical classification system, the examination encompasses 12 major categories/divisions, 44 groups, 81 classes, and 198 subclasses contributing to the national CPI (see Table A.1 in the Appendix).
    - In Table A.2 in the Appendix, the ECOICOP weight is described as the weight per division for the year 2022 (last update) in the official/offline CPI basket (CYSTAT) and in the online (Cyprus BPP) CPI basket. This comprehensive analysis aims to provide insights into the alignment between the items included in the traditional CPI basket and those offered by online retailers, accounting for 75.76% of the net total CPI expenditure weight.
2. The web-scraping framework and the CPI calculation:
    - A diverse range of online sources, including e-commerce platforms, retail websites and digital marketplaces is identified for data collection. This selection ensures the inclusion of a comprehensive array of goods and services in the CPI basket/dataset.
    - Web-scraping scripts are developed in Python to automatically extract the price information from the identified sources. These scripts navigate through web pages, locate relevant pricing data and capture the information in a structured format (CSV file) for further analysis.
    - Following each day's web scraping, two files are updated. The first file encompasses details about our pre-selected goods and services, including their name, price, retailer, ECOICOP subclass, as well as the date of scraping. The second file comprehensively documents various subclasses, their divisions, the daily average price of each subclass, the CPI per division, the General CPI, and additional columns that will be elucidated in subsequent sections.

[4] See https://github.com/maria-papz/BillionPricesCyprusScrape.

- Using the online price data of goods and services and their corresponding/associated ECOICOP weights, the CPI (base=100) is compiled according to the Laspeyres formula[5,6] :

$$CPI_t = 100 * \frac{P_t}{P_0} = 100 * \frac{\sum_{i=1}^{N} w_i * p_{it}}{\sum_{i=1}^{N} w_i * p_{i0}} \quad (1)$$

where:

  - P_t is the all-items weighted price (total cost of the basket) at the observation/current period *t*,
  - P_0 is the all-items weighted price (total cost of the basket) at the reference/base period,
  - w_i is the consumption expenditure weight of the individual item *i*,
  - p_{it} is the (average) price of the individual item *i* at period *t*,
  - p_{i0} is the (average) price of the individual item *i* at the base period, and
  - N is the total number of items in the CPI basket.

3. The comparative analysis:

- The calculated inflation rates by the Cyprus BPP are compared with the inflation indices from the CYSTAT. This approach provides insights into the reliability, accuracy, and timeliness of the web-scraping methodology.
- Discrepancies between the BPP results and traditional indices are examined to understand potential biases and limitations, contributing to a more nuanced interpretation of the findings.

3.2 Data Collection

The data collection method for this project involves *web scraping* of HTML files from different retailer websites using Python programming language. Web scraping is a technique that automates the extraction of information from web pages, and Python provides powerful libraries, such as '*Beautiful Soup'* and '*Scrapy'* that facilitate this process. Here's a step-by-step description of the data collection procedure:

1. Identifying target retailers' websites:

[5] Laspeyres Price Index, Available at: Laspeyres Price Index – Overview, Formula, and Example (corporatefinanceinstitute.com)
[6] Eurostat, Methodological Manual, 2018, Harmonised Index of Consumer Prices (HICP), Available at: HICP Methodological Manual (ec.europa.eu/eurostat)

- Compile a list of the retailer websites from which the data of the predetermined goods and services will be collected. These websites should represent a diverse range of goods and services to ensure a comprehensive and representative CPI basket/dataset of Cypriot consumers. The websites of the online retailers used for this project are listed in Table 3 below:

TABLE 3
List of Online Retailers

Retailer	URL
Adventure Without Limits (AWOL)	https://www.awol.com.cy/
Alphamega	https://www.alphamega.com.cy/
Alter Vape	https://altervape.eu/
Athlokinisi	https://athlokinisi.com.cy/
Bwell Pharmacy	https://bwell.com.cy/
Cablenet	https://cablenet.com.cy/
Consumer Protection Service	https://consumer.gov.cy/gr/
Cyprus Energy Regulation Authority (CERA)	https://www.cera.org.cy/Templates/00001/data/hlektrismos/kostos_xrisis.pdf
Cyprus Ministry of Education, Sport, and Youth	https://www.moec.gov.cy/idiotiki_ekpaidefsi/didaktra.html
Cyprus Post	https://www.cypruspost.post/uploads/2cf9ec4f5a.pdf
Cyprus Telecommunications Authority (CYTA)	https://www.cyta.com.cy/personal
Epic	https://www.epic.com.cy/en/page/start/home
E-WHOLESALE	https://www.ewsale.com/tsigaro
Electroline	https://electroline.com.cy/
European University Cyprus	https://syllabus.euc.ac.cy/tuitions/euc-tuition-fees-c.pdf
Famous Sports	https://www.famousports.com/en
FuelDaddy *(Agip, EKO, Eni, Esso, Fill n GO, Petrolina, Shell, Staroil, Total Plus)*	https://www.fueldaddy.com.cy/en
IKEA	https://www.ikea.com.cy/
Marks & Spencer	https://www.marksandspencer.com/cy/
Mazda	https://www.mazda.com.cy/home
Moto Race	https://www.motorace.com.cy/
Nissan	https://www.nissan.com.cy/
Novella Hair Mode	https://novella.com.cy/
NUMBEO	https://www.numbeo.com/cost-of-living/country_price_rankings?itemId=17&displayCurrency=EUR
Pizza Hut	https://www.pizzahut.com.cy/
Primetel	https://primetel.com.cy/en
Rio Cinemas	http://www.riocinemas.com.cy/
Sewerage Board of Limassol-Amathus (SBLA)	https://www.sbla.com.cy/Sewage-Charges
Sewerage Board of Nicosia (SBN)	https://www.sbn.org.cy/el/apoxeteftika-teli

Sewerage and Drainage Board of Larnaca (LSDB)	https://www.lsdb.org.cy/en/services/financial-information/sewage-charges/
Stephanis	https://www.stephanis.com.cy/en
Stradivarius	https://www.stradivarius.com/cy/
SupermarketCy	https://www.supermarketcy.com.cy/
The CYgar Shop	https://www.thecygarshop.com/
The Royal Cigars	https://fetch.com.cy/shop/stores/Nicosia/store/222/The%20Royal%20Cigars%20%7C%20Strovolos
Water Board of Nicosia (WBN)	https://www.wbn.org.cy/%CE%BA%CE%B1%CF%84%CE%B1%CE%BD%CE%B1%CE%BB%CF%89%CF%84%CE%AE%CF%82/%CE%B4%CE%B9%CE%B1%CF%84%CE%B9%CE%BC%CE%AE%CF%83%CE%B5%CE%B9%CF%82/
Water Board of Larnaca (LWB)	https://www.lwb.org.cy/en/charges-and-fees.html
Water Board of Limassol (WBL)	https://www.wbl.com.cy/el/water-rates
Wolt ***(Costa Coffee, Piatsa Gourounaki Nicosia, Pyxida Nicosia, Kofini Tavern Limassol, Vlachos Taverna Larnaca, Zakos Beach Restaurant Larnaca, Paphos Tavernaki, Ocean Basket, McDonald's)***	https://wolt.com/en/cyp
TOTAL	**39**

2. Understanding website structure:
    - Analyze the HTML structure of the target retailer websites to identify the location of relevant data, such as product names and prices. This involves inspecting the HTML elements using browser developer tools.
3. Python web-scraping libraries:
    - Utilize Python libraries, such as *Beautiful Soup*, *Requests*, *Tabula*, *lxml*, to send HTTP requests to the retailer websites and retrieve the HTML content of the web pages. The *Requests* library is employed for making HTTP requests, while *Beautiful Soup* is used for parsing HTML content. *Tabula* is used for reading tables in PDF files and the *lxml* library is a powerful API for parsing XML and HTML.
4. Parsing HTML content:
    - Parse the HTML content to extract the desired information. *Beautiful Soup* provides methods to navigate the HTML tree, locate specific elements and retrieve text or attributes.

5. Storing data:

   - Organize and store the extracted data and calculations in structured CSV files. The *BillionPricesProject_ProductList.csv* and *Calculations.csv* files are updated every day. The *Ref_weights.csv* file includes the ECOICOP weights of the official CPI basket. This file was initially used to calculate the weighted mean/average price of each subclass and division in the reference/base CPI basket (08/08/2023), as well as its total matched weight (75.76%). These weights are used every day to calculate the weighted average price of each subclass, the weighted average price of each division, the CPI of each division, as well as the General CPI.

   - The columns of the *Ref_weights.csv* file are the following:

     division (the division of the products), *subclass* (the subclass of the products), *weight.division* (the ECOICOP weight of each division in the official CPI basket), *weight.subclass* (the ECOICOP weight of each subclass in the official CPI basket), *matching* (1 if the subclass is included in the dataset, and 0 otherwise), *reference.mean.price* (the mean/average price of the products in each subclass in the reference basket that was created in 08/08/2023), *reference.weighted.mean.price* (the weighted mean price of each subclass in the reference basket), *reference.weighted.mean.price.division* (the weighted mean price of each division in the reference basket), *reference.weighted.mean.price.total* (the total weighted mean price of all products in the reference basket).

   - The columns of the *BillionPricesProject_ProductList.csv* file are the following:

     product_name (the name of the product), *product_price* (the price of the product), *date_time_scraped* (the date and time that the product info is scraped), *product_subclass* (the subclass of the product), *retailer* (the retailer of the product), *subclass_average* (the average price of the products in each subclass).

   - The columns of the *Calculations.csv* file are outlined below (a further explanation for the different calculations is provided in Section 3.3):

     subclass (the subclass of the products), *weight.division* (the ECOICOP weight of each division in the official CPI basket), *weight.subclass* (the

ECOICOP weight of each subclass in the official CPI basket), *matching* (1 if the subclass is included in the dataset, and 0 otherwise), *reference.mean.price* (the mean price of the products in each subclass in the reference basket that was created in 08/08/2023), *reference.weighted.mean.price* (the weighted mean price of each subclass in the reference basket), *reference.weighted.mean.price.division* (the weighted mean price of each division in the reference basket), *reference.weighted.mean.price.total* (the total weighted mean price of all products in the reference basket), *subclass_average* (the mean price of the products in each subclass), *weighted.mean.price* (the weighted mean price of each subclass), *weighted.mean.price.division* (the weighted mean price of each division), *weighted.mean.price.total* (the total weighted mean price of all products in the basket), *weight.matched* (the weight of each subclass included in the basket/dataset), *weight.matched.division* (the weight of each division included in the basket/dataset), *weight.matched.total* (the total weight of the basket), *datetime.calculated* (date and time of the calculations), *CPI_total* (the total Consumer Price Index), *CPI_division* (the Consumer Price Index of each division), *weighted_CPI_division* (the weighted Consumer Price Index of each divisions), *CPI_general* (the general Consumer Price Index), *CPI_total_inflation* (the total Consumer Price Index; Inflation in % percentage), *CPI_general_inflation* (the General Consumer Price Index; Inflation in % percentage), *division* (the division of the products), *CPI_general_lastthursday* (the General Consumer Price Index of the last Thursday of the month), *monthly_inflation_lastthursday* (the inflation calculated as the % change of the General CPIs of the last Thursdays between two consecutive months), *CPI_general_cystat* (the official general Consumer Price Index published by the Statistical Service of Cyprus each month), *CPI_mothly_inflation_cystat* (the monthly % inflation published by the Statistical Service of Cyprus).

6. Automation and schedule:

    - Automate the web data scraping and indices calculation processes by incorporating scheduling mechanisms such as YML files in GitHub. Through these tools, both the *scrape_tool.py* script, which is responsible for scraping the data of goods and services in the CPI

basket, and the *calculations.py* script, which is responsible for executing the appropriate calculations for the estimation of the CPI inflation, automatically run every day at around 9:30am (UTC time).

3.3 Calculations File for the Cyprus BPP

To conduct any Consumer Price Index (CPI) calculations, a comprehensive understanding of the basket's composition is imperative, drawing insights from the methodological details provided by both the national and European statistical services.

The CPI serves as a metric to gauge the general price level of goods and services acquired by the average household, playing a pivotal role in measuring inflation, wage adjustments, and similar economic indicators. The divisions and subclasses utilized in the CPI calculation are a subset of those outlined by the Statistical Service, encompassing 12 divisions and numerous subclasses, collectively accounting for a total weight of 100%. These divisions include Food and Non-alcoholic Beverages (19.06%), Alcoholic Beverages and Tobacco (3.47%), Clothing and Footwear (7.21%), Housing, Water, Electricity, Gas, and Other Fuels (11.19%), Furnishing, Household Equipment, and Supplies (6.44%), Health (6.37%), Transport (15.24%), Communication (4.39%), Recreation and Culture (5.75%), Education (3.92%), Restaurants and Hotels (8.36%), and Miscellaneous goods and Services (8.6%). *It is noteworthy that this project engages in web scraping across goods and services within all 12 divisions*, albeit not covering all subclasses, resulting in a *basket weight of 75.76%*. This is because online prices have a limited coverage, encompassing a significantly smaller range of retailers and product categories compared to government-conducted consumer price surveys. Particularly, the prices of numerous services remain unavailable on the web, and the diversity and quantity of retailers are constrained in comparison to the comprehensive dataset provided by official CPI data.

The initiation of web scraping for this project commenced in 08/08/2023, a date that concurrently serves as the reference date for the online CPI basket. It is on this reference date that the data collection process begins, forming the foundation upon which subsequent calculations of inflation rates are based. This approach ensures a synchronized and consistent starting point for the project, aligning the dataset with the prevailing market conditions as of the specified date.

Below, more details on the different calculations performed in the *Calculations.csv* file are provided:

- *weighted.mean.price = subclass_average * weight.subclass*
- *weighted.mean.price.division* is the sum of the *weighted.mean.price* of each subclass that belongs in the division; it corresponds to the weighted mean/average price of the division.
- *weighted.mean.price.total* is the sum of the *weighted.mean.price.division* of each division of the CPI basket; it corresponds to the total weighted mean price of the basket.
- *weight.matched = weight.subclass * matching*
- *weight.matched.division* is the sum of the *weight.matched* of each subclass that belongs in the division; it corresponds to the total weight of the division.
- *weight.matched.total* is the sum of the *weight.matched* of each division of the CPI basket; it corresponds to the total weight of the basket.
- *CPI_total = weighted.mean.price.total / reference.weighted.mean.price.total*
- *CPI_division=100*(weighted.mean.price.division/reference.weighted.mean.pri -ce.division)*
- *weighted_CPI_division = weight.matched.division * CPI_division*
- *CPI_general* is the sum of the *weighted_CPI_division* of each division in the CPI basket; it corresponds to the General Consumer Price Index
- *CPI_total_inflation* is the percentage change (%) between the current *CPI_total* and the previous *CPI_total.*
- *CPI_general_inflation* is the percentage change (%) between the current *CPI_general* and the previous *CPI_general.*

4. Results

In this section, we present the results of the Cyprus BPP. The section goes through the intricate landscape of real-time price movements, unveiling patterns, anomalies, and underlying economic phenomena.

A methodology for calculating inflation akin to the CYSTAT is employed, specifically by measuring the change in the General CPI relative to the previous month's index. The process involves computing the CPI for each division using Equation (1) and subsequently aggregating them with weighted sums to determine the General Index.

The online CPI basket for Cyprus encompasses 2,476 items, representing consumer expenditures. Notably, between 08/08/2023 and 03/11/2023, the data sample from the CPI basket (*BillionPricesProject_ProductList.csv*) seems to consist of approximately 1,000 observations/items per day, rather than the actual size of 2,476. This discrepancy is attributed to the inclusion of the daily average prices for each fuel type in Cyprus –Unleaded 95, Unleaded 98, Road diesel, Heating diesel, Kerosene– calculated directly from the Consumer Protection Service website (https://eforms.eservices.cyprus.gov.cy/MCIT/MCIT/PetroleumPrices) of the Ministry of Energy, Commerce, and Industry.

Starting from 04/11/2023 onwards, the dataset has been rectified to encompass the daily prices of the 5 fuels from all filling/gas stations in Cyprus (totalling 1,477 prices), obtained through scraping the FuelDaddy website (https://www.fueldaddy.com.cy/en). Consequently, the dataset now accurately reflects its intended size of 2,476 observations (goods and services) per day.

Figures 1 and 2, respectively, present the daily evolution of the Cyprus' General CPI and inflation from 08/08/2023 to 31/12/2023 as estimated by this work (online data).

4.1 Indices Comparability

An essential criterion for national CPIs is comparability, ensuring that *discrepancies reflect only variations in price changes or expenditure patterns rather than methodological differences*. The goal is to achieve consistent results, enabling different countries to produce identical outcomes from the same dataset of prices and weights. Variations in CPIs are naturally anticipated when prices evolve differently, or consumption patterns (i.e. expenditure weights) diverge.

FIGURE 1
Daily evolution of online General CPI in Cyprus

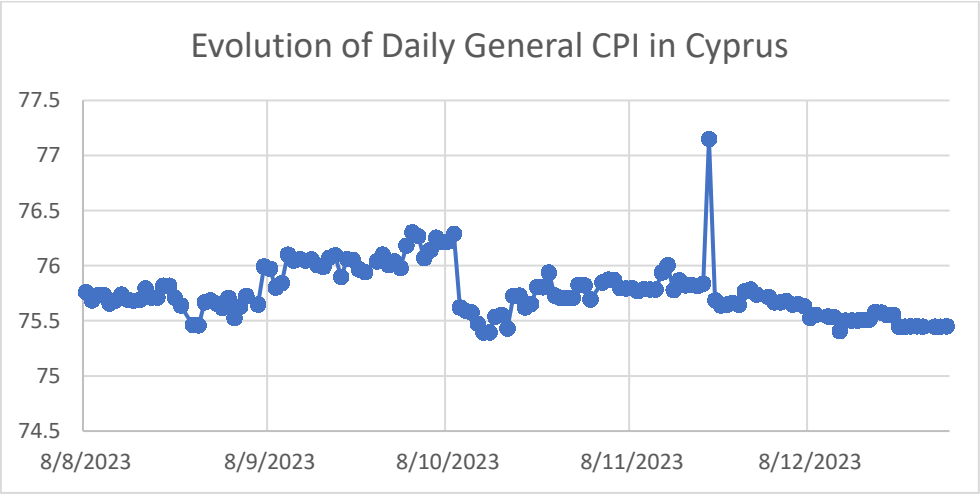
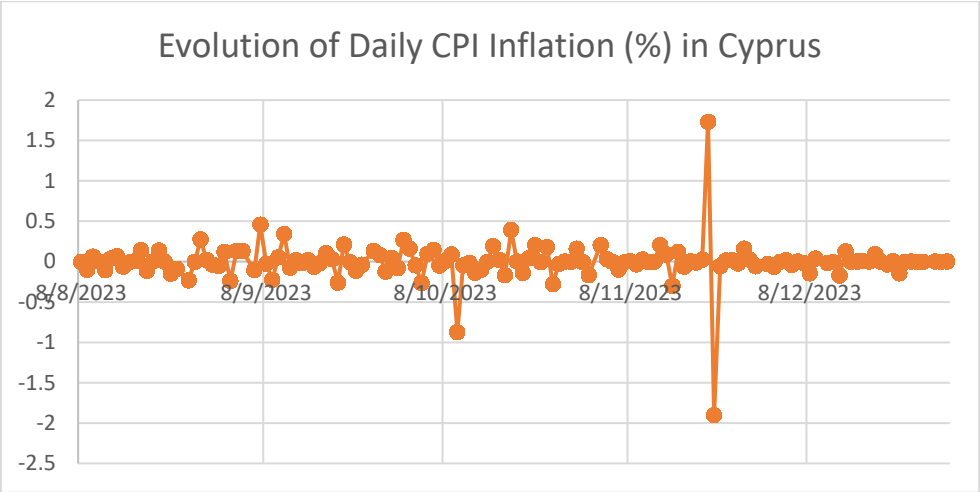


FIGURE 2
Daily evolution of online inflation in Cyprus

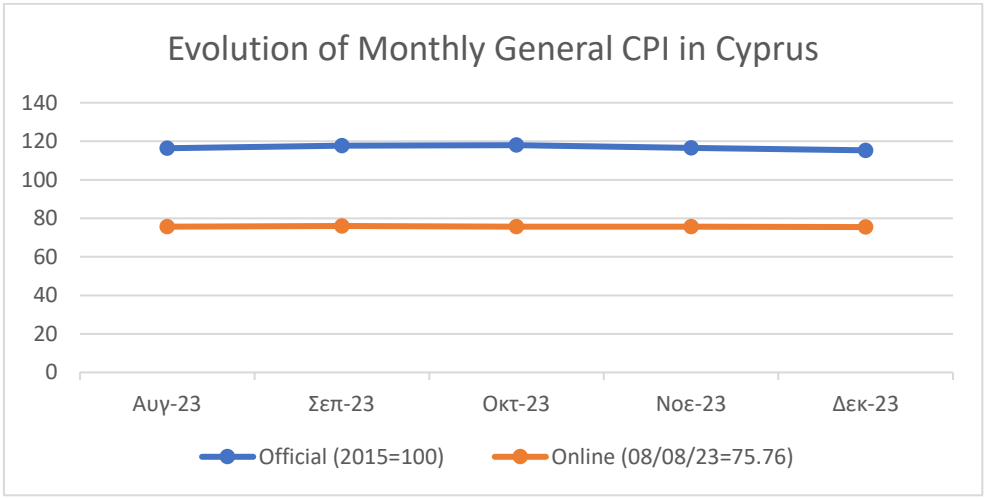


4.2 Comparison with CYSTAT

The Statistical Service of Cyprus gathers price data exclusively from urban districts of Nicosia, Larnaca, Limassol, and Paphos. Each city's monthly product price fluctuations are weighted based on their population percentages: Nicosia 42%, Limassol 30%, Larnaca 18%, and Paphos 10%. CYSTAT records prices for 805 goods and services once every month, except for some seasonal products (e.g. vegetables and fruits), meat and fuels, whose prices are collected every week (every Thursday). From January 2016 and onwards, the CPI reference period (base year) is 2015.

Figure 3 depicts the evolution of the monthly General CPI in Cyprus from August to December 2023 as estimated by this work (online data) and CYSTAT (official/offline data).

FIGURE 3
Monthly evolution of General CPI in Cyprus (offline *vs.* online)



Discrepancies between the Cyprus BPP's online and CYSTAT's offline results are noteworthy, stemming from several factors as the following:

i. The reference/base period is 08/08/2023, while CYSTAT employs the year 2015 as its base year for which the General CPI = 100.

ii. Disparities in the quantity and types (i.e. quality) of items in the CPI basket contribute to differences in the indices.[7] So, a primary distinction lies in the composition of the basket of goods and services considered by each method. Qualitative distinctions arise as CYSTAT does not disclose the specific items included in the official CPI basket.

iii. The total CPI expenditure weight differs; CYSTAT maintains 100%, whereas for this approach is 75.76%.

iv. CYSTAT's sampling frequency is monthly (except for some seasonal products, meat and fuels which is weekly), whereas Cyprus BPP gathers prices data daily – the prices of the last Thursday of each month are used as a proxy of monthly CPI.

v. CYSTAT applies population-based weights to price fluctuations in urban districts, a practice which is not adopted in this work.

Overall, the offline and online CPI baskets differ both quantitatively and qualitatively, and thus the discrepancy in inflation estimates.

However, for comparison purposes the two General CPIs must have the *same* value at the base period. So, both indices are rebased such that to have the value of 100 in 08/08/2023. The results are presented in Figure 4. It is notable that the abovementioned discrepancies in both size and direction still hold.

Finally, in Figure 5 the evolution of monthly inflation in Cyprus is presented as estimated by CYSTAT and this work. As evident, the offline and online inflation curves exhibit similar trends only from September to October 2023, albeit with varying magnitudes.

In general, the disparity between offline and online inflation measurement arises primarily due to the distinct approaches and data sources employed by each method. Offline methods typically rely on infrequent pre-existing data sets and historical

[7] In this work, the goods/services provided in the websites of the matched retailers were chosen using filters such as 'popularity' and/or 'low to high price'. Also, for the identification of the most representative basic consumer commodities (supermarket goods) to include in our basket, e.g. which specific breads, we consulted Consumer Protection Service Price Observatories, which concern the recording and monitoring of the prices of widely used consumer products sold in a large number of properties in all the provinces of Cyprus.

records, which may not capture real-time fluctuations in market prices. In contrast, online methods dynamically gather information from the internet, allowing for a more immediate and responsive reflection of current market conditions. Therefore, the divergence in values stems from the inherent differences in data gathering, frequency, and the adaptability of each method to capture the evolving nature of economic conditions.

FIGURE 4
Monthly evolution of the rebased General CPI in Cyprus (offline *vs.* online)

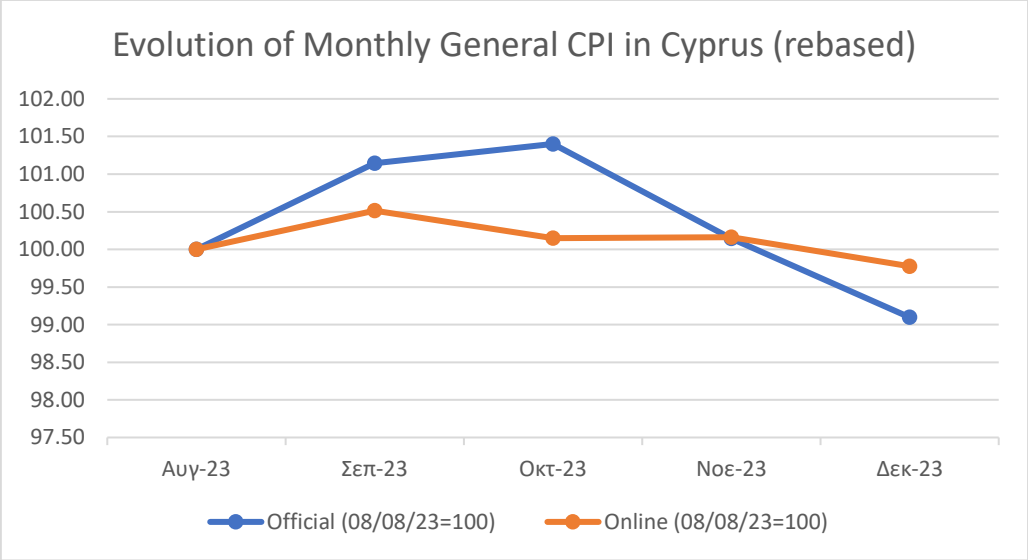
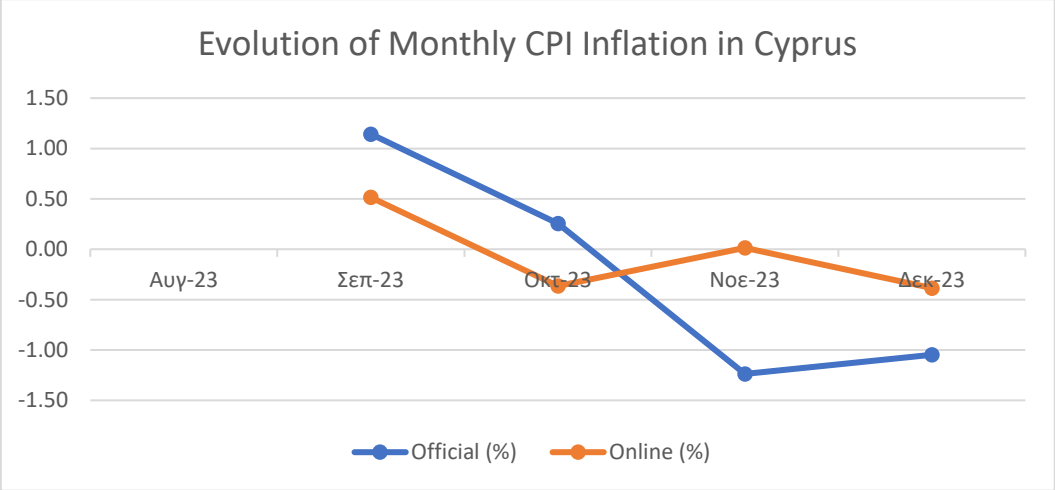


FIGURE 5
Monthly evolution of CPI Inflation in Cyprus (offline *vs.* online)



5. Discussion

The Billion Prices Project (BPP) represents a paradigm shift in the field of economic measurement, presenting a dynamic and innovative approach to understanding inflation rates in Cyprus in the contemporary digital landscape. The discussion that

follows delves into key aspects of the project, emphasizing its significance, challenges, implications, and potential avenues for further exploration.

One of the project's notable strengths lies in its ability to harness the vast potential of web-scraping technologies to capture real-time pricing data. By extracting information from a diverse range of online sources, including e-commerce platforms and digital marketplaces, the project provides a more nuanced and timely perspective on inflationary pressures. This real-time capability addresses a critical limitation of traditional methods, which often rely on periodic surveys and data collection processes that may lag the rapidly changing dynamics of digital economies.

The inclusivity of the project's methodology, encompassing a broad spectrum of goods and services, contributes to a comprehensive understanding of consumer spending patterns. The scalability and adaptability of web scraping allow for the integration of a wide array of products, reflecting the evolving nature of market demands. This inclusivity is particularly vital in a world where digital transactions and online commerce play an increasingly dominant role in shaping economic activities.

However, the project is not without its challenges and considerations. Ethical considerations surrounding web scraping, user privacy, and adherence to website terms of service underscore the need for responsible data collection practices. Striking a balance between data accessibility and ethical standards is crucial to ensure the sustainability and legitimacy of the project. Moreover, the implementation of an automation process in GitHub is a significant step towards streamlining the data collection workflow; however, ensuring its seamless operation requires ongoing monitoring and oversight. The *scrape_tool.py* script, designed to run daily, may encounter network issues, web-connectivity problems, or other unforeseen challenges that can disrupt the data-scraping process.

The comparative analysis with traditional inflation indices reveals interesting insights into the reliability and accuracy of the web-scraping methodology. Discrepancies between the results obtained from the BPP and conventional indices offer valuable information on potential biases and limitations. Understanding these differences contributes to a more informed interpretation of the data and guides future refinements to the methodology.

Looking forward, the BPP opens avenues for further exploration and refinement. As digital economies continue to evolve, adapting the methodology to capture emerging trends and technological shifts becomes imperative. Additionally, exploring the applicability of the project’s approach to different sectors, industries, or regions could

enhance its versatility and contribute to a more comprehensive understanding of inflation dynamics on a global scale.

6. Conclusion

The BPP stands as pioneering initiative that redefines the landscape of economic measurement through the innovative integration of web-scraping technologies. The project, which was first initiated in 2008, addresses the limitations of traditional inflation measurement approaches in the face of a rapidly evolving digital economy.

Through daily web scraping of HTML files from diverse retailer websites, the project captures a dynamic and real-time snapshot of pricing information, including product details and retailer specifics. This approach not only reflects the shifting patterns of consumer behavior in the digital era but also ensures a comprehensive representation of inflationary pressures within different economies.

The methodology employed in this work showcases the technical intricacies of web scraping and data processing. The use of Python based tools such as *Beautiful Soup* and *Requests* demonstrates the adaptability and efficiency of the chosen approach, enabling the extraction of relevant information from multiple sources.

The BPP's commitment to inclusivity, manifested in the broad spectrum of goods and services included in the CPI basket/dataset, offers a nuanced perspective on consumer spending patterns. The scalability and adaptability of the methodology not only capture the complexities of modern markets but also pave the way for a more responsive and accurate understanding of inflation rates.

However, recognizing the current limitations, the ongoing work aims to enrich the web-scraping process employed in this study. The goal is to collect more price data over time from a broader array of retailers, thereby constructing more accurate and comprehensive price statistics and indices. This continual refinement seeks to overcome existing challenges and improve the efficiency and reliability of the data collection methodology.

In conclusion, the BPP serves as a beacon of innovation, offering a forward-looking perspective on economic indicators and contributing to the continuous evolution of economic methodologies. In this work, an *alternative source* of data is utilized –online price data– for the first time to calculate Cyprus' inflation. It is demonstrated that online data can also be a reliable source of information for inflation measurement. Their 'big data' characteristics can greatly *simplify* measurement. Lastly, online prices have the

potential to dramatically increase the amount and quality of micro price data available for academic research.

As we move forward, the lessons learned, and advancements made by this project will undoubtedly shape the future of economic measurement in an increasingly interconnected and digitalized global economy.

Acknowledgements

The authors acknowledge and thank funding from the "Economic Modelling Hub for Economic Policy" funded by the European Commission Recovery and Resilience Plan in Cyprus.

References

Cavallo, A., and Rigobon, R., (2016), "The Billion Prices Project: Using Online Prices for Measurement and Research", *Journal of Economic Perspectives*, 30(2): 151-178.

A. Cavallo, R. Rigobon, 2008, Available at: The Billion Prices Project

Cavallo, A., (2017), "Are Online and Offline Prices Similar? Evidence from Large Multi-Channel Retailers", *American Economic Review*, 107(1): 283–303.

Eurostat, Methodological Manual, 2018, Harmonised Index of Consumer Prices (HICP), Available at: HICP Methodological Manual (ec.europa.eu/eurostat)

Laspeyres Price Index, Available at: Laspeyres Price Index – Overview, Formula, and Example (corporatefinanceinstitute.com)

Appendix

A.1 Structure of the Consumer Price Index

The coverage and classification of item indices are based on the international classification system for household consumption expenditures known as *Classification of Individual Consumption According to Purpose (COICOP)*, which was developed by the United Nations. Founded on national accounts principles, the COICOP system is the starting point for defining which expenditures, in principle, should be included in consumer price indices[8].

[8] Eurostat, Methodological Manual, 2018, Harmonised Index of Consumer Prices (HICP), Available at: HICP Methodological Manual (ec.europa.eu/eurostat)

COICOP is a hierarchical classification system that divides the basket of goods and services into:

i. Divisions (for example: 01 Food and Non-Alcoholic Beverages) – 2-digit level
ii. Groups (for example: 01.1 Food) – 3-digit level
iii. Classes (for example: 01.1.1 Bread and Cereals) – 4-digit level
iv. Subclasses (for example: 01.1.1.1 Rice) – 5-digit level

Subclasses are currently the lowest regularly published COICOP level (5 digits). Consumer price indices are produced in stages, with indices derived at each stage weighted together to produce higher-level indices. A sample of prices is collected in line with the COICOP classification system from a selection of items that are representative of consumer expenditure; prices are only collected for those items selected.

The lowest aggregate of prices, an 'elementary aggregate', covers all prices collected for one item. Item indices are first aggregated into subclass indices, which are then aggregated into class indices. Class indices can be further aggregated to form group, division, and overall indices.

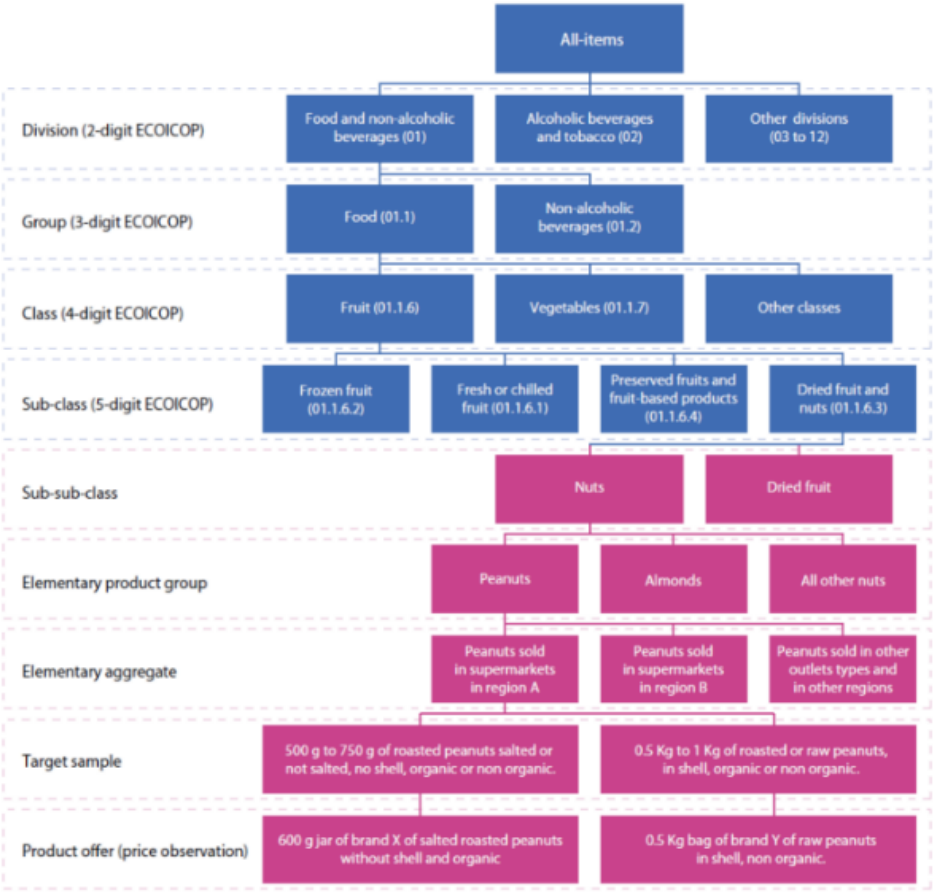
When the index is said to cover or refer to a given population, it means that the weights have been calculated to reflect the expenditure of that population. Regarding prices, the basket is not comprehensive since it does not include every possible item. However, the weights reflect all expenditure by households that is within scope, and items that are included are chosen because they make up a significant proportion of households' expenditure.

According to the HICP Methodological manual, ECOICOP is the version of COICOP used in all EU economic statistics covering consumer expenditure. It should be noted that even though ECOICOP extends to 5 digits (that is, 4 levels of classification excluding the single all-items level), each Member State must usually extend the classification to a more detailed level. This lower level is generally not for publication but is required because elementary product groups and elementary aggregates are commonly defined well below the level of ECOICOP 5-digit subclasses. It is at this level that products are sampled, which in turn determines the structure of price collection. The elementary aggregates represent the building blocks of the CPI from which the ECOICOP aggregates are calculated. Only those price indices and weights at the 5-digit subclass level and above are normally required to be transmitted to Eurostat each month. Overall, the CPI basket can be viewed as consisting of two levels:

- Level I (down to the 5-digit level) is comprised of product groups and their respective weights, which are organized according to the ECOICOP classification.
- Level II (below the 5-digit level), by contrast, uses no pre-defined classification system. The availability of detailed expenditure data by product, the sampling approach used, and resource availability will often dictate how an individual country decides to organize and structure its classification system at this lower sub-5-digit level.

Regardless of how Level II is organized, elementary aggregates must be created and exist somewhere within that Level II structure, since they are the building blocks of the CPI. At the elementary aggregate level and above, the product descriptions and the expenditure weights remain unchanged or fixed, at least in between periodic reviews. Below the level of the elementary aggregate are the product offers, which are priced each month. These may or may not be implicitly weighted. Fig. A.1 illustrates a possible stratification structure showing these two levels for a specific example (nuts).

FIGURE A.1
Classification structure for the various levels of aggregation of the Consumer Price Index



A.2 Elementary Aggregates

At the lowest level of aggregation, detailed weights are not available with the current data sources. To deal with this lack of weighting information, unweighted index formulae are used. The set of indices created using these index formulae are called 'elementary aggregates'. These combine prices into indices, treating all the products as equally important. This gives the prices an equal weight, which is the reciprocal of the number of prices in that stratum. An elementary aggregate index can be constructed in different ways. The most used unweighted index number formulae are the following:

- Jevons index (the geometric mean of price relatives or the ratio of geometric mean prices):

$$I_{Jevons}^{0,t} = \sqrt[N]{\prod_{i=1}^{N} \frac{p_{it}}{p_{i0}}} = \frac{\sqrt[N]{\prod_{i=1}^{N} p_{it}}}{\sqrt[N]{\prod_{i=1}^{N} p_{i0}}}$$

- Dutot index (the ratio of arithmetic mean prices):

$$I_{Dutot}^{0,t} = \frac{\frac{1}{N}\sum_{i=1}^{N} p_{it}}{\frac{1}{N}\sum_{i=1}^{N} p_{i0}}$$

- Carli index (the arithmetic means of price relatives):

$$I_{Carli}^{0,t} = \frac{1}{N}\sum_{i=1}^{N} \frac{p_{it}}{p_{i0}}$$

Eurostat regulations permit the use of Jevons and Dutot indices but forbid the use of the Carli index on the grounds that it does not produce indices that are comparable with other formulae, such as Dutot or Jevons. Carli index may be used exceptionally where it can be shown not to fail the comparability requirement. The regulations therefore help to ensure that differences in inflation rates between EU countries reflect underlying differences in price changes, and not simply differences in the basic formulae used to aggregate the price data.

According to HICP Methodological manual, among EU member states, 17 currently use the Jevons index in their national consumer price index (Austria, Bulgaria, Croatia, Cyprus, Denmark, Finland, France, Greece, Ireland, Italy, Luxembourg, Poland, Portugal, Romania, Slovenia, Spain and Sweden); 8 currently use the Dutot index (Belgium, Czech Republic, Estonia, Germany, Latvia, Lithuania, Malta and Slovakia); and 3 currently use a mixture of Jevons and Dutot (Hungary, Netherlands and the

United Kingdom). Beyond Europe, Australia, Canada, New Zealand, and the USA mainly use Jevons in the calculation of the national consumer price index, while Japan uses Dutot.

A.3 New Products and Product Offers – Quality Adjustments

New products and product offers frequently appear on the market, while others disappear. Their dynamics cannot be ignored without risking a degradation of the sample. When a product offer in the sample is no longer available or no longer popular, it needs to be replaced, with possibly a quality adjustment, to keep the sample representative. When making such replacements, it is essential to follow the CPI principle of comparing prices of product offers on a like-with-like basis so to reflect pure price changes.

Replacement product offers should be either identical or essentially equivalent to the one replaced, where they should be either the same or similar enough in terms of their characteristics that consumers perceive them to be equivalent. On the other hand, from the consumer perspective, a replacement product offer is not seen as equivalent, i.e. its characteristics differ from the replaced product offer, it is necessary to apply some type of quality adjustment to the price comparison. Replacements as described above can occur at any time of year. They may be forced, due to the disappearance of a specific product offer, or they can be planned, i.e. a new product model has appeared on the market which has become representative of current consumer purchases, as such the existing product-offer which is no longer representative is replaced by the new one to keep the sample representative. Re-sampling keeps the basket representative of consumer expenditure.

A.4 Descriptive Statistics

Here, the structure of the national CPI basket is presented (Table A.1), the ECOICOP weights per division in the CYSTAT (official/offline) and the (online) CPI basket (Table A.2), the distribution of the number of goods and/or services across the retailers in the online CPI basket (Table A.3), and some basic summary statistics of the online price data sets used for the calculation of the monthly inflation in Cyprus (Table A.4).

TABLE A.1

Structure of the Cyprus official CPI basket

Division of Goods and/or Services	Num. of Groups	Num. of Classes	Num. of Subclasses
1. Food & Non-Alcoholic Beverages	2	11	57
2. Alcoholic Beverages & Tobacco	2	3	6
3. Clothing & Footwear	2	5	10
4. Housing, Water, Electricity, Gas & Other Fuels	4	11	11
5. Furnishings, Household Equipment & Supplies	6	9	28
6. Health	3	6	9
7. Transport	3	9	13
8. Communication	3	0	7
9. Recreation & Culture	6	18	31
10. Education	5	0	2
11. Restaurants & Hotels	2	1	3
12. Miscellaneous Goods & Services	6	8	21
TOTAL	**44**	**81**	**198**

TABLE A.2

ECOICOP weight per division for the year 2022 in the CYSTAT (offline) and Cyprus BPP (online) CPI basket

DIVISION	ECOICOP WEIGHT (%) – OFFICIAL	ECOICOP WEIGHT (%) – ONLINE
Food & Non-Alcoholic Beverages	19.06	19.03
Transport	15.24	11.66
Housing, Water, Electricity, Gas & Other Fuels	11.19	4.94
Miscellaneous Goods & Services	8.60	4.41
Restaurants & Hotels	8.36	7.97
Clothing & Footwear	7.21	7.06
Furnishings, Household Equipment & Supplies	6.44	4.38
Health	6.37	2.14
Recreation & Culture	5.75	2.49
Communication	4.39	4.39
Education	3.92	3.82
Alcoholic Beverages & Tobacco	3.47	3.24
TOTAL	**100**	**75.76**

TABLE A.3

Distribution of the number of goods and/or services across retailers in the online CPI basket

Retailer	Number of Goods and/or Services
Adventure Without Limits (AWOL)	8
Alphamega	333
Alter Vape	3
Athlokinisi	17
Bwell Pharmacy	4
Cablenet	4
Cyprus Energy Regulation Authority (CERA)	3
Cyprus Ministry of Education, Sport, and Youth	6
Cyprus Post	6
Cyprus Telecommunications Authority (CYTA)	4
Epic	3
E-WHOLESALE	1
Electroline	3
European University Cyprus	1
Famous Sports	19
FuelDaddy ***(Agip, EKO, Eni, Esso, Fill n GO, Petrolina, Shell, Staroil, Total Plus)***	1,477
IKEA	29
Marks & Spencer	17
Mazda	1
Moto Race	14
Nissan	1
Novella Hair Mode	2
NUMBEO	1
Pizza Hut	3
Primetel	6
Rio Cinemas	6
Sewerage Board of Limassol-Amathus (SBLA)	4
Sewerage Board of Nicosia (SBN)	2
Sewerage and Drainage Board of Larnaca (LSDB)	3
Stephanis	40
Stradivarius	5
SupermarketCy	430
The CYgar Shop	1
The Royal Cigars	4
Water Board of Nicosia (WBN)	1
Water Board of Larnaca (LWB)	2
Water Board of Limassol (WBL)	2
Wolt ***(Costa Coffee, Piatsa Gourounaki Nicosia, Pixida Nicosia, Kofini Tavern Limassol, Vlachos Taverna Larnaca, Zakos Beach Restaurant Larnaca, Paphos Tavernaki, Ocean Basket Paphos, McDonald's)***	14
TOTAL	**2,476**

TABLE A.4
Summary statistics of the online prices (€) of goods and services in each dataset per month.

Dataset creation date	*N*	Min	Average	Max	Standard Deviation
31/08/2023	2,476*	0.03	119.53	20,650	1,157.52
28/09/2023	2,476*	0.03	119.53	20,650	1,157.52
26/10/2023	2,476*	0.03	119.88	20,650	1,159.23
29/11/2023	2,476	0.03	49.37	20,650	740.57
27/12/2023	2,476	0.03	52.32	20,650	763.17

Note: * From 08/08/2023 to 03/11/2023, it seems that the daily data sample size (*N*) is around 1,000 instead of 2,476, which is the actual size. This is because in the CPI basket the daily average price of the each fuel type in Cyprus (Unleaded 95, Unleaded 98, Road diesel, Heating diesel, and Kerosene) is included as calculated directly from the Consumer Protection Service website (CeGG) of the Ministry of Energy, Commerce and Industry. From 04/11/2023 onwards, however, the daily prices of the 5 fuels from all filling/gas stations in Cyprus (1,477 prices) are included as scraped from FuelDaddy, so the dataset appears to have its actual size of 2,476 observations (goods and services) per day.