

# French retail gasoline: overview of available data

Etienne Chamayou\*

CREST and Department of Economics, Ecole Polytechnique

February 20, 2015

## **Abstract:**

This note describes data collected and organized with a view to study competition and price dispersion on the French retail gasoline market.

**Keywords:** Competition, Gasoline

**JEL Classification Numbers:** L13

---

\*e-mail: *etienne.chamayou@ensae.fr*

# 1 Data sources

## 1.1 Price comparison website [prix-carburant.gouv.fr](http://prix-carburant.gouv.fr)

The main data source is the governmental price comparison website [prix-carburant.gouv.fr](http://prix-carburant.gouv.fr). The website was launched on January 2, 2007 and all gas stations having sold above  $500m^3$  of gasoline the previous year are since then required (by law) to keep prices up to date. Information available on the website includes gas station brands, location and amenities, but such information has appeared to be far less reliable than prices.

Since September 2014, price records are "open data" i.e. one can download historical price records. Nevertheless, gas station brands are not included, nor are price series of stations which have ceased activity on the website prior to 2014. This is all the more unfortunate as stations tend to be registered under several ids over time, and no effort to merge such accounts appears to have been undertaken on government's side.

From September 9, 2011 on, a script was used to collect prices and gas station brands on a daily basis. Station location and amenities were collected on a less regular basis.

## 1.2 Price comparison website [zagaz.com](http://zagaz.com)

The price comparison website [zagaz.com](http://zagaz.com) was launched on January 31, 2006 and traditionally differs from its governmental counterpart through its "crowdsourcing" philosophy. Indeed, until end 2014, information available on the website was exclusively provided by users. Since some regions were poorly covered, the website now mixes users's information and prices from the governmental website.

Though price records from the website are of limited interest, some aspects of the website are of significant interest. First, the website is way more comprehensive than the governmental website as regards the network of gas station, and way more reliable as regards the location of gas stations (the website indicates whether a user has confirmed its accuracy). Another interesting feature is that each user has a public profile exhibiting his or her contributions to the website. This offers the opportunity to build a proxy of the "choice set" (i.e. gas stations potentially patronized) of users with sufficient activity. Lastly, users' registration date is public, offering possibilities to build proxies for general consumer information.

### 1.3 INSEE data

Based on city and zip code, each gas station has been matched to its INSEE municipality, which implies that many variables are available to describe the environment in which it operates (Census data produced by INSEE and many others such as significant commuter flows).

### 1.4 Additional data sources

Data about raw product prices (Brent and wholesale diesel quotations on the Rotterdam market) were collected from UFIP and Reuters.

Data were obtained from OpenStreetMap to marginally improve and verify data about gas station characteristics and locations.

Data about gasoline price queries were finally collected from Google.

## 2 Data processing

### 2.1 Prices

Price data collected from [prix-carburants.gouv.fr](http://prix-carburants.gouv.fr) were found to have some significant shortcomings. Numerous price records were too low, high or rigid to be accurately reflect actual gas station prices. It also appeared that in many cases, diesel price was updated in place of unleaded gas price (and vice versa). Abnormal rigidity in price records can likely partly be accounted for by the presence in the data of gas stations not subject to the disclosure obligation. Prices were thus controlled and fixed or set to missing conservatively based on level, variations and rigidity.

### 2.2 Gas station characteristics

Address and gps coordinates provided on [prix-carburants.gouv.fr](http://prix-carburants.gouv.fr) are also of limited quality. Most gps coordinates were obtained by geocoding, and the accuracy of the result thus strongly depends on the quality of the address. A simple check based on INSEE code revealed a few blatant mistakes. The lack of accuracy was otherwise due to lack of information of the address (e.g "Zone industrielle XXX"). Beyond the errors induced in competition analysis, this problem hampers the ability to find gas station duplicates in the data.

## 2.3 Gas station duplicates

# 3 Descriptive statistics

## 3.1 Open data

## 3.2 Data collected on [prix-carburants.gouv.fr](http://prix-carburants.gouv.fr)

## 3.3 Data collected on [zagaz.com](http://zagaz.com)

Table 1 provides an overview of historical data released by the government about gas station prices. A total number of 7,997 gas stations registered on the website during the first year following its creation (2007). Among them, 7,470 were still posting prices in 2014. The total number of gas station IDs present in the data as of XX, 2014 is 14,462. The number of gas stations which have posted at least one price in 2014 is 11,042.

Since price series of stations having ceased activity before 2014 are not provided, there are thus c. 11,000 price series in the data, of which only c. 7,500 start in 2007

Table 1: Station entries and survival in database

Year	2007	2008	2009	2010	2011	2012	2013	2014	Total exits
2007	7997	7482	7478	7477	7476	7476	7476	7470	527
2008	0	1003	490	480	479	479	477	476	527
2009	0	0	1505	805	786	783	782	782	723
2010	0	0	0	1439	605	558	553	550	889
2011	0	0	0	0	728	442	439	437	291
2012	0	0	0	0	0	581	427	421	160
2013	0	0	0	0	0	0	717	414	303
2014	0	0	0	0	0	0	0	492	0
Total	7997	8485	9473	10201	10074	10319	10871	11042	3420

In diagonal are the numbers of stations newly registered in each year. Additional figures right from the diagonal indicate how many stations registered in year X (row) are still active in year Y (column).

## A Additional details

### Price information

- Daily files are merged recursively on station ids which leads to create a database of 10,433 individuals over 640 days, including unobserved periods which result in missing values within series.
- Since dates of price changes are known, missing periods can yet be filled in many cases (e.g: if price on day 2 is missing, it can be checked in day 3 that price hasn't changed since day 1 and if price on days 10-15 are missing, it can be seen on day 16 that the last change was made on day 13: prices for 13-15 are then input backward and forward for 10-12). TODO: stats.

### Station information:

- Matching of stations with INSEE codes: Zip codes: problem of cedex and changing zip codes. City name matching implies the generic problem of string comparison, not to mention the fact that the same name can be used in different regions and that city names sometimes change (small municipalities are regrouped). Approach: matching on zip then city name
- Matching of databases: address standardization but still remains a big issue as quite different addresses can be provided, a piece of info can be up to date in one database while a bit old in the other. It requires a multicriteria approach.
- Geocoding: address standardization
- Highway gas stations