Studies of new data sources and techniques to improve CPI compilation in Brazil

Lincoln T. da Silva[*], Ingrid L. de Oliveira[†],
Tiago M. Dantas, Vladimir G. Miranda[‡]

Abstract

The advent of new technologies is promoting deep changes in many aspects of society. National Statistical Offices (NSO) are not immune to these transformations and face challenges in how to measure the growing effect of digitalization on consumer habits in an opportune and efficient way. On the other hand, the digital revolution is also providing new opportunities for Consumer Price Index (CPI) compilers since new data sources are available to increase accuracy and development of methodologies that can be combined with the traditional ones. Here we present two case studies of usages of web data to improve CPI at the Brazilian Institute of Geography and Statistics. In the first case, we discuss the necessary steps and difficulties to implement a price scraper to replace manual collection for airfares, which are typically purchased in online platforms, by an automatic one. The main idea here is to introduce this analysis as a pilot to replace other web-commercialized components of the CPI basket. Such techniques might also be useful to deal with digital products, such as ride-sharing apps, which present some measurement issues similar to airfares. The second study deals with the use of web scraping techniques to implement hedonic models for quality adjustment in CPI. These techniques allow the extraction of product characteristics in an easy, cheap and fast way and do not rely on extensive scraping from websites. Moreover, we address the issue of using web prices to provide parameter estimation for the hedonic models considering that online and offline prices may differ. We evaluate if the models applied to both scenarios are consistent with the standard methodology.

1 Introduction

The technological development and internet popularization have profoundly changed society. Use of the internet and electronic devices have been increasingly dominating our daily life. The outcome of this process generates newly numerous and continuous data sources, creating opportunities to extract valuable information about the economy, population, and politics, for example. National Statistical Offices (NSO) are not immune to the transformations arising from the digital revolution. Their responsibility to provide reliable information to the society requires a constant search for efficient ways to produce meaningful, frequent, and high-quality statistics. The new data sources create challenges and opportunities for NSOs not only to produce statistics but also to improve existing methods of data collection.

[*]lincoln.silva@ibge.gov.br
[†]ingrid.oliveira@ibge.gov.br
[‡]vladimir.miranda@ibge.gov.br
The views expressed in this paper are those of the authors and do not necessarily reflect the views of IBGE.

Inspired by the growth in volume, variety, and speed of new data available, the United Nations (UN) created in 2014 a Global Working Group (GWG) on Big Data for Official Statistics comprising twenty-eight countries including Brazil. Its primary goal is to explore the benefits and challenges concerning the use of new data sources and technologies by the NSOs, including the Sustainable Development Goals (SDG) monitoring. Eight task teams were established, with themes varying from access and partnerships until a Global Platform creation for data, services and applications[1].

One of the task teams of the GWG concerns the use of new data sources for the compilation of prices indices. Researches on this theme led to interesting studies on the use of big data for the development and improvement of consumer price indices (CPIs) by private and public entities. On the private side, a popular approach is MIT's Billion Prices Project [Cavallo and Rigobon, 2016]. It collects prices from hundreds of online retailers on a daily basis to experiment on potential uses and improvements of new data sources for economic indicators construction. Experiments in the field have also been reported on the public side (mainly developed by NSOs), which have explored the new data sources to develop cheaper and more efficient collection practices, automatic machine-learning based tools to deal with massive data sets, and new methodologies able to handle new features of those big data [Daas et al., 2015; Polidoro et al., 2015; Breton et al., 2016; Gumundsdttir and Jónasdóttir, 2016; Hov and Johannessen, 2018; Loon and Roels, 2018; Mendonça and Evangelista, 2018].

Following this global trend, the Brazilian Institute of Geography and Statistics (IBGE) has been developing some pilot projects on the use of new data sources on its CPIs. IBGE is responsible for the compilation and publication of the official Brazilian CPIs which are structured in a framework named National System of Price Indices (SNIPC) [IBGE, 2013; Miranda et al., 2019]. The SNIPC provides a set of CPIs for distinct target populations according to different income groups, geographical areas, and periodicities. Among the indices in the SNIPC scope, the Extended Consumer Price Index (IPCA) is probably the most important one since it is the target adopted by the Brazilian central bank for the definition of its monetary policy [IBGE, 2013; Miranda et al., 2019].

In this paper, we discuss two pilot projects on the use of big data sources to improve the quality and efficiency of the CPIs produced under the SNIPC scope. In this sense, the article presents the results for two ongoing projects that have been developing at IBGE since the last year. We investigate the use of web scraping as an automatic tool used to accelerate and expand data collection for the SNIPC while reducing the collection costs.

Within this goal, the first project presented discusses the replacement of the manual collection of prices for airfares. One of the costliest processes of CPI compilation relates to the prices collection tasks. Prices collection in the SNIPC is mainly conducted by field collectors through personal visits to the commercial establishments. However, there is a small portion of products and services whose prices are already collected manually online at store websites. Among such products are airfares, skincare products, make-up, and books. Due to this particularity, these products are suitable candidates to have their collection practices replaced by automatic methods. Airfares are our first choice of evaluation since their manual collection is the most time consuming between Brazilian CPI online products. This fact is also attractive for the development of web scrapers since the time spent for the development and maintenance of the algorithms is worth if considering the gains in the amount of data extracted and the reduction in collection time. The process also allows the collection and backup of the information in a highly accurate and controlled environment.

The second case study analyses the use of web scraping tools to support the implementation of hedonics in the Brazilian official CPIs. Quality change treatment is one of the most important methodological challenges of CPI compilers. The most suitable tool to deal with the problem is

[1] An inventory of works in development by GWG members is available on https://unstats.un.org/bigdata/.

hedonic modeling. However, the implementation of hedonics on CPIs rely on a robust database containing detailed characteristics on the products of the CPI sample. The manual collection of such information is an important barrier for the implementation of hedonics by CPI compilers. Nowadays, however, detailed characteristics of the products in the sample of the CPIs can be found at the store websites. This possibility suggests combining traditional and new sources to improve the methods employed in CPIs.

The paper is structured as follows: section 2 presents a brief discussion on the methods of collection for web data; Section 3 describes the approach used for the automatic collection of airfares and the main results obtained by comparing the manual and automatic databases during the period of study; section 4 shows the experiment of using web data to support the implementation of hedonics in the SNIPC. We then discuss the preliminary results of models derived for refrigerators, using data obtained via brick and mortar and web stores; finally, in section 5 we summarize the results and present our conclusions. Despite the fact that the results presented in this paper are derived only for the IPCA, the extension to other indicators of the SNIPC is straightforward.

2 Data collection techniques

The data gathering process using the internet can be done in several manners. Two popular methods are web APIs (Application Programming Interface) and Web Scraping. The former refers to, essentially, requesting data directly from the website database, with the connection rules settled by the website owner. On the other hand, the latter concerns algorithms that convert data present in HTML to structured formats easy to understand.

Ten Bosch et al. [2018] identifies three phases for web scraping for official statistics: site analysis, data analysis and design, and production. In the site analysis phase, one should examine the website technical features regarding data availability, programmability, level of interaction needed, and legal rules. The data analysis and design phase involve evaluating the quality and applicability of the collected data regarding their use for statistics and indicators construction. Occasionally, the first two phases overlap. The data validation leads us to the third step - the production phase - when the NSOs start using the collected data to compute official statistics.

Concerning web scraping, there are still different ways of proceeding. These techniques vary both in cost and features. Three of them are frequently employed: scripting languages, point-and-click tools and as a service. When using scripting languages like Python and R, the data extraction algorithms are entirely developed. This method allows codes to be customized accordingly to the problem of interest. It requires, however, high programming skills. The use of point-and-click tools, differently, enable easy extraction of data from a website by pointing and clicking on the desired information. Non-developers can benefit from these tools while complex projects may demand more flexible means. Another possibility is the hiring of companies that offer web scraping as a service. It is a good option in the absence of in-house developers or if the data analysis phase demands a massive effort. However, data acquisition costs and lower control over the data extraction are drawbacks of this process. Hoekstra et al. [2012] presents a further discussion about the choice of technology for the Statistics Netherlands applications.

The results presented in this work rely on applications that were developed in-house using the R software. There are two main R packages for web scraping: rvest [Wickham, 2016] and RSelenium [Harrison, 2019]. The difference in their uses relates to how the websites make their data available. The rvest package is preferable when the source code contains all the target data. Since in such instances the data extraction does not require to open a browser, hence the rvest package performs quickly. However, the rvest package may fail to correctly extract the data for websites that demand

complex interactions as mouse clicks and filling in forms. These cases are better managed by RSelenium package because it permits to emulate human-like action. Some shortcomings rise since the RSelenium package requires complex coding and the extraction time is usually higher when compared with the rvest package.

We adopted both packages due to our nature of the project. For the collection of airfares, we are expected to provide the website with information of, at least, origin, destination, and departure and arrival dates. The interaction required by airlines web pages demands the use of RSelenium. On the other hand, the characteristics of a particular product are commonly found somewhere on its sales web page. It is the reason why we used rvest package to the CPI quality adjustment project.

Despite the fantastic opportunities of the web scraping for NSOs, it is important to mention that some practical issues may occur. Firstly, the code may fail unexpectedly due to, for instance, the internet speed or a web page instability. These problems are not predictable, which leads us to the second problem: the absence of human inspection. We need to develop algorithms able to prevent the collection of wrong information. Besides our effort to write reliable codes, screens shots are stored to control mistakes. Since each website has a particular data extraction code, any web page change may affect the web crawler effectiveness. Changes in fare option name, checked baggage fees or minor layout modifications, for instance, are easily handled. Major redesigns, though, may demand a labor-intensive re-coding or even make the web crawler useless. Our practical experiences are discussed further in the following sections.

3 Case study 1: automatic collection of airfares

3.1 Overview of the methods used for the compilation of airfares inflation and motivation

The rise and expansion of e-commerce have been producing significant changes in the way goods and services are transacted. Commercialization habits based on consumers visiting brick-and-mortar stores to perform a purchase are now partially or completely replaced by a platform where the stores announce their products on a web page and consumers are allowed to order in a fast and comfortable way.

Commercialization of airfares is a typical example where these new practices are massively employed. Nowadays, every air company announces its products and prices online and allows consumers to buy the tickets via website. This transaction practice has been fully incorporated by society, and this is the standard way people buy air tickets in Brazil.

Due to this fact, the prices collected for the calculation of airfares inflation in the SNIPC are already extracted from the of the most important Brazilian airline websites. As airfares can be characterized by a plethora of different destinations and product types, a determination of some representative tariffs to represent this complex universe is necessary. A compromise needs to be established in order to satisfy the requirements of the matched model approach [ILO, 2004] in which CPIs usually rely on and the resource constraints for the data collection.

In order to compile airfare inflation, IBGE follows the approach internationally recommended of defining a "standard tariff profile" whose prices should be tracked [ILO, 2004; IBGE, 2013]. To represent a "typical" consumer behavior, IBGE collects air ticket prices for domestic flights performed by an adult with checked luggage. The interval between departure and return flights are supposed to last 8 days, with a departure on a Saturday and return on Sunday of the following week. The departure date must be performed sixty days in advance of the collection date.

The routes correspond to the most visited destinations for leisure purposes with departures from each of the 16 capital cities under the scope of SNIPC [IBGE, 2013]. Only the most important

airport in each city is considered for both departure and arrival flights.

To compile the inflation of month t, IBGE uses the prices collected in each week of the month $t-2$ according to a predefined calendar. The prices corresponding to all ticket categories available for the standard profile, for the destinations of interest, by the time of the collection are eligible for extraction.

The price collection is currently performed by field collectors in each of the 16 local units in the states where IBGE compiles the index. The process is manual, and the collectors need to visit the air company website and write down the eligible prices each week. This process is costly since it is very time-demanding. Furthermore, there is no backup of the data extracted, which subjects databases to errors and demands an additional time of analysis by the head office analysts to check possible inconsistencies and perform the necessary editing.

Due to the commercialization characteristics of the airfares, based on online purchases and the costly collection process in use, this sector appears an attractive candidate to implement automatic price collection tools to speed up the procedure and to improve the quality of data.

3.2 Experiment description

In our pilot project for the automatic collection of airfares, we developed an in-house web scraper using the R software. Since every air company web page have a particular design, a different robot need to be constructed for each of them. We briefly describe the main general challenges for the development of the robots and the solutions adopted.

To extract the prices of air tickets from the airline web pages one needs to inform a set of parameters such as the origin and destination of the flights, departure and return dates, etc. After this, the tickets available are displayed to be chosen by the consumer. For automatic extraction of the prices, most of airline websites requires the robots to perform these navigation steps emulating the procedures of a person. To overcome this problem a browser automation tool is necessary.

A popular browser automation tool is Selenium, which is developed and largely adopted for automating tests of web applications carried out on a web browser. It is used, for instance, to automatically test the possible outcomes of the intense traffic on websites of e-commerce companies. Selenium is flexible and allows tests to be run on different browsers and integrates with different programming languages.

For airfares we managed to emulate the manual process by the adoption of the RSelenium package from the R software. Such package allows the connection with a Selenium server from within R, and hence allows our navigation through the air company sites until we reach the pages that contain the information of our interest.

Once we reach the pages of interest, some extra programming steps are necessary to extract the desired information and to organize the data in a structured format appropriate for use. We collect information for the flights of interest (routes and dates predefined) for all the tickets offered at the websites at the moment of the request. The extracted data contains information on the company, cities of departure and destination, dates of departure and return for all kinds of ticket's categories and the price associated with all these characteristics.

The project started on the second semester of 2017, and the scrapers have been running since January 2018. The collections follow the same calendar adopted for the manual process and are performed once a week. We restrained our analysis to data collected between January 2018 and September 2018, inclusive, in order to avoid the publication of sensitive data. However, the process was not trouble-free and some drawbacks avoided collection of information in every week of this period. The shortcomings are detailed at the end of the section.

To evaluate the efficiency of the automatic approach by the replacement of the manual process

we need to confront the results derived via both "methods". For the time interval mentioned above, the original databases contain 494,862 and 321,742 airfares for manual and automated extractions, respectively. Each row in the database represents one price for a particular combination of departure and return dates, route, air ticket category and company.

Some data cleaning was necessary to exclude "anomalies" observed in the datasets and to guarantee a "fair" comparison between the two approaches. For instance, we eliminated from the comparison the cases where prices were only available for one part of the route, either for departure or return of a given route. All the prices collected manually for weeks without a match in the automatic process were also excluded from the comparison. We also removed the prices whose collection dates did not match the ones in the official calendar. It may occur, for instance, when the collection date lies on a local holiday in one of the cities where the local units are based. In such circumstances field collectors extracts data previously or after the holiday date. Also, the date registered in the manual database corresponds to the date of the prices entry in the base, which may differ from the calendar date. However, this does not imply that the prices were not collected for the correct dates since the collectors may write down the prices for the dates defined and insert them in the base in a different moment.

After the databases pre-processing, we intersected the manual and automated databases considering a product codifier. It is composed by air company, route, departure and return dates, collection dates, and an identifier to distinguish whether the price is for a departure or a return flight. The remaining manual and automated databases include 320,213 and 305,214 air ticket prices, respectively, which correspond to 64.7% and 94.9% of the initial databases. Though the reduction in the manual database seems very significant, it mainly relies (approximately 72% of the excluded rows) on the exclusion of those weeks where the automatic process did not provide observations. For the remaining rows excluded, 14.8% correspond to the cases where the collection dates are "inconsistent" with the calendar ones, 8.7% corresponding to prices of three new areas that were inserted in the index scope in the period analyzed, which were not covered by the robots by the time of the implementation. A small portion, approximately 4.4%, was excluded due to the observation of some other kind of inconsistency information probably caused by an error when they were manually inserted by the field collectors in the database.

3.3 Results

This section compares data obtained using web scraping and the official microdata to evaluate the adherence of the two collection methods. We need to consider in the comparisons that prices are collected on a weekly basis. The "cleaned" database contains information for airfares collected in twenty-five weeks in the time interval between January 2018 and September 2018, inclusive.

Within each week and for each product (characterized by the product codifier described above), we start evaluating the number of flights collected by the manual and automatic process. The time of collection is not the same for the manual and automatic process, so we expect to observe some differences in the number of flights collected due to dynamic pricing strategies adopted by the air companies.

Figure 1 displays the differences in the number of flights obtained for each individual product codifier in a given week. The data presents the distribution of all the differences in the number of flights observed for all collections throughout the period of comparison. Negative (positive) values represent cases in which the manual process found more (fewer) flights than the robots. It is worth to mention that y-axis in Figure 1 is limited to the interval $[0, 4]$ to facilitate visualization. The null differences amount to approximately 83% of total cases.

In the center of Figure 1 the lighter dashed bar represents the cases with null difference (same

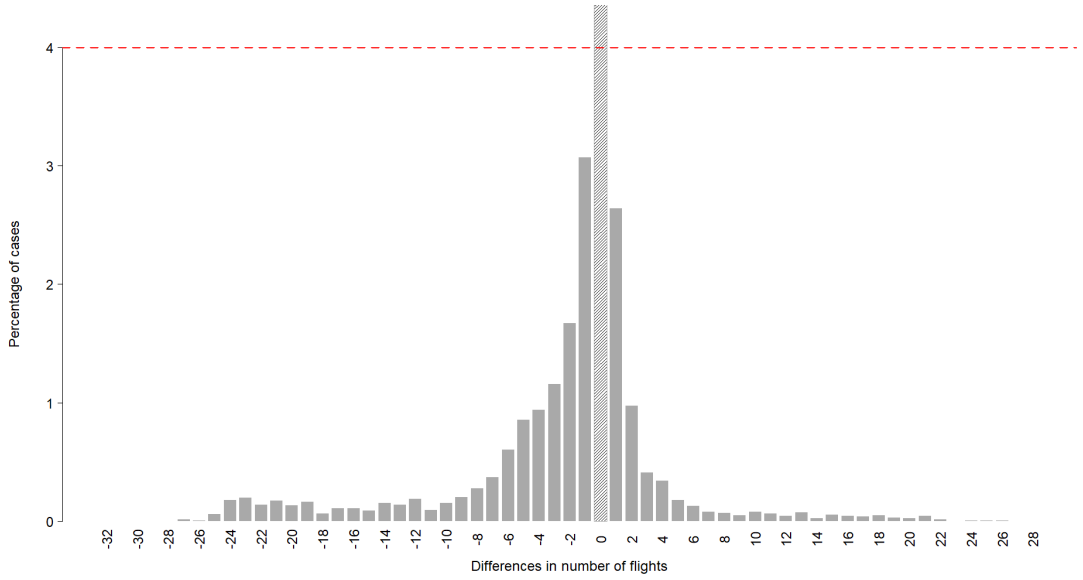


Figure 1: Frequency distribution of the differences in total of flights. Note that the barplot is limited to 4% to better illustrate the distribution of the differences in number of flights. The lighter dashed bar shows the cases with null difference, which corresponds to approximately 83% of cases analyzed. Negative (positive) values denote cases in which the manual process found more (fewer) flights than the robots.

number of flights for both collection methods). As expected, both methods collected the same number of flights in the majority portion of the cases (82.9%). The behavior of the distribution tail is predictable since we cannot ensure identical circumstances for every collection.

Further analysis reveals that the bulk of the distribution of the differences lies within the interval $[-5, 5]$, with the cumulative frequencies in this interval summing up more than 95%. These cases are probably the ones where the collection are performed under the most similar conditions. The values in the distribution tail reflect the higher differences. We suspect that these discrepancies originate from significant departures between conditions under which the experiments were performed. Also, it may suffer the influence of human errors due to the manual data entry process. Our database does not allow, however, an extensive investigation of these discrepancies due to the absence of complementary flight information such as air flight code and time of departures.

To avoid the influence of spurious effects in the following analysis, we restrict the tickets evaluated to those cases where the difference in the number of flights lie in the interval $[-5, 5]$.

We then analyze the difference in the mean prices of the airfares collected. Similar to comparisons of the number of fights, the mean values are calculated for each product codifier for both manual and automatic process. As shown in Figure 2 (right), the tickets collected have the same mean price in approximately 60% of the cases.

The distribution of non-zero mean differences is shown in Figure 2 (left). It presents a bell-shaped distribution with the differences in the mean prices mostly concentrated in small difference values, a behavior supported by the results of the cumulative density shown in Figure 2 (right). We note that 91% of the non-zero differences in the mean prices correspond to values lying in the interval between -100 and 100 Brazilian Reais. This proportion enhances to 95.8% and 99.1% if the mean prices differ in absolute value by $R\$200$ and $R\$500$, respectively. Differences exceeding R$1000 were not observed.

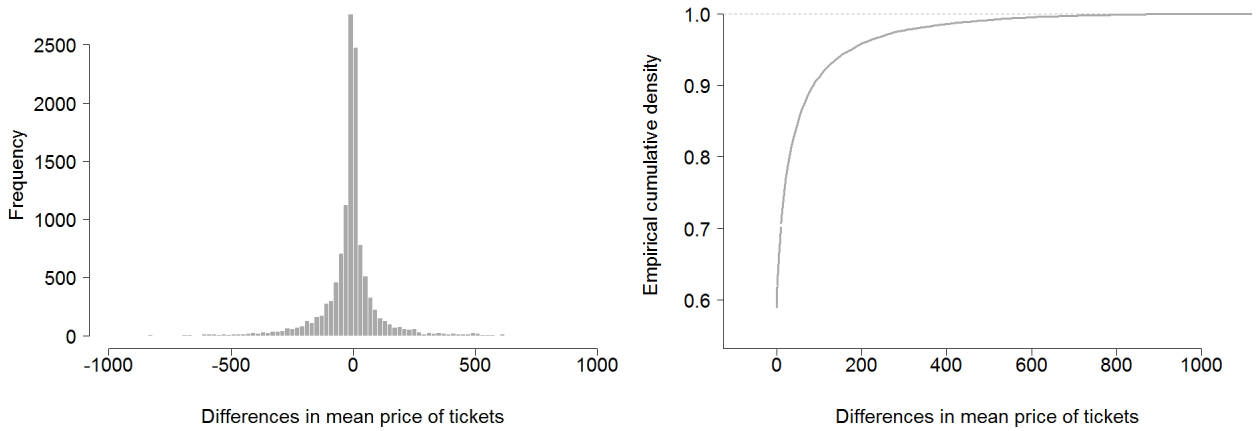


Figure 2: (left) Empirical cumulative density of the difference of mean prices calculated for the manual and automatic processes. (right) Frequency distribution of non-zero differences of mean prices calculated for the manual and automatic processes.

We further investigate the agreement between the results obtained by the two collection methods by computing an "experimental" measure of airfares' variation using the data sets available. The official inflation for airfares is calculated and published by the IBGE on a monthly basis[2]. However, we perform a slightly different exercise. Since we are limited to twenty-five weeks of data, we build a weekly index for the weeks where data is available.

Figure 3 presents the results for the variations in airfares prices during the period under study. Collection dates are omitted due to confidentiality issues. As observed, there is an excellent agreement between the series through almost all the time interval. The results for the few points with higher discrepancies are explained by the presence of a greater offer of flights and ticket categories with higher prices when field collectors executed the data collection. We suspect that the web crawler collected data over a period of promotional fares, which is expected due to the price dynamics of the sector.

3.4 Discussion

The analysis performed showed the potentialities of web scraping when replacing the manual collection process. The comparisons between the number of flights collected, the difference in mean prices and index series shown in Figures 1, 2 and 3 provide significant evidence that the web scraping collection managed to reproduce the manual one successfully.

It is also worth mentioning the reduction in the time of collection obtained using web scraping. For some air companies, web scraping reduced the time of collection in half respective the manual process. The results were derived via a single desktop machine, and hence the process can be improved via the use of more elaborate architectures. Besides time reduction, web scraping avoids human errors and, thus, contribute to the improvement in the CPI compilation. Web scraping also allows developing a robust control of the data collected since automatic screenshots can be performed allowing backup and verification of the information extracted.

[2]The detailed methodology can be found at IBGE [2013].

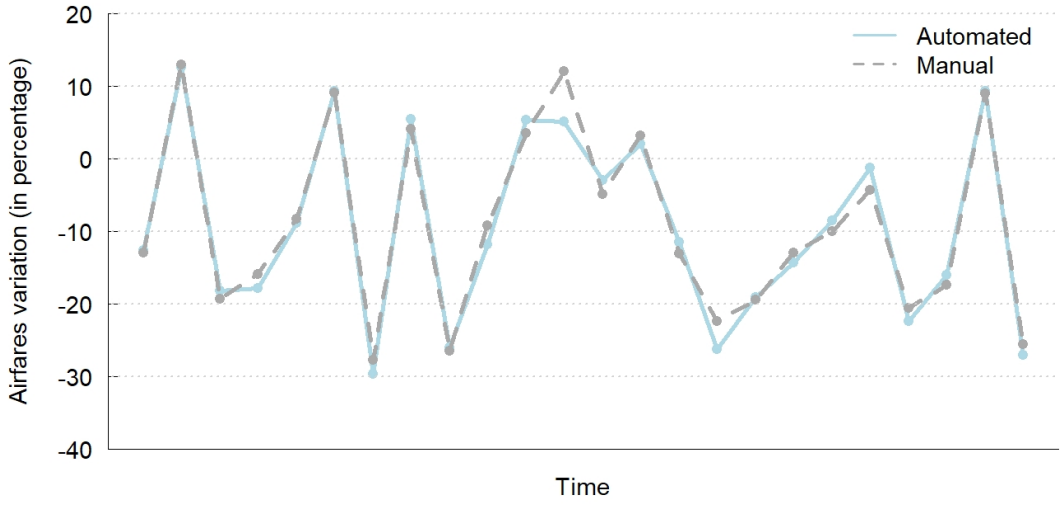


Figure 3: Comparison of weekly airfares inflation indices constructed for the manual and automated databases.

By the use of web scraping techniques for the collection of airfares, the number of routes and flights collected could be easily and cheaply extended allowing the calculation of more robust estimates of the inflation of the sector. The calculation could also be performed on a more frequent basis since prices could be collected for much smaller time intervals (each day, hour, so on).

On the other hand, important challenges that were observed during the execution of the experiment should be taken into account when trying to adopt the web scraping collection instead of manual one in the production process of the CPIs. As we previously mentioned, the automatic collection could not be performed continuously throughout all the time range analyzed. The most common problems which avoided the collection were related to website instability and issues concerning internet connection and speed.

Since websites have their own configuration, we need to custom the scraper for each of them. Besides this, the websites' design may also change without a previous warning and "breakdown" the scrapers. For these cases, the robots need to be "repaired" to deal with these changes. Along the process, one company completely changed its website layout and its navigability. It took a few days to reprogram the R code to extract data correctly. Minor website changes, however, are usually handled on the same day and do not limit the data collection.

Websites are usually not robots-friendly and once the automatic activity is detected the access to the website can be blocked. Also, each price request (manual or automatic) to the air companies websites incur in costs to the companies. Hence, massive access might lead to a non-negligible increase in the company's expenditures. Air companies can avoid this by blocking access to their website and, in critical situations, they may also proceed to legal processes. During our experiment one company blocked our access to its web page for a week, probably due to an extensive number of requests performed.

Another problem we faced regards the use of the R package. A sudden breakdown of one of the packages prevented us from collecting data for a week. It could be a significant problem if, for example, the package maintainer discontinues the package.

To deal with all these issues more tests are being performed, and a more robust tool is being developed by the IT team. We believe, however, that the implementation of the automatic collection

in the production process will not be possible without the collaboration of the air companies. Hence, we started negotiations with the principal air companies to ask for their permission to access their data. To overcome website stability issues we are considering the access of the companies data via APIs. This trajectory is much more stable than the web scraping though caution should be taken to avoid data manipulation since the companies will now have full knowledge of the data being extracted.

4 Case study 2: use of web scraping to support the implementation of hedonics at the SNIPC

4.1 Description of the problem and motivation

The standard approach used for the compilation of consumer price indices is based on two pillars [ILO, 2004]: the fixed basket and the matched model methods. The former states that the consumer habits of a target population can be represented by a basket of goods and services which are usually determined by a household budget survey performed in a given instant of time. The fixed basket also assumes that the quantities of goods consumed within the basket are not altered, that is, that consumers are immune to price changes and do not alter their consumption habits under such circumstances. The latter relies on the assumption that once the products of the basket are defined, such products will be available for pricing (ideally in the same sample of stores originally selected) through all the lifetime of the basket.

Departure from these two basic assumptions constitute what is known as "the quality change" problem [ILO, 2004; de Haan and Diewert, 2017], a theme of central importance for CPIs. Departure on the fixed basket approach occurs due to change in the quantities consumed of the goods and services of the basket throughout time. In the literature such effect is known as a "quality mix change" [de Haan and Diewert, 2017] and it can be properly treated by the use of superlative index formulae. The dynamic nature of the market is the main challenge for the application of the hypothesis of the matched model method "breakdown". Evolution of technologies, consumption habits and market pressures imply in products characteristics being partially changed overtime or in some cases to the rise of completely new products and disappearance of older ones. Those new or modified products can provide different degrees of "utility" (quality) to the consumers respective the older ones. Under such circumstances an naive attempt to compile the CPI by simply extracting the "raw" price variation between the old and new product will lead to bias in the CPI. The bias originates due to the fact that this variation is not a pure price one as products of different quality are being compared.

In the 1990s a great attention had been devoted to the problem of quality change rising from the introduction of new products in the compilation of CPIs. It was shown that this problem constitutes one of the most important sources of bias in the CPIs [Boskin et al., 1996; ILO, 2004]. Since then, many methodological improvements have been proposed in order to minimize this problem. The standard tool adopted by NSOs all over relies on hedonic modelling techniques.

The hedonic approach aims at expressing the price of a given product as a function of its attributes. One drawback for the implementation of the approach is that industries usually do not reveal how product characteristics influence its final price. Hence, one need to rely on multivariate modelling to find out the set of most important attributes that determine the price of a good. For the success of the modelling, however, it is necessary a CPIs database with rich description of product characteristics contained in the sample. The construction and maintenance of such frame of specifications are very resource intensive, provides significant increase in respondents burden, hence

poses an important barrier for the implementation of hedonics by NSOs.

Here we study the use of web scraping techniques to support the implementation of hedonics in the CPIs compiled by IBGE. We aim at circumventing the costly process of collecting the characteristics of the products by extracting such information from web stores by means of automated tools that can extract a large and rich amount of information on the products in a precise, fast, and cheap fashion.

4.2 Experiment description

The experiment cosists in the collection of prices and attributes of several household appliances such as TVs, refrigerators, computers, mobiles, etc. The prices were collected for both online and offline stores whereas the attributes are exclusive from online sources. For online collection we extract information on the prices and attributes of all household appliances of interest available in the sites of the most important Brazilian web stores. In order to extract the characteristics of the products contained in the "official" price samples which is mainly based on brick-and-mortar stores we asked the field collectors to provide only two extra information (in addition to the prices): the brand and reference of the products. With such information in hands, we identify the products in the data base obtained via the web stores and extract their characteristics.

In the following we present our preliminary results and analysis for the case of refrigerators. The results presented correspond to data obtained for a single month (February of 2019). The online data was extracted once for all refrigerators available at the collection time by means of the rvest package of the R software. The database contains information on product characteristics and prices. Transport fees are not considered. A maximum of fifteen shops per reference was considered. The offline data consists in the CPI refrigerator sample from January 15^{th} until February 15^{th} 2019.

4.3 Results for refrigerators

Before start fitting the hedonic models it is interesting to perform an exploratory analysis on the data. Table 1 summarizes the information on the price quotations, number of stores, and references found for the online and offline data. The results show that though the difference in a similar number of price quotations is found, there is a significant difference in the number of refrigerators references and the number of stores between online and offline cases. The difference in the number of models references can be explained by two factors: the first and most relevant relates to the way the official sample is built. The products eligible for composing the sample are those most commercialized by the stores in each state covered by the index, hence restricting the number of brands and references to the most popular ones. Since initially no restriction was set for the brands and references collected online, the number of varieties observed increases. The second factor relates to the presence of new models that may not be incorporated in the official sample yet and to those products commercialized exclusive online.

The difference in the number of stores relies on the fact that the offline sample is spread over different states of the country including both small local chains and local branches of big retailers. For the online sample the stores mainly correspond to big retail chains which also engaged in the e-commerce sector.

Additional information is provided by the analysis of the price distributions provided by the boxplots in Figure 4. In Figure 4a we compare the distributions associated to all the references found for the offline and online samples. Though the medians or the two sets are close, we note that the offline prices are more concentrated around the median value than the online cases. The online sample is also characterized by a large amount of extreme values leading to higher mean and

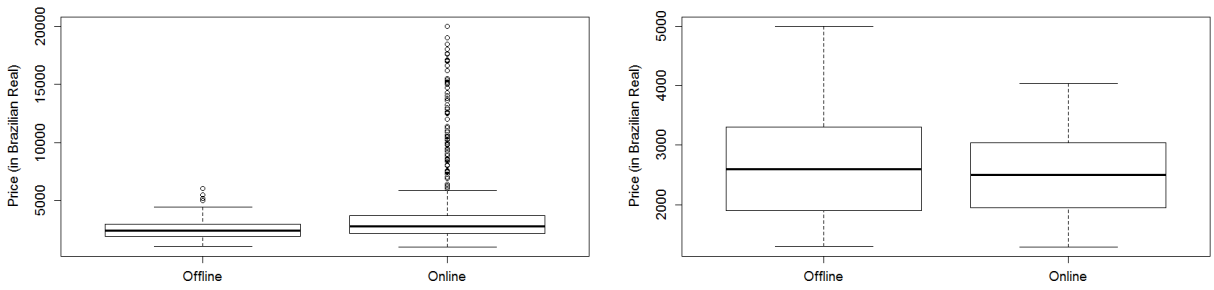
	Online	Offline
Prices	1663	1386
References	154	64
Stores	29	42

Table 1: Summary of the number of prices, models references, and stores for the refrigerators databases of the online and offline samples.

median prices for the online distribution in comparison to the offline data.

A significant change in the previous scenario is found if we move our attention to Figure 4b in which we restrict the comparisons to the references that are common to both samples, and to the stores that sell online and offline. Under these conditions we now observe that the price distribution of offline prices present a larger variability than the online ones. Another important point is that the prices for brick and mortar stores tend to be slightly more expensive than those seen online. This result is in agreement with previous studies that observed evidences that prices commercialized online in big retail chains in Brazil are cheaper than their brick and mortar stores [Cavallo, 2017].

A more refined analysis on the price differences can be performed by comparing the mean value for the online and offline prices for each product by retailer. We observe that in 87% of the cases online average prices are smaller than offline ones emphasizing the price discrimination by type of shop. Also, the online price level is approximately 11% smaller than the offline one.



(a) All references in the databases

(b) Common references in both databases

Figure 4: (a) comparison of boxplots derived for the price distributions of online and offline stores considering all references found in both databases. (b) compares the boxplot obtained for the price distributions resulting when we restrict the prices only to those corresponding to references common to both databases and to stores that commercialize both online and offline.

4.4 Modelling approach

Previous comparisons between online and offline prices for use in CPIs have been restricted to simple analysis of the differences in price levels and price evolution among those two commerce platforms [Cavallo, 2017]. We here explore if such difference is significant by means of hedonic modelling tools. We first test whether the level difference found is relevant. We then extend our study to check if the product price-determining characteristics are dependent on the nature of the stores. In other words, verify if hedonic model coefficients derived via online prices can be applied for quality adjustment treatments for offline samples.

In order to provide a "fair" comparison, we restrict the analysis to the price observations of references common to online and offline stores and to those stores selling both online and offline.

4.4.1 Hedonic model without interaction

To check if the price level differences observed are significant we start by fitting a hedonic model using online and offline prices to extract the main price determinant characteristics. We include a dummy variable which discriminates if the prices originate from the online or offline shops. If the shop dummy variable shows significant then the price level difference observed in our explanatory analysis is relevant.

In our modelling we adopted the log-lin formulation which is the one that best fitted our results. The choice of the functional form that best fit data is in agreement with international recommendations [Triplett, 2004]. The model selected was constructed by successive addition and removal of variables, based on their significance level. The final model writes:

$$log(\mathrm{Pr}) = \beta_0 + \beta_1 \mathrm{Br} + \beta_2 \mathrm{Col} + \beta_3 \mathrm{Sty} + \beta_4 \mathrm{Defr} + \beta_5 \mathrm{Cap} + \beta_6 \mathrm{Shop} \tag{1}$$

where in Eq. (1) Pr denotes the prices of the products. The price-determining variables found are:

Br = brand
Col = color finish
Sty = style
Defr = defrost type
Cap = total capacity
Shop = online or offline shop

The output of the model derived via ordinary least square (OLS) is summarized in Figure 5. As can be seen by the analysis of the p-values column, all coefficients of the model given by Eq. (1) are significant at α level of 0.01. The dummy variable (Shop) used to discriminate online and offline shops showed significant, supporting the hypothesis that the price level difference observed between online and offline shops is relevant.

However, before we bring a final conclusion it is important to check if the model fitted is consistent. Best practices states that hedonic regression coefficients must make sense [ILO, 2004]. With that said, all coefficients of the model fitted in Eq. (1) satisfy this property in the way that:

- Consul and Electrolux brands are cheaper than Brastemp one.
- stainless steel color refrigerator is more expensive than white ones.
- single door refrigerators are cheaper than top-freezer and bottom-freezer ones.
- frost free system is worth more than manual ones.
- the bigger is refrigerator total capacity the better consumer evaluates it.
- exploratory data analysis outlined that online prices are smaller than offline ones.

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      6.592e+00  2.905e-02 226.935  < 2e-16 ***
BrConsul        -1.619e-01  1.486e-02 -10.896  < 2e-16 ***
BrElectrolux    -4.476e-02  1.106e-02  -4.046 5.78e-05 ***
ColInox          1.003e-01  1.126e-02   8.909  < 2e-16 ***
StyDuplex        1.166e-01  1.717e-02   6.791 2.35e-11 ***
StyInverse       2.210e-01  2.212e-02   9.991  < 2e-16 ***
DefrFrost Free   1.615e-01  1.045e-02  15.445  < 2e-16 ***
Cap              2.684e-03  6.284e-05  42.707  < 2e-16 ***
ShopOnline      -1.094e-01  8.593e-03 -12.736  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1001 on 713 degrees of freedom
Multiple R-squared:  0.8845,    Adjusted R-squared:  0.8832
F-statistic: 682.5 on 8 and 713 DF,  p-value: < 2.2e-16
```

Figure 5: Output of the hedonic model that best fit the refrigerator's data for the online and offline prices. The model adjusted do not take into account interactions between the type of shop and product characteristics.

Further analysis of the model's residuals show that the homoscedastic assumption is being followed. We check this by the calculation of the Breusch-Pagan test. The p-value for this test was 0.019 so the null hypothesis for homoscedastic should not be rejected using α level at 0.01.

We finally evaluate the presence of multicollinearity in the variables of our model. When this assumption is violated the hedonic coefficients are sensitive to the inclusion or exclusion of other variables. A suitable measure to evaluate multicollinearity effects is the generalized variance inflation factor (GVIF) [Fox and Monette, 1992]. The GVIF results for the model given by Eq. (1) are displayed in Table 2. All the values found are close to unity indicating that the model is free of multicollinearity effects. This verdict is based on standard criteria that states that there are multicollinearity issues when GVIF exceeds 5 [James et al., 2013].

	GVIF	df	$\text{GVIF}^{1/2df}$
Br	1.41	2	1.09
Col	1.19	1	1.09
Sty	1.44	2	1.10
Defr	1.29	1	1.14
Cap	1.47	1	1.21
Shop	1.18	1	1.09

Table 2: Generalized Variance Inflation Factor - GVIF for the variables of the model fitted in Eq. (1)

Once the model showed consistent, we now have strong evidences indicating that the price level difference between online and offline shops is significant.

4.5 Hedonic model with interaction

The model given by Eq. (1) shows that online and offline price levels are different. However, this model does not elucidate if the type of shop also affects other price characteristics. For instance, if the price-determining characteristics of a product rely on the kind of shop.

In this Section we explore if such difference exists. The test is performed through the inclusion of interaction terms between the kind of shop and the variables found in the model given by Eq. (1). Then, we evaluate if any of the interaction terms is significantly different from zero. If that happens it means that the type of shop not only affect the price level but also change the hedonic characteristic coefficient.

The inclusion of interaction variables in the model follows the same procedure as the one adopted for the choice of the variables for the model Eq. (1), that is, interaction terms are included and excluded successively according to their level of significance.

None of the interactions were significant except the one between type of shop and brand. The resulting model with interaction writes:

$$log(\mathrm{Pr}) = \beta_0 + \beta_1 \mathrm{Br} + \beta_2 \mathrm{Col} + \beta_3 \mathrm{Sty} + \beta_4 \mathrm{Defr} + \beta_5 \mathrm{Cap} + \beta_6 \mathrm{Shop} + \beta_7 \mathrm{Shop} \cdot \mathrm{Br} \tag{2}$$

The output for the OLS regression of the model expressed by Eq.(2) is displayed in Figure 6, where the estimated coefficients are presented. One can note that for this new model the adjusted R^2 is equal to 0.888 which is slightly higher than 0.883 from the previous model.

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              6.608e+00  3.035e-02 217.751  < 2e-16 ***
BrConsul                -1.901e-01  1.962e-02  -9.692  < 2e-16 ***
BrElectrolux            -2.154e-02  1.448e-02  -1.488  0.13726
ColInox                  1.099e-01  1.112e-02   9.878  < 2e-16 ***
StyDuplex                8.785e-02  1.757e-02   5.000 7.24e-07 ***
StyInverse               1.956e-01  2.214e-02   8.836  < 2e-16 ***
DefrFrost Free           1.539e-01  1.036e-02  14.848  < 2e-16 ***
Cap                      2.692e-03  6.146e-05  43.804  < 2e-16 ***
ShopOnline              -7.943e-02  1.892e-02  -4.198 3.04e-05 ***
BrConsul:ShopOnline      5.489e-02  2.661e-02   2.063  0.03948 *
BrElectrolux:ShopOnline -6.629e-02  2.141e-02  -3.096  0.00204 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.09787 on 711 degrees of freedom
Multiple R-squared:  0.8899,    Adjusted R-squared:  0.8884
F-statistic: 574.9 on 10 and 711 DF,  p-value: < 2.2e-16
```

Figure 6: Output of the hedonic model that best fit the refrigerator's data for the online and offline prices. The model adjusted considers interactions between the type of shop and the product's characteristics.

Comparing both models we note that the explanatory power gain obtained with the interaction one is only marginal. Thus, the model without interaction is more parsimonious. Accordingly, the

interaction between kind of shop and other variables can be discarded and we choose the model without interaction, Eq.(1), as the best fit for our data.

This is an interesting result since it implies that a richer database of products and attributes can be exclusively extracted from web stores to construct more powerful models to be employed to treat quality changes in the offline sample.

5 Conclusions

In this paper, we discussed the pioneering ongoing projects on the use of new data sources to improve the official Brazilian CPIs produced by the National System of Consumer Price Indices (SNIPC) of the IBGE. The initial approaches focused on a parsimonious integration of the new data sources into the routines of the SNIPC.

We analyzed the use of web scraping techniques to replace the manual collection of airfares from airline websites. We briefly presented the reasons for creating an in-house web scraper using free software platforms. The initial goal of this approach was simply to evaluate if the robots were able to "reproduce" the manual collection results. The results presented showed that automatic collection succeed in this task. Such findings are promising since it opens doors for the expansion of the routes collected, enhancing the accuracy of the estimates derived for this sector in a fast, efficient and cheap way. In contrary, we also showed that the replacement of the manual by the automatic approach in the monthly production need to be treated with caution. Our experience reveals that the success of the automatic collection is sensitive to unexpected changes in the airline website architectures, anti-robot policies, and internet traffic instabilities. We argued that collaboration of the air companies is a key step to safely implement the automatic approach in the periodic routines of the CPIs.

The second part of the study investigated the use of web scraping techniques to extract detailed information on product characteristics, aiming to build and to maintain a robust database for the implementation of hedonics in CPIs. Our initial tests focused on household appliances, although the results presented here only restricts to refrigerators. Our findings revealed that the robots could provide a rich, cheap and precise source of product characteristics suitable for the implementation of hedonics in CPIs.

In addition, our explanatory analysis for the refrigerators showed that a larger variety of models could be extracted via the website stores. We also confirmed the previous findings, which showed that online prices are a bit lower than their corresponding brick and mortar similar. This finding was confirmed by the models we fitted. Based on this finding, we addressed the question if the kind of store would affect the way the product attributes influence the price of a product. We conducted experiments for a model with and without interaction between the kind of stores and the product attributes. Interaction only showed significant for one attribute. However, the explanatory power gain resulting from the model with interaction over the model without interaction was marginal. Hence, the model without interaction is the best choice to describe our data. Since the price-determining attributes do not rely on the kind of shop (a reasonable assumption), this result provides preliminary evidence that hedonic models could be entirely built by means of web data (price and attributes) and be employed on the "official" sample of the CPI. This approach is appealing since it allows the construction of hedonic models based on a wider range of products and attributes that could lead to more precise estimates for quality change treatments.

As future work, we plan to investigate whether this scenario remains the same for other household appliances present in our CPI basket. Alternative quality change treatments, like fixed effect models, will also be tested. Further, the fact that the database constructed solely by online information

can be applied for quality adjustment on offline CPI sample allows for cheaper studies on how the hedonic model coefficients change over time and the implementation of hedonic indices.

Acknowledgements

Vladimir G. Miranda and Lincoln. T. da Silva acknowledges Pedro L. Nascimento Silva for useful discussions on the modelling results. The authors acknowledges Marcio Rebello, Thiago Pereira, Andre Almeida, Emilton Aragao, and Fernando Gonçalves for useful discussions and support.

References

M. J. Boskin, E. Dulberger, Z. G. R. Gordon, and D. Jorgenson. Toward a More Accurate Measure of the Cost of Living: Interim Report to the Senate Finance Committee. United States General Accounting Office, 1996. URL https://books.google.com.br/books?id=H4W6mQEACAAJ.

R. Breton, G. Clews, L. Metcalfe, N. Milliken, C. Payne, J. Winton, and A. Woods. Research indices using web scraped data. Paper presented at the ILO/UNECE meeting of the Group of Experts on Consumer Price Indices, May 2016. URL https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.22/2016/Session_2_UK_Research_indices_using_web_scraped_data.pdf.

A. Cavallo. Are online and offline prices similar? evidence from large multi-channel retailers. American Economic Review, 107(1):283–303, January 2017. doi: 10.1257/aer.20160542. URL http://www.aeaweb.org/articles?id=10.1257/aer.20160542.

A. Cavallo and R. Rigobon. The billion prices project: Using online prices for measurement and research. Journal of Economic Perspectives, 30(2):151–78, 2016.

P. J. Daas, M. J. Puts, B. Buelens, and P. A. van den Hurk. Big data as a source for official statistics. Journal of Official Statistics, 31(2):249–262, 2015.

J. de Haan and W. E. Diewert. Quality change, hedonic regression and price index construction. Paper presented at the 15th meeting of the Ottawa Group, Eltville, Germany, May 2017.

J. Fox and G. Monette. Generalized collinearity diagnostics. Journal of the American Statistical Association, 87(417):178–183, 1992. doi: 10.1080/01621459.1992.10475190. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1992.10475190.

H. E. Gumundsdttir and L. G. Jónasdóttir. Scanner data: Initial data testing. Paper presented at the ILO/UNECE meeting of the Group of Experts on Consumer Price Indices, May 2016. URL https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.22/2016/Session_1_Iceland_Initial_data_testing.pdf.

J. Harrison. RSelenium: R Bindings for 'Selenium WebDriver', 2019. URL https://CRAN.R-project.org/package=RSelenium. R package version 1.7.5.

R. Hoekstra, O. ten Bosch, and F. Harteveld. Automated data collection from web sources for official statistics: First experiences. Statistical Journal of the IAOS, 28(3, 4):99–111, 2012.

K. N. Hov and R. Johannessen. Using scanner data for sports equipment. Paper presented at the ILO/UNECE meeting of the Group of Experts on Consumer Price Indices, May 2018. URL https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.22/2018/Norway_-_session_1.pdf.

IBGE. Sistema nacional de ndices de preos ao consumidor: mtodos de clculo. Srie relatrios metodolgicos, 14, 2013. URL https://biblioteca.ibge.gov.br/visualizacao/livros/liv65477.pdf.

ILO. Consumer Price Index Manual : Theory and Practice. International Labour Office, USA, 2004. ISBN 9789221136996. URL https://www.elibrary.imf.org/view/IMF069/01345-9789221136996/01345-9789221136996/01345-9789221136996.xml.

G. James, D. Witten, T. Hastie, and R. Tibshirani. An Introduction to Statistical Learning – with Applications in R, volume 103 of Springer Texts in Statistics. Springer, New York, 2013. ISBN 978-1-4614-7137-0. doi: 10.1007/DOI.

K. V. Loon and D. Roels. Integrating big data in the belgian cpi. Paper presented at the ILO/UNECE meeting of the Group of Experts on Consumer Price Indices, May 2018. URL https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.22/2018/Belgium.pdf.

V. Mendonça and R. Evangelista. Exploring new administrative data sources for the development of the consumer price index: The portuguese experience with actual rentals for housing. Paper presented at the ILO/UNECE meeting of the Group of Experts on Consumer Price Indices, May 2018. URL https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.22/2018/Portugal.pdf.

V. G. Miranda, P. K. da Costa, R. V. Ventura, and J. F. P. Gonçalves. Consumer prices indices at ibge: 40 years and counting. Paper submitted to the 16th meeting of the Ottawa Group, Rio de Janeiro, Brazil, 2019.

F. Polidoro, R. Giannini, R. L. Conte, S. Mosca, and F. Rossetti. Web scraping techniques to collect data on consumer electronics and airfares for italian hicp compilation. Statistical Journal of the IAOS, 31(2):165–176, 2015.

O. Ten Bosch, D. Windmeijer, A. Delden, and G. van den Heuvel. Web scraping meets survey design: combining forces. Conference on Big Data Meets Survey Science, 2018.

J. Triplett. Handbook on hedonic indexes and quality adjustments in price indexes. 2004. doi: https://doi.org/https://doi.org/10.1787/643587187107. URL https://www.oecd-ilibrary.org/content/paper/643587187107.

H. Wickham. rvest: Easily Harvest (Scrape) Web Pages, 2016. URL https://CRAN.R-project.org/package=rvest. R package version 0.3.2.