French Retail Gasoline Data Treatment

Etienne Chamayou*

CREST and Department of Economics, Ecole Polytechnique

October 15, 2014

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

1 Data sources

1.1 Price comparison website prix-carburant.gouv.fr

The main data source is the website prix-carburant.gouv.fr, created by the French Ministry for the Economy in 2006 to increase price transparency on the French retail gasoline market. It is mandatory for gas stations which have sold above $500m^3$ of gasoline the previous year to keep their prices posted on the website. Price records and gas station brands were collected from the site on a daily basis. Regarding gas station characteristics, data include the gas station address, gps coordinates and amenities.

1.2 Price comparison website zagaz.com

Another important data source is the website zagaz.com which offers the same service as prix-carburant.gouv.fr but with a "crowdsourcing" philosophy i.e. thanks to user provided data. This website is clearly not fit for a global analysis of French gas prices as some areas hardly have any active users. It is more comprehensive than the government website regarding the population of gas stations (no limit downwards), but the latter is not completely reliable as gas station which are not regularly followed by users can remain on the website. An interesting feature of the website is that each user has a public profile showing all the prices updated by the user. This offers the opportunity to build a proxy of the choice set of users with sufficient activity (though choice itself is not observed). Finally, gps coordinates used by the websites, as they are provided by users, are of better quality than those provided on prix-carburant.gouv.fr.

1.3 INSEE data

Gas station addresses have been used to match each gas station with the insee code of its municipality, which opens access to INSEE data in general. Census variables taken into account in the analysis include mainly the demography, revenue and vehicle equipment of the population. Other data of interest provided by INSEE include flows of commuters between municipalities and regroupment of municipalities in wider areas based on employment and urbanism.

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.

2 Data processing

2.1 Prices

Price data collected from 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. A frequent source of error was found to be that diesel price was updated in place of 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 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 relevaed 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.

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 if 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 changed 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 issues 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