Instituto Nacional de Estadística y Geografía

Use of Big Data in modern markets coexisting with traditional markets data

Paper written for the 16th Meeting of the Ottawa Group on Price Indices, Rio du Janeiro, Brazil May 2019

Rafael Posse and Jorge A. Reyes
rafael.posse@inegi.org.mx
jorge.reyesm@inegi.org.mx

Abstract

In countries that do not have a highly developed market where the supply and prices of products and services may change according to the type of market, challenges are present on the products for extracting prices and describing them when using methods of modern collection. Handling these data (prices, quantities and descriptions) have great potential but also limitations and challenges.

The modern market can offer multiple products and services that are still in the basket of the country and probably cover between 80 and 90 percent of the market, such as equipment and technological, electrical and electronic equipment or services in the cloud (internet). However, the traditional market can cover 80% of the market for fresh and processed foods, as well as some services. For this reason, the use of collection methods such as Web Scraping or Scanner Data will have to focus on specializing by type of market and products. The price index of the country must coexist with two worlds where are obtained the data.

The handling of the samples, the integration of data, as well as the calculation of the price indexes represents a great challenge, as it Laspeyres will not necessarily have been calculated with, when using a new collection method that attracts the universe of products or services offered in the modern market. Using scanned data presents the possibility of managing the universe of products or services in the modern market, in this case we have prices and quantities, each transaction un the POP has both data, thus we have the possibility to know the main brands of products offered and demanded in the market. With Web Scraping the same case is presented, it is feasible to attract all the product in the category, is good get the main brands offered in the market, but we don't have quantities, just prices several times per day in the period.

The article discusses the restrictions, the potential of the collection tools and the method of calculation for this type of markets and the method of data extraction used, the coexistence of different samples generate challenges solved for price variations and their indices. It concludes with the use and potential of the tools, can be applied techniques and methodologies.

Index

1. Introduction

The information currently used for the generation of price indexes (IPC, IPP), has as its main source the direct visit method for capturing prices in a sample of Points of Sale (POS) selected by sampling techniques. This activity is carried out by visiting the POSs distributed in the traditional (MT) and modern (MM) markets.

Food, fresh and processed products, are collected mainly on MT, small - format stores to detail and agricultural produce markets, where it makes up 80% of the value of such transactions. Other types of products mainly electrical and electronic products are collected in MM or organized in some of them covers more than 80% of the market value, some other services they cover 100% in the modern market such as telephone and air transport services.

The processing of mass information requires using another type of technological infrastructure and with this it is necessary to make methodological changes in the measurement and calculation of price indices. The heart of this project is to analyze how can the information be incorporated from three sources: direct visit, web scraping and scanner data, taking advantage of the potential of each and one and concluding in an index that integrates the prices of different markets using the different techniques.

The Modern and Traditional Market are key points to analyze the techniques of price extraction, not only because of the difference between them, in particular because of the techniques we will use to extract the data.

2. Conceptual framework

The ideal for price indexes is always to have the total value of sales per specific product, prices and quantities per transaction and with that build the price index for a period. This gives us a precise datum of the real average price and immediately reflects the substitution phenomenon.

One of the characteristics of the quote in the MT fieldwork is the frequency and the opportunity with which the data is acquired, which is due to the size of the sample that converts the quotes into scattered points in spacetime. While the MM can usually be accessed in two ways: one, by direct visit to your POS or online, in the latter, using an electronic catalog of physical store products by area or region of the city or country. Another option is for stores to deliver sales data, Scanner Data.

The opportunity to capture data of the products offered in the market is focused on obtaining and discriminating the data that are important for the objective pursued, in the traditional methods the information can be obtained from a reduced number of products, while, with the methods of systematic extraction, all those offered in physical stores (samples or product universes) are obtained.

In any of the collection methods we can access prices, quantities sold and detailed description of the goods or services. Finally, in any case, we can carry out product substitution and adjustment for quality changes, which depend on changes in the attributes of the products or services. Table 1 shows the differences in the registry of goods or services quoted according to the collection technique used.

Table 1. Available specifications of a product according to the collection method

Collection type			
Concepts:	**Direct Visit (fieldwork)**	**Web Scraping**	**Scanner Data**
Detailed and structured product description	Partial	yes	yes
Price	of shelf	Published	In box
Type of offer or discount and value	Register	Register	Register
Quantity			yes
Coverage of stores and products	A sample of products offered	Total, of stores and products offered	Total, of stores and products sold
Existence of product at the time of quotation	Observed	Not observed	Unidentified
Store open to the public at the time of quotation	Observed	Observed	Observed

In Mexico, as in all countries, quotes are made for: direct visit, internet, telephone or administrative records, among others. Prices are collected in different points of sale (POS) by volume of income in two types of markets: modern and traditional. Some prices are investigated by different means: telephone or internet. Some of the products and services are quoted using the e- commerce, carrying out the transaction until before payment, in this case you will find mobile telephony, tourist services, air transport, hotels and other government or regulated services.

Types of market. The INEGI carried out a redesign of the sample in 2018 using probabilistic sampling in the two types of market indicated above, for direct visit collection, with this the sample selection was refined and we obtain the results for each group of products, the Distribution was obtained from the household expenditure survey. Before a change we already worked with both types of market and our results gave rise to the partition of the market type: traditional and modern one.

Modern Market (MM)

A company that has branches or POSs distributed in a local, regional or national geographic territory. Their purchases are central and consolidated to obtain better prices, they have aggressive purchasing policies, they maintain centralized sales policies, they handle products with sizes and presentations to meet individual and family needs, they make use of information technologies to serve their customers and they have analytical capacity to segment their markets according to the types of consumers they serve and offer reduced banking services, they maintain central marketing plans with frequent strategies of offers, discounts or promotions. It has large format stores for its market. The purchase of your consumers is planned. Most of their sales are with bank cards.

Traditional Market (MT)

It usually does not comply with the points described in the MM, however, in some cases they make use of information technologies and even offer banking services, a small proportion. The majority buys from wholesalers, the inventories result from purchasing capacity and displacement. It does not have a marketing vision or apply it. Use discounts or promotions from your provider. Serves a small group of consumers. It has no management capacity to achieve discounts or higher profit margins. Most of its sales are for convenience and little planned. They do not accept transactions with bank cards, very few cases have access to information technologies. Their shops are established in small premises.

3. Motivation

The international trend and the recommendations of the main organizations that regulate international best practices lead to obtaining data that can provide us with better information for the construction of price indexes, however , the reality of the markets does not always allow us to reach the ideal , For this reason we must resolve the issues that allow us to approach it.

Working with the organized commerce data can help us to approach the ideal; this would be achieve if we had all the trades in this market and, on the other hand, that they would provide us the data per transaction daily. This would be the use of the Data Scanner, however, the reality is different, few companies accept to provide the scanned data of each transaction, most accept the extraction using Web Scraping, and the MT does not exist possibilities to do it for its lack of technology.

Let us understand better data, for price indices, such as those that reflect the phenomenon of price evolution, considering all the products sold with their actual sales prices and quantities sold, and this we understand the ideal and most accurate data.

Our main challenge is the integration of data coming from two markets and three collection methods. Prices in the modern market have high volatility due to the application of marketing strategies employed by each company, in the traditional it is almost nil. Take advantage of the potential and benefits offered by each extraction technique for the Consumer Price Index, the data themselves and the calculation methods suitable for each one.

4. Problem Statement

Find the potential of each extraction technique by market type to generate a more robust and accurate price index than can be calculated with a probabilistic sample that is limited to specific times of quotation, day and time, in a particular market, modern or traditional.

The point is to understand that the potential of Scanner Data and Web Scraping can serve to improve the accuracy of the indices because they integrate the most demanded products and brands in the market at the time of quotation and that which has little demand remains. The modern market efficiently manages brands and products, which actually buy, in addition to showing us, in the case of the Scanner data, the movement of substitutions of products observing the evolution of the quantities of each specific.

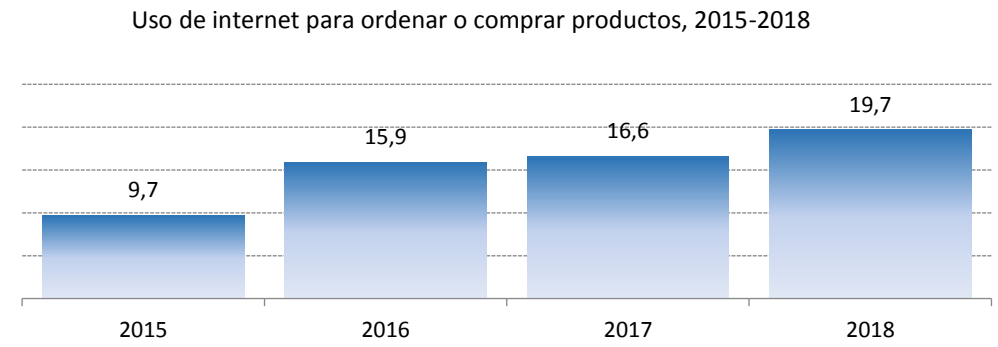
5. Electronic commerce and the use of information technologies (ICT)

5.1. Where we are?

Although in Mexico it does not have a total coverage of the internet service, some experts consider that it is covered between 60% and 70% of the population, in urban areas, considering the access to this service in computers, telephones and mobile devices with access to Internet. The INEGI presented, as a first approach to the measurement of the digital economy, the gross added value (GVA) of electronic commerce, with a participation in 2016 of 4.3% and with respect to its participation in GDP of 4.0%.

For the same year, the National Survey on Availability and Use of Information Technologies in Households (ENDUTIH) 2018, showed that 15.9% of Internet users have used it to order and buy products, increasing to 19.7% by 2018, see Figure 1. The age group between 25 and 34 years old is the one that uses the internet the most to buy products.

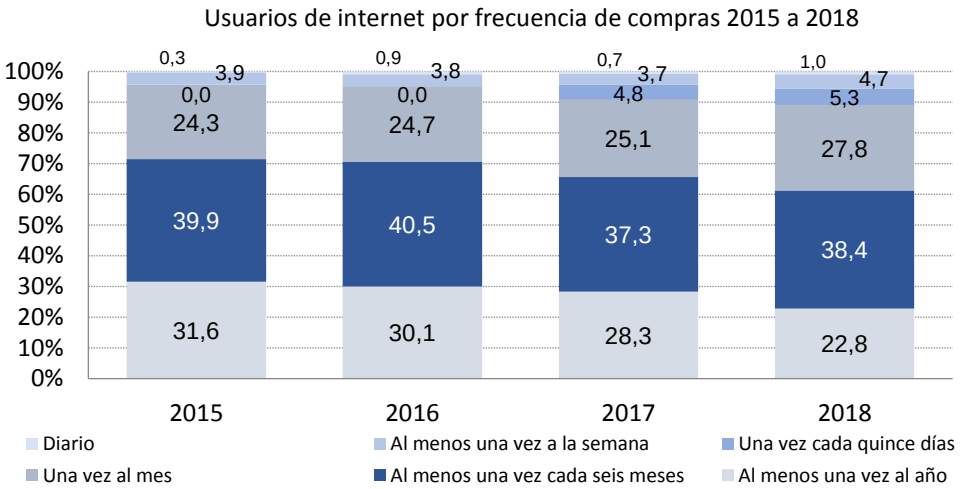
Figure 1. Use of the internet for electronic commerce



Fuente: INEGI Encuesta Encuesta Nacional sobre Disponibilidad y Uso de Tecnologías de la Información en los Hogares 2018

Total of users who buy via internet, on average, 60.3% of purchases are made to sites of national origin. The frequency with which users use the internet to make purchases, as shown in Figure 2, are purchases at least once every six months between 2015 and 2018.

Figure 2. Frequency of purchase using the Internet

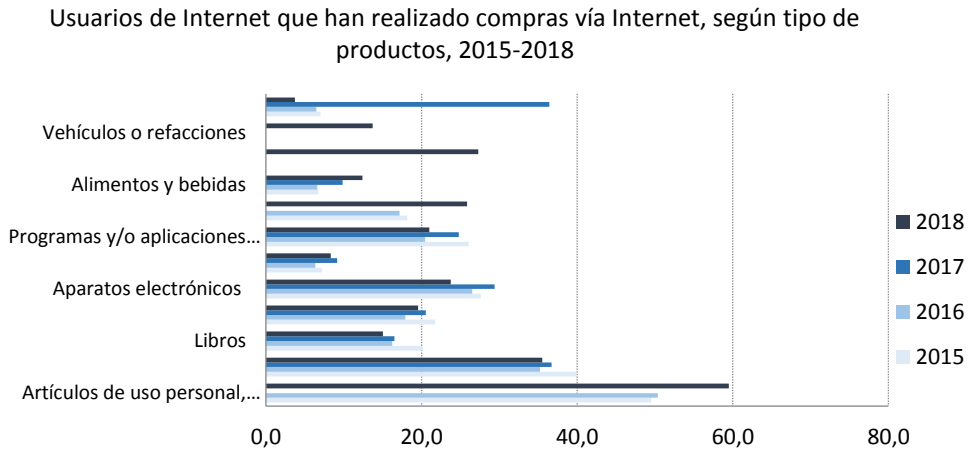


Fuente: INEGI. Encuesta Encuesta Nacional sobre Disponibilidad y Uso de Tecnologías de la Información en los Hogares 2018

The products that Mexicans buy online are: articles for personal use, electronic devices, computers, books, music, food and beverages.

Authors: Rafael Posse and Jorge A. Reyes

Figure 3. What products do Mexicans buy online from 2015 to 2018?



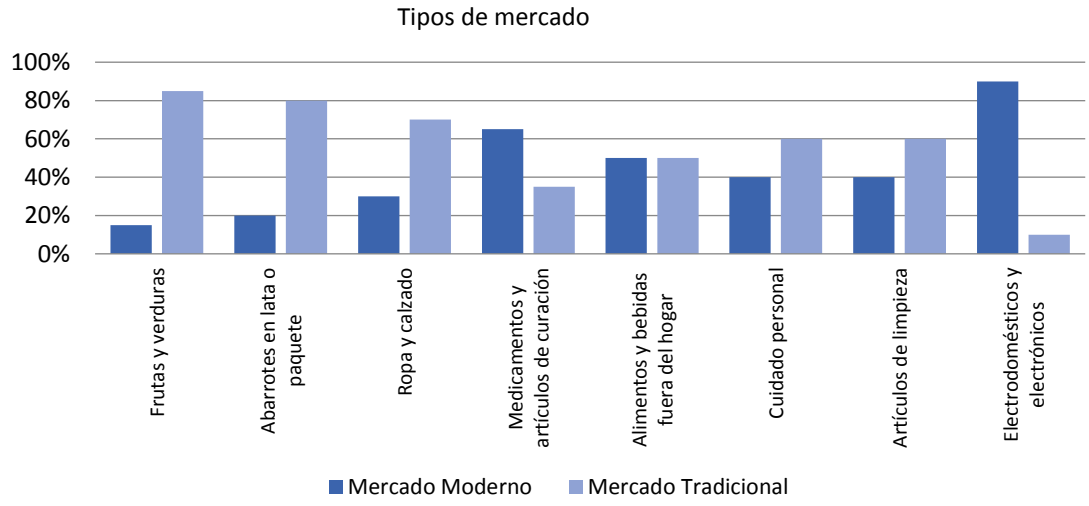
Fuente: INEGI. Encuesta Encuesta Nacional sobre Disponibilidad y Uso de Tecnologías de la Información en los Hogares 2018

Even when there is an electronic market and goods and services are acquired through this medium, it is still incipient with an average annual growth of 3%, according to studies by different experts.

6. Price quotations of the CPI in Mexico

As mentioned, INEGI uses a quotation method depending on the type of market and the product. For each type of trade, its market share will depend on the type of product, see Figure 4:

Figure 4. Groups of products by market type



To understand a little more what each of these markets represents, it should be considered that more than 56% of the population has limited purchasing capacity. In Mexico using data from the 2014 Economic Census, it shows that there are 1,832,275 retail businesses, of which 943,802 companies (52%) belong to the group of groceries: food, beverages, ice and tobacco figure 1. This is the group with the highest POP number due to purchasing capacity.

Figure 5. Structure of the Wholesale Trade that caters to retail stores

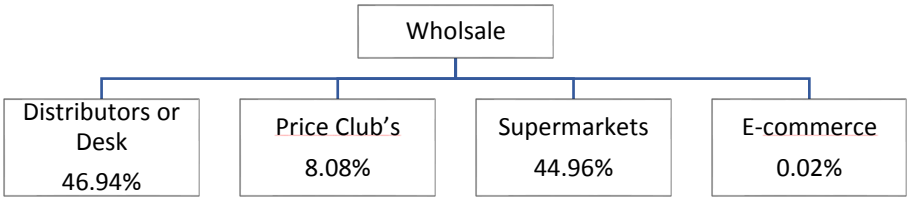
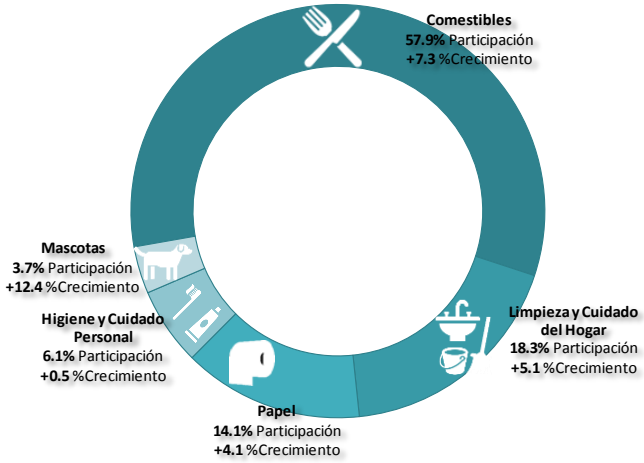
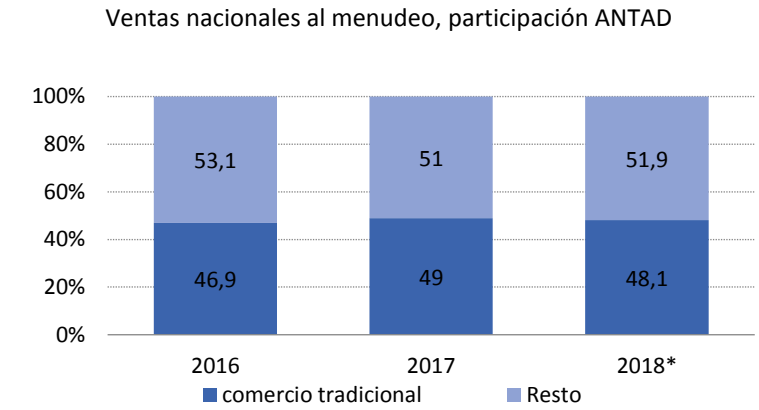


Figure 6. Trade and productive sectors to which it benefits



In a study conducted by the National Association of Self-Service Stores and Department Stores (ANTAD) in December 2018, showed that the MT went from representing 49% in 2017 to 48.1% in 2018 while the rest is covered by department stores and self-service, this only considering their participation with respect to GDP, as shown in Figure 7.

Figure 7. Modern Market Share (all product groups are included)



*** Cifras preliminares**

Fuente: Estadísticas ANTAD

7. Calculation of average prices using different data collection techniques

We will describe what happens with the three forms of price collection, analyzing that the volume of quotations can have an impact on the accuracy of the real average market price. It is observed in this case the impact of price volatility in the MM.

7.1. Price collection on fieldwork

In fieldwork operations, a small sample of a specific product (i) is usually collected in each collection period for each city in the sample. It then assumes that this observation becomes the average price of each period. A price on each visit and this can be weekly, biweekly or monthly.

$p_{i\sigma}$ = precio del ítem (i) en el número de lectura σ

1) Supposed: $p_{i1} = p_{i\sigma} \ \forall \ \sigma = 2, 3, \dots, n$

$\bar{p}_i = p_{i1}$

Table 2 shows the case of fieldwork visits with a single register in sub-period 2.

Table 2. Price recording by observation

Item	Price pi	Observ pis	Price pi	Suponed price pi1 = pis
Item (i)	pi			36
Item (i)	pi	Pi1	36	36
Item (i)	pi			36
Item (i)	pi			36
Item (i)	pi			36
Average				**36**

7.2. Collecting prices using Web Scraping

By means of this technique it is possible to considerably increase the price registration number. In this case, the reading is defined as the price obtained now of extracting information from a website from a specific source. With the greater availability of prices and their trajectory during the collection period, a more accurate average price can be calculated, without weighting, due to the lack of information regarding the quantities sold.

$p_{i\sigma}$ = precio del ítem (i) en la lectura σ

2) $\bar{p}_i = \frac{1}{n}\sum_{i=1}^{n} p_{i\sigma}$

Table 3 shows the case of web- scraping with five observations during the collection period.

Table 3. Effect of average prices of making more than one observation in the period

Item	Price pi	Observ pis	Price pi	
Item (i)	pi	pi1	36	
Item (i)	pi	pi2	36	
Item (i)	pi	pi3	18	Product Offer
Item (i)	pi	pi4	36	
Item (i)	pi	pi5	36	
Average			**32.4**	

In the previous example, the inclusion of the offer price of the third take down the average price with respect to the traditional collection of 36 pesos, to 32.4 pesos. This behavior will obey the dynamics of the market, offers, discounts or temporary increases.

7.3. Collection of prices by scanner data

With the information of the electronic points of sale, the statistical agencies can obtain prices and amounts relative to all the transactions of the reference period of a set of items. By means of this type of collection the unit values can be calculated, which are the average prices appropriate for the calculation of the elementary indices. Each unit value represents, in a single figure, the price changes of a specific item during the collection period.

The total value (V^T) referred to by the price index of an item (i) defines the set of transactions carried out by a well-defined group of economic agents with that item, at certain prices (p), and with some Weights (q) that describe your consumption pattern, over a period of time.

3) $V_i^T = \sum_{i=1}^{n} p_i q_i \quad p_i > 0,\ q_i \geq 0$

The requirement of consistency implies that the prices used for the calculation of the elementary indices are the same as those implicit in the total value of the transactions that are carried out during the reference period of time , and therefore the appropriate average prices for the calculation of the elementary indices.

We define the Total Unit Value as:

$$V_i^T = \sum_{\sigma=1}^{n} V_{i\sigma} = p_{i1}q_{i1} + p_{i1}q_{i2} + \cdots + p_{i1}q_{in}$$

4) Donde: $V_{i\sigma}$ = valor parcial del ítem (i) en la toma σ

Since the items are identical, they are also commensurable. Therefore, the total amount (Q^T) of the transactions of an item in the reference period is defined as:

$$Q_i^T = \sum_{\sigma=1}^{n} q_{i\sigma} = q_{i1} + q_{i2} + \cdots + q_{in}$$

a. Where: $q_{i\sigma}$ = item (i) volume in the quotation σ

Then the Unit Value (V^U) item 'i' is defined as:

5) $V_i^U = \frac{V_i^T}{Q_i^T}$

From which it follows that the unit value of an item is an average auto price weighted by the different quantities corresponding to each subperiod.

Substituting (2) and (3) in (4)

6) $V_i^U = \frac{V_i^T}{Q_i^T} = \frac{\sum_{\sigma=1}^{n} V_{i\sigma}}{\sum_{\sigma=1}^{n} q_{i\sigma}} = \frac{\sum_{\sigma=1}^{n} p_{i\sigma} q_{i\sigma}}{\sum_{\sigma=1}^{n} q_{i\sigma}} = \sum_{\sigma=1}^{n} p_{i\sigma} \left(\frac{q_{i\sigma}}{\sum_{\sigma=1}^{n} q_{i\sigma}}\right) = \sum_{\sigma=1}^{n} p_{i\sigma}\, s_{i\sigma}$

7) Where: $s_{i\sigma}$ = weighting of the quantity of the item (i) in the observation σ

In this way, the unit values consider the influence of the price trajectory and the quantities recorded at different times, for the formation of the average price. For this reason, it is the best estimator of the average price of item (i) for the reference period.

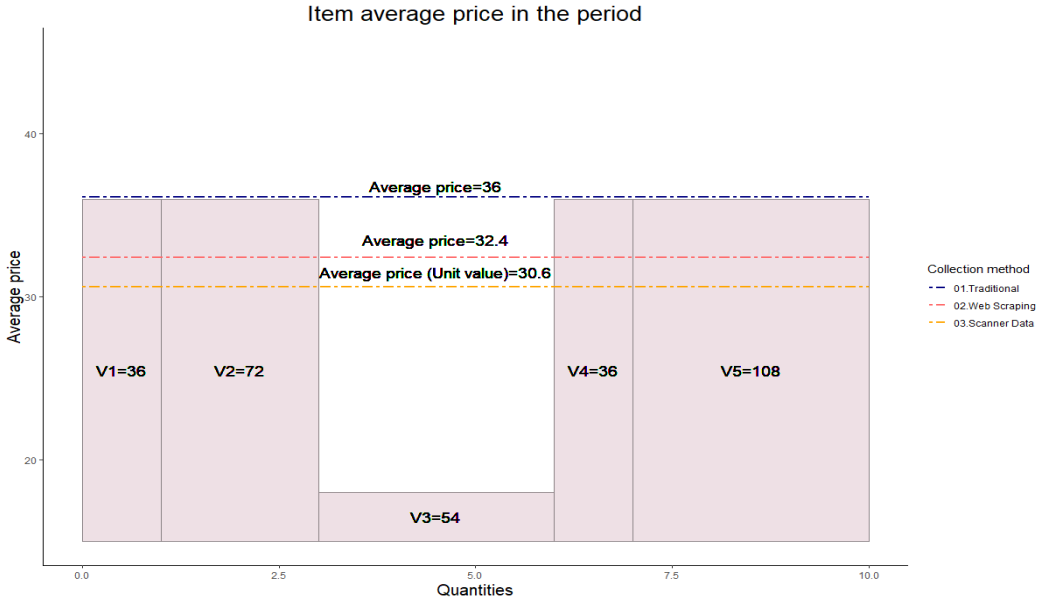
Following the example, table 4 shows how the price and quantity trajectory determines the unit value of the collection period.

Table 4. Prices using Scanner Data

Item	**Price pi**	**Observ pi**	**Quentity qi**	**Value Vi**	**Unit Value pi**	**Weighting qi/QT**	**Product Pi(qi/QT)**	
Ítem (i)	pi	pi1	1	36	36	0.1	3.6	
Ítem (i)	pi	pi2	2	72	36	0.2	7.2	
Ítem (i)	pi	pi3	3	54	18	0.3	5.4	← Price Offer
Ítem (i)	pi	pi4	1	36	36	0.1	3.6	
Ítem (i)	pi	pi5	3	108	36	0.3	10.8	
Suma			10	306				
Unit Value					30.6		30.6	

Figure 8 shows how the different quantities -expressed in the width of each column- and the partial unit values -expressed by the height of each column- determine the total unit value of the collection period.

Figure 8. Average price per collection period



Thus, unit values, considering the influence of the trajectory of prices and amounts recorded at different times, for the formation of the average price. For this reason, it is the best estimator of the average price of item (i) for the reference period.

Following the example, table 4 shows how the price and quantity trajectory determines the unit value of the collection period.

8. Extraction of data on practice

8.1. Experience when venturing into web scraping

The use of web scraping techniques we use it in the price area of INEGI to identify the offer and prices of products and services in the market. This requires a work in four stages: research and analysis of the electronic pages; analysis of data extraction; Identification of the target products or services and an identifier that indicates that it is always the same product; software tests; contrast of the data against those obtained in direct visits to the same stores and products; robot production; maintenance of the robot on the selected pages.

8.1.1. Programs and Languages for data extraction

The extraction of data began manually and using free software, which was abandoned quickly due to restrictions: number of projects, volume of data, and the number of portals. Currently the PYTHON programming language is used

8.1.2. Extraction techniques

Each virtual site on the Internet has a goal to fulfill, virtual stores in general facilitate the shopping experience, according to the product or service. At the back door of the page are a set of programs, which in addition to allowing access to products, provide security. These elements add complexity to the reading of data that can be done as: File download; Elementary extraction; Extraction by ROBOTS; Extracting the data source.

8.1.3. Human talent

It is considered necessary to have a multidisciplinary team: mathematicians, economists and computer scientists, accompanied by statisticians. It is also desirable to have a person with geo - statistical knowledge.

8.1.4. Main problems in data extraction.

There are different factors that make the extraction of data in virtual stores complex: site security, type of information display, number of products cataloged on the page, versions of the HTML code, format styles and diversity of managed products.

8.1.5. Use of Artificial Intelligence techniques

To uniquely identify a product, marketers use their own codes, so it is not always feasible to obtain a standard product identification code: SKU, GTIN, barcode; etc. It is important to always look for a unique or standard code such as the barcode.

8.1.6. Ethical considerations of the use of techniques.

In the LSNIEG[5] of the INEGI in its article 45 establishes that the Informants of the System will be obliged to provide, with veracity and opportunity, the data and reports requested by the competent authorities for statistical, census and geographic purposes, and will provide support to them.

8.2. Scanner Data[6]

To obtain the data, it is necessary to carry out a negotiation process with the companies, since the data can be by registration or aggregates. The aggregation, the average price of a period of time (prices and quantities), allows access to said information to be negotiated without breaking the confidentiality of the basic data. We can only access these data provided that they are provided directly by collaboration agreements, they are not public.

The products, prices and quantities (explicit or implicit) that you give us using Scanner Data would be ideal for the work of any person in charge of calculating the CPI, since you would have in each case both prices and quantities and the substitution effect by the consumer or the producer . Our ideal index would be in all cases this, prices and quantities per transaction or average price for a given period in addition to the quantities sold.

In any case this data of the SD would give us in addition to more information a more precise data, either using for all the transactions carried out in the country in a determined period or for a specific market and the main products that are offered in the market, of any How to use this data would give us greater precision and makes the CPI general index of the country more robust.

9. Gasoline case study.

9.1. Data

This case uses public information and direct survey in the fieldwork, we will contrast the data in three scenarios to find the technique that can give us the best accuracy with different databases.

The exercise shown was done considering this database with three different collection techniques:

a) The prices collected by INEGI fieldwork staff;

b) Collection using Web Scraping, for the same sample of gas stations that quote fieldwork;

c) Quotations with WS for all gas stations in the country.

9.1.1. Scraping process

The Web Scraping extraction is automatically programmed twice a day. Data since May 2017.

9.2. Gasoline test calculation

To compare the different quotes mentioned previously, the universe of 11,600 service stations in the country and two types of gasoline is used. The sample of petrol stations with direct collection is: 562 POV,

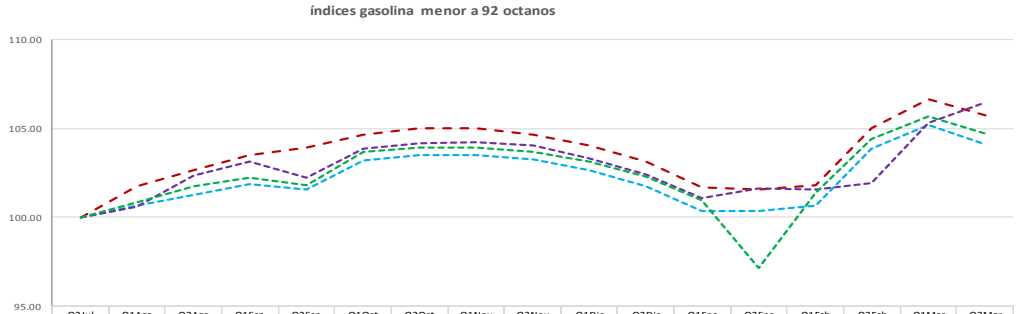
The calculation method for all these exercises was used the Laspeyres Index and using unit prices.

9.2.1. Results

In each case the index was calculated and together they are compared against the published data, in graphs 9 and 10 the corresponding indexes are shown. To compare the different quotes mentioned previously, the universe of 11,600 service stations in the country and two types of gasoline is used. The sample of petrol stations with direct collection is: 562 POV,

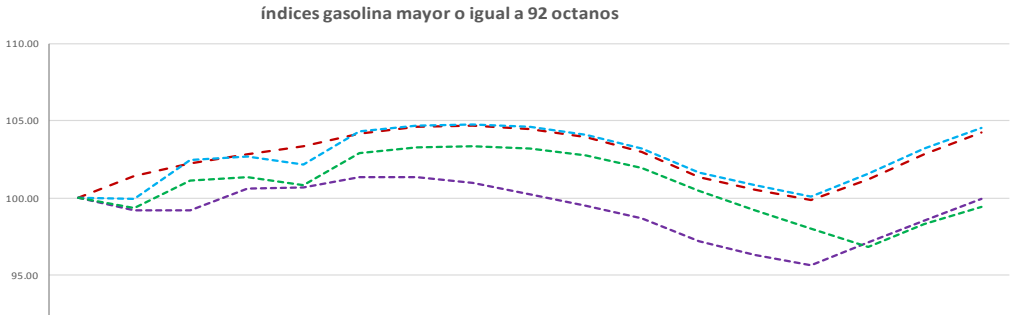
The calculation method for all these exercises was used the Laspeyres index and using unit prices.

Figure 9. Gasoline indexes lower than 92 octanes of the 2Qjul2018-2Qmar2019



	Q2Jul	Q1Ago	Q2Ago	Q1Sep	Q2Sep	Q1Oct	Q2Oct	Q1Nov	Q2Nov	Q1Dic	Q2Dic	Q1Ene	Q2Ene	Q1Feb	Q2Feb	Q1Mar	Q2Mar
						2018								2019			
Publicado	100.00	101.72	102.67	103.48	103.91	104.67	105.01	105.03	104.67	104.03	103.14	101.66	101.58	101.80	105.00	106.66	105.76
Promedio de la qna	100.00	100.63	101.27	101.84	101.54	103.20	103.46	103.49	103.24	102.62	101.73	100.37	100.32	100.65	103.85	105.20	104.10
Día de cotización	100.00	100.61	102.36	103.14	102.20	103.87	104.18	104.19	104.04	103.28	102.39	101.09	101.59	101.55	101.95	105.33	106.45
Universo de gasolineras	100.00	100.85	101.74	102.22	101.77	103.66	103.94	103.93	103.70	103.14	102.31	100.97	97.13	101.35	104.38	105.67	104.70

Figure 10. Indices greater gasolines or 92 equals octane 2Qjul2018-2Qmar2019

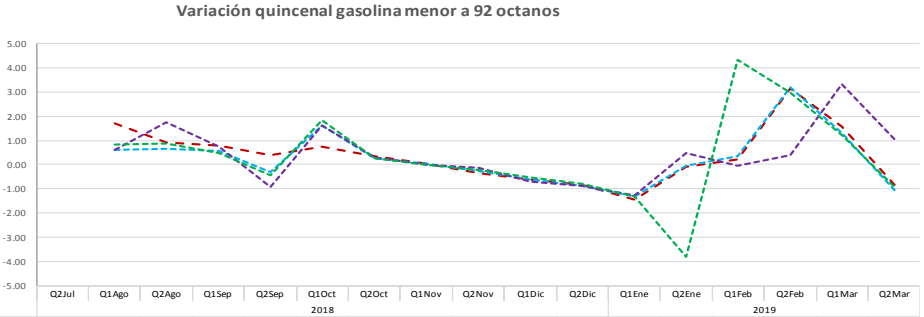


	Q2Jul	Q1Ago	Q2Ago	Q1Sep	Q2Sep	Q1Oct	Q2Oct	Q1Nov	Q2Nov	Q1Dic	Q2Dic	Q1Ene	Q2Ene	Q1Feb	Q2Feb	Q1Mar	Q2Mar
						2018								2019			
Publicado	100.00	101.39	102.22	102.83	103.37	104.15	104.58	104.66	104.45	103.93	102.96	101.37	100.56	99.84	101.23	102.85	104.23
Promedio de la qna	100.00	99.91	102.43	102.71	102.13	104.33	104.69	104.76	104.63	104.11	103.22	101.64	100.83	100.10	101.56	103.17	104.56
Día de cotización	100.00	99.15	99.16	100.63	100.67	101.36	101.34	100.96	100.21	99.51	98.66	97.18	96.32	95.62	97.13	98.53	99.91
Universo de gasolineras	100.00	99.36	101.10	101.35	100.80	102.90	103.25	103.31	103.18	102.73	101.93	100.43	99.22	98.04	96.85	98.27	99.40

The published index, dotted line, in all graphs, corresponds to the quotes made directly. The average of the fortnight, blue, day of quotation, purple and universe of gas stations, green, correspond to the results obtained with web scraping.

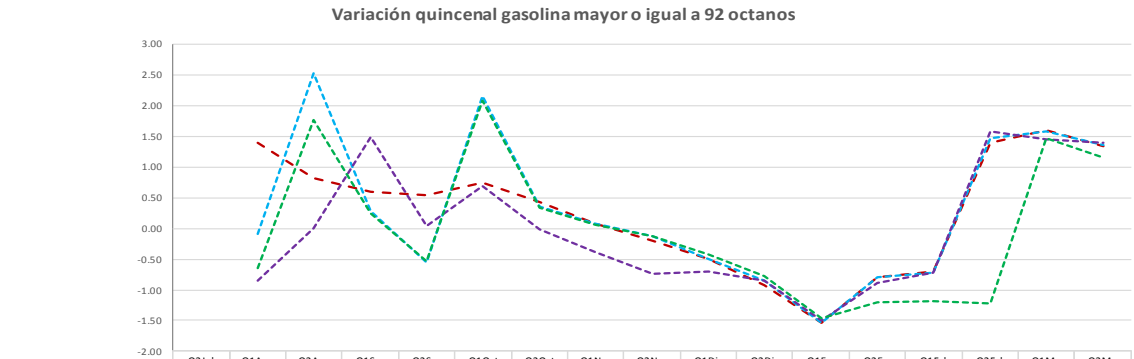
The variations of our test calculations and the published gasoline index are shown in graphs 11 and 12.

Figure 11. Gasoline variation less than 92 octanes of the 2Qjul2018-2Qmar2019



	Q2Jul	Q1Ago	Q2Ago	Q1Sep	Q2Sep	Q1Oct	Q2Oct	Q1Nov	Q2Nov	Q1Dic	Q2Dic	Q1Ene	Q2Ene	Q1Feb	Q2Feb	Q1Mar	Q2Mar
						2018								2019			
Publicado		1.72	0.94	0.79	0.41	0.73	0.33	0.02	-0.35	-0.61	-0.86	-1.44	-0.08	0.21	3.15	1.58	-0.85
Promedio de la qna		0.63	0.63	0.57	-0.30	1.64	0.25	0.02	-0.24	-0.60	-0.87	-1.33	-0.05	0.33	3.18	1.30	-1.05
Día de cotización		0.61	1.74	0.75	-0.91	1.63	0.30	0.01	-0.15	-0.72	-0.87	-1.26	0.50	-0.04	0.38	3.32	1.06
Universo de gasolineras		0.85	0.89	0.47	-0.44	1.85	0.27	0.00	-0.22	-0.54	-0.81	-1.31	-3.80	4.34	2.99	1.23	-0.92

Figure 12. Gasoline variation greater than or equal to 92 octanes of the 2Qjul2018-2Qmar2019



	Q2 Jul	Q1 Ago	Q2 Ago	Q1 Sep	Q2 Sep	Q1 Oct	Q2 Oct	Q1 Nov	Q2 Nov	Q1 Dic	Q2 Dic	Q1 Ene	Q2 Ene	Q1 Feb	Q2 Feb	Q1 Mar	Q2 Mar
						2018								2019			
Publicado		1.39	0.81	0.59	0.53	0.75	0.42	0.08	-0.20	-0.50	-0.93	-1.54	-0.80	-0.71	1.38	1.60	1.34
Promedio de la qna		-0.09	2.52	0.27	-0.56	2.15	0.35	0.07	-0.13	-0.50	-0.85	-1.54	-0.80	-0.72	1.46	1.58	1.35
Día de cotización		-0.85	0.00	1.49	0.04	0.69	-0.02	-0.38	-0.74	-0.70	-0.86	-1.50	-0.89	-0.73	1.58	1.44	1.39
Universo de gasolineras		-0.64	1.76	0.25	-0.55	2.08	0.34	0.06	-0.13	-0.43	-0.78	-1.47	-1.21	-1.19	-1.21	1.47	1.16

According to the previous graphs, the approximations that the three exercises have with respect to the published one, is minimal, not being determinant, we can conclude that:

It is worth mentioning that this is a preliminary exercise, which shows results that at first sight may seem trivial, but through which we are acquiring a better experience in the collection of prices, search for statistical methods, which allow us to guarantee the quality of the data, with the aim of reaching a level that provides reliability to decision-making, creating a solid, comprehensive and complete knowledge. This experience is used in the medium term to continue integrating other products to this methodology.

Conclusions

Some important benefits of using web scraping and scanner data:

a) Increase in the precision of the elementary indexes. This results from the great availability of huge volumes of data, which allow better and more robust estimates of prices, average or relative, for the reference period, as such volumes approaches to the universe. With the availability of the quantities (in the scanner data case), unit values can be calculated periodically, which in theory are the better inputs for a price index of homogeneous goods or services.

b) Provide important information and hints for market research. Since it is possible to track the price path through time (and quantities in the scanner data), the different phases of the product cycle identified and represent different brands and models. This information is very helpful to determine which and how many items to include in an elementary aggregate.

c) We find that organizing fieldwork work better, making the total of proportional quotes per day and market type, generates greater precision than doing it per day at random.

d) The use of the WS allows an improvement in the organization and precision of the quotes, always including the same products in all its brands and presentations, considering the best-selling products, if they have a page with the prices of products and updated every day.

e) A challenge is the Quality Adjustment, we must develop more the Artificial Intelligence technique in Web Scraping and Scanner Data to compare thousands of products and make them comparable, applying some of the existing techniques for this, always in an automated way. We do various tests.

f) We found that the products that make up groups such as appliances and appliances, computers and personal care, have the same behavior in the trend and price variations in MM and MT. We analyze basically follow the modern market and complement with a small sample of the MT.

g) In Web Scraping we send letters requesting permits and notifying our access to the sites of companies that offer their products on their websites daily, in order to be within the best international practices and ethics practices in data extraction. In some cases, they give us files with daily price data.

h) We use a new measurement or calculation technique in a product or groups of products or market if we have experienced enough in the data laboratory to understand its behavior, quality, accuracy and integration in the general index.

Concluding, average prices, or price relatives calculated from a mix of price quotations coming from the three different price collection methods are indeed better estimations than all price quotations obtained by traditional fieldwork. How better? It depends on the proportions of price quotations coming from the MM. These proportions are, in fact, implicit weights. So the problem is to choose the correct mix of items to price in an elementary aggregate, according with the market characteristics (MM and MT proportions) and the consumption patterns (the places where consumers buy).

Bibliography

- Frances Krsinich , Tatauranga Aotearoa, Statistics New Zeland , "Price Indexes from Online Data Using the Fixed-effects Window-splice (FLEWS) Index", Paper presented at the Ottawa Group, Tokyo, Japan, 20-22 May 2015.
- INEGI, Modern and Traditional Market According to the ENIGH 2016.
- International Monetary Fund, "Update of the Consumer Price Index Manual", chapter 10: Scanner Data
- Ingolf Boettcher, Automatic Data Collection on the Internet (web scraping), May 18, 2015.
- Law of the National System of Statistical and Geographic Information.
- Ragnhild Nygaard, Statistisk Sentralbyr , Statistics Norway, The use online prices in the Norwegian Consumer Price Index, paper written for the 14 th meeting of the Ottawa Group, Tokyo, Japan, 20-22 May 2015.
- ANTAD: https://antad.net/indicadores/indicantad/
- Study on online sales Mexico 2019: https://www.amvo.org.mx/estudio-venta-online
- Study of Electronic Commerce in Mexico 2018:
- https://www.asociaciondeinternet.mx/en/component/remository/func-startdown/95/lang,es-es/?Itemid=
- National Survey on Availability and Use of Information Technologies in Homes 2018:
- https://www.inegi.org.mx/programas/dutih/2017/default.html#Tabulados
- Glossary Of Statistical Terms: https://stats.oecd.org/glossary/
- INEGI, National Index of Consumer Prices, methodological document, Base second fortnight of July 2018.
- http://www.beta.inegi.org.mx/app/biblioteca/ficha.html?upc=702825104177
- INEGI, Economic Censuses 2014: https://www.inegi.org.mx/programas/ce/2014/
- INEGI, Electronic commerce.
- http://www.beta.inegi.org.mx/contenidos/programas/vabcoel/2018/metodologias/Info_coel.pdf
- ISCAM, Profile of the Wholesale Sector Abarrotero, 2018: https://www.anam.mx