Web scraping of a booking platform: exploring new data and methodology for the hotel service consumer price index.

Adrien Montbroussous
with the work of Camille Freppel and Ombéline Guillon
Institut National de la Statistique et des Études Économiques (INSEE)
French National Institute of Statistics and Economic Studies

Abstract

To improve the quality of the hotel price index, a new method to collect data, web scraping, could be helpful. This paper presents an experiment using solely one website. One of the main goals is to improve the coverage of the index, especially in tourist areas such as ski resorts and coasts. Other goals are enhancing the sample size and seizing consumer behavior better. It will also be interesting to take into account nights booked in advance. Currently, price collectors go to the hotels and ask for the cost of a two-person room with breakfast for the night of the collection. After describing the online data collection and its challenges, the study focuses on an index computation using homogenous classes. Classes are made homogeneous enough that the rooms inside them are considered substitutable. At the lowest level, prices are aggregated using the Jevon index formula in each class. These micro-indexes are then aggregated with the Laspeyres index formula at the higher levels.

Keywords : Indices, web scraping, tourism, homogenous classes

Paper for the 17th International Conference of the Ottawa Group
Rome, June 2022

Introduction

More than 50 % of consumer spending is services-related. However, there are several kinds of services and ways of booking or buying. The development of the internet and booking platforms has accelerated this trend. It's never been easier than today to book a flight or a train ticket or even a hotel room. The consumer price index has succeeded in taking into account the new consumptions habits by adapting its methodology. Indeed, the share of products surveyed online is increasing and web scraping robots have been developed to collect prices in the transport sector (train, airplane, Maritim). Moreover, real-time tariffing optimization (yield management) has been generalized in several sectors (like tourism, transportation, and Hostellerie), which leads to challenging the current index to better catch the price evolution of the prices of these moving price services.

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 Adrien Montbroussous replaced since September 1, 2021.

Contents

1 Current Price Index and Study Issues

1.1 Current Price Index and Study Issues The CPI measures the inflation using a 3 level sampling plan : the product variety, the agglomeration, and the type of outlets

The CPI is a tool measuring the evolution of average prices of goods and services consumed by households on the french territory. It is a synthesis measure of a fixed basket of products pure price changes (with constant quality). The CPI is an annually chained Laspeyres index[1]. It synthesizes nearly 30,000 elementary micro-indexes – an elementary index generally representing the crossing of a variety and an agglomeration. At this level, the formula for calculating a micro-index is an unweighted geometric mean of price relatives for heterogeneous varieties[2] (Jevons index) and an average price ratio for homogeneous varieties (Dutot index). The weights used for the aggregations at the higher level represent the share of the expenditure associated with the item concerned within the total household consumption expenditure covered by the CPI. To remain representative of household consumption, the weightings are updated each year and are obtained, in particular, from the annual assessments of household consumption expenditure measured by the National Accounts Department. The CPI sampling plan is characterized by the following 3 levels:

- geographical criterion: readings are taken in 99 urban areas with more than 2,000 inhabitants in mainland France and in 4 overseas departments;
- variety: a sample of more than 1,100 families of products and services, called "varieties" is defined to represent the heterogeneity of products within the 303 product groups. Variety is the elementary base level for tracking goods and services and calculating the index. The list of varieties remains today confidential and only a few average prices of a sample of products and services are published at this level;
- type of outlet: a sample of around 30,000 outlets, stratified by type, is constructed to represent the diversity of goods and services by brand, retailer, and consumer purchasing method (including the internet). We thus distinguish, for example, hypermarkets from supermarkets, which are also distinguished from markets.

From these three criteria, a sample of nearly 160,000 products is built. Added to this are 80 million products tracked through checkout data and more than 500,000 prices collected on the internet. The sample is updated annually to incorporate changes in consumer behavior and, in particular, to introduce new goods or services. The last base change was is 2015. It was characterized by the selection of new agglomerations, resulting from the recent results of the population census, and by the optimization of the number of statements by crossing variety x agglomeration. If a product disappears during the year, it is replaced by a similar product, and a quality adjustment is made to correct the differences in characteristics between the replaced product and the replacement.

[1] Arithmetic average of elementary indices weighted by the values of the reference period.

[2] A variety grouping products whose price levels are relatively dispersed.

1.2 The current hotel price index and collection limits

1.2.1 The current method of calculating the price index for hotel nights

Hotel nights are tracked under COICOP[3] Function 11 - Restaurants and Hotels, in particular within item 11.2.0.1.1 Room rental which represents 0.8% of consumption in the CPI basket[4]. In Metropolitan France, six different varieties are monitored:

- Overnight at a 5-star hotel with 43 price collections;
- Overnight at a 4-star hotel with 143 price collections;
- Overnight at a 3-star hotel with 206 price collections;
- Overnight at a 2-star hotel with 195 price collections;
- Overnight at a 1-star hotel with 39 price collections;
- Overnight at an unranked hotel with 34 price collections;

These prices are collected in 658 hotels. Prices for Hotel nights are collected in the overseas departments as well. The metropolitan varieties are so-called homogeneous varieties, i.e. hotels of the same comfort (number of stars) are not substitutable for the consumer within an agglomeration. In other words, when the price of a room in a 4-star hotel increases, the customer does not transfer his consumption to another 4-star hotel in the same agglomeration. This can be explained by the fact that the consumer is not aware of the different prices charged in other shops in the area and/or that the room of another 4-star hotel in the same area does not allow him to satisfy the same need (different location for example)

[3] Classification of Individual Consumption by Purpose. International nomenclature that breaks down household consumption by units of need.

[4] These weighting data are reviewed annually and come for year N from semi-final national accounts data for year N-2. For 2021 and 2022, the estimates of the quarterly accounts for the year N-1 have also been used.

Link to consumer theory

In the microeconomic framework of consumer theory, the household seeks to minimize its spendings under the constraint of reaching a certain level of utility, the evolution of the prices of a basket of goods between two periods, therefore corresponds to the evolution of the budget that he must devote to the purchase of a set of goods while maintaining its constant usefulness between the two periods. It is then possible to link the formulas for calculating the micro-indexes of the varieties x agglomerations of the CPI by using the usual utility functions between two instants (0 and 1) and n goods:

- Leontief utility: zero elasticity of substitution between goods – in the case of the CPI, substitution between outlets [a]
  $I^{1,0}(p^0,p^1) = \frac{\frac{1}{n}\sum_{i=1}^{n} p_i^1}{\frac{1}{n}\sum_{i=1}^{n} p_i^0}$
  The Dutot index used in the context of homogeneous varieties testifies to the non-substitutability of outlets within the agglomeration. In other words, the consumer carries out his price arbitrations among different stores.

- Cobb-Douglas utility: unitary elasticity of substitution between goods – in the case of between-outlet CPI.
  $I^{1,0}(p^0,p^1) = \prod(\frac{p_i^1}{p_i^0})^{\frac{1}{\pi}}$
  The Jevons index used in the context of heterogeneous varieties testifies to a substitution behavior between points of sale in the agglomeration characterized by a unitary elasticity. In other words, the consumer carries out his price arbitrations at the scale of the agglomeration.

It should be noted that apart from the elementary consumer markets on which substitutions exist (i.e. on which the consumer capitalize on competition), there are no substitutions.

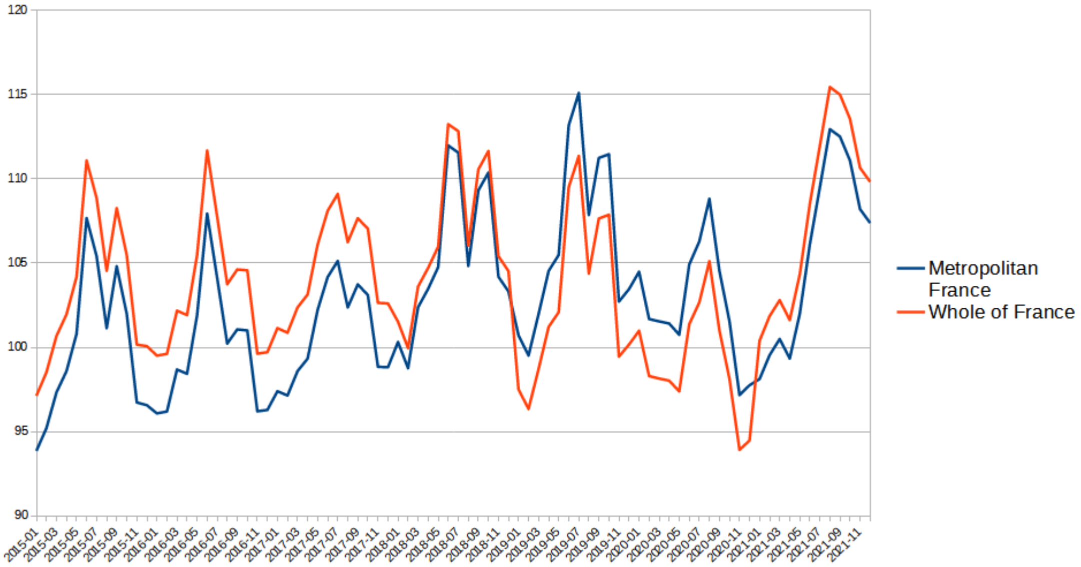
[a] Indeed, in the current configuration of the CPI calculation, a single price for a variety of goods data is collected at a particular outlet.

Prices are collected in the field, in the agglomerations (of more than 2,000 inhabitants) defined within the scope of the CPI. Prices are collected from Monday to Friday. Hotel's rooms prices are collected for the night of the day of the price collector's visit. Only 17 prices concerning 4-star hotels are collected one month in advance. This collection is original due to the application constraints since these 17 price collections are made by the price managers in the office on the internet, the prices are collected in a spreadsheet before being entered the following month at the same time as the prices collected by the price collectors on field. In order to ensure the constant quality of the product, it is defined more precisely as one night for two people, two breakfasts included. In addition, the interviewers fill out a form to ensure that the characteristics of the hotel and the room are the same each month (without imposed quotas):

- location of the hotel (city center versus outskirts);
- type of hotel (independent, franchise chain, voluntary chain);
- room comfort (classic, standard, superior, luxury, privilege, other).

Between December 2015 and December 2021, the prices of hotel nights increased by 11.2% in France. They increased by an average of 1.8% each year (see graph n°1).

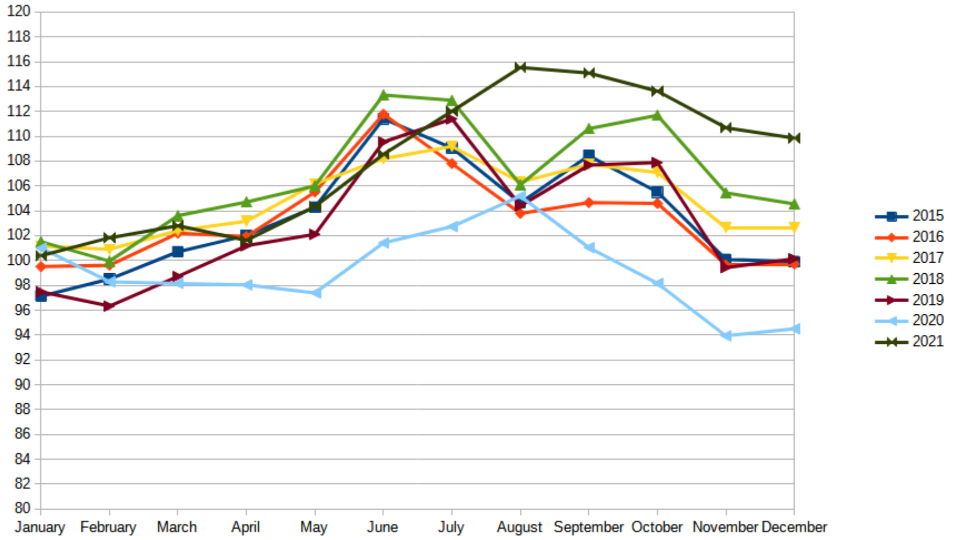
Figure 1: Room rental price index between January 2015 and December 2021



Source : CPI.
Reading note: Base 2015

The seasonality observed between 2015 and 2021 is heterogeneous, some characteristics can nevertheless be identified: price increases from February to June then from August to October and price decreases in August and November (see graph n°2). The profiles for the years 2020 and 2021 are atypical due to the health crisis, in addition to the particular periods linked to successive lockdowns, there is a rise in prices between the months of July and August.

Figure 2: Room rental price index between 2015 and 2021 (base 100= December N-1)



Source: CPI.
Field: Metropolitan France
Reading note: For example, the indices for the year 2021 are in base 100 = December 2020.

1.2.2 The limits of the current method

The current method has several limitations:

- Limit n°1 – the absence of other reservation methods: other channels for reservations of nights sold separately exist[5], such as reservations on the Internet, which are also gaining in importance[6]. These online reservations can be made via the hotel's website or via a platform (for example Booking, Hotels.com, Trivago, Expedia, Last Minute Travel, etc.)[7];

- Limit n°2 – the absence of advance hotel reservations: the current index does not take into account rooms booked in advance by consumers. However, hotels can adjust their prices according to the booking of the rooms, and the prices can vary according to the anteriority of purchase;

- Limit n°3 – the representativeness of the sample (size and geography): the collection of prices in the field is strictly limited to agglomerations of more than 2,000 sampled inhabitants. A collection on the Internet would make it possible to overcome this framework and to have a better representation of the coast and the mountainous areas. Furthermore, this sample of hotels remains limited (4% of hotels in France).

- Limit n°4 – aggregation of changes by agglomeration: the Dutot indices calculated by agglomeration are then aggregated with the weight of household food consumption expenditure. Household consumption expenditure on hotel nights is a priori different from that of food consumption. For example, Paris represents 32% of hotel nights in 2019[8] and 22% of food consumption expenditure in 2019.

- Limit n°5 – the representativeness of weekends and school holidays: the collection of hotel prices in the field is carried out according to the CPI calendar which differs from the calendar months[9]. In concrete terms, the current index does not make it possible to correctly monitor weekend (Saturday and Sunday nights prices) since the price collectors do not collect weekend prices. Furthermore, this calendar does not ensure that all school holidays are taken into account (for example, the prices for the Christmas holidays are not recorded in December, the prices for the month of August are not exhaustive due to one week off for the prices collectors).

- Limit n°6 – the case of full hotels: the collection of prices for the same day has the disadvantage in high season of encountering full hotels: in this case, if the hotel is able to give the price of the room following a possible withdrawal during the day, this price is noted, otherwise the price is charged.

[5] Other modes of marketing are not discussed: (i) booking as part of travel packages where overnight stays are bundled with other travel services as this consumption segment would be classified in another COICOP post; (ii) the reservation within the framework of company arrangements for which preferential rates are negotiated by companies and administrations does not concern the scope of the CPI.

[6] According to a study by Phocuswright[1], the share of online reservations in the turnover of French hoteliers rose from 26% in 2011 to 34% in 2015.

[7] In Europe, 70% of hotel bookings made online come from platforms, with the remaining 30% being bookings made directly on hotel websites [2].

[8] 31 % of hotel nights excluding business customers.

[9] Each month, the CPI is based on 20 field data collection days spread over the working days of 4 consecutive weeks. A calendar month consists of 28 to 31 days and does not correspond to a whole number of weeks. Consequently, each year, INSEE adapts the field data collection calendar (CPI calendar) so that the 48 weeks of data collection coincide as closely as possible with the calendar months. This adaptation consists in fixing weeks without collection, on average 4 per year.

1.3 Yield management, used in the hotel sector, is the subject of international recommendations and experiments using webscraping

1.3.1 International recommendations on yield management

New forms of dynamic pricing (yield management, revenue management or even real-time pricing) appeared in the 1980s in the case of airlines in the United States following the deregulation of the market. This method, consisting of optimizing the prices of a service in real time according to demand in a market segment, has nowadays become a marketing technique commonly used in air or rail transport, car rentals, entertainment and in the hotel industry for example. The positive consequences of using this concept are both on the producer's side with an increase in turnover, and on the consumer's side with the possibility of lower prices without impacting the quality of service. Such a price adjustment is an essential tactical asset for companies operating in a highly competitive environment and for which price is the primary variable of choice for the user. For these services, very frequent price changes may occur, depending on the date of the service. Prices may also vary depending on the date of purchase. Thus, for the same service consumed at a given time, the price may be different, which complicates the calculation of the consumer price index, if it is intended to be representative of all these prices. The IMF's consumer price index manual[3] and the draft Eurostat recommendation [4] relating mainly to air transport and holiday stays[10], but whose preliminary version related to all services at volatile prices or whose price depends on prior booking[11]

recommend:

- to collect prices reflecting price volatility by distributing price collections over the month;
- to constitute a sample of prices representative of the behavior of the buyers, by integrating in particular the anteriority of the purchase, the various types of classes and conditions of reimbursement of the ticket;
- to include in the CPI the price collected previously for the month during which the consumption of the service begins and not when the service is reserved.

1.3.2 Experiments using webscraping

For several years, the collection of prices on the Internet has been developed in national statistical institutes in order to describe the consumption of goods and services on the Internet. In addition, several countries, in particular Belgium, the Netherlands, Italy, Germany or the United Kingdom, have automated these collections using computer webscraping programs, thus reducing collection costs compared to physical readings while by obtaining a greater volume of data. This webscraping method is particularly effective in the context of:

- pricing that does not depend on where the service is purchased;
- a limited number of websites offering these services, which limits the number of robots to develop;
- goods or services whose order is made online via a form. The information collected this way is structured and the characteristics of the product collected are well defined.

[10] These recommendations can be found in chapter 12.5 Flights and package holidays of the final version of the Eurostat manual of November 2018.[5]

[11] "Recommendation on the treatment of tangible services purchased in advance and/or priced flexibly", March 2018.

Faced with the growing importance of this technique, Eurostat adopted in November 2020 a manual of good practices on webscraping [6]. Nevertheless, the advantages of this technique are to be considered with regard to certain limits:

- the importance of communication from statistical institutes to the reservation platforms to avoid blocking of IP addresses[12];
- the cost of collection can vary greatly depending on the frequency of changes of the website which can lead to regular maintenance of the webscraping robot. These changes may be easily detectable in case of an interuption of the collection[13], others issues require regular monitoring of the data to spot a missing variable despite full data collection[14]

  .

In France, webscraping is already used by the Direction Générale de l'Aviation Civile (DGAC) to calculate the air transport price index and by the consumer price division to calculate the rail transport price index since 2020. In the hotel sector, Belgium has been using this technique since 2015. Italy and the Netherlands have also carried out work on this subject more recently.

[12] An information letter concerning the investigation related to the collection of prices within the framework of the CPI and the webscraping of the website has been sent to the online booking platform

[13] A change to the booking platform site in November 2020 prevented the collection of prices for a week. Another significant change in October 2021 caused collection to stop for a week.

[14] A change to the site since the end of March 2021 has prevented the collection of data on breakfasts. A change to the site in October 2021 prevented the continued collection of a confirmation variable on the number of room occupants as well as a variable to uniquely identify each hotel. For the identifier of the hotels, another could be reconstituted but the same hotel sometimes has several different ones.

1.4 The choice of platform webscraping

The various limits, recommendations and European experiments have led to considering the calculation of a price index from the webscraping of a large amount of data from a booking platform. The collection of these prices will make possible to take into account a different reservation channel, a reservation in advance of the consumer, to overcome the agglomerations of the CPI and the collection calendar. As part of this exploratory work, we chose as a platform booking.com which is the first website visited in the field of travel and tourism with 13.7 million monthly visits and 6.5 million unique visits[15]. In addition, several European countries, including Italy, Belgium and the Netherlands, have more or less recently started work on the same platform. Finally, in order to measure a potential coverage bias on the platform, a matching of 2 and 3 star hotels monitored as part of the metropolitan CPI and web-scraped establishments on the platform was carried out: only 5% of the hotels monitored in the CPI framework are not found on the platform. This proxy shows good visibility of hotels on the platform even if the effect of the health context is not measurable (more hotels seeking to be visible on the platform for example). The price index resulting from this choice is not intended to be representative of all the different booking methods. An analysis on other platforms can be carried out later (Belgium has started to experiment with webscraping of the Expedia platform after having worked on the Booking platform). In addition, the platform's pricing policies may differ depending on whether or not the hotel has joined the Preferred Partner Program. Indeed, the work of M. Cure, A. Cazaubiel, B. Johansen and T. Vergé [7] show that joining this type of program leads to an increase in sales and prices[16].

Limit of interpretations of data from webscraping in 2020 and 2021

The years 2020 and 2021 were marked by various restrictions (national lockdowns from March 17 to May 11, 2020, from October 30 to December 15, 2020, from April 3 to May 3, 2021 in mainland France, local lockdowns, administrative closures, travel restrictions, curfew) following the health crisis. The collection rate (share of normal or pseudo-normal observations[a]) of prices in the hotel sector shows a significant impact of the health crisis on price collection over the period from March 2020 to June 2021, the pre-crisis level is still not reached again.

Table 1: Rate of collection of hotel room prices in 2019, 2020, 2021

	2019	2020	2021
January	89.4%	86.6%	56.7%
February	91.4%	93.5 %	59.4 %
March	94.0%	70.7 %	62.3 %
April	91.4%	0.6 %	56.7 %
May	88.4%	19.8 %	65.0 %
June	92.0%	45.3%	75.2%
July	90.0 %	72.5 %	83.0 %
August	89.1%	73.1 %	78.0 %
September	92.0 %	83.8 %	81.6 %
October	91.2%	85.9%	81.3%
November	88.4%	56.6 %	79.6 %
December	87.6%	52.7%	79.2%

Source: CPI field records.

The price changes obtained using the webscraping of the online booking platform should therefore be interpreted with caution.

[a] The product is not on the shelves but its price is still displayed at the usual location of its sale. In the case of the hotel industry, the pseudo-observation code was used during the health crisis for statements by telephone or internet (only on the hotel's website).

[15] Analysis of internet and webmobile traffic of the similarweb.com website

[16] Apart from the level of prices, their evolution could, a priori, also differ.

2 Building a database from webscraping

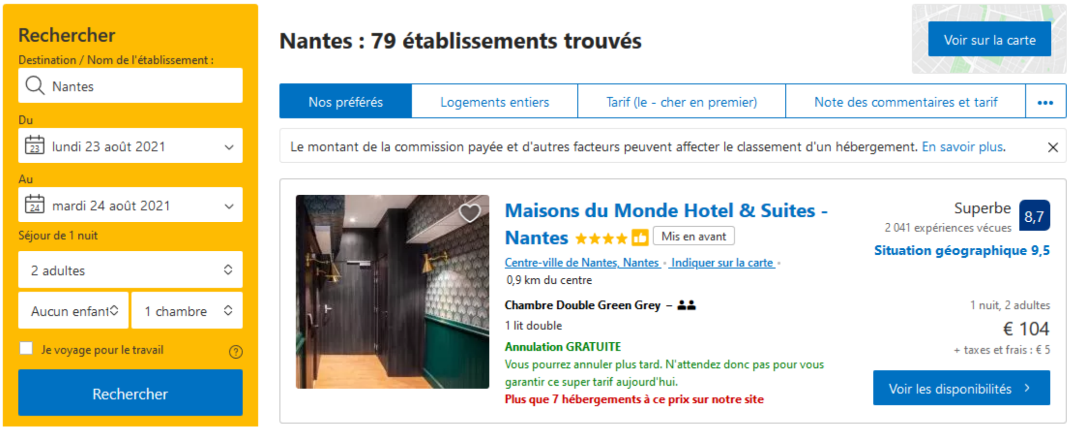
The collection of data by webscraping began by being set up using two approaches: collection over the whole of France restricted by filters on the options and collection without filters over a restricted geographical area.

2.1 Definition of a collection protocol

The choice was made to develop the platform's webscraping robot using the Python language because it is a good compromise between speed of execution and accessibility to statisticians who are not computer experts.

The program collects data for a given destination for a stay in a hotel (specific filter on the platform) of two adults for one night at 0, 30 and 60 days prior, as well as 10 and 20 days since December 2021. He sends a request to the URL of the online booking platform with the destination, the date and the conditions chosen for the room and he receives in return a file in HTML format describing a list of hotels (see illustration n° 3) from which the relevant information must be extracted.

Figure 3: Example of the result of a query to obtain the hotels available on the platform site



Rechercher

Destination / Nom de l'établissement : Nantes

Du: lundi 23 août 2021

Au: mardi 24 août 2021

Séjour de 1 nuit

2 adultes

Aucun enfant

1 chambre

□ Je voyage pour le travail

Nantes : 79 établissements trouvés

Le montant de la commission payée et d'autres facteurs peuvent affecter le classement d'un hébergement. En savoir plus.

Maisons du Monde Hotel & Suites - Nantes ★★★★ Mis en avant

Centre-ville de Nantes, Nantes • Indiquer sur la carte •

0,9 km du centre

Chambre Double Green Grey –

1 lit double

Annulation GRATUITE

Vous pourrez annuler plus tard. N'attendez donc pas pour vous garantir ce super tarif aujourd'hui.

Plus que 7 hébergements à ce prix sur notre site

Superbe 8,7

2 041 expériences vécues

Situation géographique 9,5

1 nuit, 2 adultes

€ 104

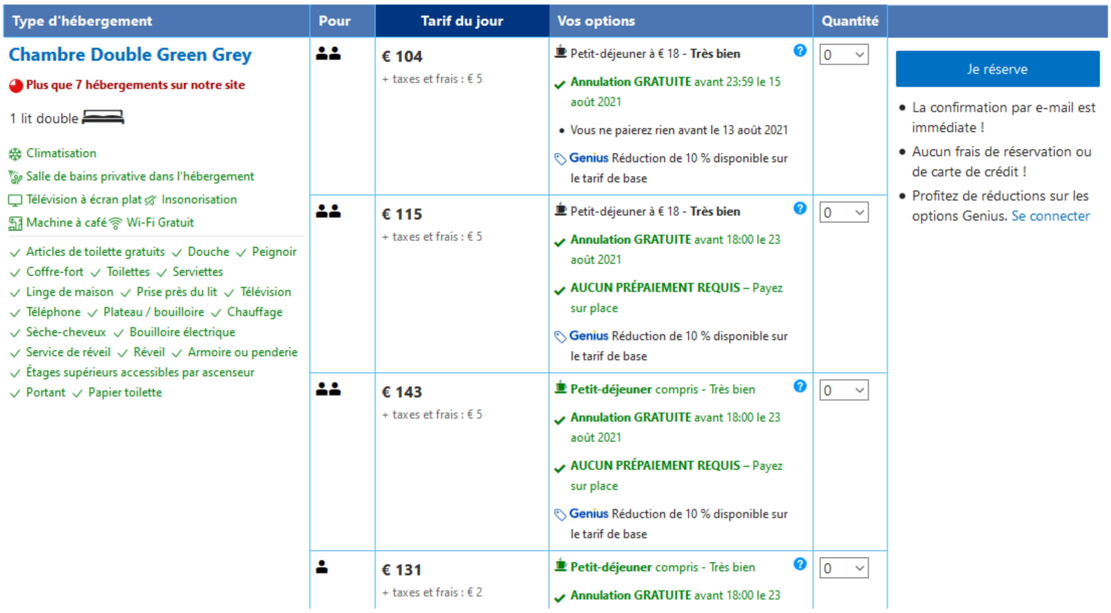
+ taxes et frais : € 5

Rechercher
Voir sur la carte
Nos préférés
Logements entiers
Tarif (le - cher en premier)
Note des commentaires et tarif
•••
Voir les disponibilités ›

Note: This query displays all the hotels available on the night of August 23 in Nantes for 2 adults. For each hotel, only one room is highlighted by the platform. This is the one used in approach n°1.

The results obtained on the first page with the list of hotels for a destination contain in particular the link to access the specific page of each of these hotels (see illustration n°4). On the hotel page, it is possible to display all the rooms offered, with their different conditions.

Figure 4: Example of the result of a query to obtain all the rooms available on the platform



Type d'hébergement	Pour	Tarif du jour	Vos options	Quantité	
Chambre Double Green Grey Plus que 7 hébergements sur notre site 1 lit double Climatisation Salle de bains privative dans l'hébergement Télévision à écran plat Insonorisation Machine à café Wi-Fi Gratuit ✓ Articles de toilette gratuits ✓ Douche ✓ Peignoir ✓ Coffre-fort ✓ Toilettes ✓ Serviettes ✓ Linge de maison ✓ Prise près du lit ✓ Télévision ✓ Téléphone ✓ Plateau / bouilloire ✓ Chauffage ✓ Sèche-cheveux ✓ Bouilloire électrique ✓ Service de réveil ✓ Réveil ✓ Armoire ou penderie ✓ Étages supérieurs accessibles par ascenseur ✓ Portant ✓ Papier toilette	👤👤	**€ 104** + taxes et frais : € 5	Petit-déjeuner à € 18 - **Très bien** ✓ **Annulation GRATUITE** avant 23:59 le 15 août 2021 • Vous ne paierez rien avant le 13 août 2021 **Genius** Réduction de 10 % disponible sur le tarif de base	0	Je réserve • La confirmation par e-mail est immédiate ! • Aucun frais de réservation ou de carte de crédit ! • Profitez de réductions sur les options Genius. Se connecter
	👤👤	**€ 115** + taxes et frais : € 5	Petit-déjeuner à € 18 - **Très bien** ✓ **Annulation GRATUITE** avant 18:00 le 23 août 2021 ✓ **AUCUN PRÉPAIEMENT REQUIS** – Payez sur place **Genius** Réduction de 10 % disponible sur le tarif de base	0	
	👤👤	**€ 143** + taxes et frais : € 5	**Petit-déjeuner** compris - Très bien ✓ **Annulation GRATUITE** avant 18:00 le 23 août 2021 ✓ **AUCUN PRÉPAIEMENT REQUIS** – Payez sur place **Genius** Réduction de 10 % disponible sur le tarif de base	0	
	👤	**€ 131** + taxes et frais : € 2	**Petit-déjeuner** compris - Très bien ✓ **Annulation GRATUITE** avant 18:00 le 23	0	

Note: This query displays all available rooms for the Maisons du monde Hotel & Suites – Nantes for the night of August 23.

In order to minimize the number of requests on the platform, two approaches have been considered:

- Approach n°1 – "collection of hotels with filters": the program collects hotels throughout France (metropolitan France and overseas departments) with the use of filters (breakfast included, free cancellation). This means that all the rooms in a hotel are not collected, but only the one highlighted by the online booking platform (we only retrieve the results from the first page). It can be assumed that the room thus pre-selected is the most sold in general. This collection began from October 19, 2020 for a large number of regions, then for the whole of France from December 22, 2020. As of August 5, 2021, 1,038 requests have been sent to the platform. online booking.

- Approach n°2 - "collection of hotels without filter with room": the program collects all the rooms offered in 4 areas (Morbihan, Savoie, Avignon and Paris 5th). This approach requires a collection of all the hotels in these areas upstream to relaunch a request on the page of each of these hotels. Concretely, for these four zones, this can lead to up to 25 times more requests sent to the online booking platform (a results page containing a maximum of 25 hotels): hence the need to limit this collection geographically. Approach n°2 is supplemented by a collection of all the hotels on the front page throughout the territory, which makes it possible to constitute a repository of hotels. This collection began on October 13, 2020. As of August 2, 2021, 1,981 requests were sent to the online booking platform for the first stage nationwide, then 1,237 requests sent to the website for the second step of searching for all available rooms in 4 areas.

The robot was initially launched manually only on weekdays in October 2020 and then launched daily using the tools made available by the Innovation and Technical Instruction Division (DIIT) of INSEE (Onyxia platform). From November 2020 a "Docker image" of this robot, containing all the necessary libraries and the program input destination file is created. This image is executed each night in order to collect the prices of hotel nights following the two alternating processes (each approach is collected every other day) and the three anticipations. The Innovation platform has been closed and its tools migrated to the SSPCloud platform[17]. There was a transition period during this migration with manual collection between May 27 and August 30. The robot was migrated to the SSPCloud platform with the help of the DIIT, which we thank very much and is now open source.

The results obtained on the first page with the list of hotels for a destination contain in particular the link to access the specific page for each of these hotels (see illustration n°2). On the hotel page, it is possible to display all the rooms offered, with their different conditions. The data extracted from these two approaches are presented in the following table (cf. table n°2).

Table 2: Main variables collected using the robot developed in Python

Approach n°1 - Base hotels with filters	Approach n°2 - Base hotels with rooms
Page Rank	
Hotel name	Hotel name
	Hotel ID
Star Rating	Star Rating
Customer Rating	
Cancellation method (here filter on free cancellation)	Cancellation methods
Prepayment required	Prepayment required
Breakfast (and other meals included or not – here, breakfast is necessarily included)	
Room name	Room name
	Room identifier
Room Capacity	Room Capacity
Room price	Room price
	Breakfast prices
Price of taxes (tourist tax and booking fees)	

[17] hiips://datalab.sspcloud.fr

2.2 Comparisons of the scope of web-scraped hotels according to the two approaches, the tourist accommodation occupancy survey and the CPI hotel sample

The automated collection of hotels' rooms prices throughout the territory according to the two approaches makes it possible to constitute two repositories of establishments. In order to study possible biases in coverage, these two benchmarks can be compared with that compiled thanks to the monthly survey of frequentation in collective tourist accommodation conducted by INSEE. The update of the number of establishments is carried out by the survey managers continuously during the survey and occasionally based on other sources, in particular Sirene, Atout France, and the regional tourism committees. The tourism division of the Montpellier establishment transmitted an extraction of this reference system as part of this work on March 22, 2021, which includes 17,909 establishments.

Before being able to compare these four data sources, a cleaning common to the two databases resulting from the webscraping was carried out. First of all, a restriction of the field of observations is made in order to be limited to hotels[18] (the filters present on the platform not being sufficient) and to remove pensions or half-pensions (see table n°3).

Table 3: Track of the two webscraped database cleaning operations

	Approach n°1 - Base hotels with filters	Approach n°2 - Base hotels with rooms
Deletion of lines relating to establishments that are not hotels	84 725	347 169
Deletion of lines relating to pensions or half-pensions	37,379	74,067
Final total number of observations	1,100,676	2,719,009
Final total number of establishments	6,226	15,777

Source: Databases from the webscraping of the platform, as of July 30, 2021.
Field: Whole of France.

In addition, a common variable is also created in order to identify the mode of operation of the hotel (chain or independent) from the textual analysis of the hotels' labels in order to isolate the chains. This variable may therefore tend to underestimate chains if a string is misidentified. Due to the construction method of this variable, it requires monitoring to identify cases of the appearance of new franchises and thus update this variable.

The regional comparison of establishments shows that there is little difference between the tourism benchmark and all the hotels collected without using a filter (maximum 1.5 point difference for Auvergne-Rhône Alpes) while significant differences are present between the tourism benchmark, the CPI sample and all the hotels collected with the use of filters. Indeed, the use of the breakfast included and free cancellation filters leads to over-representation of Île-de-France (25.8% of establishments against 14.8% in the tourism benchmark, see table no. 4)

[18] Text analysis on room names and hotel names on the words: studio, loft, dormitory, caravan, house, duplex, bungalow, apartment, hostels and lodges.

Table 4: Comparison of the proportion of distinct establishments by French region for webscraped databases with filters, without filters and the tourism repository

	Tourist attendance survey	CPI	Base with filters	Base without filter
Ile-de-France	15 %	21 %	26,1 %	15,6 %
Centre-Val de Loire	3,9 %	2,4 %	3,5 %	3,8 %
Bourgogne-Franche-Comté	4,9 %	3 %	3,4 %	4,6 %
Normandie	4,7 %	4,3 %	4,3 %	5 %
Hauts-de-France	4 %	5,6 %	5 %	4,3 %
Grand Est	7,5 %	6,7 %	7,3 %	7,1 %
Pays de la Loire	4,1 %	4,9 %	3,9 %	4,2 %
Bretagne	5,1 %	4,9 %	3,9 %	5,1 %
Nouvelle-Aquitaine	10,2 %	8,5 %	6,9 %	10,4 %
Occitanie	10,7 %	7,8 %	7,9 %	10,8 %
Auvergne-Rhône-Alpes	15,5 %	15,8 %	12,8 %	14 %
Provence-Alpes-Côte d'Azur	11,9 %	15,2 %	13 %	12,3 %
Corse	2,4 %	0 %	2 %	2,9 %

Source: Databases from webscraping, as of July 30, tourism repository as of March 22.
Field: Metropolitan France.

The comparison according to the operating methods shows that without the use of filters, the platform tends to reference more independent hotels (72.2% compared to 63.4% of hotels in the tourism repository). The use of the breakfasts included and free cancellation filters leads, on the contrary, to an over-representation of the chains (49.1% compared to 36.6% of the hotels in the tourism reference system, cf. table n°5)

Table 5: Comparison of the proportion of distinct establishments according to hotel classification for webscraped databases with filters, without filters and the tourism repository

Hotel classification	Tourism benchmark	Base without filter	Base with filter
Unclassified	27.7%	18.7 %	10.6 %
1 star	2.2%	2.2 %	2.4 %
2 stars	22.0%	23.7%	14.2%
3 stars	34.2%	38.6%	44.5%
4 stars	11.6%	14.2%	24.3%
5 stars	2.4%	2.5%	4.1%

Source: Databases from webscraping, as of July 30, 2021, tourism repository as of March 22, 2021.
Field: Whole of France.

These comparisons therefore show the existence of a bias in the coverage of hotels existing in the territory compared to those referenced on the online booking platform. This bias is more important when using webscraping using filters (breakfast included, free cancellation). These comparisons would benefit from being improved by carrying out a collection of all the hotels on the platform without a reservation date. Indeed, depending on the month of the year, some hotels may be closed for example. It would be necessary to plan the webscraping of hotels throughout France without imposing a date. It will also be interesting to compare this data with data from several booking platforms collected by Eurostat. Finally, this comparison of the different benchmarks could also be used for business statistics[19].

In the rest of this paper, only approach n°1 - the base of hotels with filters (i.e. hotels in the whole of France with a room highlighted by the online booking platform, with free cancellation and breakfast included) will be studied. Alternate day collection for each approach was discontinued in September to move to daily collection with filters.

[19] For example, Italy has matched the data from the webscraping of the webscraped platform with the administrative register of tourist accommodation establishments in the Emilia-Romagna region alone. The objective is to complete and enrich the information on tourist accommodation establishments already surveyed as part of traditional statistical surveys, but also to understand the degree of coverage of these surveys. The work of the Italians [8], [9] shows in particular that the coverage of tourist villages (100 %) and rented houses (78.7 %) on the online booking platform is almost total unlike campsites (13%), hostels (24.3%) and rented rooms (29.9%). Finally, certain types of establishments only appear on the online booking platform: chalets, boats, inns, lodges, motels, villas, homestays.

2.3 Cleaning and enrichment of the database of hotel nights with breakfast included and free cancellation pre-selected by the online booking platform

In order to allow its full exploitation, certain variables of the database with filters have been cleaned in order to extract only the numeric information from the webscraped field. These are the variables price, taxes, maximum number of occupants and room capacity. Finally, the field of observations is reduced to only rooms with a capacity of two people[20], observations associated with a price and dates of stay from December 2020. Thus, the usable database consists of 1,625,328 observations for $5,980^*$ hotels as of December 31, 2021.

Table 6: Distribution of the number of observations and the number of hotels according to the months of the overnight stay between December 2020 and December 2021

Month	Number of observations	Number of hotels
December 2020	32,119	2,802
January 2021	59,048	4,151
February 2021	85,255	4,575
March 2021	127,326	4,788
April 2021	149,570	4,996
May 2021	163,489	5,053
June 2021	148,015	5,219
July 2021	142,217	5,286
August 2021	109,384	5,195
September 2021	137,650	5,367
October 2021	138,380	$5,304^*$
November 2021	144,027	$5,088^*$
December 2021	144,836	$4,376^*$

Source: Database with filters from webscraping, as of December 31, 2021.
Field: Whole of France.
**: The counts for the months of October, November and December of the number of hotels are actually underestimated due to the disappearance of the identifier collected following a change in the platform in mid-October. There has been a rematch for hotels that have already appeared in previous months, but new entrants are not usable at this time.*

[20] This could only be done for data up to mid-October, as the variable used cannot be collected since

3 Analysis of prices collected by webscraping

This part, in addition to providing a quick overview of room prices, offers two studies on the prices of hotel nights collected by webscraping using filters: a first on the analysis of price determinants and a second analysis only on the observations (hotels x rooms)[21] whose prices are collected 0, 30 and 60 days prior in order to identify the main price evolution profiles and to validate or not the anteriority chosen at the start of the experimentation.

3.1 Overview of average room prices between December 2020 and December 2021

The rooms highlighted by the online booking platform with breakfast included and free cancellation are offered on average at 128.5 € over the period from December 2020 to December 2021. The prices range between 11 € and 10,025 €. The more stars the hotel has, the higher the average room rate. Non-classified hotels form a heterogeneous category of hotels and room prices are on average equivalent to those of 2 stars (see table n°8).

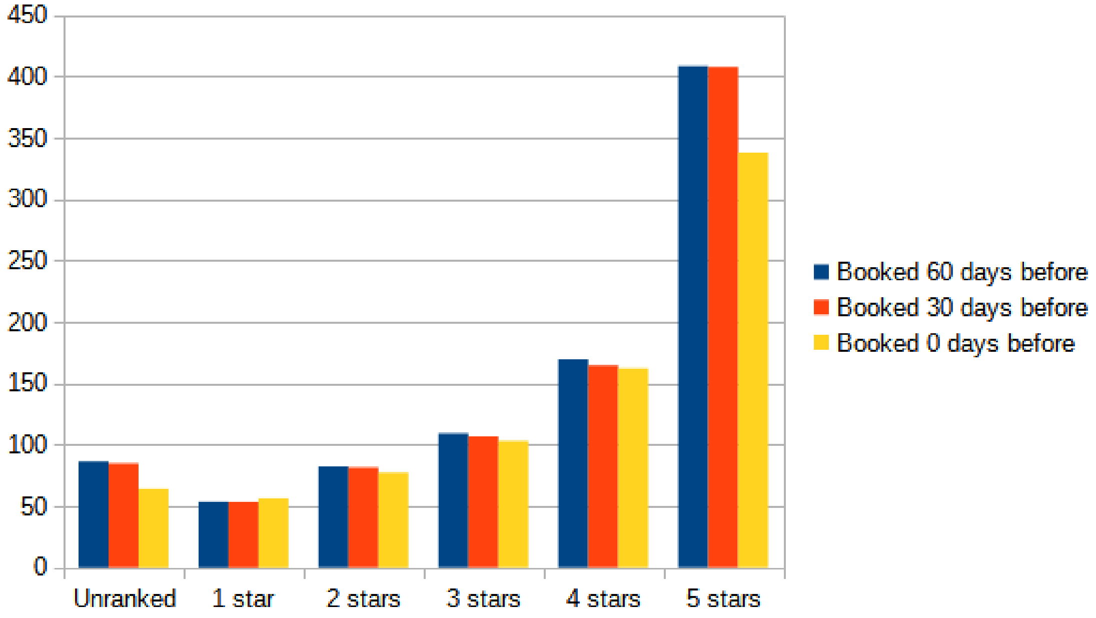
Table 7: Comparison of average hotel prices x rooms according to hotel ranking

Hotel classification	Unclassified	1 star	2 stars	3 stars	4 stars	5 stars
Average price (in €)	81	54	81	107	166	402

Source: Database with filters from webscraping, as of December 31.
Field: Whole of France.
Note: average prices are calculated from a geometric mean. Taking into account prior reservations, average prices tend to decrease the closer we get to a reservation for the same day regardless of the classification of the hotel except for 1-star hotels (see graph n°5).

[21] This is the name of the room highlighted by the platform for web-scraped hotels.

Figure 5: Evolution of average hotel prices x rooms according to hotel classification and reservation history



Source: Base with filters from webscraping, as of December 31.
Field: Whole of France.
Note: Average prices are calculated from a geometric mean.

The hotel classification according to the number of stars is the most discriminating factor of the price of overnight stays according to the decision tree built on the database with filters[22] with the data until July 30, 2021. This tree (Figure reffigures:arbredecision) was constructed using the threshold of 500 minimum observations as the stopping criterion. This tree highlights the specificity of 5-star hotels and even more so those located in Île-de-France, Provence-Alpes-Côte d'Azur and Pays de la Loire. Other criteria are strongly discriminating:

- the hotel operating model: independent hotels offer rooms on average at a higher price (144 €) than chains (96 €);

- the tourist area by often isolating the provincial urban area (modality "4") and the other spaces (modality "5"): hotels located in an urban provincial area offer rooms on average at a lower price (95 €) than those located on the coast (123 €) or in mountain ranges (145 €);

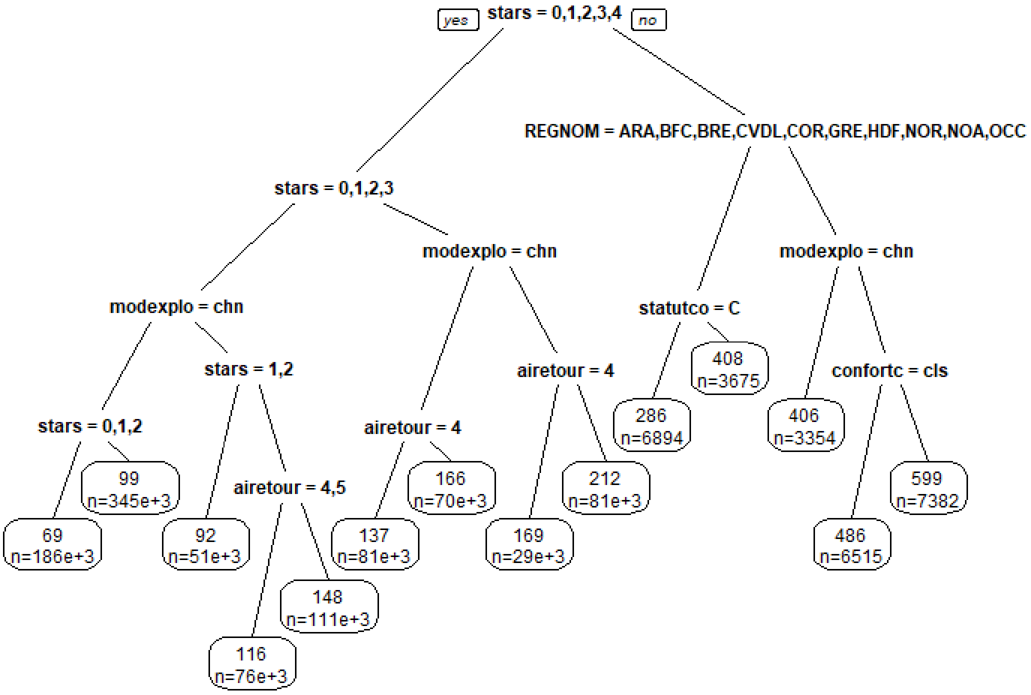
- the French region: Room prices by region vary from single to double on average. Prices are higher on average in Corsica (142 €) and Île-de-France (139 €). Conversely, prices in Hauts de France (91 €), Centre Val de Loire (91 €) are lower on average;

- the status of the municipality: hotels located in a central municipality offer rooms on average at a higher price (120 €) than municipalities in the suburbs (94 €);

- the comfort of the room: superior rooms are on average offered at 152 € while classic rooms are on average offered at 105 €;

[22] The decision tree is used here only in an exploratory approach by analyzing the decision conditions present on each node.

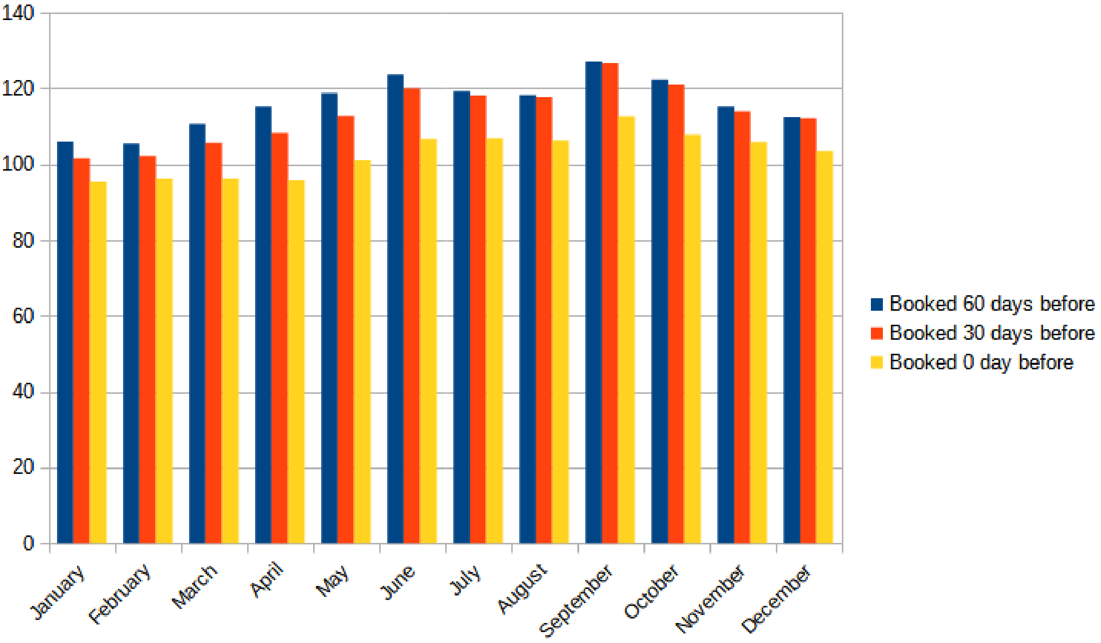
Figure 6: Decision tree whose variable of interest is the price of hotel nights



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

Prices increase on average for overnight stays between December and June regardless of the date of booking, while they drop on average for overnight stays taking place in July and August. Thereafter, there is a further increase between August and September followed by a decrease until December. In addition, prices tend to fall on average the closer we get to the date of the overnight stay (see graph n°7).

Figure 7: Evolution of the average prices of hotels x rooms according to the calendar months of the night and the anteriority of reservation



Source: Base with filters from webscraping, as of December 31.
Field: Whole of France.
Note: average prices are calculated from a geometric mean.

3.2 Analysis of room price determinants using a hedonic model

The analysis of the determinants of room prices consisted in studying, using a hedonic model, the main factors influencing the price of overnight stays.

A hedonic price model

The hedonic price model is a linear model that links the price of a good or service to characteristics. The hedonic price method[10] assumes that the price of a product depends on its characteristics.
Let p_n^t be the price of service n at period t, K the number of characteristics measured by the $z_{n,k}^t$, β_0^t and β_k^t respectively the constant and the parameters of the characteristics to be estimated, a hedonic price model is presented as follows in log-linear form:
$\log p_n^t = \beta_0^t + \sum_{k=1}^{K} \beta_k^t z_{n,k}^t + \epsilon_n^t$

The price per night (more exactly the log(price)) is modeled with the following explanatory variables:

- variables related to the location of the hotel: French region, French department, tourist area, status of the municipality;
- of hotel-related variables: hotel star rating, mode of operation (chain or independent), room comfort
- of the variables linked to the calendar: day of the night, month of the night, school holidays or not, public holiday or not, anteriority of the reservation

The models are built on two databases as of July 30, 2021: those containing reservations 30 and 60 days prior (792,289 observations) and those containing all reservations (60, 30 days prior and for the same day , i.e. 897,896 observations).

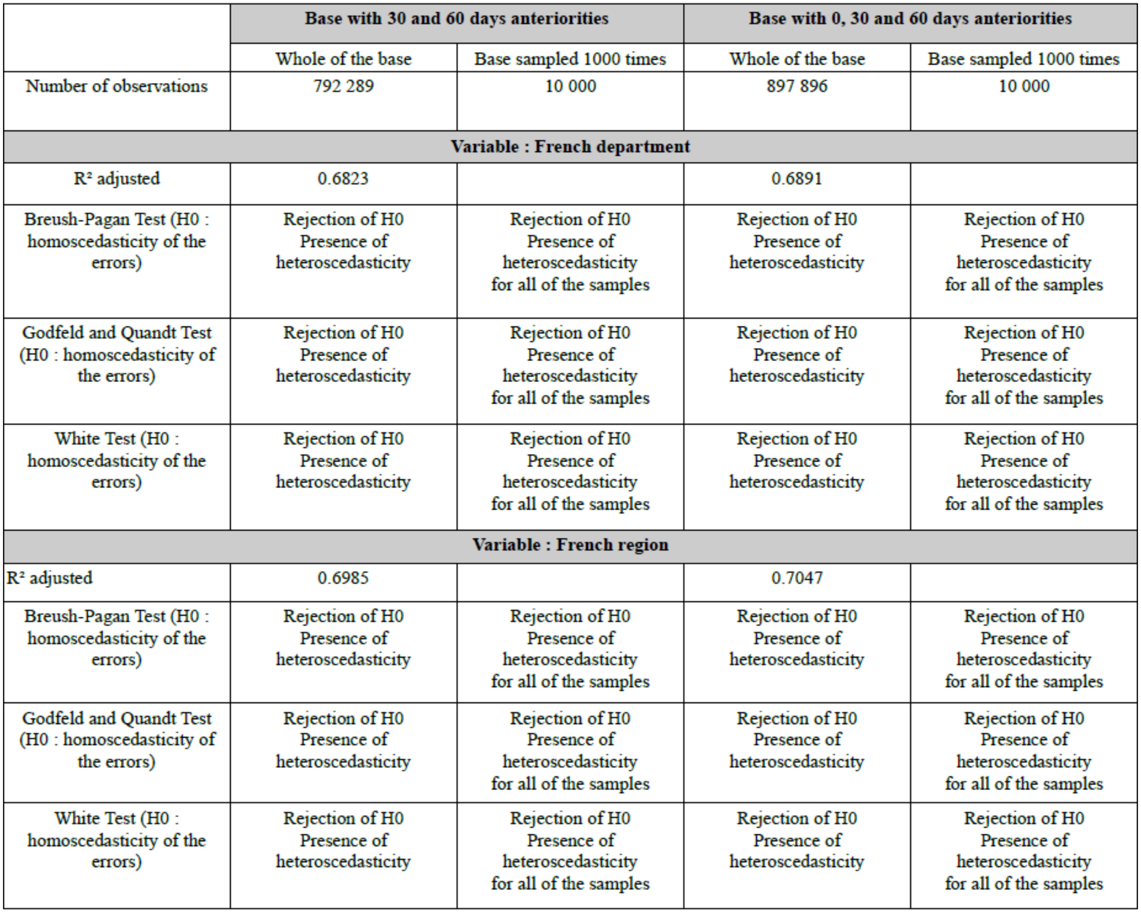
Finally, the models are built on all of the two databases and then on samples of 10,000 observations randomly drawn from each of the databases 1,000 times to ensure the robustness of the parameter estimates. The hedonic model is constructed using a stepwise variable selection method. It starts from the empty model and at each step, when the variable that leads to the lowest AIC has been added to the model, it will remove the variable from the model that causes the AIC to decrease as much as possible, if such a variable exists. The algorithm stops when the addition of any variable does not decrease the AIC. The model retained by this method retains all the variables tested with the exception of the region. This selection of variables is also obtained with the ascending method (forward) or the descending method (backward). These models nevertheless present heteroscedasticity which is corrected by estimating the parameters using the generalized least squares method. For reasons of simplicity, only the coefficients resulting from the regression with the region variable instead of the department variable are presented here (cf. figure n°8), the others are in the appendix (cf. table n°24 and 25 in the Appendix).

Analysis of the regression coefficients shows that:

- 1-star hotels stand out with lower prices than unrated hotels, all other things being equal; 3, 4 and 5 star hotels stand out with higher prices;
- independent hotels charge higher prices than chains, all other things being equal;
- overnight stays are less expensive during school holidays or public holidays, all other things being equal. The prices of these nights are also lower on Friday, Saturday compared to Sunday than on weekdays. These effects seem to show a possible price differentiation according to the targeted clientele (professional or for personal reasons);
- overnight prices are higher for reservations 60 days prior to 30 days prior; prices are lower for 30-day prior reservations compared to the price for same-day reservations and prices for 60-day prior reservations are slightly higher than for same-day reservations, all other things being equal;
- prices for overnight stays from April to August are higher than December prices, all other things being equal;
- the prices of overnight stays in central, isolated and non-urban municipalities are higher than in the suburbs, all other things being equal. Prices for overnight stays in mountain ranges are higher than in Île-de-France, all other things being equal, while prices on the coast, in urban areas in the provinces and other areas are lower than in in Ile-de-France. Finally, prices are higher in Provence-Alpes-Côte d'Azur (PACA) compared to the Auvergne Rhône-Alpes region (ARA), all other things being equal. Conversely, prices in the Pays de la Loire, Center Val de Loire, Hauts de France, Occitanie, Grand Est, Brittany, Bourgogne Franche-Comté are lower than the ARA region[23].

[23] The analysis with the department variable instead of the region variable shows that the prices in the departments of Paris, Hauts de Seine, Yvelines, Seine et Marne, Vaucluse, Haute-Saône, Var are higher than in Aube, all other things being equal. Conversely, prices in the Territoire de Belfort, Creuse and Gers departments are lower. The examples given are those with the most extreme coefficients in absolute value.

Figure 8: Main results of tested regressions (R^2 adjusted, heteroscedasticity test)



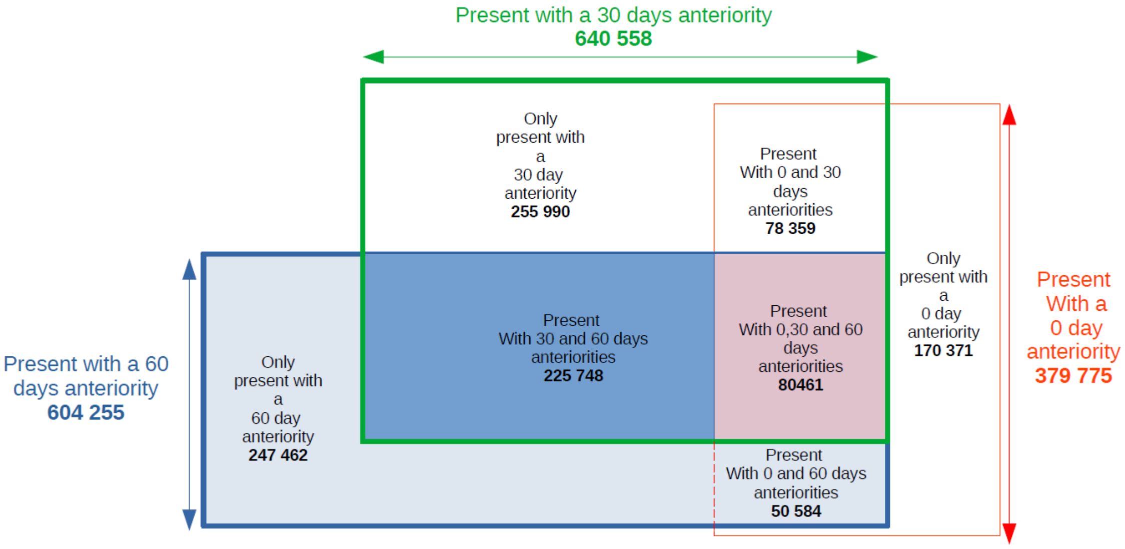
	Base with 30 and 60 days anteriorities		**Base with 0, 30 and 60 days anteriorities**	
	Whole of the base	Base sampled 1000 times	Whole of the base	Base sampled 1000 times
Number of observations	792 289	10 000	897 896	10 000
Variable : French department				
R^2 adjusted	0.6823		0.6891	
Breush-Pagan Test (H0 : homoscedasticity of the errors)	Rejection of H0 Presence of heteroscedasticity	Rejection of H0 Presence of heteroscedasticity for all of the samples	Rejection of H0 Presence of heteroscedasticity	Rejection of H0 Presence of heteroscedasticity for all of the samples
Godfeld and Quandt Test (H0 : homoscedasticity of the errors)	Rejection of H0 Presence of heteroscedasticity	Rejection of H0 Presence of heteroscedasticity for all of the samples	Rejection of H0 Presence of heteroscedasticity	Rejection of H0 Presence of heteroscedasticity for all of the samples
White Test (H0 : homoscedasticity of the errors)	Rejection of H0 Presence of heteroscedasticity	Rejection of H0 Presence of heteroscedasticity for all of the samples	Rejection of H0 Presence of heteroscedasticity	Rejection of H0 Presence of heteroscedasticity for all of the samples
Variable : French region				
R^2 adjusted	0.6985		0.7047	
Breush-Pagan Test (H0 : homoscedasticity of the errors)	Rejection of H0 Presence of heteroscedasticity	Rejection of H0 Presence of heteroscedasticity for all of the samples	Rejection of H0 Presence of heteroscedasticity	Rejection of H0 Presence of heteroscedasticity for all of the samples
Godfeld and Quandt Test (H0 : homoscedasticity of the errors)	Rejection of H0 Presence of heteroscedasticity	Rejection of H0 Presence of heteroscedasticity for all of the samples	Rejection of H0 Presence of heteroscedasticity	Rejection of H0 Presence of heteroscedasticity for all of the samples
White Test (H0 : homoscedasticity of the errors)	Rejection of H0 Presence of heteroscedasticity	Rejection of H0 Presence of heteroscedasticity for all of the samples	Rejection of H0 Presence of heteroscedasticity	Rejection of H0 Presence of heteroscedasticity for all of the samples

Source: Base with filters from the webscraping of the online booking platform, as of July 30, 2021.
Field: Metropolitan France.

3.3 Analysis of selected prior art

In this part, only the observations (hotels x rooms) whose prices are collected 0, 30 and 60 days prior for the same day are studied, this concerns 80,461 observations (see illustration n°9).

Figure 9: Distribution of hotels x rooms according to booking priority


Present with a 30 days anteriority
640 558
Only present with a 30 day anteriority
255 990
Present With 0 and 30 days anteriorities
78 359
Only present with a 0 day anteriority
170 371
Present With a 0 day anteriority
379 775
Present with a 60 days anteriority
604 255
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
Present With 0 and 60 days anteriorities
50 584

Source: Base with filters from webscraping, as of December 31, 2021.
Field: Whole of France.

66% of hotels x rooms have the same price whether the reservation was made 60 or 30 days in advance, while only 22% have the same price for a reservation 30 days and 0 days ago. 16% of hotels x rooms are at the intersection of these two groups, i.e. these hotels practice a stable price regardless of the anteriority of the reservation of the night (see table n°8). A priori, the 60-day booking priority seems less relevant but will require an analysis outside of the health crisis period.

Table 8: Distribution of observations according to their pricing profile

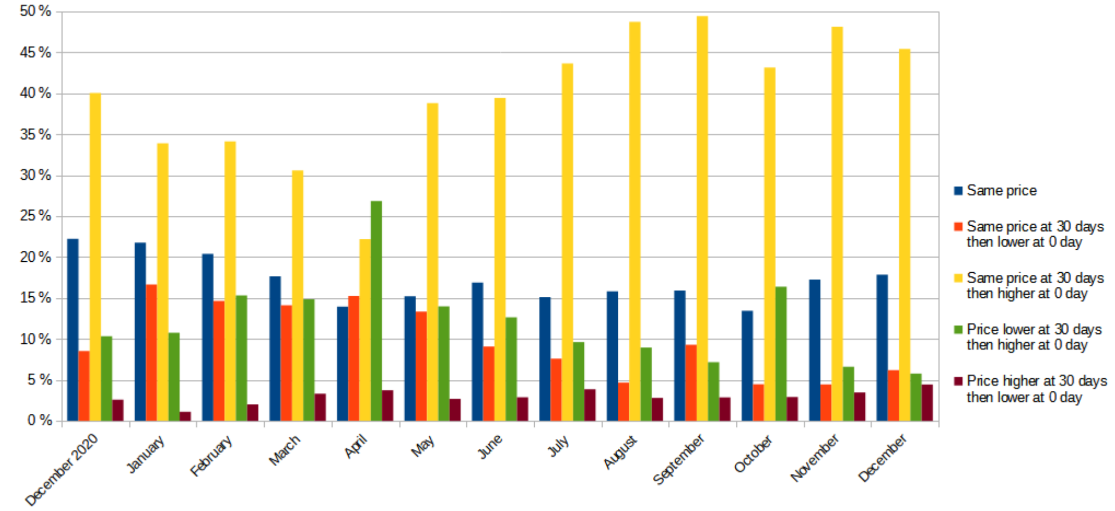
	Distribution of observations
Same prices at 60 days, 30 days and 0 days	16 %
Identical prices at 60 and 30 days then down to 0 days	10%
Identical prices at 60 and 30 days then upwards at 0 days	40%
Price down to 30 days then down to 0 days	4%
Price down at 30 days then up at 0 days	12%
Price down at 30 days then identical at 0 day	3%
Price up at 30 days then up at 0 days	8%
Price up at 30 days then down at 0 days	3%
Price up at 30 days then identical at 0 day	3%

Source: Database with filters from webscraping, as of December 31, 2021.
Field: France

This study shows that prices tend to increase for a reservation for the same day compared to that made at 30 days (60% of observations), but few are up both between 60 and 30 days and then between 30 and 0 days of seniority (8 %). Conversely, 17% of observations show a price drop between a reservation made 30 days later and a reservation made the same day.

Moreover, this rise in prices between 30 and 0 days tends to be confirmed more fully from May, in particular for the profile of hotels x rooms which practiced a stable price between a reservation 60 and 30 days prior then a increase for a reservation the same day (32% on average each month between December and April, 40% between May and July and 47% between August and December, see graph n°10). As a reminder, the third lockdown took place from April 3 to May 3, 2021 in France, which may explain the atypical nature of the month of April, which deserves to be analyzed on a another year. Conversely, the profile of hotels x rooms offering a stable price when booking 60 and 30 days in advance then a drop for a same-day booking seems to be less present lately (14% on average each months between December to April, 10% between May and July, and only 6% between August and December). Finally, the profile of hotels x rooms practicing price stability regardless of age is more or less stable over time (19% on average each month between December and April, 15% between May and July and 16% between August and December).

Figure 10: Change in share of certain pricing profiles over calendar months of overnight stay



Source: Base with filters from webscraping, as of December 31, 2021.
Field: Whole of France, only the observations (hotels x rooms) present at the three booking anteriorities.

This approach based solely on the observations (hotels x rooms) present in the three prior reservations leads to a result that contrasts with the panorama of average prices (part 1 of this chapter). Indeed, the panorama showed a drop in prices for reservations at 0 days. This decline is therefore driven by new entrants.

4 Discussion around the construction of a new price index

This part will propose the construction of a price index according to the method of homogeneous classes. This index will then be compared to the current CPI. Two tests will then be conducted to measure (i) the impact of considering the civil calendar versus using the CPI calendar; (ii) the impact of taking into account the 2020 consumption weightings impacted by the health crisis compared to the 2019 data.

Throughout this chapter, the analysis will focus solely on metropolitan France.

4.1 The homogeneous classes approach: unweighted geometric means of rooms within classes aggregated by an arithmetic Laspeyres formula

The approach of constructing a fixed-basket index was ruled out due to a high imputation rate over time[24]. It would average 45% over the period from January to July 2021.

Imputation rate under a fixed basket approach

The fixed basket approach consists of monitoring a sample of hotels x rooms defined in December 2020 on a given day throughout 2021. As part of the CPI field collection, the product is monitored for a given day with a tolerance of more or less three days due to the practical organization of price rounds by the investigators. This way of proceeding makes it possible to ensure that we are indeed measuring changes on average over a month, to guarantee the opening of the point of sale on that day and to neutralize any "day of the week" effects on the prices. Concretely, if the product must be collected every Thursday of week 1 of the CPI calendar, a collection on Wednesday of week 1 of the CPI calendar for a month of the year is authorized.
From all the hotels x rooms collected in December 2020, it is proposed to calculate an imputation rate for the months of January to July 2021, the date of the first analysis on the subject which has not been updated. For this, the CPI collection calendar is retained, the products must always be collected either on weekends or during the week for the correct week of the CPI calendar (week 1, 2 3 or 4) and the correct reservation anteriority (0 , 30, 60 days).

Table 9: Imputation rate of the number of hotels x rooms under a fixed basket approach according to the months of the CPI calendar

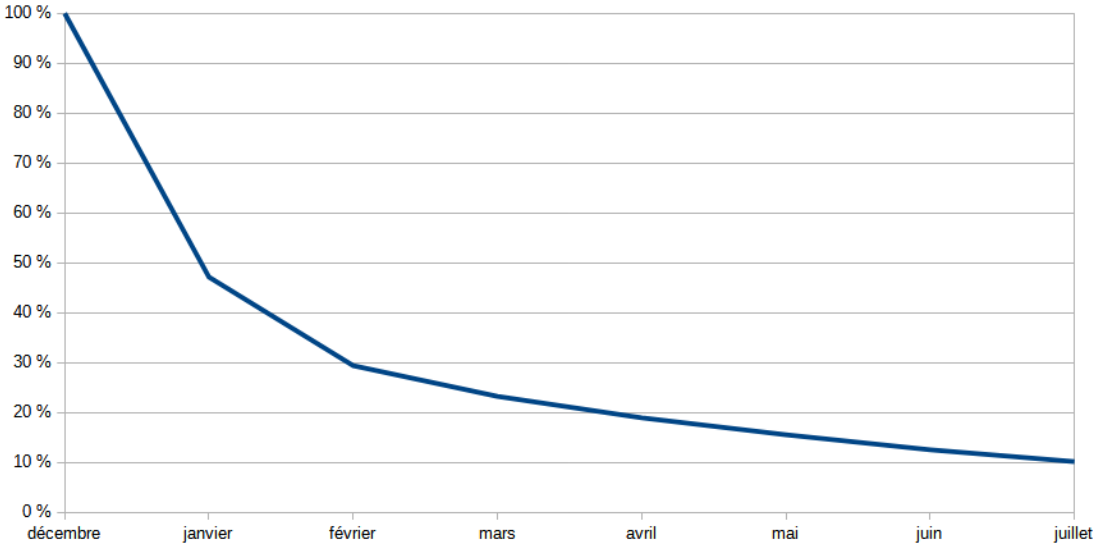
	January	February	March	April	May	June	July
Imputation rate	53 %	49 %	43 %	43 %	41 %	43 %	46 %

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

[24] This approach could nevertheless be further investigated by selecting a sample of "well-monitored, well-sold" rooms, which would limit the imputation rate. Moreover, this approach would require recoding the current CPI chain in R (case of replacements, recalculation of the base price, etc.), which would require a significant IT investment.

In January, 53% of products (hotels x rooms) are to be charged or replaced. It is interesting to complete this data with the rate of those present throughout the year compared to the sample defined in December 2020, which would make it possible to choose between imputation or replacement.

Figure 11: Evolution of the attendance rate of hotels x rooms present as part of a fixed-basket approach according to the months of the CPI calendar between december 2020 and july 2021



Source : Base with filters from webscraping, as of July 30.
Field: Whole of France.
*Reading note: 30 % of the hotels * rooms in the sample present in December are still present in February (i.e. present in December, January and February).*

Another approach has been adopted consisting of constructing classes that are sufficiently refined and homogeneous to consider that hotel rooms are substitutable for the consumer within these group[25]. This approach has a double advantage: being easier to set up and requiring few imputations.

4.1.1 Aggregation and stratification

Within these homogeneous classes, it is assumed that the consumer has the possibility of substituting products between them because the different rooms allow him to satisfy the same needs while being aware of the different prices charged for all the hotels[26] (all available rooms are displayed online on one or more pages during a request on the platform). **This hypothesis of substitutability of rooms within these homogeneous strata leads to adopting an unweighted geometric mean of prices (Jevons index)**. Within a stratum s, the

[25] This approach is called "monthly chaining and replenishment means" (MCR) in the Eurostat manual or "class mean imputation" in the IMF manual. It makes it possible to take into account the high turnover of products (particularly in the electronics sector) by authorizing a complete monthly renewal of the sample. The textbooks advocate using this method on large enough samples, and ensuring that products are not discounted too frequently or consistently released at a discount (at the end of the life cycle).

[26] En reality, only for hotels present on the online booking platform.

micro-index is therefore:

$$\boxed{I^m_{Jevons,s} = \frac{\prod_{i \in s,m} (price^m_i)^{1/n_m}}{\prod_{i \in s,0} (price^0_j)^{1/n_0}}}$$

n_0 (resp. n_m) is the number of observations of stratum s at month 0 (resp. at month m)

This level of granularity chosen for the micro-index reflects the assumption that certain parameters on which the price depends are important for the consumer. They were selected with the help of the analysis of price determinants[27]:

- the geography apprehended by the intersection of the region (outside Île-de-France) and tourist areas and by the intersection of Île-de-France and the status of the communes of Île-de-France (centre, suburb, isolated, outside urban unit);
- the comfort of the hotel measured by its classification: unclassified, 1, 2, 3, 4 or 5 stars;
- the hotel's operating model: chain or independent;
- the comfort of the room (classic / superior);
- the different anteriorities 0, 30 and 60 days because reserving an overnight stay two months in advance entails more uncertainties or constraints than at the last minute for the consumer. Consumers using one of these three prior art therefore have different profiles;
- both weekend/week periods allow control of calendar effects to be of constant utility.

For the following, stratum s is defined as the following intersection: **geographical area x hotel comfort x operating model x room comfort x anteriority x period**. The micro-indices are then aggregated by an **arithmetic Laspeyres type index**. This is indeed the classic index used in the CPI above a certain level of aggregation. It reflects a basket-type approach that eliminates the effects of consumption structure: an arithmetic mean of the elementary price indices is calculated by fixing the weights over the course of a year.

$$\boxed{I^m_{hotels} = \overset{\substack{géographic\\area}}{\sum_g} \overset{\substack{number\\of stars}}{\sum_s} \overset{\substack{exploitation\\model}}{\sum_e} \overset{comfort}{\sum_c} \overset{anteriority}{\sum_a} \overset{period}{\sum_p} w_g * w_s * w_e * w_c * w_a * w_p * I^m_{s(g,s,e,c,a,p)}}$$

w_g is the weight of the tourist area g, w_s of the number of stars s, w_e of the operating model e, w_c of the comfort of room c , w_a of anteriority a and w_p of period p

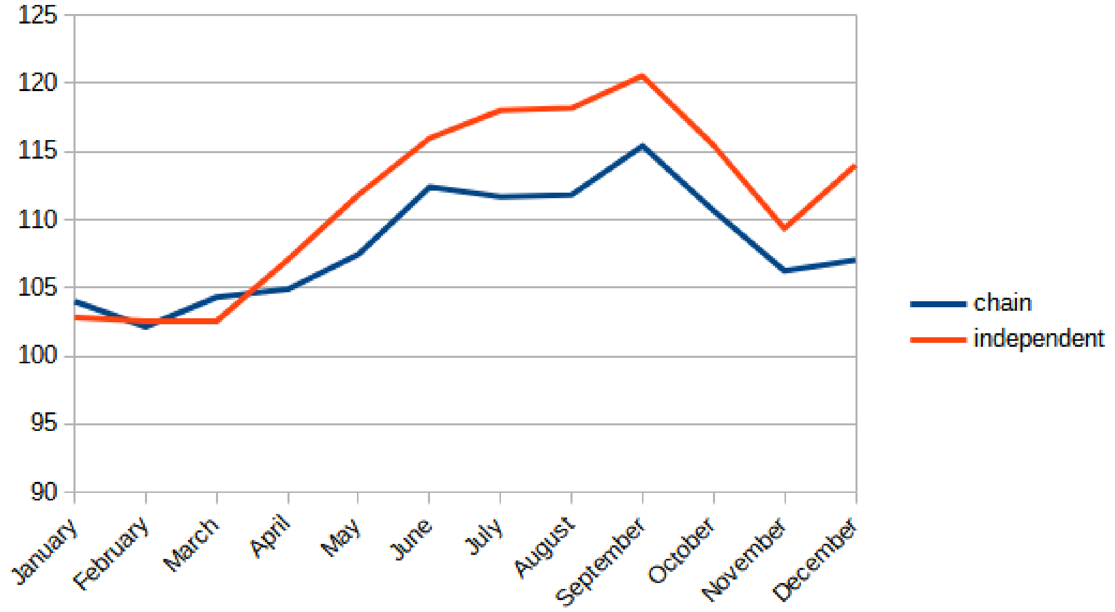
4.1.2 Price indices according to the different variables

When all the prices of the same stratum (*geographical area * hotel comfort * operating model * room comfort * anteriority * period*) are absent or when only one price within the same

[27] (i) this analysis of determinants focused on price levels and not price variations; (ii) not all the determinants have been kept, an operational choice takes place between the size of the classes (the more precise they are, the more closely they allow consumer behavior to be traced), and the volume of imputations (the smaller the classes are, the higher is potentially the number of empty classes).

stratum is available, the imputation is performed at the micro-index level. In order to analyze the different price dynamics according to the characteristics of the hotels and rooms, we calculated indices by variable.

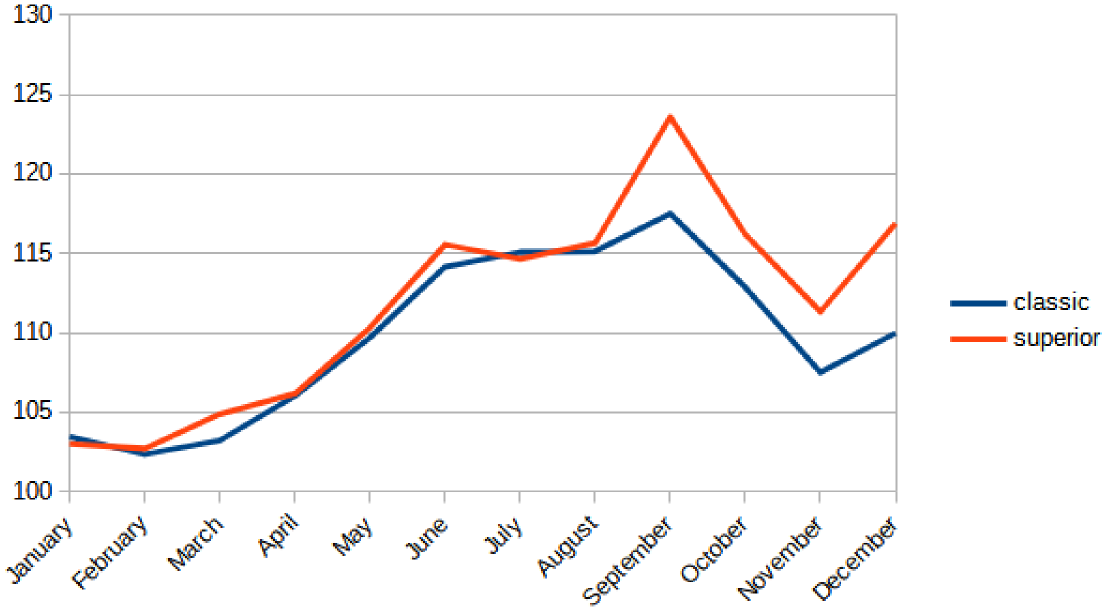
Figure 12: Price indices for hotel nights without allocation according to operating mode between January and December 2021 (base 100 = December 2020)



Source : *Database with filters from webscraping, as of December 31.*
Field: Metropolitan France, only the observations (hotels x rooms) present at the 30 and 60 day booking dates.

The indices differentiated according to the comfort of the room (see figure n°13 have a very similar dynamism at the beginning of the year, the higher increase in superior comfort rooms over the second part of the year (peak in September) may be due to a greater recovery in business travel.

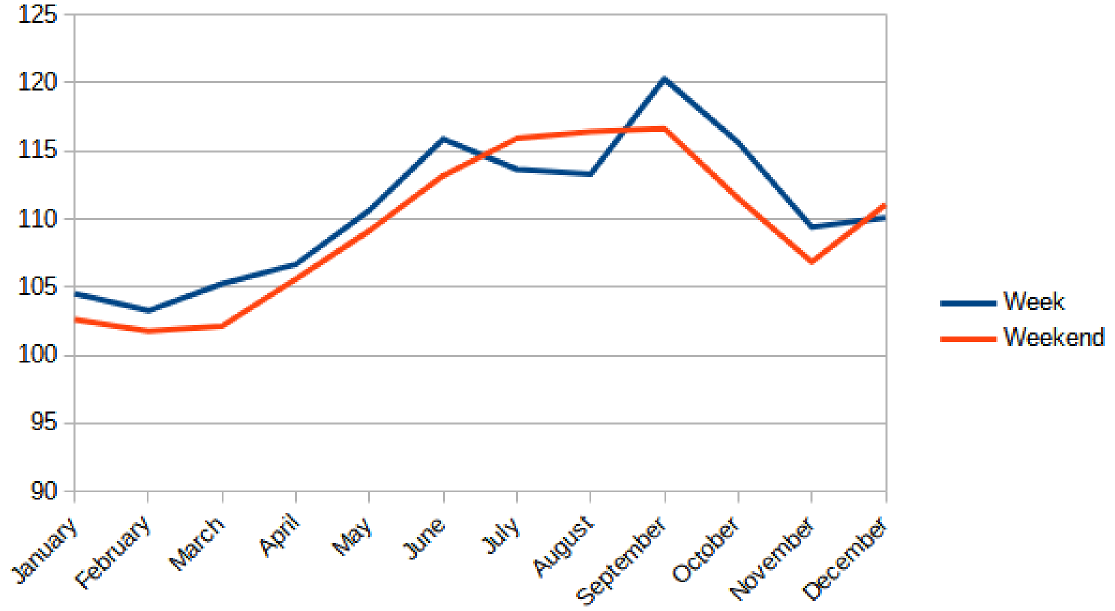
Figure 13: Price indices for hotel nights without allocation according to room comfort between January and December 2021 (base 100 = December 2020)



Source : Database with filters from webscraping, as of December 31.
Field: Metropolitan France, only the observations (hotels x rooms) present at the 30 and 60 day booking dates.

The analysis of price behavior according to the period of the night (see figure n°14) has the particularity of analyzing a field that is not currently taken into account in the price index: weekend reservations. We see that prices are stable on weekends between February and March, while during the week there is a fairly significant increase, an explanatory track may be the school holiday period. The summer period also presents different behaviors in the evolution of prices between the week and the weekend: weekend prices are on the rise between June and July while weekly prices are falling. Stability is observed for the weekend thereafter until September, while for weekly prices, prices rise sharply between August and September, which can be explained by a resumption of overnight stays for professional reasons.

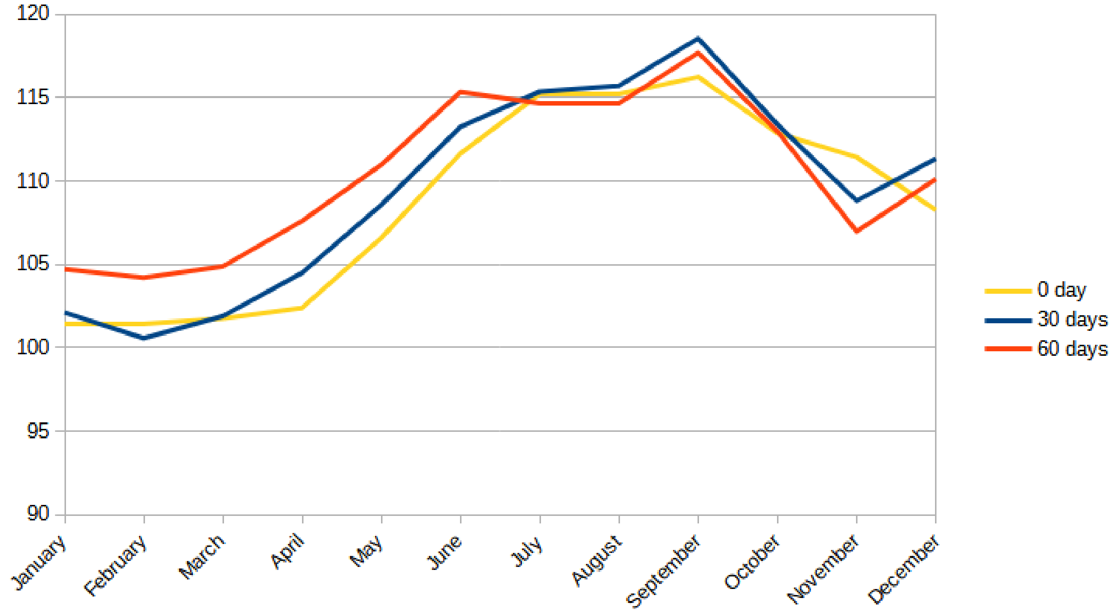
Figure 14: Price indices for hotel nights without allocation according to the period of the night between January and December 2021 (base 100 = December 2020)



Source : Database with filters from webscraping, as of December 31.
Field: Metropolitan France, only the observations (hotels x rooms) present at the 30 and 60 day booking dates.

Except for the month of November, the evolutions of prices according to anteriority are quite similar (see figure n°15). anteriority) increased a month earlier in the first half.

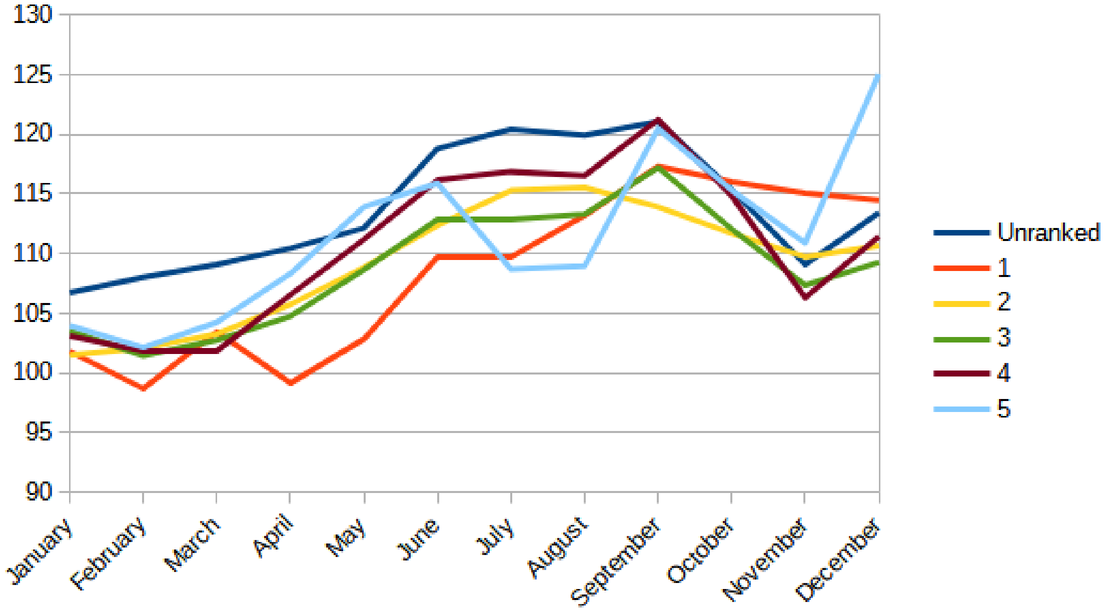
Figure 15: Price indices for hotel nights without charge according to prior reservation (base 100 = December 2020)



Source : Database with filters from webscraping, as of December 31.
Field: Metropolitan France, only the observations (hotels x rooms) present at the 30 and 60 day booking dates.

If the average prices of unclassified hotels and 2-star hotels are relatively close, we see that the price dynamics are not, they are perhaps closer to 3 stars (see figure n°16. We can notice the atypical profile of the 5 stars whose prices fall in the summer and have a pronounced increase in September and December (an explanatory factor may be the fact that many 5 star hotels are located in the Parisian region).

Figure 16: Price indices for hotel nights without allocation according to the number of stars of the hotel between January and December 2021 (base 100 = December 2020)

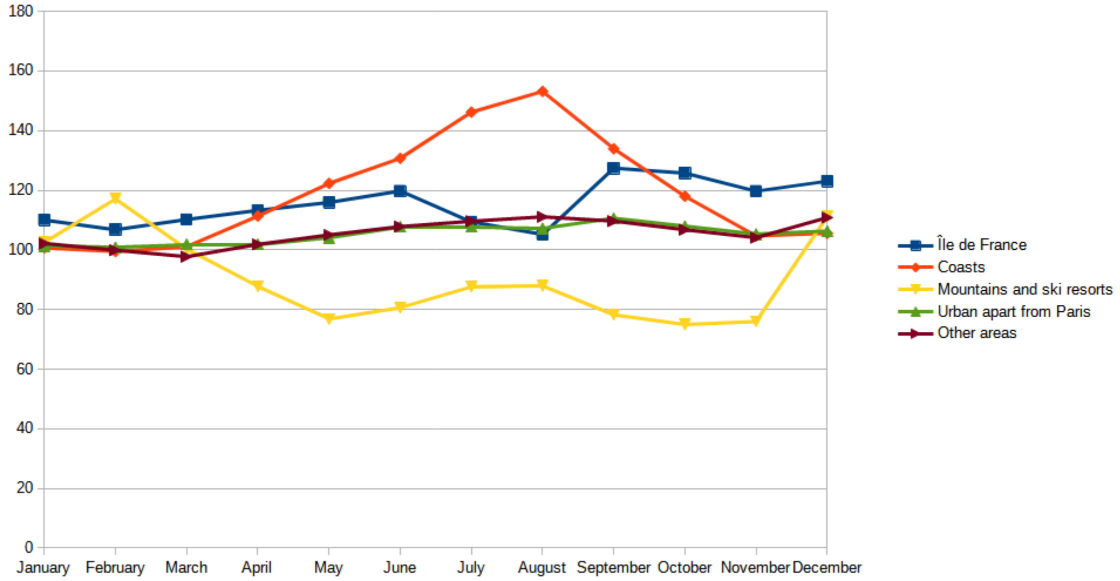


Source : *Database with filters from webscraping, as of December 31, 2021.*
Field: Metropolitan France, only the observations (hotels x rooms) present at the 30 and 60 day booking dates.

For the indices according to tourist areas[28] (see figure n°17), prices have distinct seasonal changes depending on the tourist area: a rise in prices in the late spring/summer periods for the coastline, an upward momentum for ski resorts at the start and end of the year. The particularity for ski resorts is that the base month (December) is one of the months when prices are the highest, which explains why the index drops below 80 during the year. As far as Île de France is concerned, the index drops from June to August and rises again in September. An explanatory factor is the drop in overnight stays for professional reasons during the summer period.

[28] From the end of October, the loss of collection of a variable due to a change in the platform led to problems for the matching of new hotels to the geographical reference (this represents 5213 observations on 427 243).

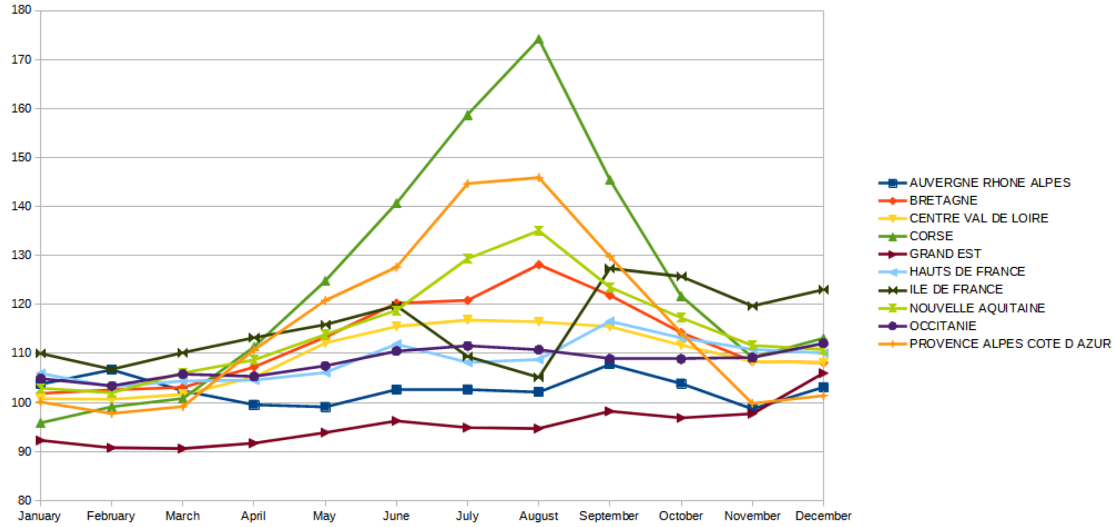
Figure 17: Price indices for hotel nights without imputation by tourist area between January and December 2021 (base 100 = December 2020)



The analysis of the indices by region (see figure n°18) has certain limitations because all the regions are not of the same size (in terms of population and surface area) and sometimes cover several tourist areas of different kinds. However, we can notice that Corsica has a price dynamism close to the coastal zones' one seen in the figure n°17, with a significant increase in summer.

Source: Base with filters from webscraping, as of December 31, 2021.
Field: Metropolitan France, only the observations (hotels x rooms) available at the 30 and 60 days booking dates.

Figure 18: Price indices for hotel nights without imputation by region between January and December 2021 (base 100 = December 2020)



Source: Base with filters from webscraping of the online booking platform, as of December 31, 2021.
Field: Metropolitan France, only observations (hotels x rooms) present in the 30 and 60 days reservation priors.

When no micro-index by geographical area x hotel comfort x operating model x seniority x period is available over a month, the average change in rooms is calculated for all the criteria combined. To calculate a micro-index, the number of rooms may vary from day to day and during the month either due to unavailability or due to a computer problem (the robot could not correctly collect the information). In another way, we can say that the absent prices are imputed implicitly within the micro-index by the average of the other prices of the same stratum.

4.1.3 Weightings

The weightings were determined using data provided by the tourism division from the monthly visitor accommodation survey[29]. Three sets of data on the number of rooms occupied were transmitted for the years 2019 and 2020 (year impacted by the health crisis).

Weighting set n°1 – raw 2019 data
The data for the year 2019 make it possible to obtain weights on the number of rooms occupied for the intersection of region x tourist area x number of stars (1-2 stars, 3 stars, 4-5 stars, unclassified). Several assumptions were then added:

- in order to distinguish the weights for 1 and 2 star hotels and for 4 and 5 star hotels within each region, a distribution key has been defined at the level of each region from the hotel repository maintained by the Montpellier tourism center (not based on consumption);

[29] A distortion of the distribution of web-scraped hotels compared to the hotel stock of the tourism repository had been highlighted. Retaining the weights from the tourist frequentation survey (based on the hotel stock of the reference system) makes it possible to adjust to this survey. Rooms with breakfast and free cancellable are assumed to be a good representative of the room set. Belgium also uses this weighting method based on tourist attendance despite the identified coverage bias.

- in order to distribute the weight of Île-de-France according to the different statuses of the municipalities (central city, suburb, isolated city, outside urban unit), a distribution key has been defined from the hotel repository (and not based on consumption);
- the weight of chains and independent hotels is calculated at the regional level and therefore applies uniformly regardless of the tourist area or the classification of hotels according to the number of stars.

Weighting set n°2 – 2019 data for personal reasons
The set of weightings n°1 could be improved by taking into account only the rooms occupied by customers for personal reasons (ie excluding business customers). The rooms occupied solely for personal reasons (see table n°10) are less numerous in proportion in Île-de-France, in Hauts-de-France and in Pays de la Loire compared to to all occupied rooms. On the contrary, rooms occupied solely for personal reasons are more numerous for Corsica and the Provence Alpes Côte d'Azur region.

Weighting set n°3 – raw 2020 data
The data for the year 2020 does not show the region x tourist area x number of stars intersection, but the region x tourist area intersection. An additional hypothesis was therefore added by applying the weight of the hotel classification (1-2 stars, 3 stars, 4-5 stars, unclassified) of each region regardless of the tourist area. The health crisis that began in 2020 had a particularly significant impact on tourist attendance in Île-de-France (high proportion of accommodation frequented by non-resident tourists, in the high end in particular[30]) . This distortion can be observed from the distribution of the number of rooms occupied in metropolitan France (see table n°10):

- lower weight for Île-de-France because many hotels remained closed, especially during the summer;
- greater weight for less urban or coastal regions (Burgundy Franche-Comté, Normandy, Pays de la Loire, Brittany, New Aquitaine, Auvergne Rhône-Alpes). These regions regained a better level of attendance during the summer of 2020.

[30] Furthermore, also falling business tourism due to the cancellation of face-to-face meetings and the cancellation of major events (outside the CPI scope).

Table 10: Breakdown of occupied rooms in 2019 and 2020 by region

Region	2019 (1)	2019 (2)	2020 (1)
Ile-de-France	32 %	31 %	22 %
Center	3%	3 %	3 %
Burgundy – Franche-Comté	3 %	3 %	4 %
Normandy	4%	4 %	5 %
Hauts de France	5 %	3 %	5 %
Great East	7%	7 %	7 %
Pays de la Loire	4 %	3 %	5 %
Britain	4%	4 %	5 %
New Aquitaine	8 %	8 %	9 %
Occitania	8 %	8 %	8 %
Auvergne-Rhône Alpes	11 %	11 %	14 %
Provence Alpes Côte d'Azur	11 %	13 %	11 %
Corsica	1%	2%	1%

Source: Monthly visitor accommodation survey.
Note: (1) raw data, (2) data for rooms occupied for personal reasons.
Field: Metropolitan France.

Weighting sets no. 1 and no. 3 pose a problem because of the inclusion of overnight stays for professional reasons.

Finally, other weightings are not available in the monthly tourist accommodation survey, so we had to opt for conventions due to a lack of information:

- the weight of the comfort of the room: the distribution is calculated from webscraped observations of the online booking platform (whole of France, without the use of filters). A weight of 85
- the weight of anteriority: equal distribution of weights;
- the weight of the period: 40% for the week, 60% for the weekend in order to overweight the weekend.

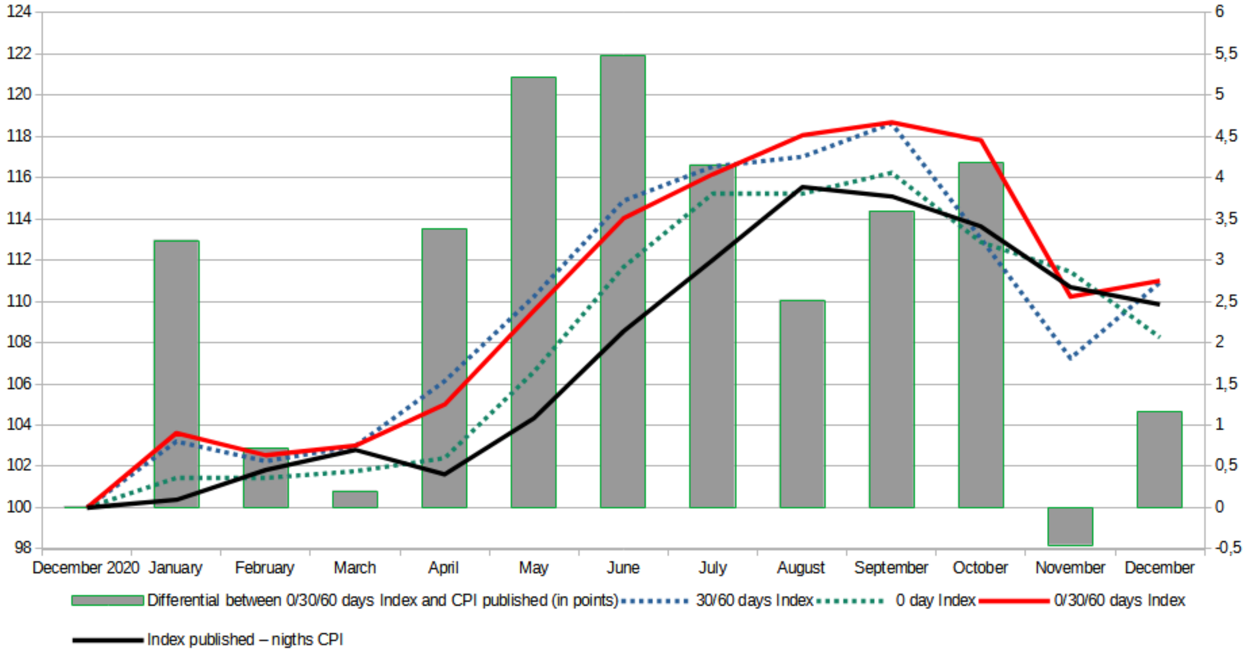
4.2 Comparison of current indices and indices by homogeneous classes

4.2.1 Comparison of proposed hotel indices

Three indexes are therefore obtained: a price index for same-day bookings (index n°1), a price index for 30 or 60-day bookings (index n°2), and a price index for bookings all anteriorities combined (index n°3). Due to the gradual widening of the geographical scope as part of the webscraping of the online booking platform, not all regions are present in all strata in December (base month). The 30 and 60 day anteriorities do not cover the Bourgogne-Franche Comté, Normandy and Pays de la Loire regions. Hotels in these areas are still covered by same-day reservations. The three indices show equivalent upward momentum over the period from May to July compared to the current index but differ at the start of 2021 (see graph n°19). Indeed, the evolution of prices measured by these three indices seems more dynamic in January than the current index, the three indices decrease in February (against an increase in the current index), increase to a lesser extent than the index (with the exception of the index for same-day reservations), then increase more significantly in April while the current index falls. Subsequently, only the price index for reservations 60, 30 days in advance and reservations for the same day will be studied because it corresponds to the index which makes it possible to take into account the diversity of pricing profiles. highlighted previously[31]. For the end of the year, there is a difference in momentum between the indices with the web scraped data which increases and the CPI index which decreases between November and December. This can be explained by the fact that the CPI index does not include the last weeks of December with the end of year celebrations and school holidays. The next part will focus on analyzing in more detail the impact of the chosen schedule.

[31] Ideally, prices should be collected only a few days before the date of the overnight stay (D-2 for example). Same-day bookings here are a proxy for this booking date.

Figure 19: Comparison of different price indices for hotel nights between December 2020 and December 2021



Source : *Database with filters from webscraping, 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 spread is calculated as the difference between the 0/30/60 day index – the CPI index*

4.2.2 Impact of the selected schedule

A test is carried out to estimate the impact of a modification of the calendar chosen: calendar months (calendar not fixed) versus CPI months (calendar fixed but not taking into account weekends and imperfectly school holidays).

Table 11: CPI Collection Schedule

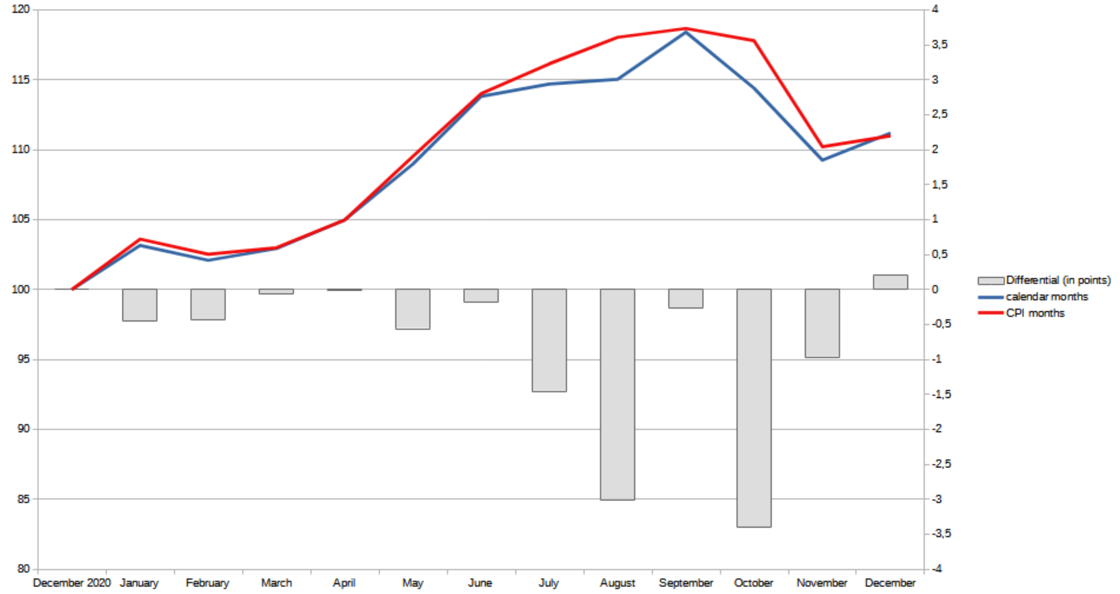
Month	Start date	End date
December 2020 (base month)	23-11-2020	20-12-2020
January 2021	04-01-2021	29-01-2021
February 2021	01-02-2021	26-02-2021
March 2021	01-03-2021	26-03-2021
April 2021	29-03-2021	23-04-2021
May 2021	03-05-2021	28-05-2021
June 2021	31-05-2021	25-06-2021
July 2021	28-06-2021	23-07-2021
August 2021	02-08-2021	27-08-2021
September 2021	30-08-2021	24-09-2021
October 2021	27-09-2021	22-10-2021
November 2021	25-10-2021	19-11-2021
December 2021	22-11-2021	17-12-2021

Let's first look at the indices calculated with the 2019 and 2020 weightings for the 3 anteriorities combined (see figure n°20). The average difference in absolute value is 0.92 points between the two indices. The price index for hotel room nights calculated using the CPI calendar includes:

- January: the index with the CPI calendar is more dynamic, it does not include the prices of the first days of January nor the last (which are a Friday, a Saturday, two Sundays, a Monday and a Tuesday). In addition, the base month for the index with the civil calendar incorporates the last week of December with end-of-year holidays and school holidays with higher average prices.
- July: slightly more dynamic rise in the index with the CPI calendar than with the calendar month. The July CPI calendar incorporates one price week in June and has one priceless week in July.
- August: very dynamic rise in the index with the CPI calendar, it does not include the prices of the first day of August nor the last three (which are a Friday, two Saturdays and two Sundays).
- September: The index with the CPI calendar is not very dynamic while the one using the calendar month is very dynamic. The September CPI calendar has one week less in September (last week of September).
- October: less marked drop in the CPI compared to the calendar month: The CPI includes prices for the last week of September but not the last week of October.
- November: very marked drop in the index with the CPI calendar, also marked drop but a little less in the index with the civil calendar. The index with the CPI calendar includes the prices of the last week of October but not the two weeks of the end of November.
- December: marked rise in the index with the calendar calendar, less marked rise in the index with the CPI calendar which includes 2 weeks of November and 2 weeks of December while the calendar month takes the rise of the 15-day holiday of Christmas.

It should be noted that the robot's collection every other day according to the two approaches (without filter and all the rooms on a restricted geographical area, with filters with the room put forward by the platform and on the whole territory) makes it possible to cover the different days in two weeks and therefore provides a good approximation of the index. However, this modus operandi may have impacts for weekends particularly reserved during school holidays (such as the last of July). It was stopped in September when the choice was made to maintain only the collection with filters and this every day.

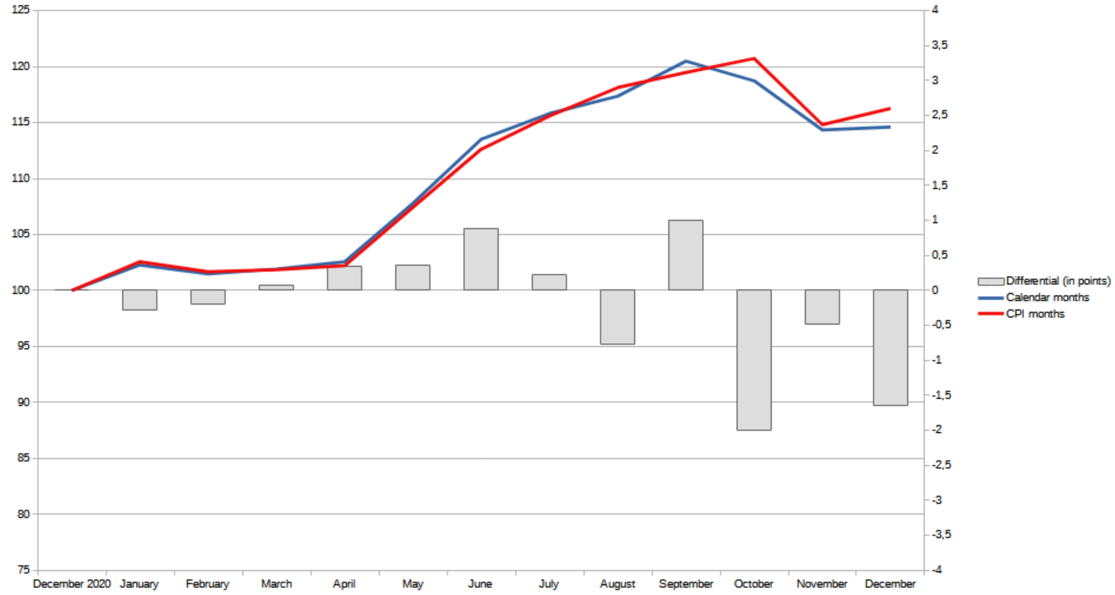
Figure 20: Comparison of the evolution of the prices of hotel nights taking into account anteriorities 0, 30 and 60 days according to the civil calendar and according to the CPI calendar



Sources: Author's calculations, database with filters from webscraping, as of December 31, 2021, monthly tourist accommodation visitor survey.
Field: Metropolitan France.
Note: the weights used here are those of the year 2019 for the consumption of rooms for personal reasons. The differential is calculated as the difference between the calendar calendar index – the CPI calendar index.

The calendar currently used for the CPI is used with field collection for the same day, a focus on the index calculated only with data 0 days old (see figure n°21) will make it easier to analyze this "calendar effect". The CPI calendar differs from the civil calendar mainly towards the end of the year. If we look at the evolution of prices, the index with the CPI calendar increases from April to October while the calendar calendar index increases from April to September and begins to decrease from October. This shift can be explained by the fact that the last week of the calendar month of October is not included in the CPI month (cf. table n°11). Between November and December the index with the CPI calendar increases while the calendar index remains stable.

Figure 21: Comparison of the evolution of the prices of hotel nights taking into account of anteriority 0 days according to the civil calendar and according to the CPI calendar

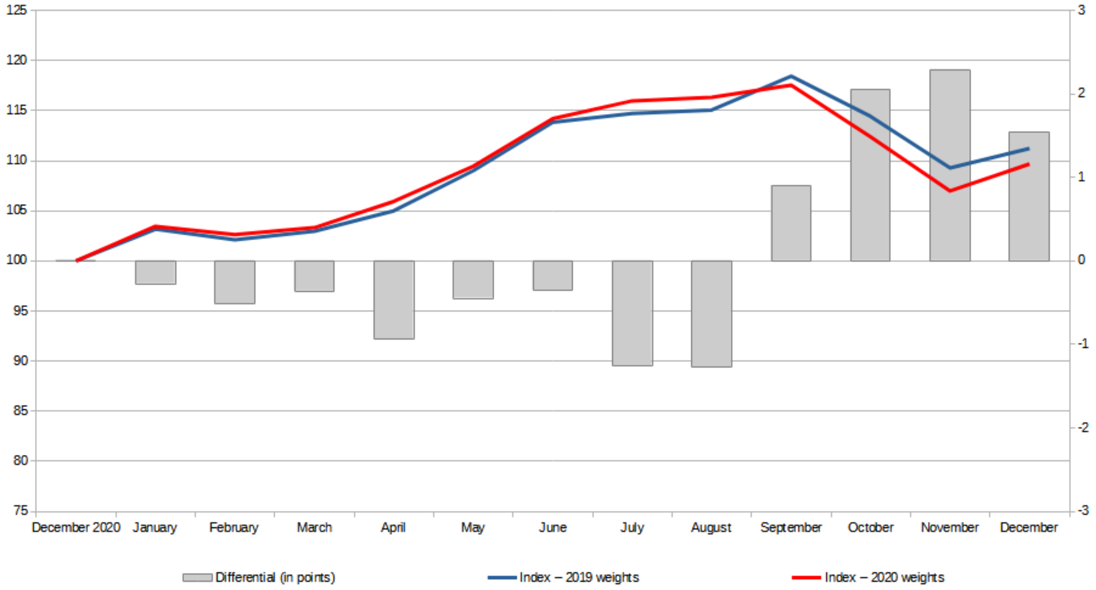


Sources: Author's calculations, database with filters from webscraping, as of December 31, 2021, monthly tourist accommodation visitor survey.
Field: Metropolitan France.
Note: the weights used here are those of the year 2019 for the consumption of rooms for personal reasons. The differential is calculated as the difference between the calendar calendar index – the CPI calendar index.

4.2.3 Impact of weights

A test is carried out to estimate the effect of the health crisis on the weightings used in the calculation of the index.
A first comparison is made between the attendance data for the year 2019 and the attendance data for the year 2020 (see graph n°22). The average difference in value is 0.1 point between the two indices and the average difference in absolute value is 1.01 point. There is a significant compensation between the periods of January to August and September to December. The largest difference is measured in October (+2.1 points). The main difference between the 2019 and 2020 attendance data - the lower weight for Île-de-France (cf. 4.1.3) - and the dip in summer with a rebound in September for this region (see figure n°17) explain the difference.

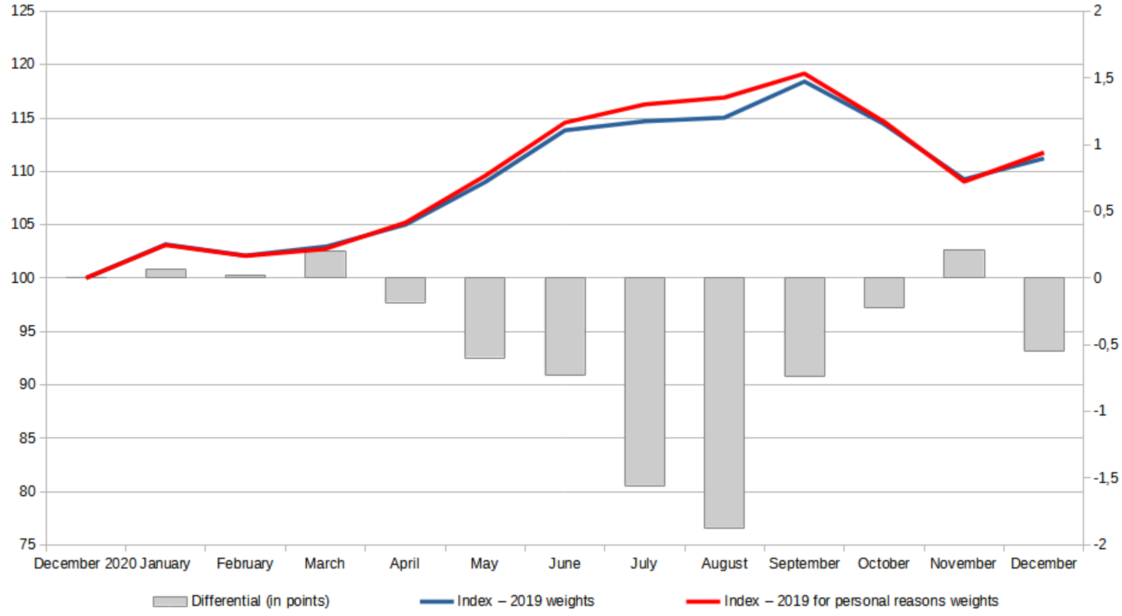
Figure 22: Comparison of the evolution of the prices of hotel nights taking into account 0, 30 and 60 days anteriorities according to the use of the 2019 and 2020 weightings



<u>Sources</u>: Author's calculations, base with filters from webscraping, as of December 31, 2021, monthly tourist accommodation visitor survey.
<u>Field</u>: Metropolitan France.
<u>Note</u>: The calendar used here is the civil calendar. The differential is calculated as the difference between the 2019 weighting index – 2020 weighting index.

A second comparison is made between the attendance data for the year 2019 and the attendance data for the year 2019, only taking into account the rooms reserved for personal reasons (see graph n°23) . It can be seen that the two indices are very close with the exception of the summer period (June, July, August and September) when the index with the weightings for personal reasons is higher by 1.2 points on average.

Figure 23: Comparison of the evolution of the prices of hotel nights taking into account 0, 30 and 60 days anteriorities according to the use of the 2019 and 2019 weightings only with rooms reserved for personal reasons



Sources: Author's calculations, base with filters from webscraping, as of December 31, 2021, monthly tourist accommodation visitor survey.
Field: Metropolitan France.
Note: The calendar used here is the civil calendar. The differential is calculated as the difference between the 2019 weighting index – 2019 weighting index for personal reasons.

5 Conclusion

This study focuses on the construction of a price index for hotel nights taking into account real-time pricing using the web scraping of a large booking platform. The index under study is based on an aggregation of elementary indices of sufficiently homogeneous classes by a Laspeyres formula. The main challenge is therefore defining these homogeneous classes and ensuring their relevance for understanding price changes and identifying calendar effects on the price level. For this, an essential prerequisite is the analysis of the price determinants. This prerequisite remains to be completed with the help of other considerations:

- the availability of consumption data at the level of these classes to be able to aggregate them;
- the integration of time dimensions (period[32] and anteriority) and their meaning:
- the integration of the consumption period to better seize the consumer's utility according to his departure date. The integration can be in the construction of the classes or by weighting the prices within a class according to the period;
- integrate the anteriority class dimension in the construction of classes or weight within a class the prices according to the anteriority class;
- the choice of the aggregation mesh can also rely on an operational arbitration considering the number of missing micro-indices to be imputed in the end. The more detailed this mesh is, the more this method has the disadvantage of generating an important number of missing micro-indices and being dependent on classes with few prices[33].

This experience of both web scraping and the creation of an index tries to take into account yield management. It could be a starting point for other studies on other services concerned with real-time pricing.

[32] Weekday, weekday/weekend, public holiday, school holidays.

[33] These classes require greater vigilance, particularly in cleaning and price validation.

References

[1] Phocuswright. European online travel overview. 2013.

[2] Autorité de la concurrence. Décision 15-d-06 du 21 avril 2015 sur les pratiques mises en œuvre par les sociétés booking.com b.v., booking.com france sas et booking.com customer service france sas dans le secteur de la réservation hôtelière en ligne. 2015.

[3] FMI. Consumer price index manual – concepts and methods. 2020.

[4] Eurostat. Recommendation on the treatment of flights and package holidays. 2018.

[5] Eurostat. Hicp methodological manual. 2018.

[6] Eurostat. Practical guidelines on web scraping for the hicp. 2020.

[7] M. Cure, A. Cazaubiel, B. Johansen, and T. Vergé. Paying for prominence and consumer prices prices : Evidence from booking' preferred partner program (non publié). 2021.

[8] Istat. Developing software for web scraping : the italian experience on portals offering tourist accomodation. *Presentation at the NTTS 2021 (New techniques and technologies for statistics)*, 2021.

[9] Istat. Methods and analysis for combining web scraping data with data on tourist accommodations survey. *Presentation at the NTTS 2021 (New techniques and technologies for statistics)*, 2021.

[10] R. Le Saout and B. Vignolles. La méthode des prix hédoniques, principes et illustration à partir du prix des terrains à bâtir. 2017.

[11] Berthier J-P. Introduction à la pratique des indices statistiques. 2005.

[12] A. Chauvet-Peyrard. Les indices de prix, de la théorie à la pratique. 2005.

[13] Alain Sauvant. Le « yield management » une question à 1,4 milliard de dollars, document de présentation. 2013.

[14] P. Sillard and L. Wilner. Indices de prix à utilité constante et substitutions intermensuelles. 2015.

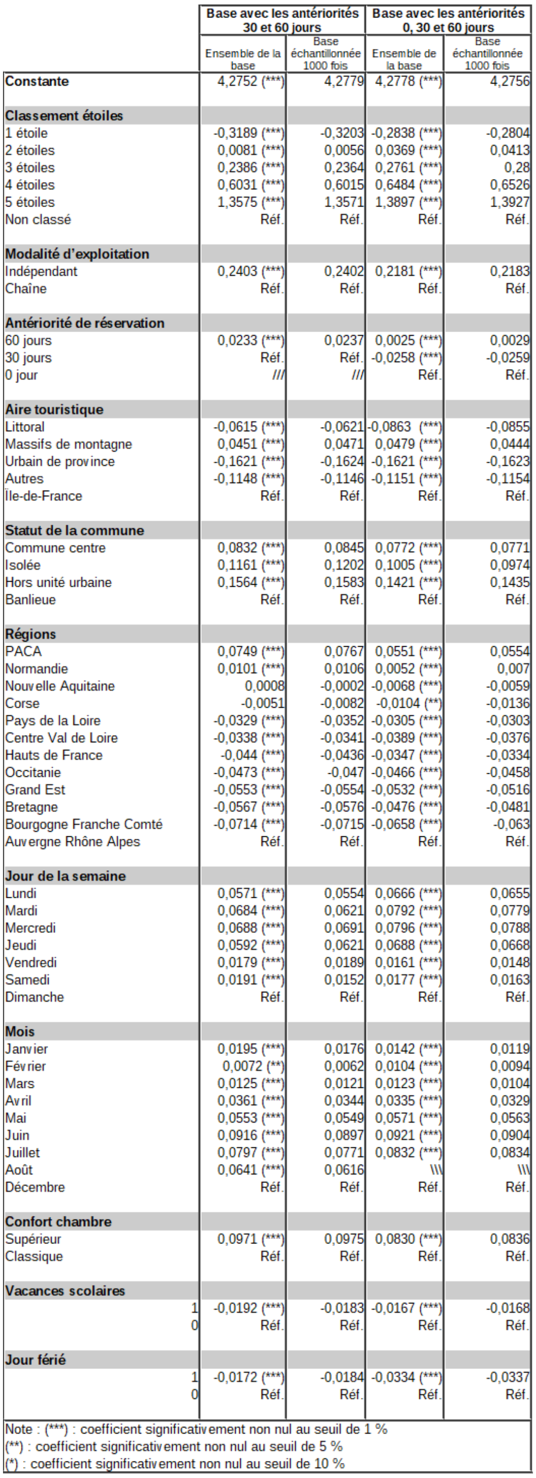
[15] P. Sillard and L. Jaluzot. Échantillonnage des agglomérations de l'ipc pour la base 205, document de travail. 2016.

[16] P. Sillard. Document de travail, indices de prix à la consommation. 2017.

[17] C. Freppel, O. Guillon, and A. Montbroussous. Indices des prix à la consommation des nuitées hôtelières : L'expérience du webscraping d'une plateforme de réservation en ligne, 2022.

6 Appendices

Figure 24: Analysis of the determinants of the prices of hotel nights in mainland France (regression coefficients)

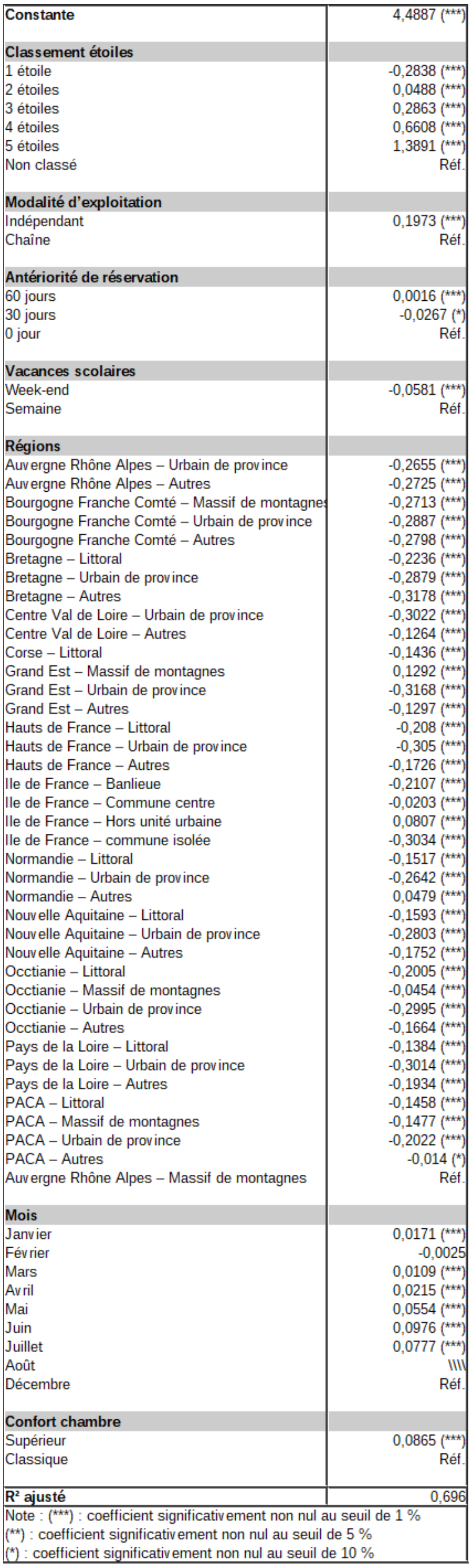


	Base avec les antériorités 30 et 60 jours		Base avec les antériorités 0, 30 et 60 jours	
	Ensemble de la base	Base échantillonnée 1000 fois	Ensemble de la base	Base échantillonnée 1000 fois
Constante	4,2752 (***)	4,2779	4,2778 (***)	4,2756
Classement étoiles				
1 étoile	-0,3189 (***)	-0,3203	-0,2838 (***)	-0,2804
2 étoiles	0,0081 (***)	0,0056	0,0369 (***)	0,0413
3 étoiles	0,2386 (***)	0,2364	0,2761 (***)	0,28
4 étoiles	0,6031 (***)	0,6015	0,6484 (***)	0,6526
5 étoiles	1,3575 (***)	1,3571	1,3897 (***)	1,3927
Non classé	Réf.	Réf.	Réf.	Réf.
Modalité d'exploitation				
Indépendant	0,2403 (***)	0,2402	0,2181 (***)	0,2183
Chaîne	Réf.	Réf.	Réf.	Réf.
Antériorité de réservation				
60 jours	0,0233 (***)	0,0237	0,0025 (***)	0,0029
30 jours	Réf.	Réf.	-0,0258 (***)	-0,0259
0 jour	///	///	Réf.	Réf.
Aire touristique				
Littoral	-0,0615 (***)	-0,0621	-0,0863 (***)	-0,0855
Massifs de montagne	0,0451 (***)	0,0471	0,0479 (***)	0,0444
Urbain de province	-0,1621 (***)	-0,1624	-0,1621 (***)	-0,1623
Autres	-0,1148 (***)	-0,1146	-0,1151 (***)	-0,1154
Île-de-France	Réf.	Réf.	Réf.	Réf.
Statut de la commune				
Commune centre	0,0832 (***)	0,0845	0,0772 (***)	0,0771
Isolée	0,1161 (***)	0,1202	0,1005 (***)	0,0974
Hors unité urbaine	0,1564 (***)	0,1583	0,1421 (***)	0,1435
Banlieue	Réf.	Réf.	Réf.	Réf.
Régions				
PACA	0,0749 (***)	0,0767	0,0551 (***)	0,0554
Normandie	0,0101 (***)	0,0106	0,0052 (***)	0,007
Nouvelle Aquitaine	0,0008	-0,0002	-0,0068 (***)	-0,0059
Corse	-0,0051	-0,0082	-0,0104 (**)	-0,0136
Pays de la Loire	-0,0329 (***)	-0,0352	-0,0305 (***)	-0,0303
Centre Val de Loire	-0,0338 (***)	-0,0341	-0,0389 (***)	-0,0376
Hauts de France	-0,044 (***)	-0,0436	-0,0347 (***)	-0,0334
Occitanie	-0,0473 (***)	-0,047	-0,0466 (***)	-0,0458
Grand Est	-0,0553 (***)	-0,0554	-0,0532 (***)	-0,0516
Bretagne	-0,0567 (***)	-0,0576	-0,0476 (***)	-0,0481
Bourgogne Franche Comté	-0,0714 (***)	-0,0715	-0,0658 (***)	-0,063
Auvergne Rhône Alpes	Réf.	Réf.	Réf.	Réf.
Jour de la semaine				
Lundi	0,0571 (***)	0,0554	0,0666 (***)	0,0655
Mardi	0,0684 (***)	0,0621	0,0792 (***)	0,0779
Mercredi	0,0688 (***)	0,0691	0,0796 (***)	0,0788
Jeudi	0,0592 (***)	0,0621	0,0688 (***)	0,0668
Vendredi	0,0179 (***)	0,0189	0,0161 (***)	0,0148
Samedi	0,0191 (***)	0,0152	0,0177 (***)	0,0163
Dimanche	Réf.	Réf.	Réf.	Réf.
Mois				
Janvier	0,0195 (***)	0,0176	0,0142 (***)	0,0119
Février	0,0072 (**)	0,0062	0,0104 (***)	0,0094
Mars	0,0125 (***)	0,0121	0,0123 (***)	0,0104
Avril	0,0361 (***)	0,0344	0,0335 (***)	0,0329
Mai	0,0553 (***)	0,0549	0,0571 (***)	0,0563
Juin	0,0916 (***)	0,0897	0,0921 (***)	0,0904
Juillet	0,0797 (***)	0,0771	0,0832 (***)	0,0834
Août	0,0641 (***)	0,0616	\\\	\\\
Décembre	Réf.	Réf.	Réf.	Réf.
Confort chambre				
Supérieur	0,0971 (***)	0,0975	0,0830 (***)	0,0836
Classique	Réf.	Réf.	Réf.	Réf.
Vacances scolaires				
1	-0,0192 (***)	-0,0183	-0,0167 (***)	-0,0168
0	Réf.	Réf.	Réf.	Réf.
Jour férié				
1	-0,0172 (***)	-0,0184	-0,0334 (***)	-0,0337
0	Réf.	Réf.	Réf.	Réf.

Note : (***) : coefficient significativement non nul au seuil de 1 %
(**) : coefficient significativement non nul au seuil de 5 %
(*) : coefficient significativement non nul au seuil de 10 %

Source: Database with filters from webscraping, as of July 30, 2021.
Field: Metropolitan France.

Figure 25: Analysis of the determinants of the prices of hotel nights in mainland France (regression coefficients)



Constante	4,4887 (***)
Classement étoiles	
1 étoile	-0,2838 (***)
2 étoiles	0,0488 (***)
3 étoiles	0,2863 (***)
4 étoiles	0,6608 (***)
5 étoiles	1,3891 (***)
Non classé	Réf.
Modalité d'exploitation	
Indépendant	0,1973 (***)
Chaîne	Réf.
Antériorité de réservation	
60 jours	0,0016 (***)
30 jours	-0,0267 (*)
0 jour	Réf.
Vacances scolaires	
Week-end	-0,0581 (***)
Semaine	Réf.
Régions	
Auvergne Rhône Alpes – Urbain de province	-0,2655 (***)
Auvergne Rhône Alpes – Autres	-0,2725 (***)
Bourgogne Franche Comté – Massif de montagnes	-0,2713 (***)
Bourgogne Franche Comté – Urbain de province	-0,2887 (***)
Bourgogne Franche Comté – Autres	-0,2798 (***)
Bretagne – Littoral	-0,2236 (***)
Bretagne – Urbain de province	-0,2879 (***)
Bretagne – Autres	-0,3178 (***)
Centre Val de Loire – Urbain de province	-0,3022 (***)
Centre Val de Loire – Autres	-0,1264 (***)
Corse – Littoral	-0,1436 (***)
Grand Est – Massif de montagnes	0,1292 (***)
Grand Est – Urbain de province	-0,3168 (***)
Grand Est – Autres	-0,1297 (***)
Hauts de France – Littoral	-0,208 (***)
Hauts de France – Urbain de province	-0,305 (***)
Hauts de France – Autres	-0,1726 (***)
Ile de France – Banlieue	-0,2107 (***)
Ile de France – Commune centre	-0,0203 (***)
Ile de France – Hors unité urbaine	0,0807 (***)
Ile de France – commune isolée	-0,3034 (***)
Normandie – Littoral	-0,1517 (***)
Normandie – Urbain de province	-0,2642 (***)
Normandie – Autres	0,0479 (***)
Nouvelle Aquitaine – Littoral	-0,1593 (***)
Nouvelle Aquitaine – Urbain de province	-0,2803 (***)
Nouvelle Aquitaine – Autres	-0,1752 (***)
Occtianie – Littoral	-0,2005 (***)
Occtianie – Massif de montagnes	-0,0454 (***)
Occtianie – Urbain de province	-0,2995 (***)
Occtianie – Autres	-0,1664 (***)
Pays de la Loire – Littoral	-0,1384 (***)
Pays de la Loire – Urbain de province	-0,3014 (***)
Pays de la Loire – Autres	-0,1934 (***)
PACA – Littoral	-0,1458 (***)
PACA – Massif de montagnes	-0,1477 (***)
PACA – Urbain de province	-0,2022 (***)
PACA – Autres	-0,014 (*)
Auvergne Rhône Alpes – Massif de montagnes	Réf.
Mois	
Janvier	0,0171 (***)
Février	-0,0025
Mars	0,0109 (***)
Avril	0,0215 (***)
Mai	0,0554 (***)
Juin	0,0976 (***)
Juillet	0,0777 (***)
Août	\\\\
Décembre	Réf.
Confort chambre	
Supérieur	0,0865 (***)
Classique	Réf.
R² ajusté	0,696

Note : (***) : coefficient significativement non nul au seuil de 1 %
(**) : coefficient significativement non nul au seuil de 5 %
(*) : coefficient significativement non nul au seuil de 10 %

Source: Database with filters from webscraping, as of July 30, 2021.
Field: Metropolitan France.