French Retail Gasoline Data Treatment

Etienne Chamayou[*]

CREST and Department of Economics, Ecole Polytechnique

May 22, 2014

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

1 Data processing

Price and station information: Duplicate detection. Not trivial: typically occurs when brand changes... new record with new address/brand... use price information Investigation of brand changes more detailed in descriptive statistics.

Location: same corner gas station... crucial for consumer information / advantage in location Duplicates: picture of competition, potentially interesting if stations closed can be found in data. Brand changes: Connected to duplicates... important policy implications Highway gas stations: different market

The database which lies behind the website is first essentially recreated, using the same gas station identification numbers as it is contained in webpages (zip code + 3 digits which appear to merely reflect the registration order of gas stations with the same zip code). Two observations follow: there appears to be a significant turnover in the data, some prices appear to reflect mistakes by managers (extraordinary variations in price).

Since there are very few gas stations which are newly opened each year, it must be that gas stations either stop providing information because they are not/no more committed to do so or create a new account and are then registered under a different identification number... hence the need to reconcile them.

Treatment of prices:

– 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

- Abnormal values are then looked for based on prior observation. This allows to observe that some prices are input as 0.156 instead of 1.56 and correct rather than transform in missing values. TODO: stats
- Abnormal price variations are then looked for based on observation. Quite often it appears that the variation must have been the consequence of a mistake given the size and the following reverse change. Such changes are eliminated. When no converse changes can be found due to missing data, suspect periods are also switched to missing.
- Abnormal price durations are finally examined based on prior observation. As mentioned in XXX, a duration of 1 month is considered suspect, 1 month and 1/2 highly suspect, 2 months almost certainly an error.

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