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

Rafael Posse and Jorge A. Reyes

Mayo 2019

INEGI

Introduction

The traditional way to obtain market prices is to collect them by direct visit to the information sources. We use sampling techniques and we do it in any selected PDV, there we identify the target product.

Obtaining large volumes of data requires using other collection techniques and this leads to other techniques for managing this data. We went from a small sample to data lakes. This has an important impact on methodological changes in the measurement, processing and calculation of price indices.

The heart of this paper is to analyze how can this lake of information be incorporated from three sources: direct visit, web scraping and scanner data, taking advantage of the potential of each one concluding in a more accurate index that integrate prices of those markets using different techniques.

Conceptual framework

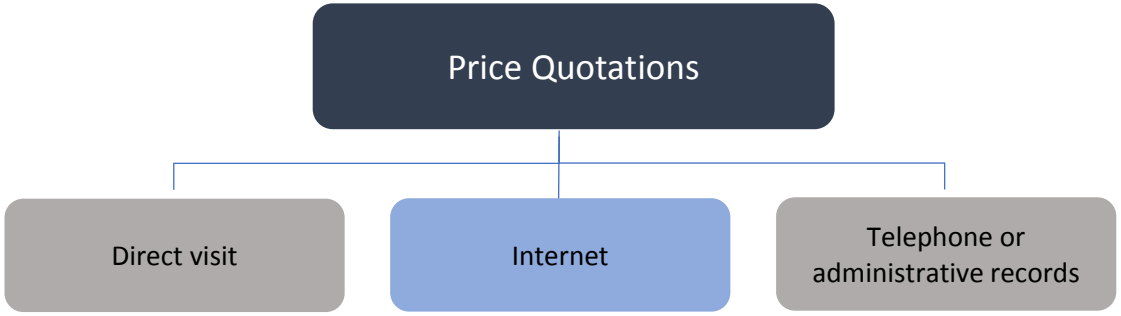
The ideal for price indexes is always to have the total value of sales per specific product, prices and quantities per transaction

In the traditional methods you get the information from a sample, while, with the methods of new collection techniques all those products offered in physical stores can be obtained, you have the option: samples or universe.

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

Table 1. Available specifications of a product according to the collection method			
	Collection type		
Concepts	**Direct Visit (field)**	**Web Scraping**	**Scanner Data**
Detailed and structured product description	Partial	yes	yes
Price	On shelf	Published	Scanned in cashier
Type of offer or discount and value	On shelf	Published	Applied in cashier
Quantity			yes
Coverage of stores and products	A sample of POS	Access to the Universe	Access to the Universe
Existence of product at the time of quotation	Observed	Not observed	Not buy (register)
The POS is open at the moment of collect prices	Phisical Observed	Virtual access	yes

Conceptual framework



To day prices are collected in different points of sale (POS)

The INEGI carried out a redesign of the sample in 2018 using probabilistic sampling in the two types of market indicated above. The distribution was obtained from the last Household Expenditure Survey.

Modern Market (MM)

A company that has branches or POS's distributed in a local, regional or national geographic territory. Their purchases are central and consolidated to obtain better prices, they have aggressive prices policies, they maintain centralized sales policies and use aggressive Marketing.

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.

Traditional Market (MT)

The majority of POS 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, give services to a small group of consumers.

Motivation

The international trend and recommendations from the main organizations that regulate the best practices that lead to get data that provide us better information for the price indexes.

01

We can get better data for price indices, understand it as those that reflect better the phenomenon of price evolution, considering all the sold products, their sales prices and quantities sold.

02

Our main challenge is the usability of data 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. How we can explote the potenciality for the price founts?

Problem Statement

Why to find the potential of each extraction technique by market type to generate a more robust and accurate price index than those that can be get prices using a probabilistic sample?

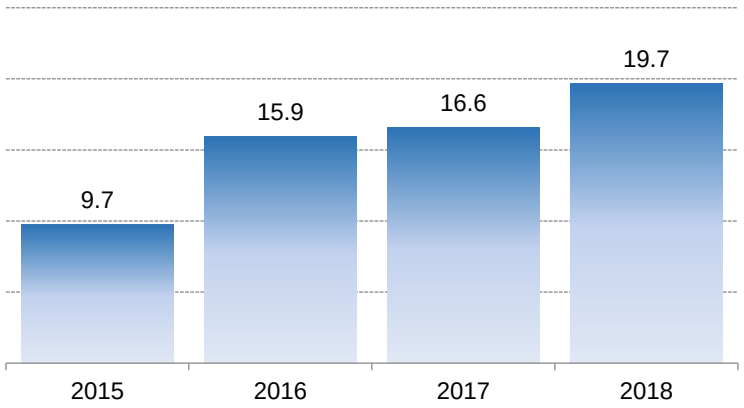
INEGI

ELECTRONIC COMMERCE AND THE USE OF INFORMATION TECHNOLOGIES (ICT)

Where we are?

The INEGI have a first approach to the measurement of the digital economy, the gross added value (GAV) for the electronic commerce, with a participation in 2016 of 4.3% and with respect to its participation in GDP of 4.0%[1].

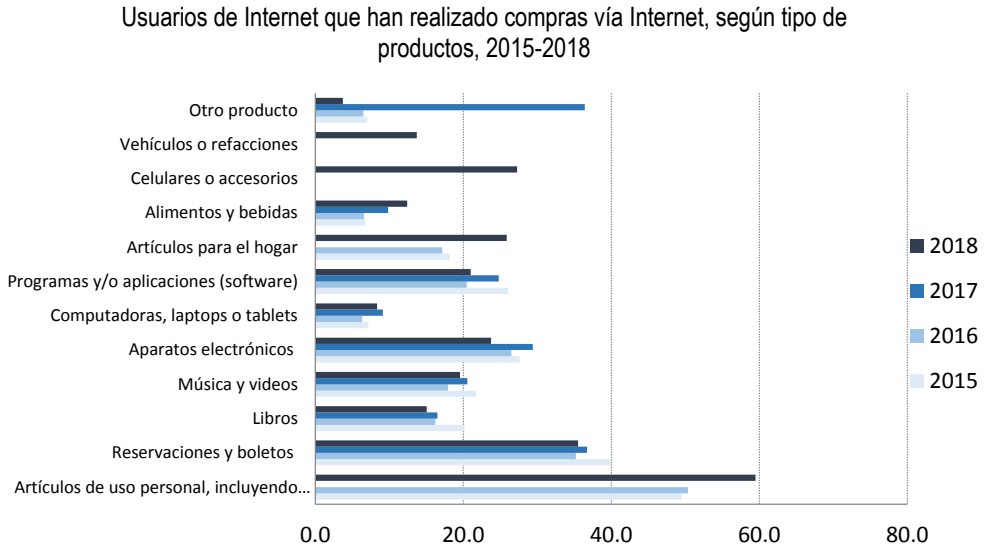
Use of the internet for electronic commerce



ENDUTIH[2] 2018, showed that 15.9% of Internet users have used it to order and buy products, increasing to 19.7% by 2018.

Total of users who buy via internet, on average, 60.3% of purchases are made to sites of national origin.

Frequency of purchase using the Internet



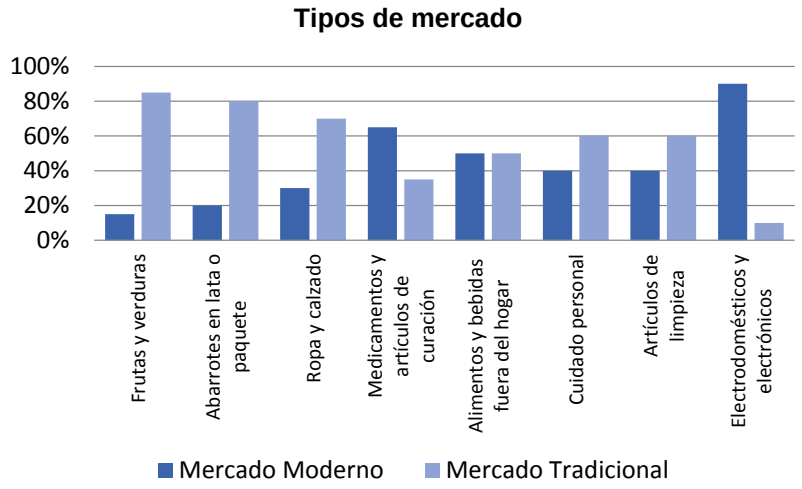
The frequency with which users use the internet to make purchases, are purchases at least once every six months between 2015 and 2018.

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

[1] INEGI, Electronic commerce.
[2] National Survey on Availability and Use of Information Technologies in Homes (ENDUTIH).

MEXICO CPI PRICE QUOTATIONS

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.



More than 56% of the population in Mexico has limited purchasing capacity.

In Mexico there are 1,832,275 retail POS, of which 943,802 (52%) belong to groceries: food, beverages, ice and tobacco[3].

46.94%

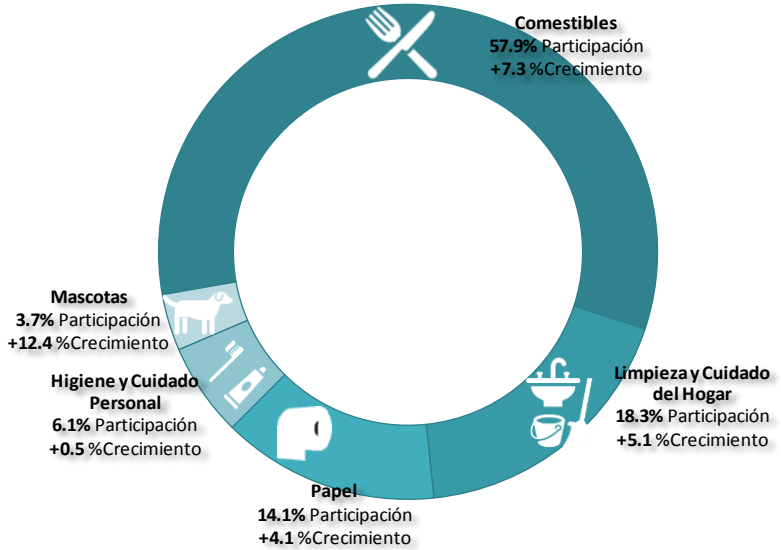
8.08%

Structure of the Wholesale Trade that caters retail stores[4].

44.96%

0.02%

Trade and productive sectors which it benefits



[3] INEGI, Economic Censuses 2014
[4] ISCAM, Perfil del Sector Mayorista Abarrotero, 2018

Calculation of average prices according to data collection technology

01

Fieldwork price collection

In field 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.

02

Collecting prices using Web Scraping

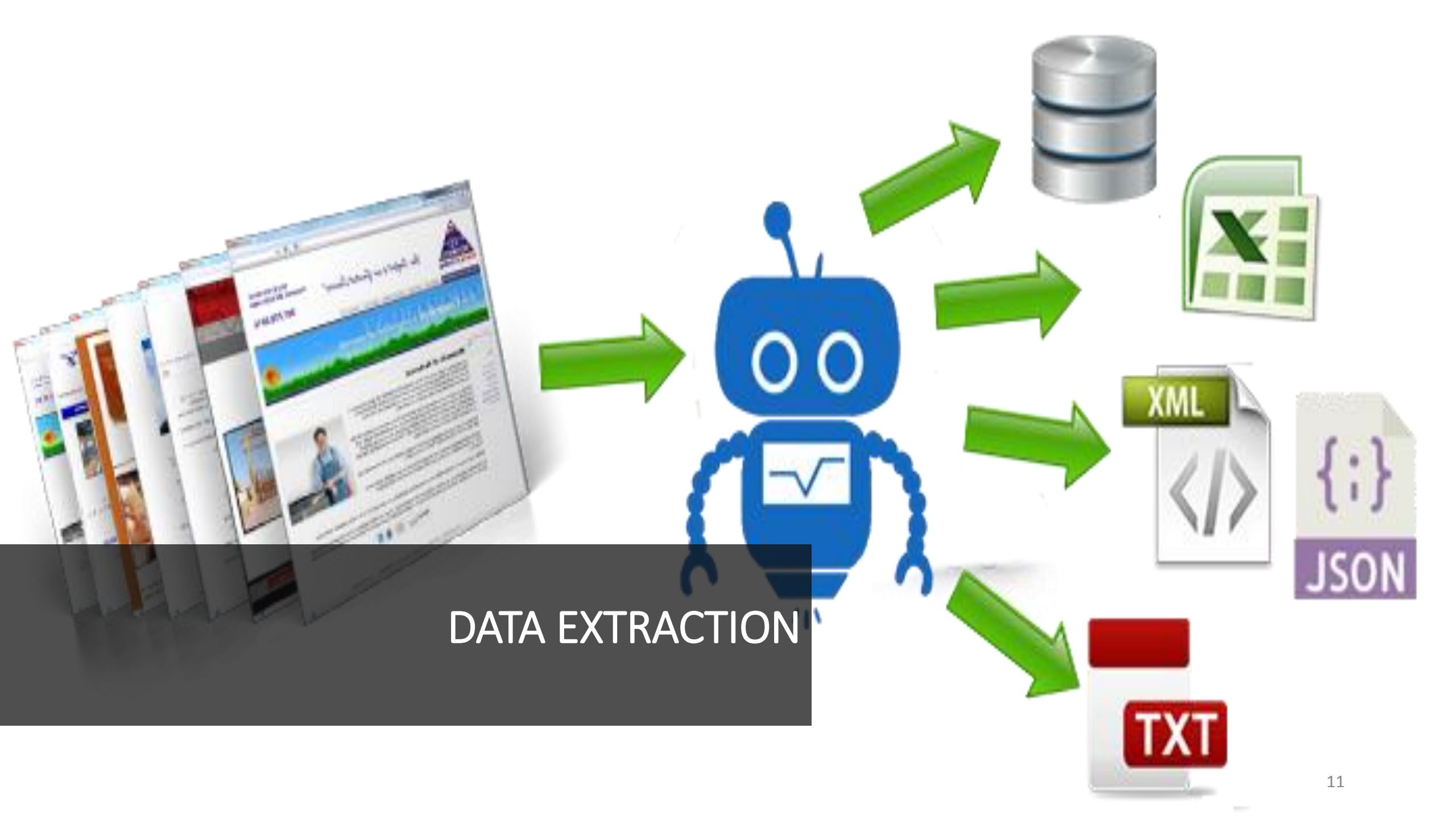
By means of this technique it is possible to considerably increase the price registration number. In this case, the reading is defined as the price obtained now of extracting information from a website from a specific source.

03

Collecting prices using Web Scraping

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.

INEGI



DATA EXTRACTION

Experience when venturing into web scraping

Research and analysis of the electronic pages

Exploration of virtual stores.

Analysis of data extraction

The extraction of data began manually and using free software.

Currently the PYTHON programming language is used.

Identification of the target products or services.

An identifier that indicates that it is always the same product

It is important to always look for a unique or standard code such as the barcode.

Software tests

For some products the price area has data records since 2016, and as of 2017.

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.

.

In the LSNIEG5 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 purposes statistical, census and geographic purposes. and will provide support to them

Main problems in data extraction

1) site security
2) type of information display
3) number of products cataloged on the page
4) versions of the HTML code
5) format styles and diversity of managed products.

Scanner data

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.

We can only access these data provided that they are provided directly by collaboration agreements, they are not public.

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.

Few companies accept to provide the scanned data of each transaction, most accept the extraction using Web Scraping.

INEGI

CASE STUDY

Gasoline

Data: public information and fieldwork survey, we will face the data in three scenarios to find the technique that can give us best accuracy, depending on the technique used

Collections techniques

The prices collected by INEGI fieldwork staff.

Collection using Web Scraping, for the same sample of gas stations that quote fieldwork.

Quotations with WS for all gas stations in the country.

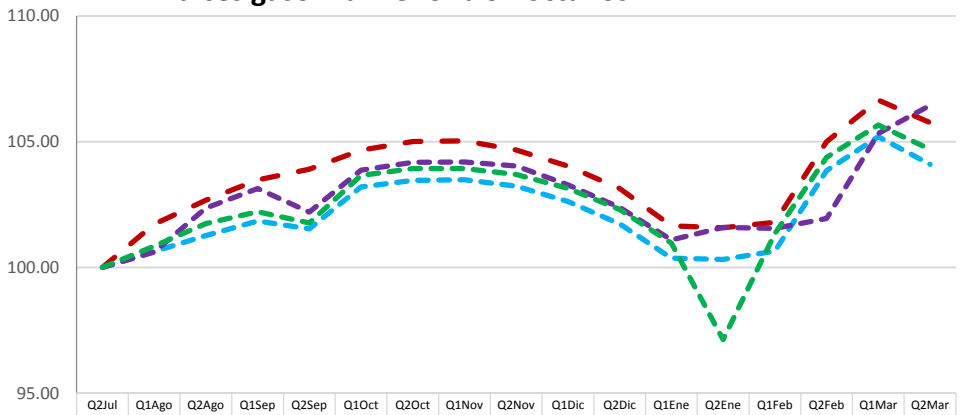
To compare the different quotes mentioned previously, the universe of 11,600 service stations in the country and two types of ɔe of trade, its market gasoline used for calculations. The sample of gas stations using fieldwork survey: 562 POS.

Cálculo

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

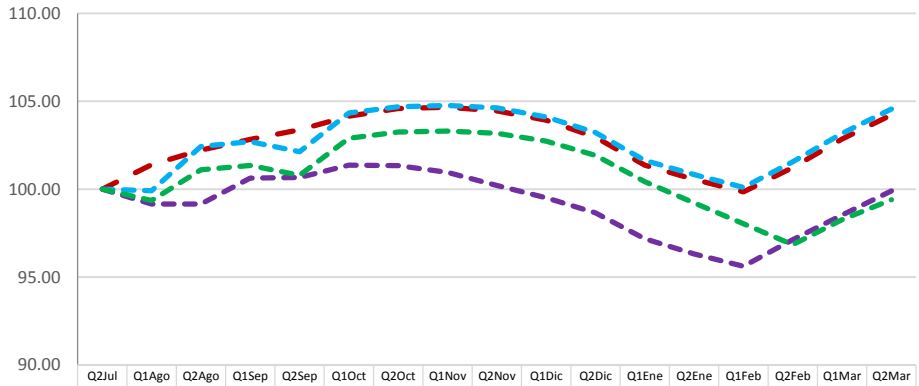
In each case the index was calculated and compared against the published data.

índices gasolina menor a 92 octanos



	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

índices gasolina mayor o igual a 92 octanos

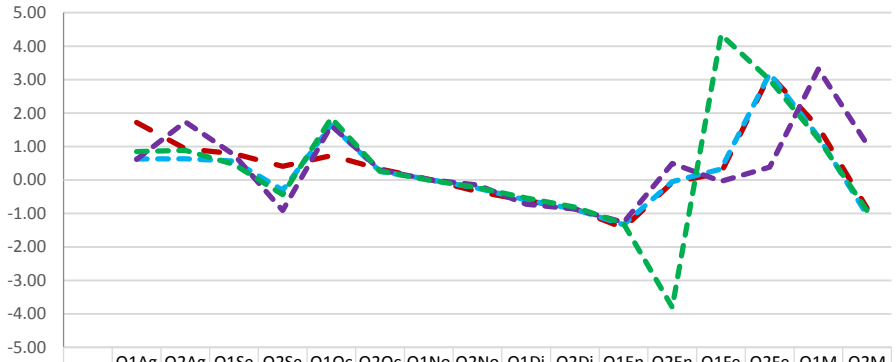


	Q2Jul	Q1Ago	Q2Ago	Q1Sep	Q2Sep	Q1Oct	Q2Oct	Q1Nov	Q2Nov	Q1Dic	Q2Dic	Q1Ene	Q2Ene	Q1Feb	Q2Feb	Q1Mar	Q2Mar
						2018								2019			
Publicado	100.00	101.39	102.22	102.83	103.37	104.15	104.58	104.66	104.45	103.93	102.96	101.37	100.56	99.84	101.23	102.85	104.23
Promedio de la qna	100.00	99.91	102.43	102.71	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, red 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. ..

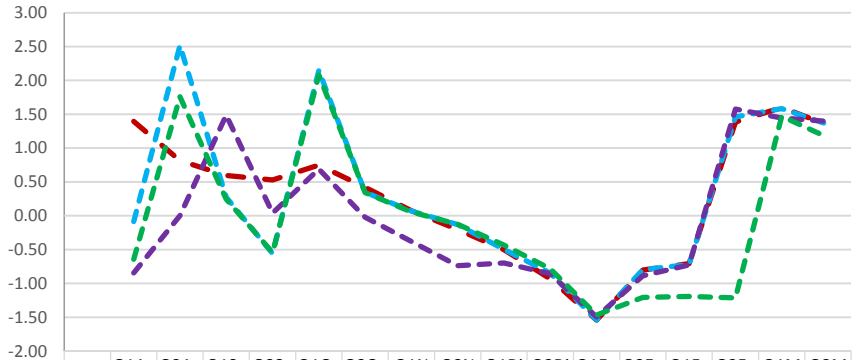
The variations and gasoline indexes are shown in the next figures

Variación quincenal gasolina menor a 92 octanos



	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

Variación quincenal gasolina mayor o igual a 92 octanos



	Q2Jul	Q1Ago	Q2Ago	Q1Sep	Q2Sep	Q1Oct	Q2Oct	Q1Nov	Q2Nov	Q1Dic	Q2Dic	Q1Ene	Q2Ene	Q1Feb	Q2Feb	Q1Mar	Q2Mar
		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, 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

Conclusions

Mexico Benefits for using web scraping and scanner data techniques::

- Increase in the precision of the indexes. This results from the availability of huge volumes of data, which allow accurate and robust estimates of prices average, 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; this for each elementary aggregate.

- Provided: important information and hints for market research. Since it is possible to track the price trend through time (adding quantities in the scanner data case). The different phases of the product cycle identified and represent different brands and models. This information is very helpful to determine the optimum for how many and which items to include in an elementary aggregate.

- We find that help us to organizing field work better, if make the total quotes proportionalper day and market type, give us a better accuracy than doing it by day random.

- The WS use 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; always that they have a web site with the prices of products and updated at less once a day.

Conclusions (continue)

- 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 several and diferent tests for it.

- 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.

- 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.

- 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

In practical terms, average prices, and relative prices 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 they are? 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, in an elementary aggregate, according with the market characteristics (MM and MT proportions) and the consumption patterns (the places where consumers buy). Now get your CPI.

INEGI

Thanks
GREAT DAY!

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

Rafael Posse and Jorge A. Reyes

Mayo 2019

INEGI