Hotel services consumer price index :

Web scraping of a booking platform: exploring new data and methodology

Adrien Montbroussous
with the work of Camille Freppel and Ombéline Guillon

Institut National de la Statistique et des Études Économiques
French National Institute of Statistics and Economic Studies

Thursday 9th June, 2022

Plan

1. **Context and Web scraping**
2. **Firsts results and reflection elements**
3. **Homogeneous classes index**
4. **What's next ?**

Plan

1. **Context and Web scraping**

2. Firsts results and reflection elements

3. Homogeneous classes index

4. What's next ?

Context of the experiment

This project is the subject of a European grant and was started by Camille Freppel (formerly in the services sector) in collaboration with Ombéline Guillon from the methodology section of the consumer prices division, whom I replaced since September, 1st, 2021.

Context : Hotel services index

- Room rent represents 0,8 % of the consumption in the French CPI basket in 2021.
- Prices are collected on the field by collectors in agglomerations with more than 2000 inhabitants selected in the CPI sampling plan.
- Prices are collected from Monday to Friday, once per month (Monday of the first week of January => Monday of the first week of February) for the night of the collection (if the hotel is full, the price is estimated by the owner). The CPI months are composed of 4 weeks and do not always match calendar months.
- To assure constant quality, the product surveyed is a night **for 2 persons with breakfast included**

Context : yield management, quality improvement leads

In the current Methodology for the hotel service consumer price index

- Booking in advance is not taken into account.
- Online consumption is not represented.
- Some tourist areas are not well represented
- Prices aren't collected for Saturday and Sunday nights.

Yield management and volume of data

Some prices are volatile and depend of the anteriority of booking.

Studying yield management's pricing strategies requires a consequent volume of data.

How to get this kind of data?

- Scanner data
- Application Programming Interface (API)
- web scraping

Web scraping : define a collection protocol

Daily web scraping of the platform using Python. Code automatically launched since august thanks to our SSPCloud platform (docker + gitlab.ci).

- Requests sent for 0, 30 and 60 days anteriorities.
- Brutal web scraping : we are parsing the HTML page.
- Data collection for whole of France with breakfast included and free cancellation filters.

Web scraping : limits

Dependence to the platform

- Some interruption of the data collection due to changes of the platform (one week interruption in October 2021 for instance).
- Maintenance of the cleaning codes is required because the platform is evolving : disappearance or format change of some variables.

Plan

1. Context and Web scraping

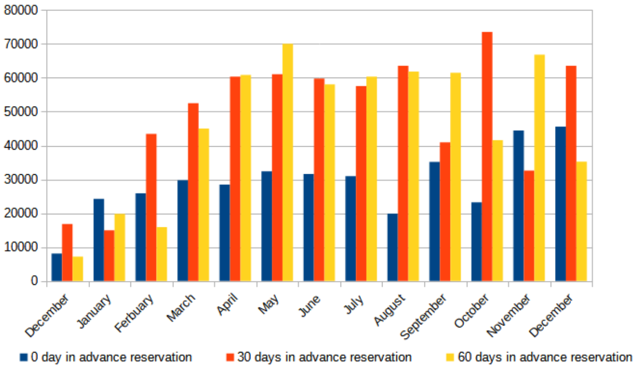
2. **Firsts results and reflection elements**

3. Homogeneous classes index

4. What's next ?

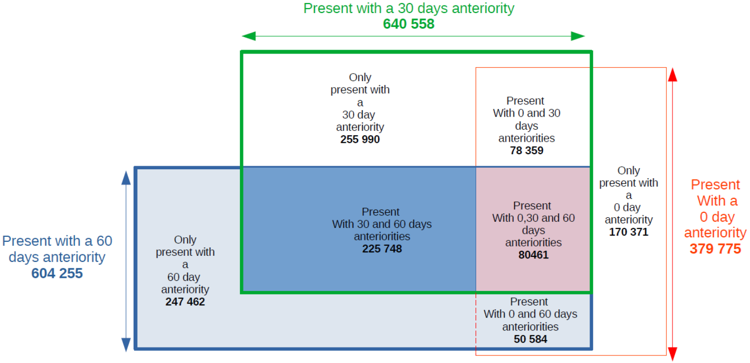
Data collected

Figure: Distribution of the observations according to the calendar month and booking anteriority



Note : Data in December 2020, for a 60 and 30 booking in January 2021 and 60 days in advance aren't with a full coverage geographic scope. A collect interruption that lasted one week in October explains the diminutions at 0 days in October 2021, 30 days in November 2021 and 60 days in December 2021

An offer policy hard to apprehend



Present with a 30 days anteriority
640 558

Present with a 60 days anteriority
604 255

Present With a 0 day anteriority
379 775

Only present with a 30 day anteriority
255 990

Present With 0 and 30 days anteriorities
78 359

Only present with a 60 day anteriority
247 462

Present With 30 and 60 days anteriorities
225 748

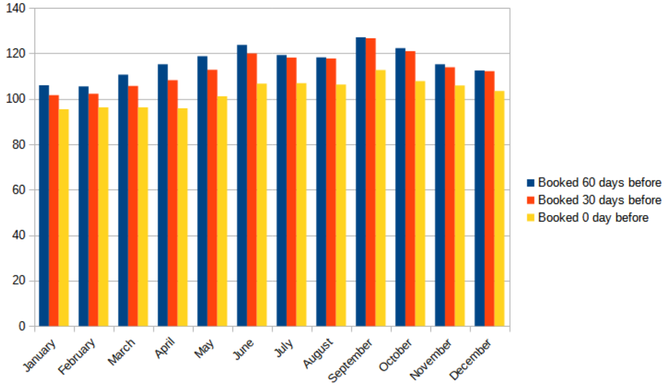
Present With 0,30 and 60 days anteriorities
80461

Only present with a 0 day anteriority
170 371

Present With 0 and 60 days anteriorities
50 584

Price according to the booking anteriority

Figure: Evolution of the average price according to the booking anteriority

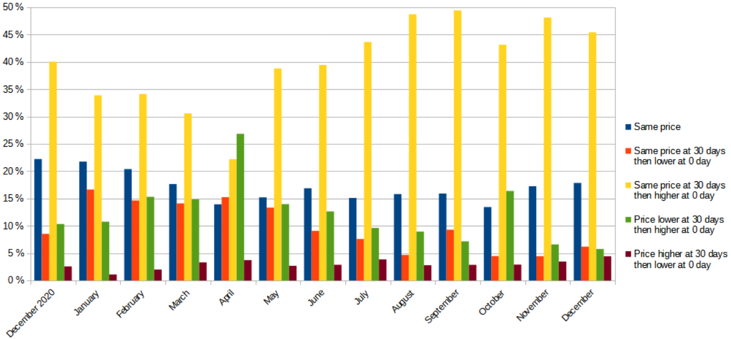


Source : Base with filter from web scraping, as of December, 31st 2021. Field : Whole of France.

Note : average prices are calculated with a geometric mean.

Pricing Profiles

Figure: Evolution of the pricing profiles distribution between December 2020 and December 2021

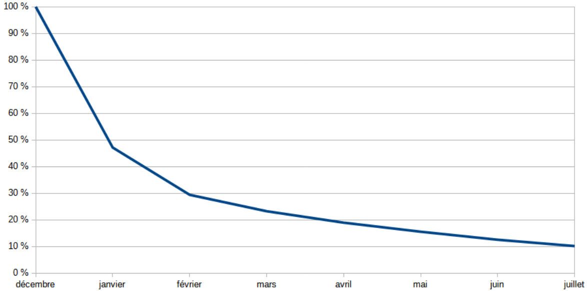


Source : Base with filter from web scraping, as of December, 31st 2021. Field : Whole of France.

Hypothesis : a diminution carried by new products

Fixed basket index ?

Figure: Evolution of the attendance rate of hotels x rooms as part of a fixed-basket approach



Source : Base with filters from web scraping, as of 30 July 2021. Field : Whole of France.

- product : hotel x room at one particular week day of one week selected and followed for each month (ex Monday of first week)
- potentially a lot of replacements / imputations

Plan

1. Context and Web scraping

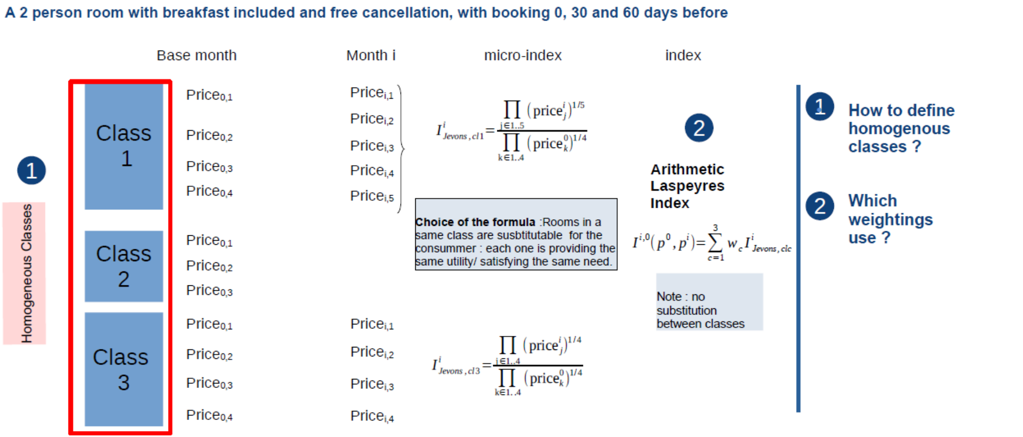
2. Firsts results and reflection elements

3. **Homogeneous classes index**

4. What's next ?

Principle of the method

scope : room for 2 persons with breakfast included and free cancellation, with booking at 0, 30 and 60 days in advance



A 2 person room with breakfast included and free cancellation, with booking 0, 30 and 60 days before

Base month

Month i

micro-index

index

1

Homogeneous Classes

Class 1

Price$_{0,1}$
Price$_{0,2}$
Price$_{0,3}$
Price$_{0,4}$

Price$_{i,1}$
Price$_{i,2}$
Price$_{i,3}$
Price$_{i,4}$
Price$_{i,5}$

$$I^i_{Jevons,cl1}=\frac{\prod_{j\in 1..5}(\text{price}^i_j)^{1/5}}{\prod_{k\in 1..4}(\text{price}^0_k)^{1/4}}$$

Class 2

Price$_{0,1}$
Price$_{0,2}$
Price$_{0,3}$

Choice of the formula :Rooms in a same class are susbtitutable for the consummer : each one is providing the same utility/ satisfying the same need.

Class 3

Price$_{0,1}$
Price$_{0,2}$
Price$_{0,3}$
Price$_{0,4}$

Price$_{i,1}$
Price$_{i,2}$
Price$_{i,3}$
Price$_{i,4}$

$$I^i_{Jevons,cl3}=\frac{\prod_{j\in 1..4}(\text{price}^i_j)^{1/4}}{\prod_{k\in 1..4}(\text{price}^0_k)^{1/4}}$$

2

Arithmetic Laspeyres Index

$$I^{i,0}(p^0,p^i)=\sum_{c=1}^{3} w_c I^i_{Jevons,clc}$$

Note : no substitution between classes

1 **How to define homogenous classes ?**

2 **Which weightings use ?**

Analysis of the price determining characteristics

- Data cleaning and imputation
  - consequent step for web scraped data
  - City name cleaning in order to match with a geographic referential
  - Variable creation by text analysis, room type for instance (superior, classic), hotel chain, etc.
- Data analysis
  - Use of **regression trees and hedonic models** to find price determining characteristics : number of stars, independence of the hotel, day, month, type of room, geographical information...
  - Some results are unexpected : prices are lower during scholar holidays and weekends all other things equaled.

Classes retained

1 Géographical criterion

- Cross of the french regions (exception of IDF) and tourist area, for the IDF we are using the city status (urban , ...)
- Limit : unlikely, in short term it will be better to use the crossing of french department and tourist area
- ***Metropolitan France only***

2 Type of hotel

- Number of stars
- Chain / independent

3 Type of room

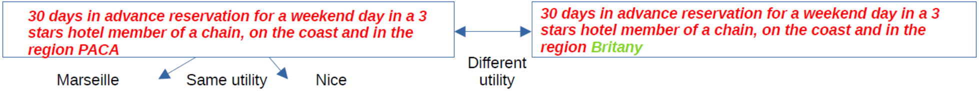
- Room comfort

4 Anteriority

- Booking the room 60 days before brings more uncertainty and constraints than at last minute
- Consumer using one of these 3 anteriorities have different profiles

5 Period

- Week-end vs weekday



30 days in advance reservation for a weekend day in a 3 stars hotel member of a chain, on the coast and in the region PACA

Marseille

Same utility

Nice

Different utility

30 days in advance reservation for a weekend day in a 3 stars hotel member of a chain, on the coast and in the region Britany

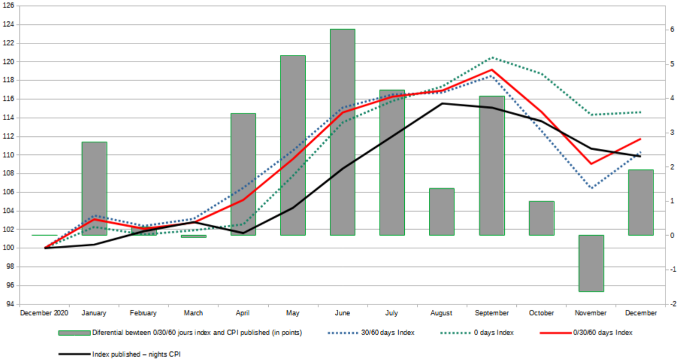
Weightings choice : 3 possibilities

To weight micro indexes, three data set exported from the tourist establishments register were conceivable :

- Weightings set n°1 – 2019 data (ex : 32 % of the room booked were in the Parisian region (IDF))
- Weightings set n°2 – 2019 data, only with the room booked for personal reasons (ex : 31 % of the room booked were in the Parisian region (IDF)) - **Kept for the experiment after a comparison of results**
- Weightings set n°3 – 2020 data (ex : 22 % of the room booked were in the Parisian region (IDF))

Indexes with webscraped data vs indexes published

Figure: Comparison of price indexes for hotel nights between December 2020 and December 2021



Source : Database with filters from web scraping, as of December 31. Field : Metropolitan France.

Note : The calendar here is for the calendar month (except for the CPI), the weights used here are for the year 2019 for personal room consumption (except for the CPI, 2019 weights corrected by specific treatment taking into account the impact of the health crisis). The differential is calculated as the difference between the 0/30/60 day index – the CPI index

Plan

1. Context and Web scraping
2. Firsts results and reflection elements
3. Homogeneous classes index
4. **What's next ?**

Conclusions and limits

- For the following of the study
  - Analyse data with 10 and 20 days anteriorities (collected since December 2021).
  - Study with a more robust basis month (there was a lockdown in December 2020 in France.)
  - Look if there is a substitution between anteriorities.
  - Run a new year of computation.

Thanks for your attention.

Comments and questions welcomed
adrien.montbroussous@insee.fr

Insee Measuring, understanding