



Some Applications of Web Scraping in the CPI, The case of Gasoline

June 2023

EXTRACTION OF BIG DATA VOLUMES



Web Scraping



Scanner Data



Citizens' Statistics



1

In Big Data we work with: Volume, Speed, Variety, Veracity and Value.

2

It provides sufficient infrastructure to generate consistent data series

3

It helps to maintain the continuity of the data collection flow, with basic quality standards

4

Leverage data based on robust analysis, for its implementation in the Price Index

RESEARCH AND DEVELOPMENT

WEB SCRAPING IN CPI

The beginning

Since 2018, the appropriate methods and techniques for the recovery of products and prices have been investigated on the websites of different commercial chains.

Learning

It started with household appliances and white goods. The extraction is carried out in department stores and supermarkets such as: Liverpool, Coppel, Famsa, Wal Mart, Soriana and La Comer.

Final stage

The leading drill is gasoline. Data is available for all gas stations in the country since August 2018. Work is underway to incorporate this method into the official calculation.

What's Next


In 2021, extractions of some services were incorporated:

- LP Gas
- Air transport
- Telecommunications

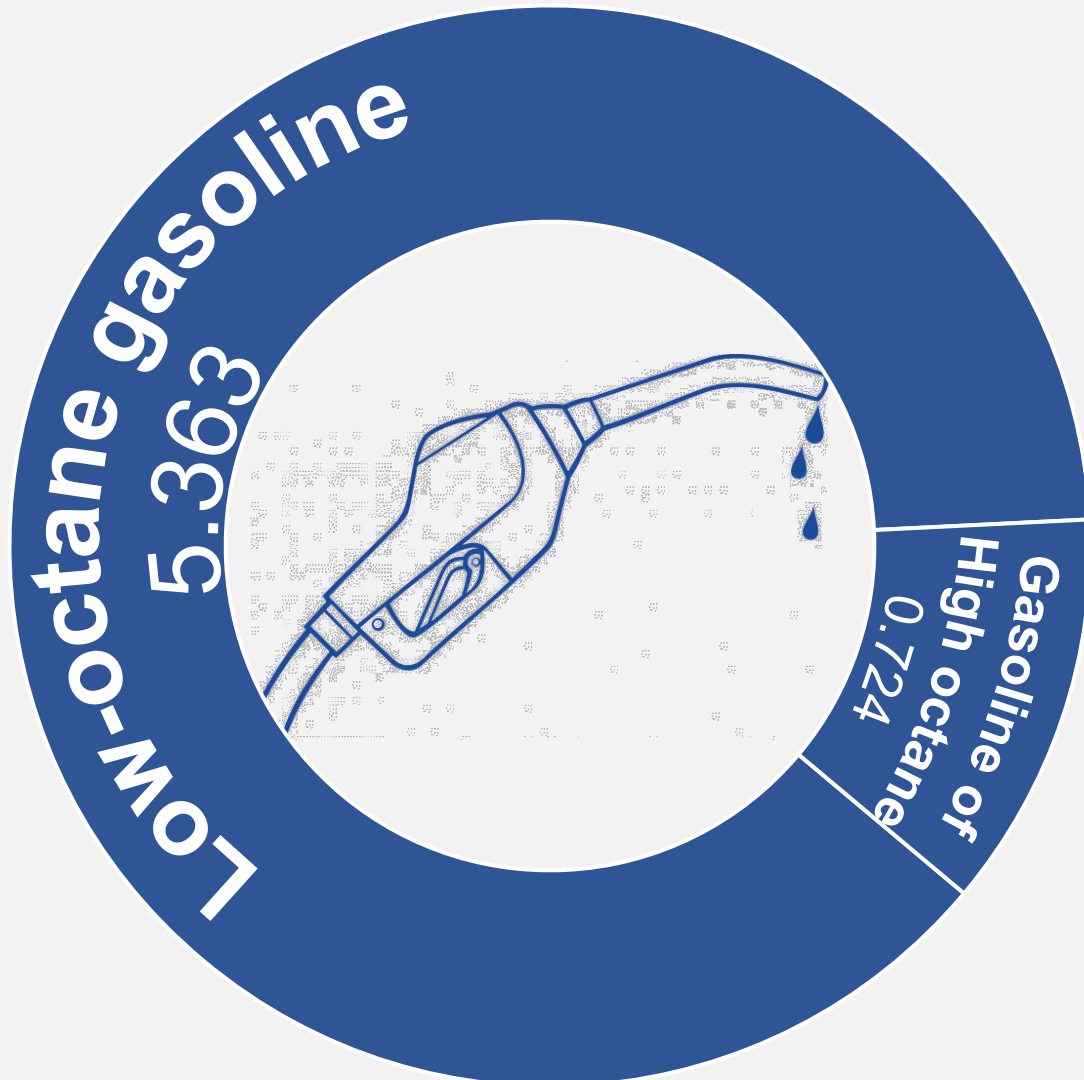
In 2021, the extraction will be extended to all the generic foods in the basket offered in supermarkets

At the door

By the second half of 2021, extraction would have been extended to all generic food from the basket offered in the chains. New store chains will be incorporated, and data extraction will be distributed by entity.



SUBSYSTEM
GASOLINE



SUBSYSTEM GASOLINE

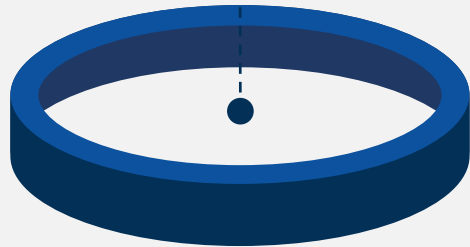


it consists of two generics:

- High-octane gasoline (92 octane)
- Low-octane gasoline (87 octane)

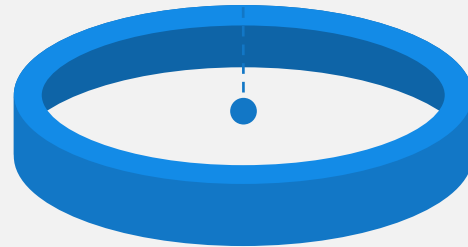
The sample of gas stations is directed (not probabilistic) consists of 570 service stations in in which prices are quoted once a week, by direct visit.

WEB SCRAPING GASOLINE PRICES



Service stations

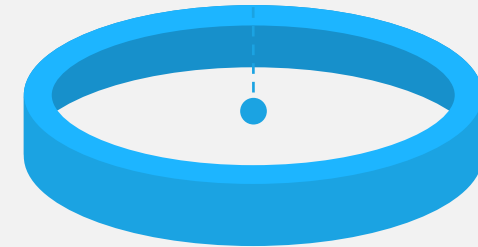
Every day the prices of 12,545 existing gas stations are extracted.



Quoted prices

On average, 24,245 gasoline prices are obtained in the extraction:

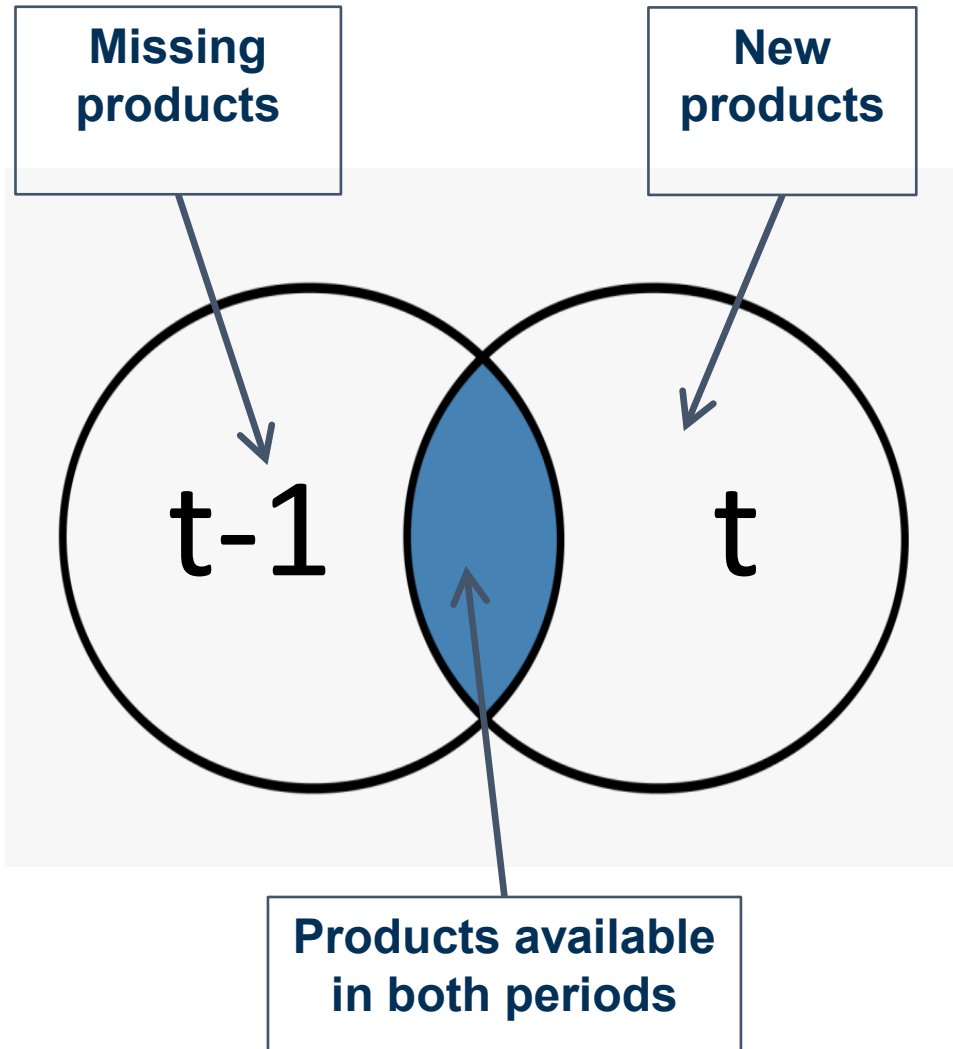
- 12,542 low octane.
- 11,703 high-octane.



Twice a day

The extraction of the information is done automatically at 8:00 and 16 hrs, bringing about 48,490 daily prices.

INTEGRATION METHODOLOGY



Matched model method.

The approach of the method is to estimate the price change between two periods using the products available in both time periods only, excluding the prices of new and missing products.

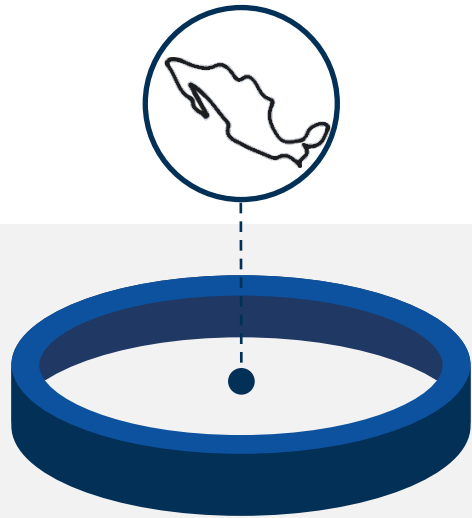
However, it is strengthened by the large surplus of the census of products represented in the complete set of data extracted from the web.



CPI REGIONS

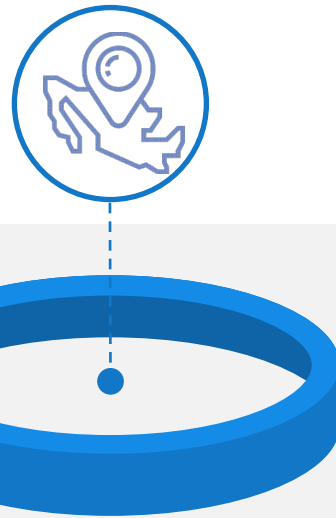


GEOGRAPHICAL DISTRIBUTION SCRAPING GASOLINE



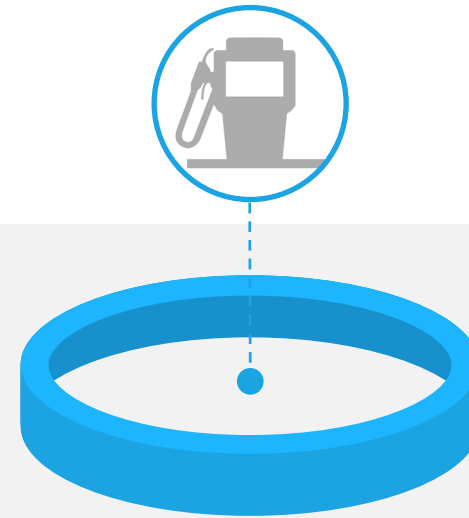
Geographic areas

55



Regions

7

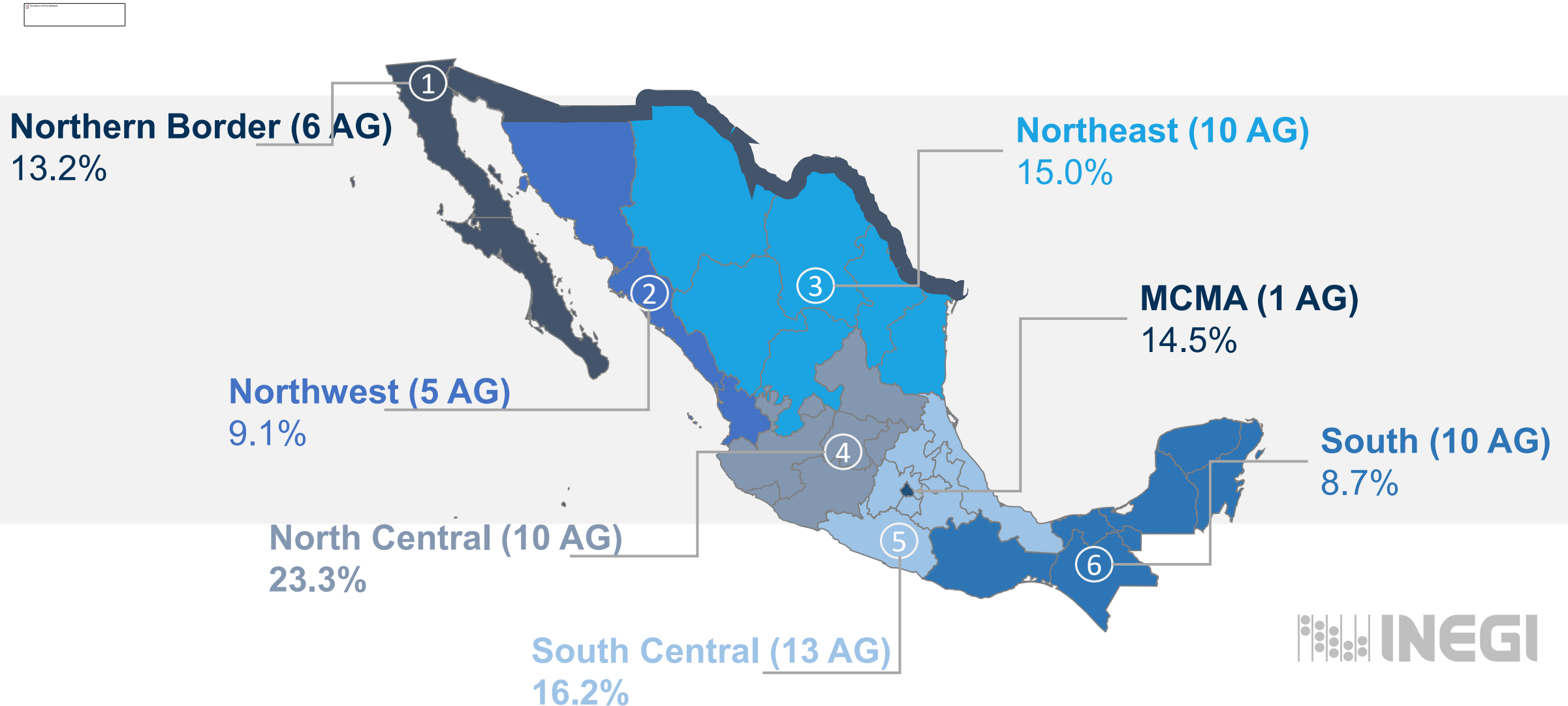


Service stations

1. Northern Border	1,369
2. Northwest	1,101
3. Northeast	1,808
4. North Central	2,299
5. South Central	1,501
6. South	1,019
7. MCMA	1,001

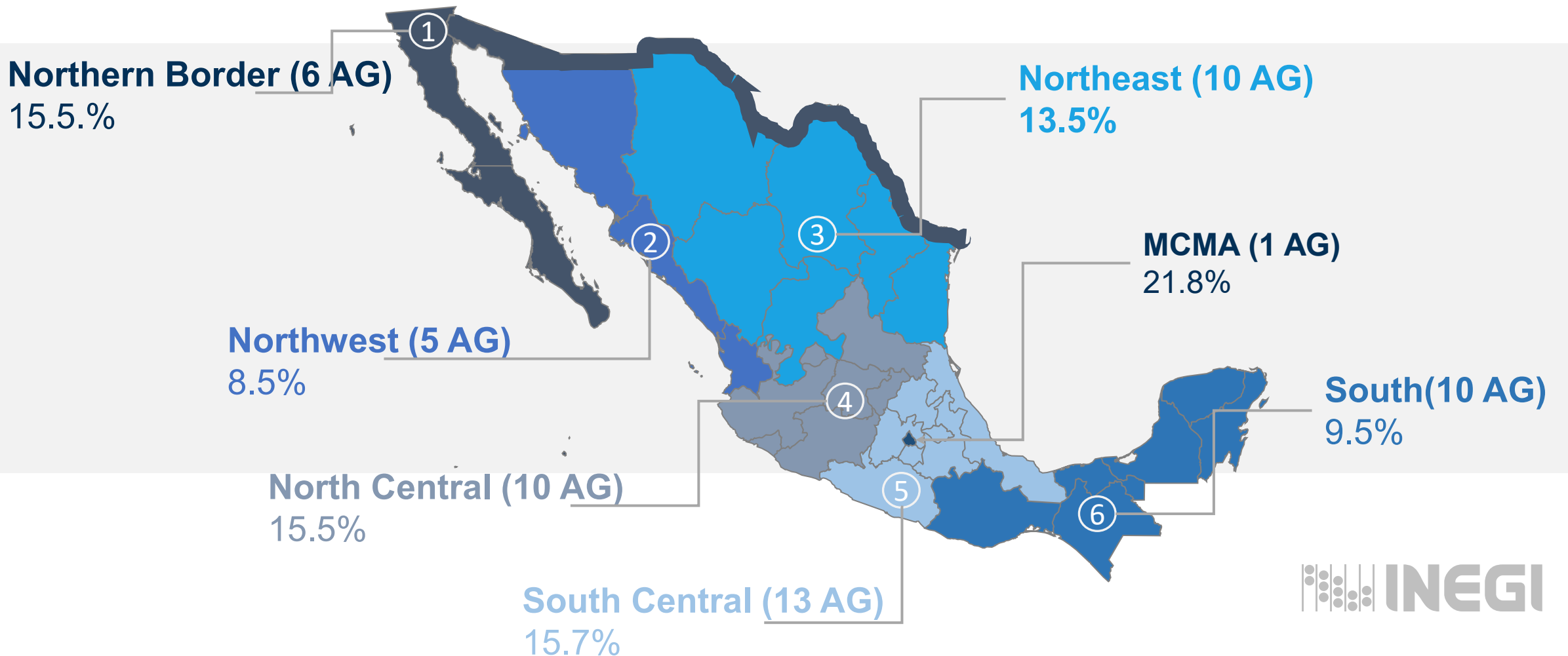
CPI REGIONS AND THEIR WEIGHTING

LOW-OCTANE GASOLINE



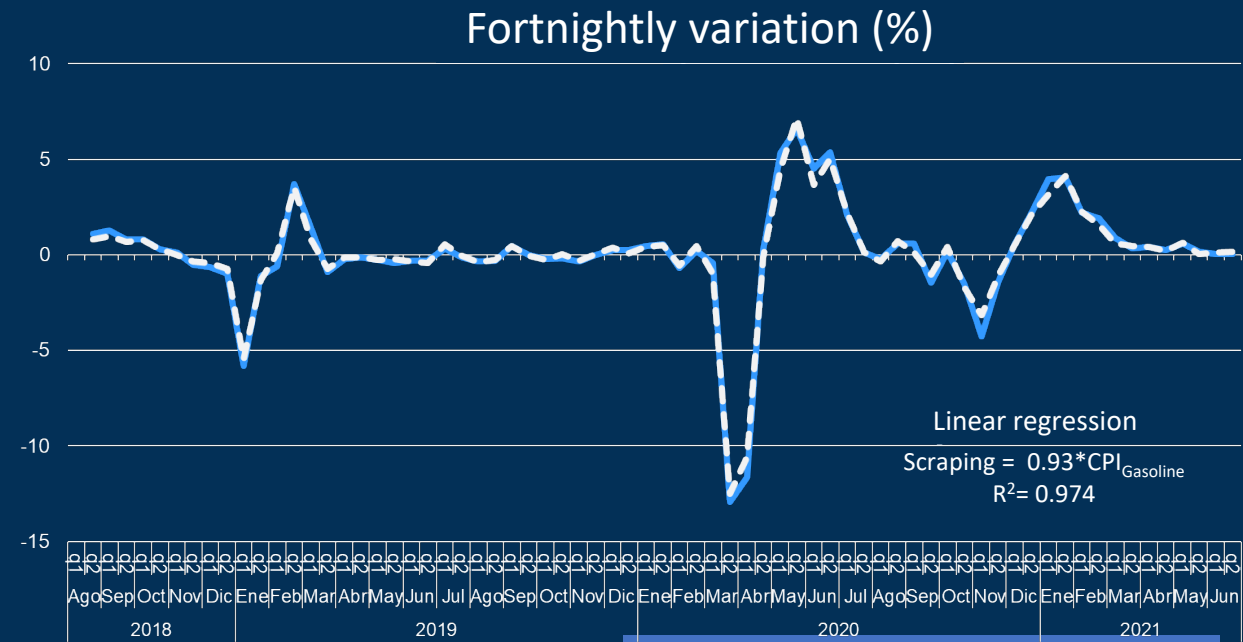
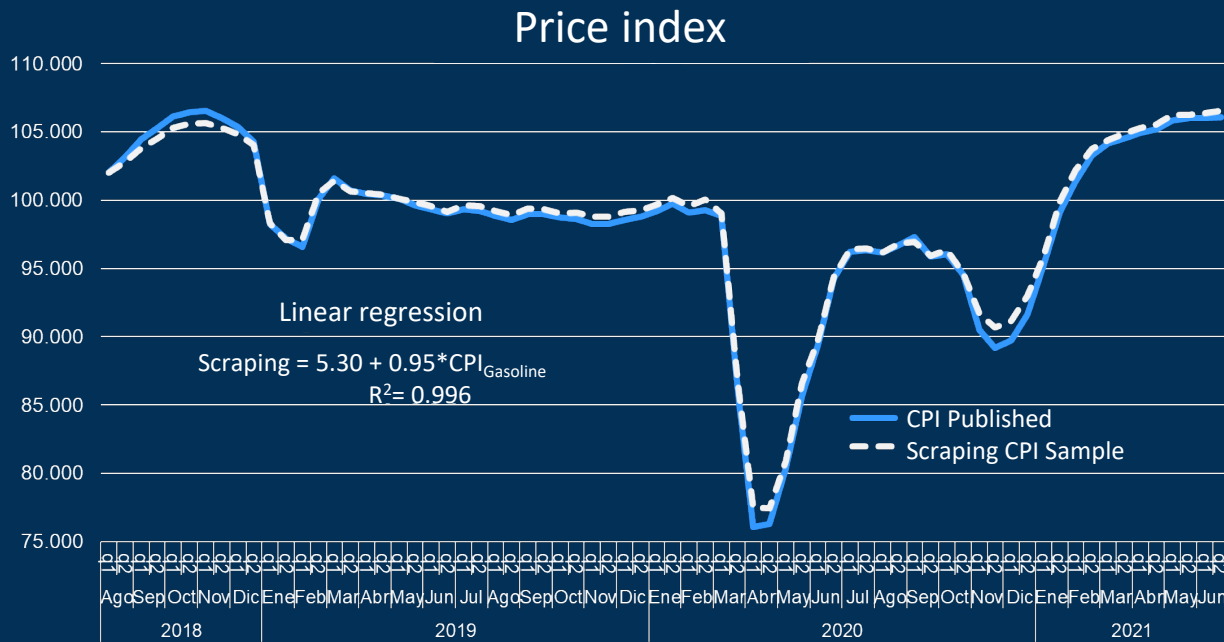
CPI REGIONS AND THEIR WEIGHTING

HIGH-OCTANE GASOLINE



PERFORMANCE: UNIVERSE OF GAS STATIONS ON THE NORTHERN BORDER

LOW-OCTANE GASOLINE



Scraping:

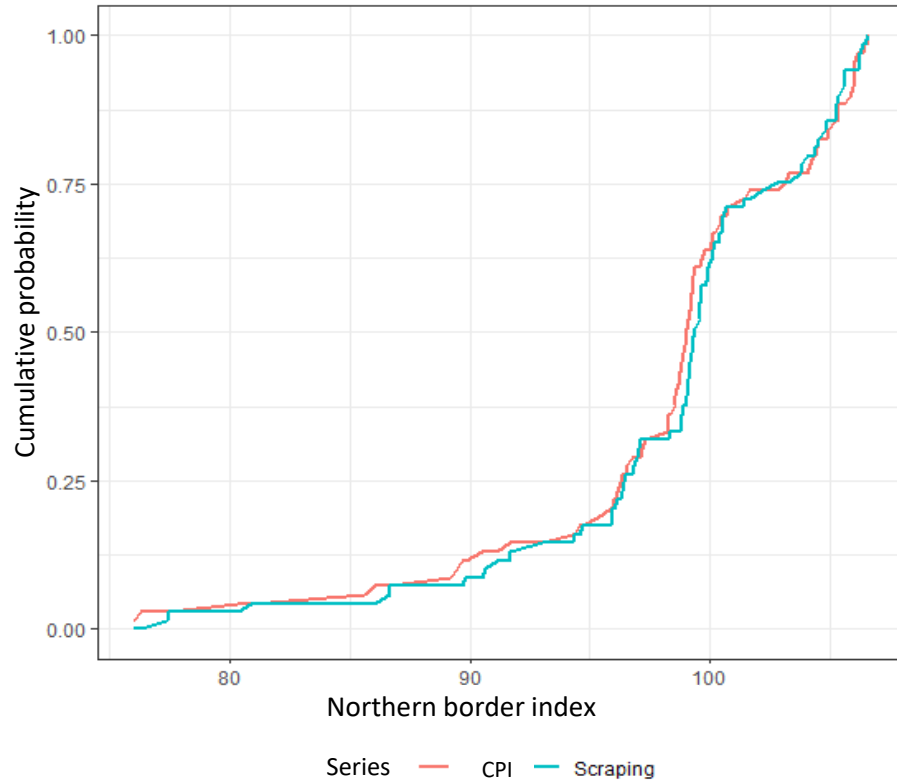
- All gas stations in the cities of the region
- Daily quote, twice daily
- Matched Model Method
- Regional weighted index

Mann-Whitney: Does not reject equal means, paired sample (p-value 0.825)



STATISTICAL TESTS LOW-OCTANE GASOLINE NORTHERN BORDER CONSUMER PRICE INDEX

Empirical cumulative distribution function for low-octane gasoline

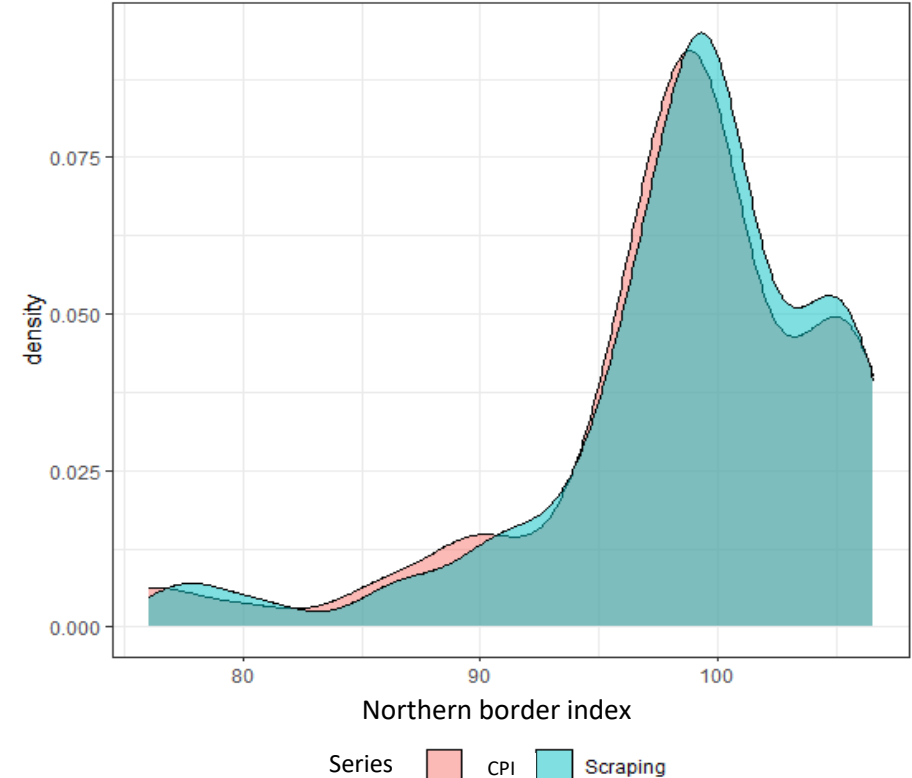


Kolmogorov-Smirnov:
Does not reject the same
distribution (with p-value
of 0.604)

Levene: Does not reject
equality of variance
(with p-value of 0.766)

Cucconi: Does not reject
the same trend and
dispersion (with p-value
of 0.911)

Empirical density function low octane



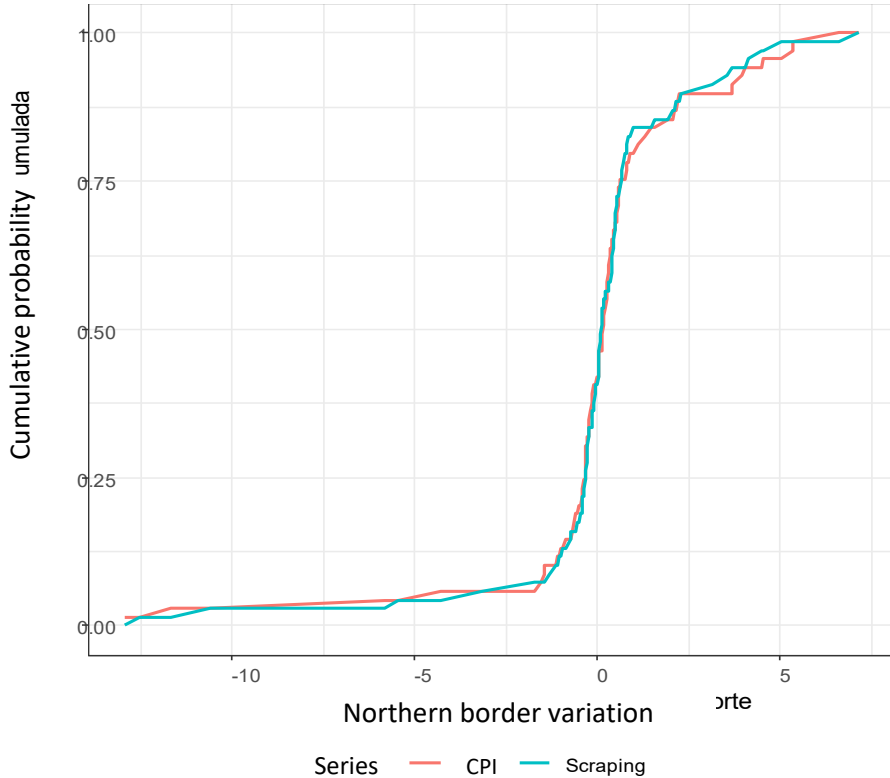
Scraping:

- All gas stations in the cities of the region
- Daily quote
- Matched Model Method
- Regional weighted index

STATISTICAL TESTS LOW-OCTANE GASOLINE NORTHERN BORDER FORTNIGHTLY VARIATION



Empirical cumulative distribution function for low-octane gasoline

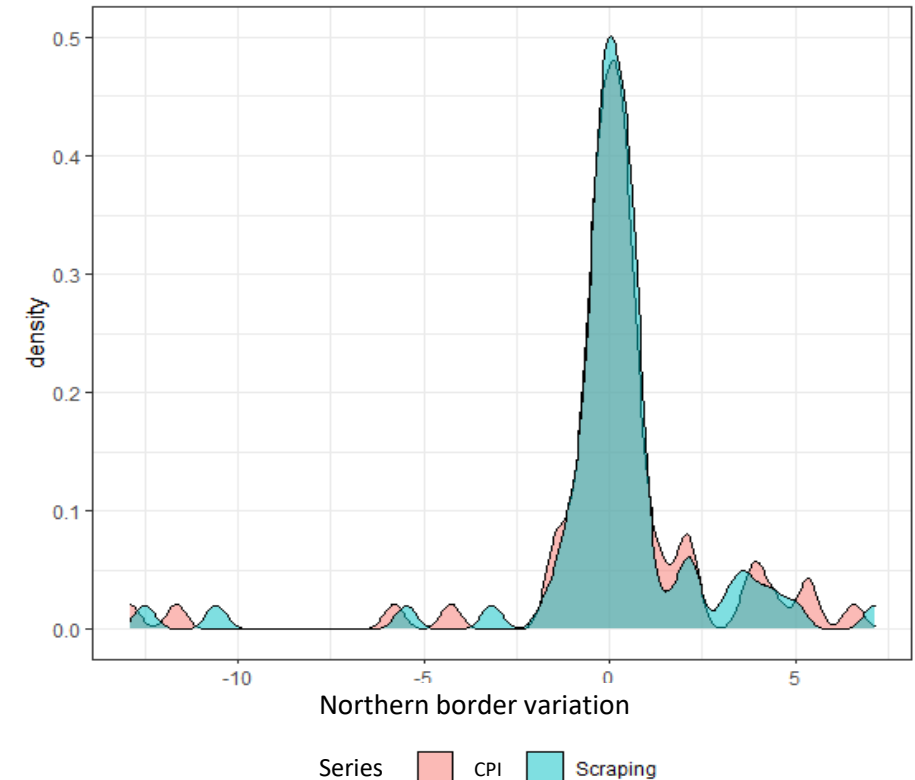


Kolmogorov-Smirnov:
Does not reject the same distribution (with p-value of 0.999)

Levene: Does not reject equality of variance (with p-value of 0.805)

Cucconi: Does not reject the same trend and dispersion (with p-value of 0.778)

Empirical density function, low octane gasoline



Scraping:

- All gas stations in the cities of the region
- Daily quote
- Matched Model Method
- Regional weighted index



LOW-OCTANE GASOLINE STATISTICAL TESTS

Indexes

For the seven regions of the CPI, equality tests were carried out for:

- Distribution
- Variance
- Trend and dispersion
-

In none of the cases can they be rejected.

Variations

For the seven regions of the CPI, equality tests were conducted for:

- Media
- Distribution
- Variance
- Trend and dispersion
-

In none of the cases can they be rejected.

Both procedures describe the process that generates prices in the same way, indices describe overlapping curves.

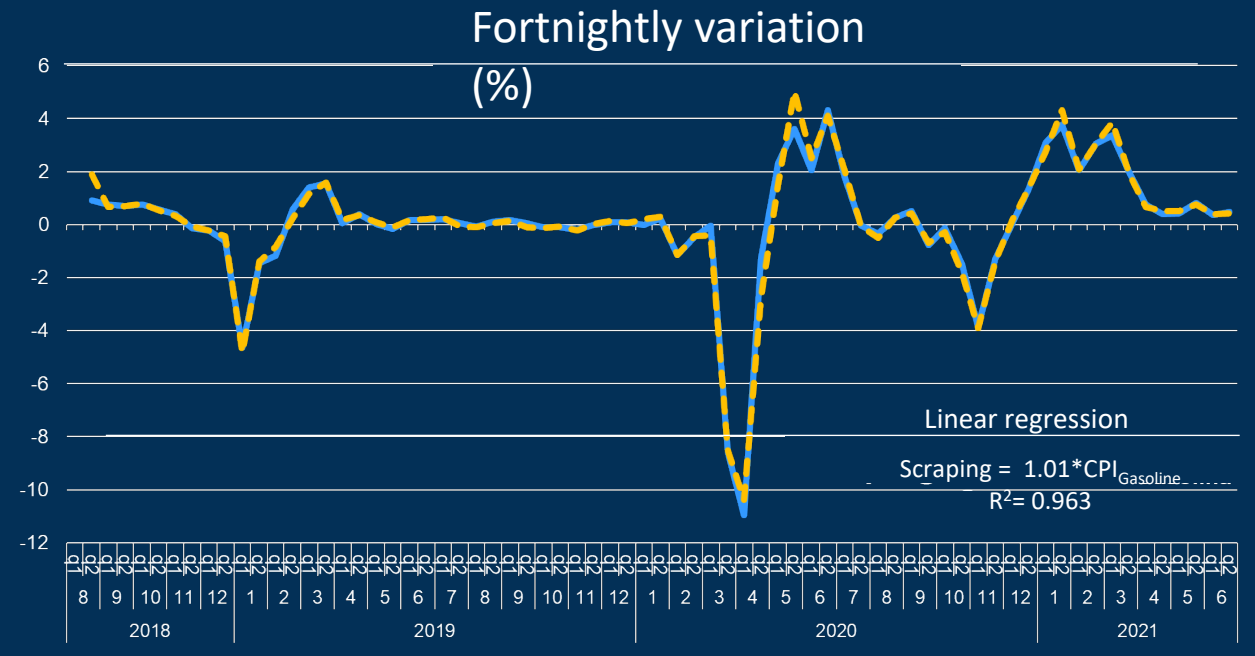
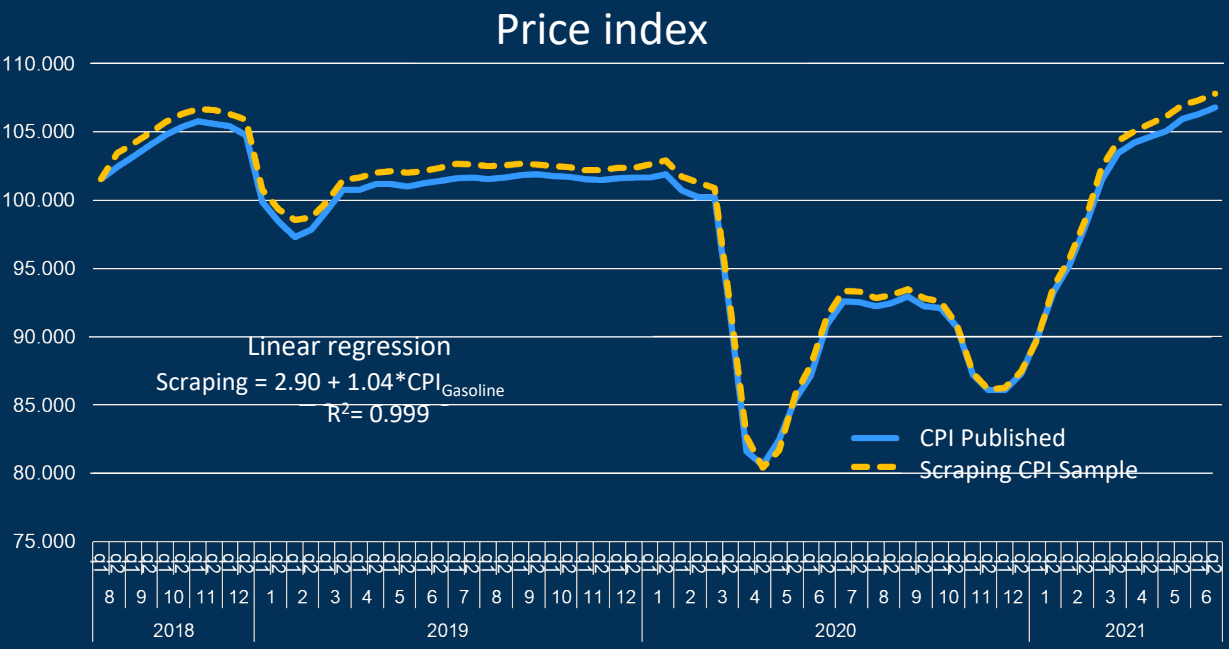
On the other hand, the dynamics behind the price developments come from the distribution itself.





HIGH OCTANE
GASOLINE

PERFORMANCE: SAMPLE OF GAS STATIONS ON THE NORTHERN BORDER HIGH-OCTANE GASOLINE



Scraping:

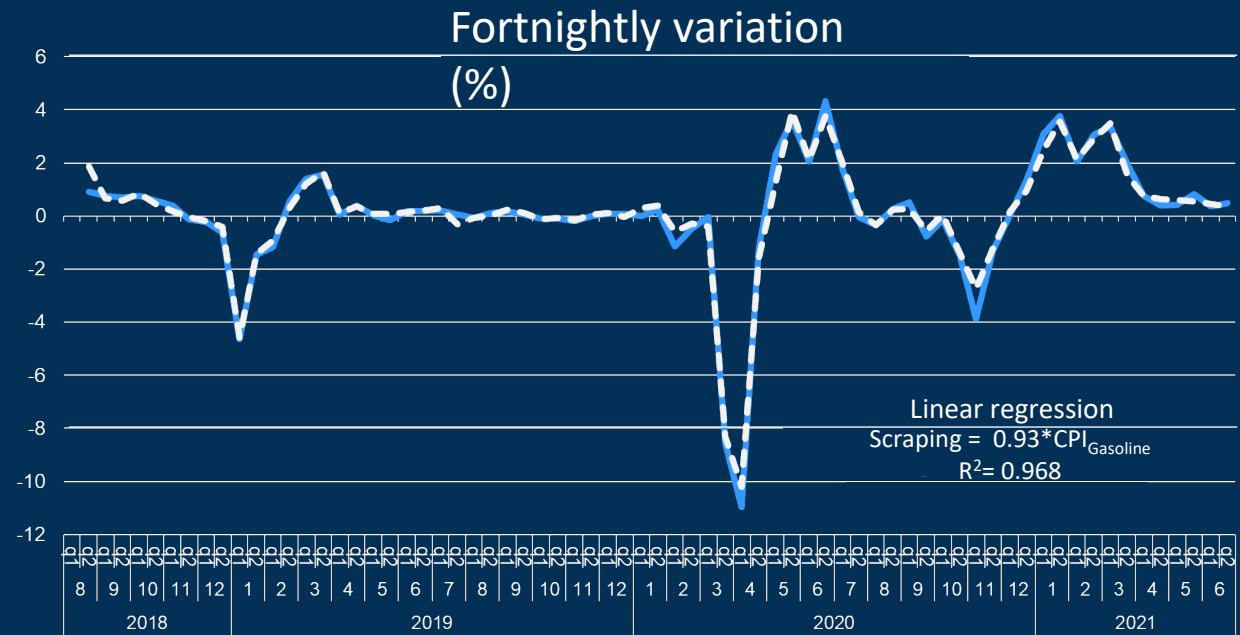
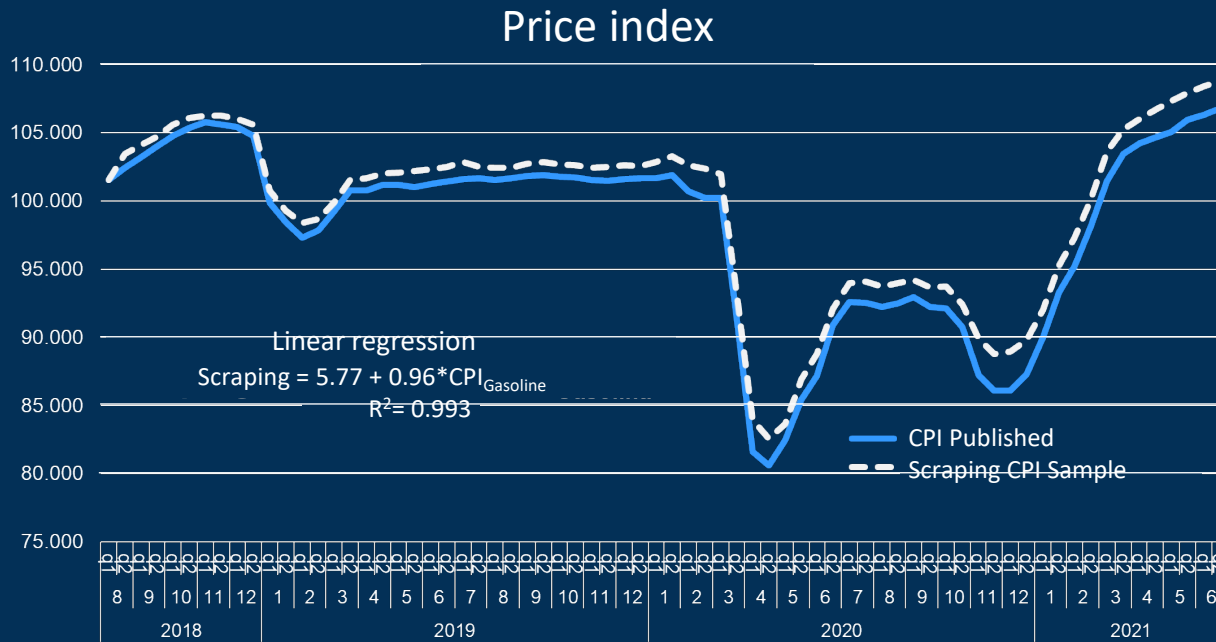
- Same gas stations of the sample of the region
- Daily quote, twice daily
- Matched Model Method
- Regional weighted index

Mann-Whitney: Does not does not reject equal means, paired sample (p-value 0.834)



PERFORMANCE : UNIVERSE OF GAS STATIONS ON THE NORTHERN BORDER

HIGH-OCTANE GASOLINE



Scraping:

- All gas stations in the cities of the region
- Daily quote, twice a day
- Matched Model Method
- Regional weighted index

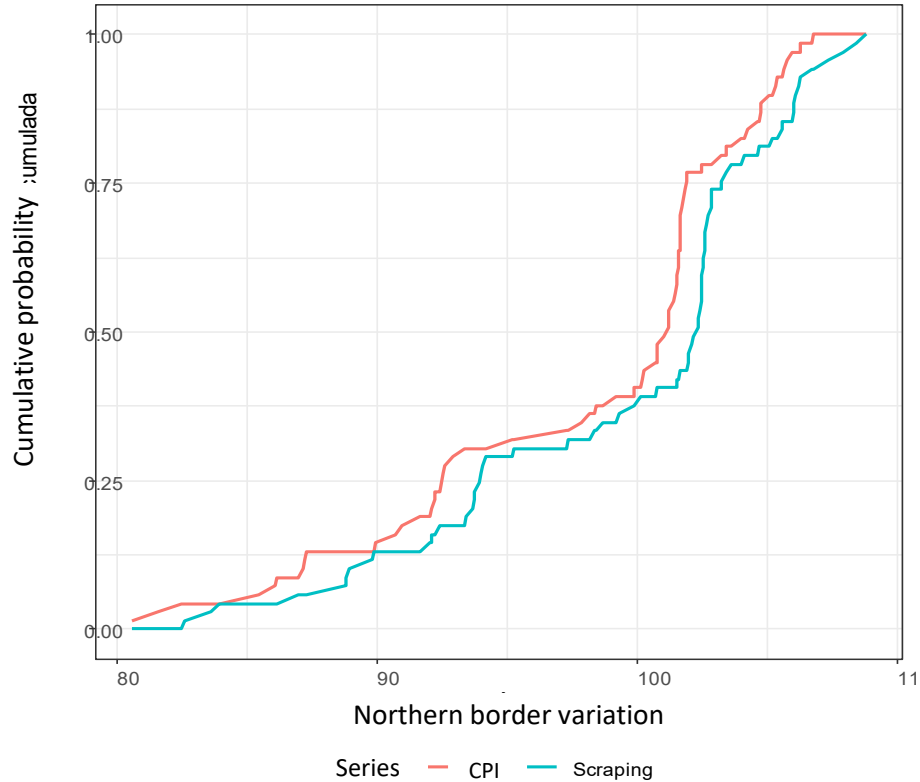
Mann-Whitney: Does not reject equal means, paired sample (p-value 0.376)



STATISTICAL TESTS GASOLINE HIGH OCTANE NORTHERN BORDER PRICE INDEX



Empirical cumulative distribution function for high-octane gasoline

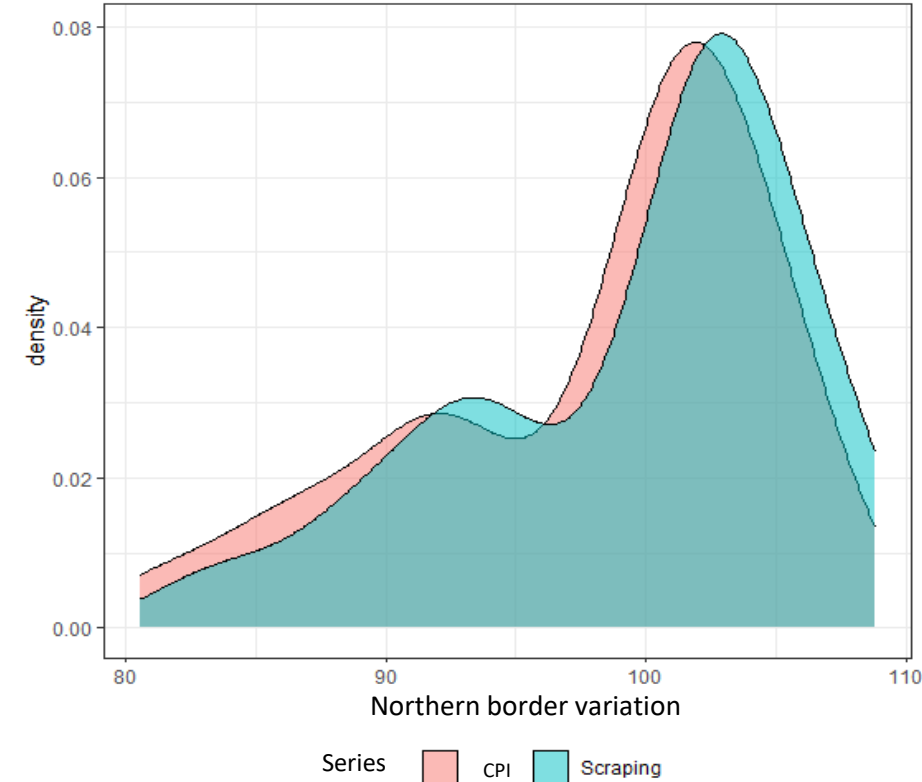


Kolmogorov-Smirnov:
Rejects equal distribution
(with p-value of 0.001)

Levene: Does not reject
equality of variance
(with p-value of 0.822)

Cucconi: Does not reject
the same trend and
dispersion (with p-value
of 0.120)

Empirical density function, High-octane gasoline



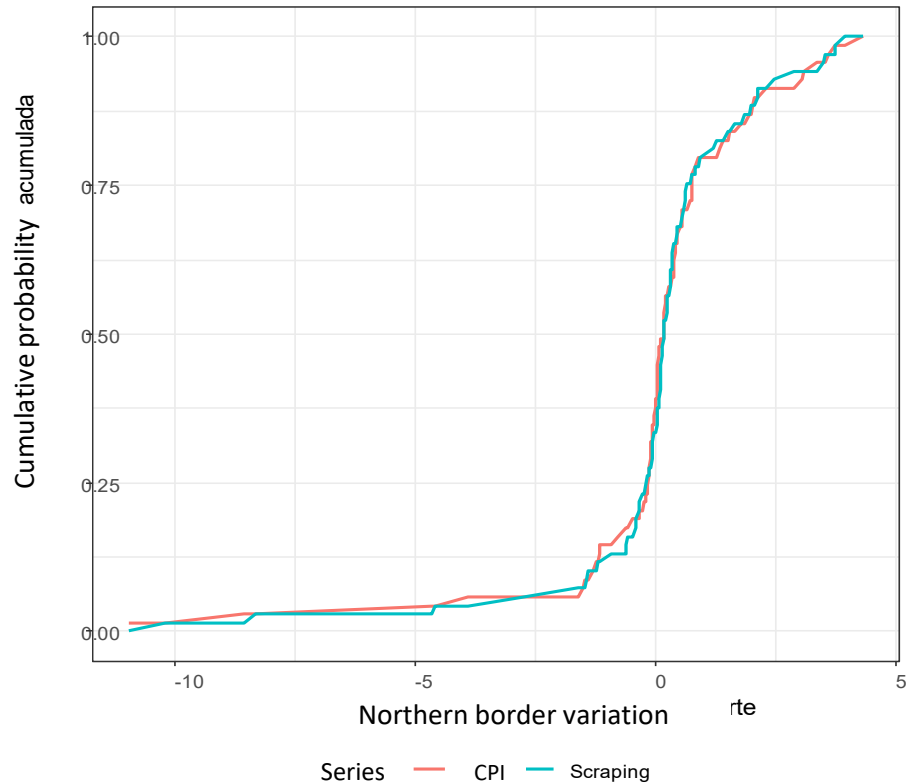
Scraping:

- All gas stations in the cities of the region
- Daily quote
- Matched Model Method

STATISTICAL TESTS GASOLINE HIGH OCTANE NORTHERN BORDER FORTNIGHTLY VARIATION



Empirical cumulative distribution function for high-octane gasoline

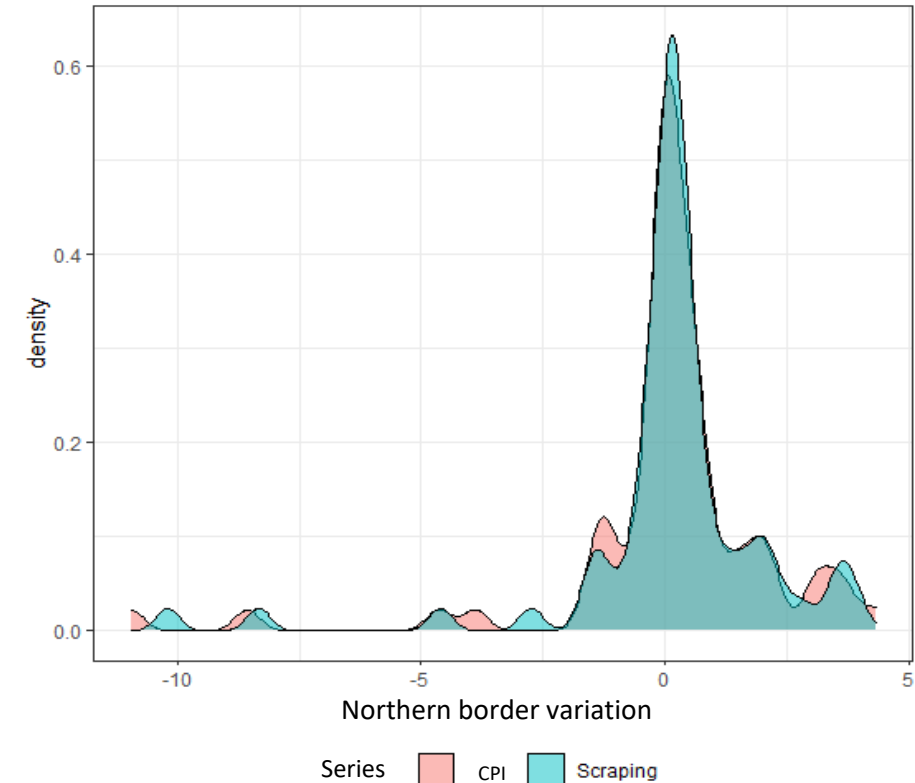


Kolmogorov-Smirnov:
Does not reject equal
distribution (with p-value
of 0.959)

Levene: Does not reject
equality of variance
(with p-value of 0.786)

Cucconi: Does not reject
the same trend and
dispersion
(with p-value of 0.925)

Empirical density function, High-octane gasoline



Scraping:

- All gas stations in the cities of the region
- Daily quote
- Matched Model Method
- Regional weighted index



HIGH-OCTANE GASOLINE STATISTICAL TESTS

Indexes

For the seven CPI regions, equality tests were carried out for:

- Distribution
- Variance
- Trend and dispersion
-

In Region 2, none can be rejected.
In the regions:1, 4, 5, 6 and 7, the equality of distribution is rejected.
In Region 3 the equality of distribution and trend are rejected.

Variations

For the seven CPI regions, equality tests were carried out for:

- **Mean**
- **Distribution**
- **Variance**
- **Trend and dispersion**
-

In neither cases can they be rejected.

Both procedures describe the process that generates prices equivalently, indices describe parallel or overlapping curves (with the exception of region 3).

On the other hand, the dynamics behind the evolution of prices come from the same distribution.

CONCLUSION



With the statistical tests, it is concluded that it is possible to incorporate the prices obtained by web scraping for generic gasoline into the CPI production calculation.

The prices that the Comisión Reguladora de Energía (CRE) publishes represent the universe of gas stations in the 55 geographical areas, using them increases the accuracy in the measurement.

Conociendo
México

800 111 46 34

www.inegi.org.mx

atencion.usuarios@inegi.org.mx



INEGI Informa

GRACIAS



List of products obtained with Web Scraping

	Generic
1	Fresh Cheese
2	Avocado
3	Rice
4	Courgette
5	Shrimp
6	Pork
7	Beef
8	Onion
9	Chile Poblano
10	Dry Chile
11	Chile Serrano
12	Peach
13	Bean
14	Guava
15	Egg
16	Tomato
17	Milk powder
18	Lettuce and cabbage
19	Lemon
20	Apple

	Generic
21	Melon
22	Orange
23	Nopales
24	Other canned fruits
25	Other fruits
26	Other dried legumes
27	Other vegetables and legumes
28	Other fresh chilies
29	Other seafood
30	Other Cheeses
31	Potato and other tubers
32	Papaya
33	Cucumber
34	Pear
35	Fish
36	Pineapple
37	Banana
38	Chicken
39	Yellow Cheese

	Generic
40	Manchego and Chihuahua cheese
41	Oaxaca and asadero cheese
42	Watermelon
43	Grape
44	Yogurt
45	Carrot
46	Green tomato
47	Green beans
48	Chayote Squash
49	Chicken and salt concentrates
50	Pasteurized and fresh milk

[Go Back](#)