Journées de méthodologie statistique de l'Insee 2022

Indices des prix à la consommation des nuitées hôtelières : l'expérience du webscraping d'une plateforme de réservation en ligne

Adrien MONTBROUSSOUS (*), Camille FREPPEL (*), Ombéline GUILLON(*)

(*) Insee, Direction des statistiques démographiques et sociales

adrien.montbroussous@insee.fr

Mots-clés. (6 maximum) : Indices, webscraping, prix à la consommation, tourisme, classes homogènes.

Domaines. Collecte, analyse des données et data science

Résumé

L'indice des prix à la consommation pour les nuitées hôtelières est calculé à partir de relevés effectués sur le terrain par des enquêtrices et enquêteurs de l'Insee qui relèvent le prix d'une chambre pour une nuitée le jour même pour 2 personnes avec petit-déjeuner. L'indice des prix à la consommation pour les nuitées hôtelières est un indice à panier fixe avec différents niveaux d'agrégation : on relève les prix pour un même produit chaque mois dans un point de vente, à la même date (par exemple le premier mardi du mois). Afin d'améliorer l'indice et de s'émanciper de certaines limites de la méthode actuelle (couverture de la consommation en ligne partielle, champ géographique restreint, pas de prise en compte de la réservation en avance, etc.), la piste de recourir à une autre source de données a été explorée.

Cette étude propose ainsi de challenger l'indice actuel en explorant une méthode de collecte innovante, le webscraping. Le développement d'un robot de collecte automatisée d'une plateforme de réservation en ligne permet notamment de tenir compte de l'antériorité de la réservation de la nuitée et de s'affranchir des agglomérations définies dans le cadre de l'indice des prix à la consommation en améliorant la couverture des zones touristiques (zones littorales et massifs montagneux). Cette technique repose néanmoins sur le bon vouloir de la plateforme, une stratégie de collecte doit donc être mise en place afin d'adresser un nombre de requêtes idoine au site Internet de vente. Ces données collectées par le robot sont brutes et nécessitent un nettoyage ; par exemple la valeur pour une caractéristique n'est pas forcément décrite de la même manière entre deux observations. Dans notre étude, plusieurs antériorités de réservation ont été retenues lors de la construction du robot de collecte (réservation à 60 jours d'avance, à 30 jours d'avance et pour le jour-même).

Après une analyse des déterminants du prix des nuitées hôtelières, une réflexion sur la méthode à adopter pour élaborer un indice de prix à partir des données webscrapées sur la plateforme a été menée. L'indice à panier fixe est comparé à un indice construit à partir de classes suffisamment homogènes pour considérer que les chambres sont substituables à l'intérieur de ces strates.

La question des pondérations disponibles et pertinentes à utiliser entre également en compte pour choisir la méthode et les différents niveaux d'agrégations.

Nous comparerons les résultats de l'indice calculé avec les données de la plateforme de réservation en ligne avec l'indice publié. Les données ont commencé à être collectées en décembre 2020, la particularité étant qu'avec la crise sanitaire et les différents confinements il est difficile d'avoir du recul sur certaines données et comportements de consommation. Ce papier se concentrera sur les données allant de décembre 2020 à décembre 2021.

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.

Introduction

Plus de 50 % des dépenses de consommation des ménages concernent les services, or ces derniers se caractérisent par une multitude de formes, de modes de réservation et/ou d'achat. Le développement d'Internet et de ses plateformes de réservation a accéléré cette tendance. Il n'a jamais été aussi facile qu'aujourd'hui de réserver un billet d'avion, de train ou une nuitée hôtelière sur internet. L'indice des prix à la consommation a su s'adapter pour certains de ces services en prenant en compte les nouvelles habitudes de consommation des ménages. En effet, la part des relevés sur internet est croissante et des robots de prix ont été développés dans le secteur du transport (aérien, ferroviaire, maritime). Par ailleurs, des modèles de tarifications en temps réels (yield management) se sont généralisés dans certains secteurs (tourisme, transport, hôtellerie notamment), ce qui amène à challenger la collecte des prix afin de refléter au mieux les évolutions de prix de ces services à prix flottants.

L'objectif de cette étude est dans un premier temps d'étudier la pertinence de recourir au webscraping (programmes informatiques qui permettent d'extraire des informations d'un site internet) dans l'indice des prix des services de l'hôtellerie pour compléter l'indice actuel. L'analyse consistera ensuite à étudier les déterminants de prix des nuitées hôtelières pour enfin discuter de la méthodologie de construction d'un nouvel indice de prix.

Ce projet fait l'objet d'une bourse (grant) européenne et a été commencé par Camille Freppel (anciennement sectoriel des services) en collaboration avec Ombéline Guillon de la section méthodologie de la division des prix à la consommation, qu'Adrien Montbroussous a remplacé depuis le premier septembre 2021.

Table des matières

1 Indice des prix actuel et enjeux de l'étude

1.1 L'IPC est l'instrument de mesure de l'inflation dont le plan de sondage repose sur trois niveaux : le type de produit, l'agglomération et la forme de vente

L'IPC est l'instrument de mesure de l'évolution du niveau moyen des prix des biens et services consommés par les ménages sur le territoire français. C'est une mesure synthétique de l'évolution de prix d'un panier fixe de produits, à qualité constante. L'IPC est un indice de Laspeyres [1] chaîné annuellement. Il synthétise près de 30 000 micro-indices élémentaires – un indice élémentaire représentant en général le croisement d'une variété et d'une agglomération. À ce niveau, la formule de calcul d'un micro-indice est une moyenne géométrique non pondérée de rapports de prix pour les variétés hétérogènes [2] (indice de Jevons) et un rapport de prix moyen pour les variétés homogènes (indice de Dutot). Les pondérations utilisées pour les agrégations au niveau supérieur représentent la part des dépenses associées au poste concerné au sein de l'ensemble des dépenses de consommation des ménages couvertes par l'IPC. Afin de demeurer représentatif de la consommation des ménages, les pondérations sont actualisées chaque année et sont obtenues, notamment, à partir des évaluations annuelles des dépenses de consommation des ménages mesurées par le département de la comptabilité nationale. Le plan de sondage de l'IPC est caractérisé par les 3 niveaux suivants :

- critère géographique : les relevés sont effectués dans 99 agglomérations de plus de 2 000 habitants en France métropolitaine et dans 4 départements d'outre-mer ;
- variété : un échantillon de plus de 1 100 familles de produits et de services, appelées « variétés », est défini pour représenter l'hétérogénéité des produits au sein des 303 groupes de produits. La variété est le niveau de base élémentaire pour le suivi des biens et des services et pour le calcul de l'indice. La liste des variétés reste, à ce jour, confidentielle et seuls quelques prix moyens d'un échantillon de produits et de services sont publiés à ce niveau ;
- forme de vente : un échantillon d'environ 30 000 points de vente, stratifié par forme de vente est construit pour représenter la diversité des biens et services par marque, enseigne et mode d'achat des consommateurs (y compris internet). On distingue ainsi par exemple les hypermarchés des supermarchés qui sont également distingués des marchés.

De ces trois critères résulte un échantillon de près de 160 000 produits. À cela s'ajoutent 80 millions de produits suivis dans le cadre des données de caisses et plus de 500 000 prix collectés sur internet. La mise à jour de l'échantillon est annuelle afin d'intégrer les évolutions des comportements de consommation et, notamment, introduire des nouveaux biens ou services. Le dernier changement de base date de 2015. Il s'est caractérisé par le tirage de nouvelles agglomérations, issues des résultats récents du recensement de la population et par l'optimisation du nombre de relevés par croisement variété x agglomération. En cas de disparition d'un produit en cours d'année, celui-ci est remplacé par un produit proche et un ajustement qualité est effectué afin de corriger les différences de caractéristiques existantes entre le produit remplacé et le remplaçant.

1. Moyenne arithmétique d'indices élémentaires pondérée par les valeurs de la période de référence.
2. Une variété regroupant des produits dont les niveaux de prix sont relativement dispersés.

1.2 L'indice des prix nuitées hôtelières actuel et les limites liées à la collecte

1.2.1 La méthode actuelle de calcul de l'indice des prix des nuitées hôtelières

Les nuitées hôtelières sont suivies dans le cadre de la fonction 11 - Restaurants et hôtels de la COICOP [3], en particulier au sein du poste 11.2.0.1.1 Location de chambre qui représente 0,8 % de la consommation du panier de l'IPC [4]. En métropole, six variétés différentes sont suivies :

- Nuitée dans un hôtel 5 étoiles avec 43 relevés ;
- Nuitée dans un hôtel 4 étoiles avec 143 relevés ;
- Nuitée dans un hôtel 3 étoiles avec 206 relevés ;
- Nuitée dans un hôtel 2 étoiles avec 195 relevés ;
- Nuitée dans un hôtel 1 étoile avec 39 relevés ;
- Nuitée dans un hôtel non classé avec 34 relevés.

Ces relevés sont effectués dans 658 hôtels. Les nuitées hôtelières sont également suivies dans les DOM. Les variétés métropolitaines sont des variétés dites homogènes, c'est à dire que les hôtels de même confort (nombre d'étoiles) ne sont pas substituables pour le consommateur au sein d'une agglomération. Autrement dit, lorsque le prix d'une chambre d'un hôtel 4 étoiles augmente, le client ne reporte pas sa consommation vers un autre hôtel 4 étoiles de la même agglomération. Cela peut se justifier par le fait que le consommateur n'a pas connaissance des différents prix pratiqués dans les autres points de vente de l'agglomération et/ou que la chambre d'un autre hôtel 4 étoiles de la même agglomération ne lui permette pas de satisfaire le même besoin (localisation différente par exemple).

3. Classification of Individual Consumption by Purpose. Nomenclature internationale qui permet de décomposer la consommation des ménages par unités de besoin.

4. Ces données de pondérations sont revues annuellement et proviennent pour l'année N des données semi-définitives des comptes nationaux de l'année N-2. Pour 2021 et 2022, les estimations des comptes trimestriels pour l'année N-1 ont également été mobilisées.

Lien avec la théorie du consommateur

Dans le cadre microéconomique de la théorie du consommateur, le ménage cherche à minimiser son coût sous la contrainte d'atteindre d'un certain niveau d'utilité, l'évolution des prix d'un panier de biens entre deux périodes correspond donc à l'évolution du budget qu'il doit consacrer à l'achat d'un ensemble de biens en maintenant son utilité constante entre les deux périodes. Il est alors possible de relier les formules de calcul des micro-indices des variétés x agglomérations de l'IPC en recourant à des fonctions d'utilité usuelles entre deux instants (0 et 1) et n biens :

— Utilité de Léontief : élasticité de substitution entre les biens nulle – dans le cas de l'IPC, substitution entre points de vente [a]

$I^{1,0}(p^0, p^1) = \frac{\frac{1}{n}\sum_{i=1}^{n} p_i^1}{\frac{1}{n}\sum_{i=1}^{n} p_i^0}$

L'indice de Dutot utilisé dans le cadre des variétés homogènes témoigne d'une non substituabilité des points de vente au sein de l'agglomération. En d'autres termes, le consommateur effectue ses arbitrages de prix à l'échelle des points de vente.

— Utilité de Cobb-Douglas : élasticité de substitution unitaire entre les biens – dans le cas de l'IPC entre points de vente.

$I^{1,0}(p^0, p^1) = \prod(\frac{p_i^1}{p_i^0})^{\frac{1}{\pi}}$

L'indice de Jevons utilisé dans le cadre des variétés hétérogènes témoigne d'un comportement de substitution entre points de vente de l'agglomération caractérisé par une élasticité unitaire. En d'autres termes, le consommateur effectue ses arbitrages de prix à l'échelle de l'agglomération.

Il est à noter qu'en dehors des marchés élémentaires de consommation sur lesquelles des substitutions existent (ie sur lesquelles le consommateur fait jouer la concurrence), il n'y a pas de substitutions.

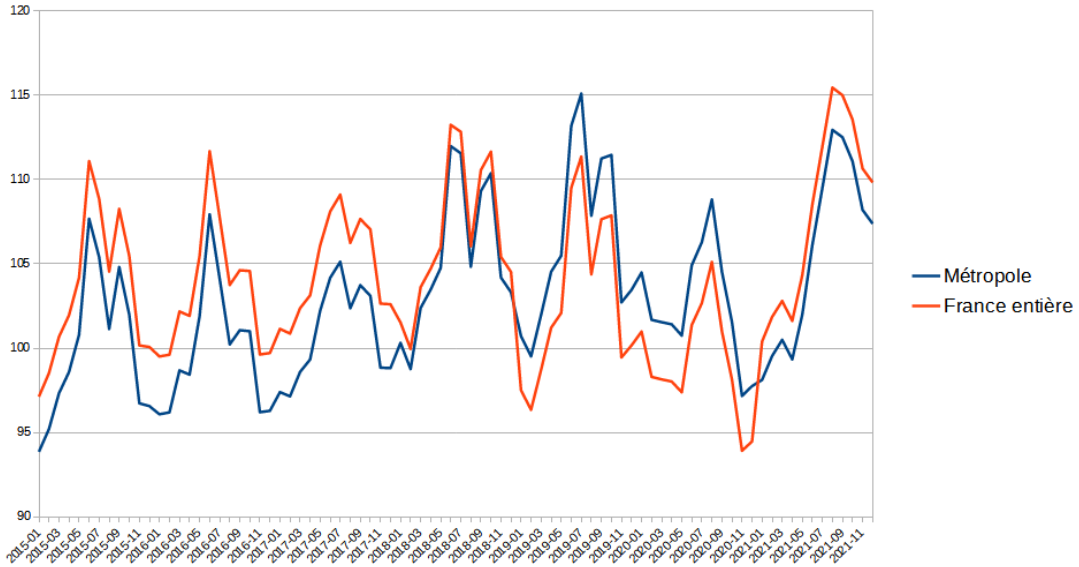
a. En effet, dans la configuration actuelle du calcul de l'IPC, un seul prix pour une variété de produits donnée est relevé dans un point de vente particulier.

La collecte des prix est effectuée sur le terrain, dans les agglomérations (de plus de 2 000 habitants) définies dans le cadre de l'IPC. Les prix sont collectés du lundi au vendredi et dans le cas des prix des hôtels pour la nuitée du jour du passage de l'enquêteur. Seuls 17 relevés concernant les hôtels 4 étoiles sont collectés avec un mois d'avance. Cette collecte est originale du fait des contraintes applicatives puisque ces 17 relevés sont effectués par les gestionnaires prix en bureau sur internet, les prix sont collectés dans un tableur avant d'être saisis le mois suivant en même temps que les prix collectés par les enquêteurs sur le terrain. Afin de s'assurer de la qualité constante du produit, celui-ci est défini plus précisément comme une nuitée pour deux personnes, deux petits-déjeuners compris. Par ailleurs, les enquêteurs remplissent un formulaire permettant de s'assurer que les caractéristiques de l'hôtel et de la chambre sont les mêmes chaque mois (sans quotas imposés) :

— localisation de l'hôtel (centre-ville versus périphérie) ;
— type d'hôtel (indépendant, chaîne franchisée, chaîne volontaire) ;
— confort de la chambre (classique, standard, supérieure, luxe, privilège, autre).

Entre décembre 2015 et décembre 2021, les prix des nuitées hôtelières ont augmenté de 11,2 % en France. Ils ont augmenté en moyenne de 1,8 % chaque année (cf. graphique n°1).

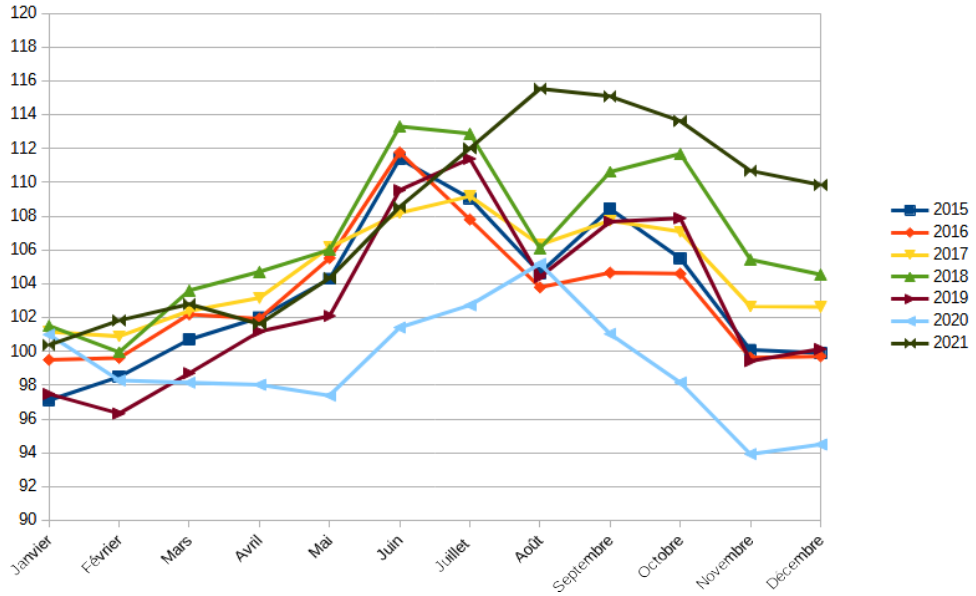
FIGURE 1 – Indice des prix du poste Location de chambre entre janvier 2015 et décembre 2021



Source : IPC.
Note de lecture : Base 2015

La saisonnalité observée entre 2015 et 2021 est hétérogène, certaines caractéristiques peuvent néanmoins être dégagées : des hausses de prix de février à juin puis de août à octobre et des baisses de prix en août et novembre (cf. graphique n°2). Les profils des années 2020 et 2021 sont atypiques en raison de la crise sanitaire, en plus des périodes particulières liées aux confinements successifs, on constate une hausse des prix entre les mois de juillet et août.

FIGURE 2 – Indice des prix du poste Location de chambre entre 2015 et 2021 (base 100= décembre N-1)



Source : IPC.
Champ : Métropole
Note de lecture : Par exemple, les indices de l'année 2021 sont en base 100 = décembre 2020.

1.2.2 Les limites de la méthode actuelle

La méthode actuelle présente plusieurs limites :

— Limite n°1 – l'absence d'autres modes de réservation : d'autres canaux de réservations de nuitées vendues isolément [5] existent comme la réservation sur internet et qui par ailleurs prennent de l'importance [6]. Ces réservations en ligne peuvent s'effectuer via le site internet de l'hôtel ou via une plateforme (par exemple Booking, Hotels.com, Trivago, Expedia, Last Minute Travel, etc) [7] ;

— Limite n°2 – l'absence de réservation d'hôtel par anticipation : l'indice actuel ne prend pas en compte les réservations de chambre avec anticipation des consommateurs. Or les hôtels peuvent ajuster leur prix suivant le remplissage des chambres, et les prix peuvent varier en fonction de l'antériorité d'achat ;

— Limite n°3 – la représentativité de l'échantillon (taille et géographie) : la collecte des prix sur le terrain est strictement limitée aux agglomérations de plus de 2 000 habitants échantillonnées. Une collecte sur internet permettrait de s'affranchir de ce cadre et d'avoir une meilleure représentation du littoral et des massifs montagneux. Par ailleurs, cet échantillon d'hôtels reste limité (4 % des hôtels en France).

— Limite n°4 – l'agrégation des évolutions par agglomération : les indices de Dutot calculés par agglomération sont ensuite agrégés selon le poids de la dépense de consommation alimentaire des ménages. La dépense de consommation des ménages en nuitées hôtelières est a priori différente de celle de la consommation alimentaire. À titre d'exemple, Paris représente 32 % des nuitées hôtelières en 2019 [8] et 22 % de la dépense de la consommation alimentaire en 2019.

— Limite n°5 – la représentativité du week-end et des vacances scolaires : la collecte des prix des hôtels sur le terrain est réalisée selon le calendrier IPC qui diffère du mois calendaire [9]. Concrètement, l'indice actuel ne permet pas de suivre correctement les tarifs du week-end (nuitée du samedi) et du dimanche puisque les enquêteurs ne collectent pas de prix le week-end. Par ailleurs, ce calendrier n'assure pas une prise en compte de l'ensemble des vacances scolaires (par exemple les prix des vacances de Noël ne sont pas relevés en décembre, les prix du mois d'août ne sont pas exhaustifs du fait d'une semaine de congés des enquêteurs).

— Limite n°6 – le cas des hôtels complets : la collecte des prix pour le jour même a l'inconvénient en haute saison de rencontrer des hôtels complets : dans ce cas, si l'hôtel est capable de donner le prix de la chambre suite à un éventuel désistement dans la journée ce prix est relevé sinon le prix est imputé.

5. D'autres modes de commercialisation ne sont pas évoqués : (i) la réservation dans le cadre de forfaits de voyages au sein desquels les nuitées sont assemblées avec d'autres services de voyage car ce segment de consommation serait classé dans un autre poste de la COICOP ; (ii) la réservation dans le cadre de dispositifs d'entreprises pour lesquels des tarifs préférentiels sont négociés par les entreprises et les administrations ne concerne pas le champ de l'IPC.

6. D'après une étude du cabinet Phocuswright[1], la part des réservations en ligne dans le chiffre d'affaires des hôteliers français est passée de 26 % en 2011 à 34 % en 2015.

7. En Europe, 70 % des réservations d'hôtels faites en ligne proviennent de plateformes, les 30 % restant étant des réservations effectuées directement sur les sites des hôtels [2].

8. 31 % des nuitées hôtelières hors clientèle d'affaires.

9. Chaque mois, l'IPC repose sur 20 jours de collecte terrain répartis sur les jours ouvrés de 4 semaines consécutives. Un mois calendaire comprend de 28 à 31 jours et ne correspond pas à un nombre entier de semaines. En conséquence, chaque année, l'Insee adapte le calendrier de collecte terrain (calendrier IPC) de façon à ce que les 48 semaines de collecte coïncident au mieux avec les mois du calendrier. Cette adaptation consiste à fixer des semaines sans collecte, en moyenne au nombre de 4 par an.

1.3 Le yield management, à l'oeuvre dans le secteur de l'hôtellerie, fait l'objet de recommandations internationales et d'expérimentations à l'aide du webscraping

1.3.1 Les recommandations internationales sur le yield management

De nouvelles formes de tarifications dynamiques (yield management ou revenue management ou encore tarification en temps réel) sont apparues dans les années 1980 dans le cas des compagnies aériennes aux États-Unis suite à la déréglementation du marché. Cette méthode, consistant à optimiser en temps réel les prix d'un service en fonction de la demande sur un segment de marché, est devenue de nos jours une technique marketing couramment utilisée dans les transports aériens ou ferroviaires, les locations de voiture, les spectacles et dans l'hôtellerie par exemple. Les conséquences positives de l'utilisation de ce concept sont à la fois du côté du producteur avec une hausse du chiffre d'affaires, et du côté du consommateur avec à la clé la possibilité d'une baisse des prix sans impact sur la qualité de service. Un tel ajustement des prix est un atout tactique essentiel des entreprises opérant dans un environnement fortement concurrentiel et dont le prix est la première variable de choix pour l'usager. Pour ces services, des variations de prix très fréquentes peuvent survenir, en fonction de la date du service. Les prix peuvent également varier suivant l'antériorité de l'achat. Ainsi, pour un même service consommé à un moment donné, le prix peut être différent, ce qui complique le calcul de l'indice des prix à la consommation, s'il se veut représentatif de l'ensemble de ces prix. Le manuel de l'indice des prix à la consommation du FMI[3] et le projet de recommandation d'Eurostat [4] portant surtout sur les transports aériens et séjours de vacances [10] mais dont la version préliminaire portait sur l'ensemble des services à prix volatils ou dont le prix dépend de l'antériorité de réservation [11] préconisent :

— de collecter des prix reflétant la volatilité des prix en répartissant les relevés sur le mois ;
— de constituer un échantillon de prix représentatif du comportement des acheteurs, en intégrant notamment l'antériorité de l'achat, les différents types de classes et de conditions de remboursement du billet ;
— d'intégrer dans l'IPC le prix collecté avec antériorité pour le mois au cours duquel la consommation du service débute et non pas lorsque le service est réservé.

1.3.2 Des expérimentations à l'aide du webscraping

Depuis plusieurs années, la collecte des prix sur Internet s'est développée dans les instituts nationaux statistiques afin de décrire la consommation des biens et services sur Internet. Par ailleurs, plusieurs pays, notamment la Belgique, les Pays-Bas, l'Italie, l'Allemagne ou le Royaume-Uni ont automatisé ces collectes par des programmes informatiques de webscraping en réduisant ainsi les coûts de collecte par rapport aux relevés physiques tout en obtenant un plus grand volume de données. Cette méthode de webscraping est particulièrement efficace dans le cadre :

— de tarifs ne dépendant pas de l'endroit où le service est acheté ;
— d'un nombre limité de sites Internet proposant ces services, ce qui permet de limiter le nombre de robots à développer ;
— de biens ou services dont la commande se fait en ligne via un formulaire car l'information collectée est structurée, les caractéristiques du produit collecté sont bien définies.

Face à l'importante croissante de cette technique, Eurostat a adopté en novembre 2020 un manuel de bonnes pratiques sur le webscraping [6]. Néanmoins, les avantages liés à cette technique sont à mettre au regard de certaines limites :

10. Ces recommandations figurent au chapitre 12.5 Flights and package holidays de la version définitive du manuel d'Eurostat de novembre 2018.[5]

11. « Recommendation on the treatment of tangible services purchased in advance and/or priced flexibly », mars 2018.

— l'importance de la communication des instituts statistiques auprès des sites pour éviter le blocage des adresses IP [12] ;
— le coût de collecte peut être très variable suivant la fréquence des changements du site internet de vente qui peuvent entraîner des maintenances régulières du robot de webscraping. Ces changements peuvent être facilement détectables en cas d'arrêt du robot de collecte [13], d'autres nécessitent une surveillance régulière des données pour repérer une variable manquante malgré une collecte entière des données [14].

En France, le webscraping est déjà utilisé par la direction générale de l'aviation civile (DGAC) pour calculer l'indice des prix du transport aérien et par la division des prix à la consommation pour calculer l'indice des prix du transport ferroviaire depuis 2020. Dans le cadre de l'hôtellerie, la Belgique recourt à cette technique depuis 2015. L'Italie et les Pays-Bas ont également mené des travaux sur ce sujet plus récemment.

12. Une lettre d'informations concernant l'enquête liée à la collecte des prix dans le cadre de l'IPC et le webscraping du site internet a été adressée à la plateforme de réservation en ligne

13. Un changement du site de la plateforme de réservation en novembre 2020 a empêché la collecte des prix pendant une semaine. Un autre changement important en Octobre 2021 a provoqué un arrêt de la collecte pendant une semaine .

14. Un changement du site depuis fin mars 2021 a empêché de recueillir les données sur les petits-déjeuners. Un changement sur le site en octobre 2021 a empêché de continuer à recueillir une variable de confirmation sur le nombre d'occupants des chambres ainsi qu'une variable permettant d'identifier uniquement chaque hôtel. Pour l'identifiant des hôtels, un autre a pu être reconstitué mais le même hôtel en a parfois plusieurs différents.

1.4 Le choix du webscraping de la plateforme

Les différentes limites, les recommandations et les expérimentations européennes ont conduit à envisager le calcul d'un indice des prix à partir du webscraping d'une large quantité de données d'une plateforme de réservation. La collecte de ces prix permettra de prendre en compte un canal de réservation différent, une réservation par anticipation du consommateur, de s'affranchir des agglomérations de l'IPC et du calendrier de collecte. Dans le cadre de ce travail exploratoire, nous avons choisi comme plateforme le premier site internet visité dans le domaine du voyage et du tourisme avec 13,7 millions de visites mensuelles et 6,5 millions de visites uniques [15]. Par ailleurs, plusieurs pays européens, dont l'Italie, la Belgique et les Pays-Bas, ont débuté plus ou moins récemment des travaux sur la même plateforme. Enfin, afin de mesurer un potentiel biais de couverture sur la plateforme, un appariement des hôtels 2 et 3 étoiles suivis dans le cadre de l'IPC métropolitain et les établissements webscrapés sur la plateforme a été mené : seuls 5 % des hôtels suivis dans le cadre de l'IPC ne sont pas retrouvés sur la plateforme. Ce proxy montre une bonne visibilité des hôtels sur la plateforme même si l'effet du contexte sanitaire est non mesurable (plus d'hôtels cherchant à être visibles sur la plateforme par exemple). L'indice des prix issus de ce choix n'a pas vocation à être représentatif de l'ensemble des différents modes de réservations. Une analyse sur d'autres plateformes pourra être menée par la suite (la Belgique a commencé à expérimenter le webscraping de la plateforme Expedia après avoir travaillé sur la plateforme Booking). Par ailleurs, les politiques tarifaires de la plateforme peuvent différer selon que l'hôtel ait rejoint ou non le programme de partenariat (Preffered Partner Program). En effet, les travaux de M. Cure, A. Cazaubiel, B. Johansen et T. Vergé [7] montrent que l'adhésion à ce type de programme conduit à une augmentation des ventes et des prix [16].

Limite des interprétations des données issues du webscraping en 2020 et en 2021

Les années 2020 et 2021 ont été marquées par diverses restrictions (confinements nationaux du 17 mars au 11 mai 2020, du 30 octobre au 15 décembre 2020, du 3 avril au 3 mai 2021 en métropole, confinements locaux, fermetures administratives, limitation des déplacements, couvre-feu) suite à la crise sanitaire. Le taux de collecte (part des observations normales ou pseudo-normales [a]) des prix du secteur de l'hôtellerie montre un impact important de la crise sanitaire sur la collecte des prix sur les périodes mars 2020 à juin 2021, le niveau d'avant crise n'étant toujours pas retrouvé.

TABLE 1 – Taux de collecte des prix des nuitées hôtelières en 2019, 2020, 2021

	2019	2020	2021
Janvier	89,4 %	86,6 %	56,7 %
Février	91,4 %	93,5 %	59,4 %
Mars	94,0 %	70,7 %	62,3 %
Avril	91,4 %	0,6 %	56,7 %
Mai	88,4 %	19,8 %	65,0 %
Juin	92,0 %	45,3 %	75,2 %

	2019	2020	2021
Juillet	90,0 %	72,5 %	83,0 %
Août	89,1 %	73,1 %	78,0 %
Septembre	92,0 %	83,8 %	81,6 %
Octobre	91,2 %	85,9 %	81,3 %
Novembre	88,4 %	56,6 %	79,6 %
Décembre	87,6 %	52,7 %	79,2 %

Source :Relevés terrain de l'IPC.

Les évolutions de prix obtenues à l'aide du webscraping de la plateforme de réservation en ligne seront donc à interpréter avec précaution.

a. Le produit n'est pas en rayon mais son prix est toujours affiché à l'emplacement habituel de sa vente. Dans le cas de l'hôtellerie, le code de pseudo-observation a été employé lors de la crise sanitaire pour les relevés par téléphone ou par internet (uniquement sur le site internet de l'hôtel).

15. Analyse de trafic internet et webmobile du site internet similarweb.com

16. Outre le niveau des prix, l'évolution de ces derniers pourrait, a priori, également différer.

2 Construction d'une base de données à partir du webscraping

La collecte des données par webscraping a commencé par être mise en place suivant deux approches : une collecte sur France entière restreinte par des filtres sur les options et une collecte sans filtre sur une zone géographique restreinte.

2.1 Définition d'un protocole de collecte

Le choix a été fait de développer le robot de webscraping de la plateforme à l'aide du langage Python car il s'agit d'un bon compromis entre rapidité d'exécution et accessibilité à des statisticiens n'étant pas experts en informatique.

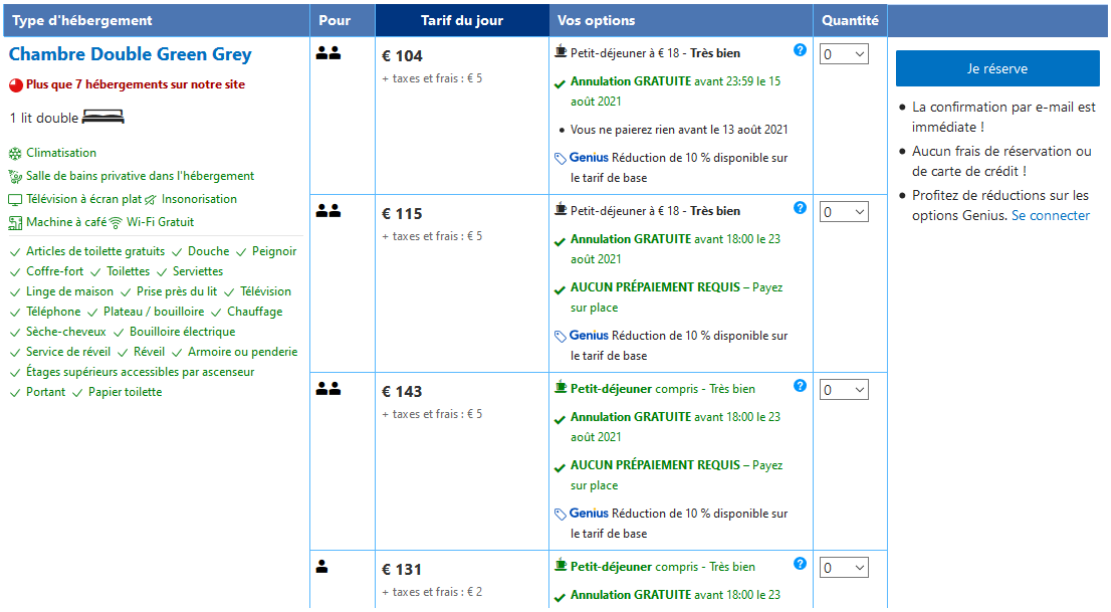
Le programme recueille les données pour une destination donnée pour un séjour dans un hôtel (filtre spécifique sur la plateforme) de deux adultes d'une nuit à 0, 30 et 60 jours d'antériorité, ainsi que 10 et 20 jours depuis décembre 2021. Il envoie une requête à l'url de la plateforme de réservation en ligne avec la destination, la date et les conditions choisies pour la chambre et il reçoit en retour un fichier au format HTML décrivant une liste d'hôtels (cf. illustration n°3) dont il faut extraire l'information pertinente.

FIGURE 3 – Exemple de résultat d'une requête pour obtenir les hôtels disponibles sur le site de la plateforme

Note : Cette requête permet de visualiser tous les hôtels disponibles la nuit du 23 août à Nantes pour 2 adultes. Pour chaque hôtel, une seule chambre est mise en avant par la plateforme. C'est celle retenue dans l'approche n°1.

Les résultats obtenus sur la première page avec la liste d'hôtels pour une destination contiennent notamment le lien pour accéder à la page spécifique de chacun de ces hôtels (cf. illustration n°4). Sur la page de l'hôtel, il est possible d'afficher toutes les chambres proposées, avec leurs différentes conditions.

FIGURE 4 – Exemple de résultat d'une requête pour obtenir toutes les chambres disponibles sur la plateforme



Note : Cette requête permet de visualiser toutes les chambres disponibles pour l'hôtel Maisons du monde Hotel & Suites – Nantes pour la nuit du 23 août.

Afin de minimiser le nombre de requêtes sur la plateforme, deux approches ont été envisagées :

— Approche n°1 – « collecte hôtels avec filtres » : le programme collecte des hôtels sur tout le territoire (métropole et DOM) avec l'utilisation de filtres (petit-déjeuner inclus, annulation gratuite). Cela signifie que toutes les chambres d'un hôtel ne sont pas collectées mais uniquement celle mise en avant par la plateforme de réservation en ligne (on récupère uniquement les résultats de la première page). On peut supposer que la chambre ainsi présélectionnée est la plus vendue en général. Cette collecte a débuté à partir du 19 octobre 2020 pour un grand nombre de régions, puis pour la France entière à partir du 22 décembre 2020. à la date du 5 août 2021, ce sont 1 038 requêtes qui ont été envoyées sur la plateforme de réservation en ligne.

— Approche n°2 - « collecte hôtels sans filtre avec chambre » : le programme collecte l'intégralité des chambres proposées sur 4 zones (Morbihan, Savoie, Avignon et Paris 5ème). Cette approche nécessite une collecte de l'ensemble des hôtels de ces zones en amont pour relancer une requête sur la page de chacun de ces hôtels. Concrètement, pour ces quatre zones, cela peut entraîner jusqu'à 25 fois plus de requêtes envoyées sur la plateforme de réservation en ligne (une page de résultats contenant au maximum 25 hôtels) : d'où la nécessité de limiter géographiquement cette collecte. L'approche n°2 est complétée par une collecte de l'ensemble des hôtels en première page sur l'ensemble du territoire, ce qui permet de constituer un référentiel d'hôtels. Cette collecte a débuté à partir du 13 octobre 2020. À la date du 2 août 2021, ce sont 1 981 requêtes envoyées sur la plateforme de réservation en ligne pour la première étape sur l'ensemble du territoire puis 1 237 requêtes adressées au site internet pour la deuxième étape de recherche de toutes les chambres disponibles sur 4 zones.

Le robot a été dans un premier temps lancé manuellement uniquement en semaine en octobre 2020 puis a été lancé quotidiennement grâce aux outils mis à disposition par la division innovation et instruction technique (DIIT) de l'Insee (plateforme Onyxia) à partir de novembre 2020. Une « image Docker » de ce robot, contenant l'ensemble des librairies nécessaires et le fichier des

destinations en entrée du programme est créée. Cette image est exécutée chaque nuit afin de recueillir les prix des nuitées hôtelières suivant les deux processus en alternance (chaque approche est recueillie un jour sur deux) et les trois antériorités. La plateforme Innovation a été fermée et ses outils migrés sur la plateforme du SSPCloud [17]. Il y a eu un période transitoire de collecte manuelle entre le 27 mai et le 30 août. Le robot a été migré vers la plateforme du SSPCloud avec l'aide de la DIIT, que nous remercions grandement et est maintenant en open source.

Les résultats obtenus sur la première page avec la liste d'hôtels pour une destination contiennent notamment le lien pour accéder à la page spécifique de chacun de ces hôtels (cf. illustration n°2). Sur la page de l'hôtel, il est possible d'afficher toutes les chambres proposées, avec leurs différentes conditions. Les données extraites de ces deux approches sont présentées dans le tableau suivant (cf. tableau n°2).

TABLE 2 – Principales variables collectées à l'aide du robot développé en Python

Approche n°1 - Base hôtels avec filtres	Approche n°2 - Base hôtels avec chambres
Rang de la page	
Nom de l'hôtel	Nom de l'hôtel
	Identifiant de l'hôtel
Nombre d'étoiles	Nombre d'étoiles
Notation clients	
Modalité d'annulation (ici filtre sur l'annulation gratuite)	Modalités d'annulation
Prépaiement requis	Prépaiement requis
Petit-déjeuner (et autres repas inclus ou non – ici, le petit-déjeuner étant forcément inclus)	
Nom de la chambre	Nom de la chambre
	Identifiant de la chambre
Capacité de la chambre	Capacité de la chambre
Prix de la chambre	Prix de la chambre
	Prix du petit déjeuner
Prix des taxes (taxe de séjour et frais de réservation)	

17. https ://datalab.sspcloud.fr

2.2 Comparaisons du champ des hôtels webscrapés selon les deux approches, l'enquête de fréquentation dans les hébergements touristiques et l'échantillon d'hôtels de l'IPC

La collecte automatisée des hôtels sur l'ensemble du territoire selon les deux approches permet de constituer deux référentiels d'établissements. Afin d'étudier d'éventuels biais de couverture, ces deux référentiels peuvent être comparés à celui constitué grâce à l'enquête mensuelle de fréquentation dans les hébergements collectifs touristiques menée par l'Insee. La mise à jour du parc d'établissements est effectuée par les gestionnaires de l'enquête en continu lors de l'enquête et ponctuellement en s'appuyant sur d'autres sources, notamment Sirene, Atout France, et les comités régionaux du tourisme. Le pôle tourisme de l'établissement de Montpellier a transmis une extraction de ce référentiel dans le cadre de ces travaux à date du 22 mars 2021 qui comporte 17 909 établissements.

Avant de pouvoir comparer ces quatre sources de données, un premier nettoyage commun aux deux bases de données issues du webscraping a été réalisé. Tout d'abord une restriction du champ des observations est opérée afin de se limiter à des hôtels [18] (le filtre présent sur la plateforme n'étant pas suffisant) et de retirer les pensions ou demi-pensions (cf. tableau n°3).

Table 3 – Suivi des deux opérations de nettoyage des bases de données webscrapées

	Approche n°1 - Base hôtels avec filtres	Approche n°2 - Base hôtels avec chambres
Suppression des lignes relatives à des établissements qui ne sont pas des hôtels	84 725	347 169
Suppression des lignes relatives à des pensions ou demi-pensions	37 379	74 067
Nombre total final d'observations	1 100 676	2 719 009
Nombre total final d'établissements	6 226	15 777

Source : Bases de données issues du webscraping de la plateforme, à la date du 30 juillet 2021.
Champ : France entière.

Par ailleurs, une variable commune est également créée afin d'identifier le mode d'exploitation de l'hôtel (chaîne ou indépendant) à partir de l'analyse textuelle des libellés des hôtels afin d'isoler les chaînes. Cette variable peut donc avoir tendance à sous-estimer les chaînes dans les bases de données si une chaîne de caractères est mal identifiée. Du fait du mode de construction de cette variable, cela nécessite une veille pour identifier les cas d'apparition de nouvelles franchises et ainsi mettre à jour cette variable.

La comparaison régionale des établissements montre qu'il y a peu de différence entre le référentiel tourisme et l'ensemble des hôtels collectés sans utilisation de filtre (au maximum 1,5 point d'écart pour l'Auvergne-Rhône Alpes) tandis que des différences importantes sont présentes entre le référentiel tourisme, l'échantillon de l'IPC et l'ensemble des hôtels collectés avec utilisation de filtres. En effet, le recours aux filtres petit-déjeuner inclus et annulation gratuite conduit à sur-représenter l'Île-de-France (25,8 % des établissements contre 14,8 % dans le référentiel tourisme, cf. tableau n°4)

18. Analyse textuelle sur les noms de chambres et noms d'hôtels sur les mots : studio, loft, dortoir, roulotte, maison, duplex, bungalow, appartement, auberges et gîtes.

TABLE 4 – Comparaison de la proportion d'établissements distincts selon les régions pour les bases de données webscrapées avec filtres, sans filtre et le référentiel tourisme

	Enquête de fréquentation touristique	IPC	Base avec filtres	Base sans filtre
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 : Bases de données issues du webscraping, à la date du 30 juillet, référentiel tourisme à la date du 22 mars.
Champ : France métropolitaine.

La comparaison selon les modalités d'exploitation montre que sans l'utilisation de filtres, la plateforme a tendance à référencer plus d'hôtels indépendants (72,2 % contre 63,4 % des hôtels dans le référentiel tourisme). Le recours aux filtres petits-déjeuners inclus et annulation gratuite conduit, au contraire, à sur-représenter les chaînes (49,1 % contre 36,6 % des hôtels dans le référentiel tourisme, cf. tableau n°5)

TABLE 5 – Comparaison de la proportion d'établissements distincts selon la modalité d'exploitation pour les bases de données webscrapées avec filtres, sans filtre et le référentiel tourisme

Modalité d'exploitation	Référentiel tourisme	Base sans filtre	Base avec filtre
Chaîne	36,6 %	27,8 %	49,1 %
Indépendant	63,4 %	72,2 %	50,9 %

Source : Bases de données issues du webscraping, à la date du 30 juillet 2021, référentiel tourisme à la date du 22 mars 2021.
Champ : France entière.

La comparaison selon le nombre d'étoiles montre que sans l'utilisation de filtres, la plateforme a tendance à référencer un peu plus d'hôtels 3 étoiles (4,4 points d'écart par rapport au référentiel tourisme), 4 étoiles (2,6 points d'écart) et 2 étoiles (1,7 point d'écart) et moins d'hôtels non classés (-9,0 point d'écart). Le recours aux filtres petit-déjeuner inclus et annulation gratuite exacerbe ces écarts et conduit à sur-représenter nettement les hôtels 3 et 4 étoiles au détriment des hôtels non classés. À noter qu'avec les filtres, les hôtels 2 étoiles sont moins représentés (cf. tableau n°6).

TABLE 6 – Comparaison de la proportion d'établissements distincts selon le classement hôtelier pour les bases de données webscrapées avec filtres, sans filtre et le référentiel tourisme

Classement hôtelier	Référentiel tourisme	Base sans filtre	Base avec filtre
Non classé	27,7 %	18,7 %	10,6 %
1 étoile	2,2 %	2,2 %	2,4 %
2 étoiles	22,0 %	23,7 %	14,2 %
3 étoiles	34,2 %	38,6 %	44,5 %
4 étoiles	11,6 %	14,2 %	24,3 %
5 étoiles	2,4 %	2,5 %	4,1 %

Source : Bases de données issues du webscraping, à la date du 30 juillet 2021, référentiel tourisme à la date du 22 mars 2021.
Champ : France entière.

Ces comparaisons montrent donc l'existence d'un biais de couverture des hôtels présents sur le territoire par rapport à ceux référencés sur la plateforme de réservation en ligne. Ce biais est d'autant plus important en recourant aux webscraping à l'aide de filtres (petit-déjeuner inclus, annulation gratuite). Ces comparaisons gagneraient à être améliorées en réalisant une collecte de tous les hôtels de la plateforme en s'affranchissant d'une date de réservation. En effet, selon les mois de l'année, certains hôtels peuvent être fermés par exemple. Il faudrait prévoir le webscraping des hôtels France entière sans imposer de date. Il sera également intéressant de comparer ces données aux données de plusieurs plateformes de réservations récupérées par Eurostat. Enfin, cette comparaison des différents référentiels pourrait aussi servir aux statistiques d'entreprises. [19]

Dans la suite du rapport, seule l'approche n°1 - la base des hôtels avec filtres (i.e. hôtels sur le périmètre France entière avec une chambre mise en avant par la plateforme de réservation en ligne, avec annulation gratuite et petit-déjeuner inclus) sera étudiée. La collecte un jour sur deux pour chaque approche a été arrêtée en septembre pour passer à une collecte quotidienne avec filtres.

19. Par exemple, l'Italie a apparié les données issues du webscraping de la plateforme webscrapée avec le registre administratif des établissements d'hébergement touristique sur la seule région Emilie-Romagne. L'objectif est de compléter et d'enrichir les informations sur les établissements d'hébergement touristique déjà enquêtés dans le cadre des enquêtes statistiques classiques mais aussi de comprendre le degré de couverture de ces enquêtes. Les travaux des Italiens [8], [9] montrent notamment que la couverture des villages touristiques (100 %) et des maisons louées (78,7 %) sur la plateforme de réservation en ligne est quasi-totale contrairement aux campings (13 %), auberges (24,3 %) et chambres louées (29,9 %). Enfin certains types d'établissements n'apparaissent que sur la plateforme de réservation en ligne : chalets, bateaux, auberges, lodges, motels, villas, hébergements chez l'habitant.

2.3 Nettoyage et enrichissement de la base de données des nuitées hôtelières avec petit déjeuner inclus et annulation gratuite présélectionnées par la plateforme de réservation en ligne

Afin de permettre sa pleine exploitation, certaines variables de la base de données avec filtres, ont été nettoyées afin de n'extraire que l'information chiffrée du champ webscrapé. Il s'agit des variables prix, taxes, nombre maximal d'occupants et capacité de la chambre. Enfin, le champ des observations est réduit aux seules chambres dont la capacité est de deux personnes [20] aux observations associées à un prix et aux dates de séjour à partir de décembre 2020. Ainsi, la base de données exploitable est constituée de 1 625 328 observations pour 5980* hôtels à la date du 31 décembre 2021.

TABLE 7 – Répartition du nombre d'observations et du nombre d'hôtels selon les mois de la nuitée entre décembre 2020 et décembre 2021

Mois	Nombre d'observations	Nombre d'hôtels
Décembre 2020	32 119	2 802
Janvier 2021	59 048	4 151
Février 2021	85 255	4 575
Mars 2021	127 326	4 788
Avril 2021	149 570	4 996
Mai 2021	163 489	5 053
Juin 2021	148 015	5 219
Juillet 2021	142 217	5 286
Août 2021	109 384	5 195
Septembre 2021	137 650	5 367
Octobre 2021	138 380	5 304*
Novembre 2021	144 027	5 088*
Décembre 2021	144 836	4 376*

Source : Base avec filtres issue du webscraping, à la date du 31 décembre 2021.
Champ : France entière.
** : Les comptages pour les mois d'octobre, novembre et décembre du nombre d'hôtel sont en réalité sous évalués du fait de la disparition de l'identifiant collecté suite à un changement de la plateforme mi octobre. Il y a eu un réappareillement pour les hôtels déjà apparus lors des mois précédents mais les nouveaux entrants ne sont pas exploitables pour l'instant.*

20. Cela n'a pu être fait que pour les données jusqu'à la mi octobre, la variable utilisée ne pouvant plus être collectée depuis

3 Analyse des prix collectés par webscraping

Cette partie, en plus de dresser un rapide panorama des prix des chambres, propose deux études sur les prix des nuitées hôtelières collectés par webscraping à l'aide de filtres : une première sur l'analyse des déterminants de prix et une seconde analyse sur les seules observations (hôtels x chambres) [21] dont les prix sont collectés à 0, 30 et 60 jours d'antériorité afin d'identifier les principaux profils d'évolution des prix et de confirmer ou non les antériorités sélectionnées en amont de la collecte.

3.1 Panorama des prix moyens des chambres entre décembre 2020 et décembre 2021

Les chambres mises en avant par la plateforme de réservation en ligne avec le petit-déjeuner inclus et annulation gratuite sont proposées en moyenne à 128,5 € sur la période de décembre 2020 à décemnbre 2021. Les prix s'échelonnent entre 11 € et 10 025 €. Plus l'hôtel a d'étoiles, plus le prix de la chambre est élevé en moyenne. Les hôtels non classés forment une catégorie d'hôtels hétérogènes et les prix des chambres sont en moyenne équivalents à ceux des 2 étoiles (cf. tableau n°8).

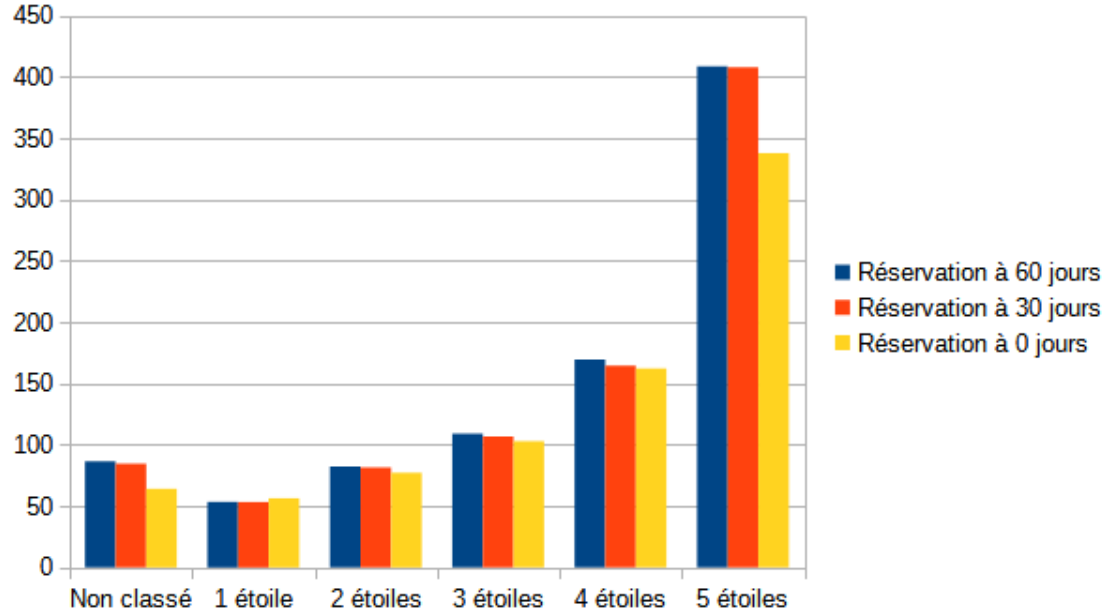
TABLE 8 – Comparaison des prix moyens des hôtels x chambres selon le classement hôtelier

Classement hôtelier	Non classé	1 étoile	2 étoiles	3 étoiles	4 étoiles	5 étoiles
Prix moyen (en €)	81	54	81	107	166	402

Source : Base avec filtres issue du webscraping, à la date du 31 décembre.
Champ : France entière.
Note : les prix moyens sont calculés à partir d'une moyenne géométrique. En tenant compte des antériorités de réservation, les prix moyens ont tendance à diminuer plus on se rapproche d'une réservation pour le jour même quel que soit le classement de l'hôtel sauf pour les hôtels 1 étoile (cf. graphique n°5).

21. Il s'agit du nom de la chambre mise en avant par la plateforme pour les hôtels webscrapés.

FIGURE 5 – Évolution des prix moyens des hôtels x chambres selon le classement hôtelier et l'antériorité de réservation



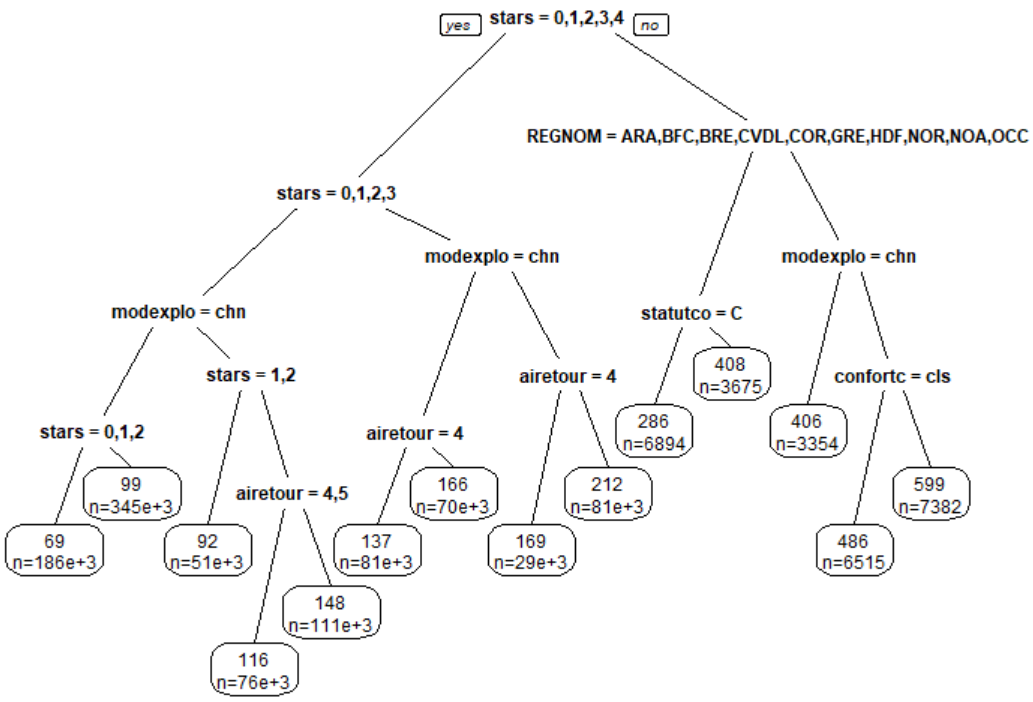
Source : Base avec filtres issue du webscraping, à la date du 31 décembre.
Champ : France entière.
Note : les prix moyens sont calculés à partir d'une moyenne géométrique.

Le classement hôtelier selon le nombre d'étoiles est le facteur le plus discriminant du prix des nuitées d'après l'arbre de décision construit sur la base de données avec filtres [22] avec les données jusqu'au 30 juillet 2021. Cet arbre (Figure 6) a été construit en utilisant le seuil de 500 observations minimum comme critère d'arrêt . Cet arbre met en évidence la spécificité des hôtels 5 étoiles et plus encore ceux situés en Île-de-France, Provence-Alpes-Côte d'Azur et Pays de la Loire. D'autres critères sont fortement discriminants :

- le modèle d'exploitation de l'hôtel : les hôtels indépendants offrent des chambres en moyenne à un prix plus élevé (144 €) que les chaînes (96 €) ;
- l'aire touristique en isolant souvent l'urbain de province (modalité « 4 ») et les autres espaces (modalité « 5 ») : les hôtels situés dans une zone urbaine de province offrent des chambres en moyenne à un prix moins élevé (95 €) que ceux situés en littoral (123 €) ou dans les massifs montagneux (145 €) ;
- la région : Les prix des chambres selon la région varie du simple au double en moyenne. Les prix sont plus élevés en moyenne en Corse (142 €) et en Île-de-France (139 €). À l'inverse, les prix en Hauts de France (91 €), Centre Val de Loire (91 €) sont plus bas en moyenne ;
- le statut de la commune : les hôtels situés dans une commune centre offrent des chambres en moyenne à un prix plus élevé (120 €) que les communes en banlieue (94 €) ;
- le confort de la chambre : les chambres supérieures sont en moyenne proposées à 152 € alors que les chambres classiques sont en moyenne proposées à 105 € ;

22. L'arbre de décision est utilisé ici uniquement dans une approche exploratoire en analysant les conditions de décision présentes sur chaque nœud.

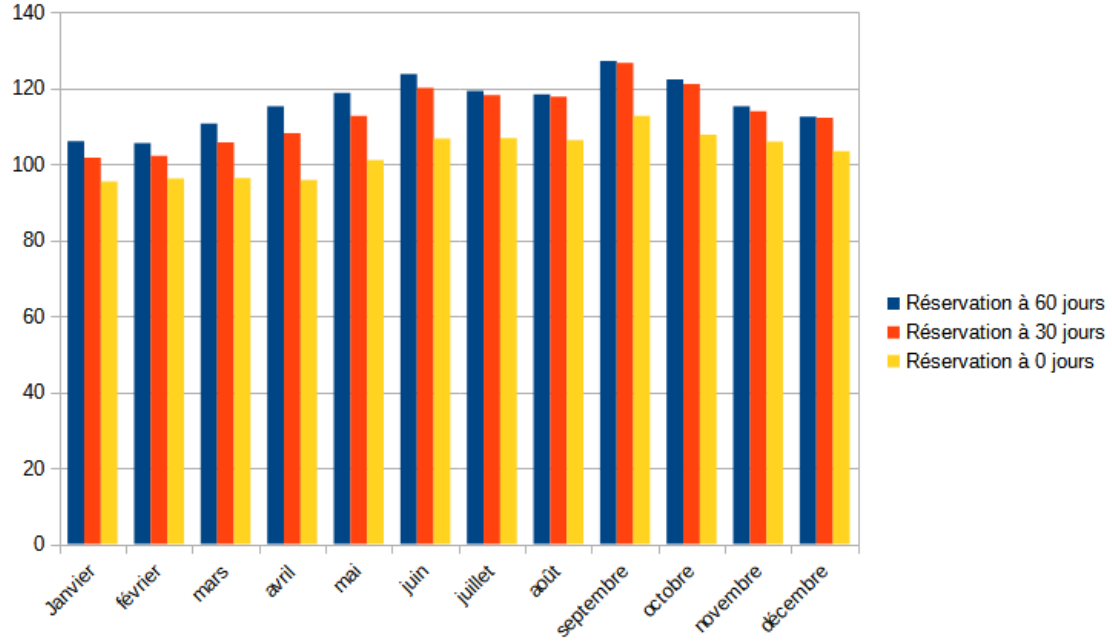
FIGURE 6 – Arbre de décision dont la variable d'intérêt est le prix des nuitées hôtelières



Source : Base avec filtres issue du webscraping, à la date du 30 juillet 2021. Champ : France entière.

Les prix augmentent en moyenne pour les nuitées entre décembre et juin quelle que soit l'antériorité de réservation, alors qu'ils baissent en moyenne pour les nuitées ayant lieu en juillet et août. Par la suite, il y a une nouvelle hausse entre août et septembre suivie de baisse jusqu'en décembre. Par ailleurs les prix ont tendance à baisser en moyenne plus on s'approche de la date de la nuitée (cf. graphique n°7).

FIGURE 7 – Évolution des prix moyens des hôtels x chambres selon les mois civils de la nuitée et l'antériorité de réservation



Source : Base avec filtres issue du webscraping, à la date du 31 décembre.
Champ : France entière.
Note : les prix moyens sont calculés à partir d'une moyenne géométrique.

3.2 Analyse des déterminants des prix des chambres à l'aide d'un modèle hédonique

L'analyse des déterminants des prix des chambres a consisté à étudier grâce à un modèle hédonique les principaux facteurs ayant une influence sur le prix des nuitées.

Un modèle hédonique des prix

Le modèle hédonique des prix est un modèle linéaire qui lie le prix d'un bien ou d'un service à des caractéristiques. La méthode des prix hédoniques[10] part du principe que le prix d'un produit dépend de ses caractéristiques.
Soient p_n^t le prix du service n à la période t, K le nombre de caractéristiques mesurées par les $z_{n,k}^t$, β_0^t et β_k^t respectivement la constante et les paramètres des caractéristiques à estimer, un modèle hédonique des prix se présente de la manière suivante sous forme log-linéaire :
$\log p_n^t = \beta_0^t + \sum_{k=1}^{K} \beta_k^t z_{n,k}^t + \epsilon_n^t$

Le prix de la nuitée (plus exactement le log(prix)) est modélisé avec les variables explicatives suivantes :

- des variables liées à l'implantation de l'hôtel : région, département, aire touristique, statut de la commune ;
- des variables liées à l'hôtel : nombre d'étoiles de l'hôtel, mode d'exploitation (chaîne ou indépendant), confort de la chambre
- des variables liées au calendrier : jour de la nuitée, mois de la nuitée, vacances scolaires ou non, jour férié ou non, antériorité de la réservation

Les modèles sont construits sur deux bases de données au 30 juillet 2021 : celles comportant les

réservations à 30 et 60 jours d'antériorité (792 289 observations) et celles comportant toutes les réservations (à 60, 30 jours d'antériorité et pour le jour-même, soit 897 896 observations).

Enfin, les modèles sont construits sur l'ensemble des deux bases de données puis sur des échantillons de 10 000 observations tirées aléatoirement dans chacune des bases 1 000 fois pour s'assurer de la robustesse des estimations des paramètres. Le modèle hédonique est construit à l'aide d'une méthode de sélection de variables pas à pas (stepwise). Elle part du modèle vide et à chaque étape, lorsque la variable qui conduit à l'AIC le plus faible a été ajoutée au modèle, elle va enlever la variable du modèle qui fait décroître au maximum l'AIC, si une telle variable existe. L'algorithme s'arrête lorsque l'ajout d'aucune variable ne fait décroître l'AIC. Le modèle retenu par cette méthode retient l'ensemble des variables testées à l'exception de la région. Cette sélection de variables est également obtenue avec la méthode ascendante (forward) ou la méthode descendante (backward). Ces modèles présentent néanmoins de l'hétéroscédasticité qui est corrigée par une estimation des paramètres par la méthode des moindres carrés généralisés. Pour des raisons de simplicité, seuls les coefficients issus de la régression avec la variable région à la place de la variable département sont présentés ici (cf. figure n°8), les autres sont en annexe (cf. tableau n°24 et 25 en Annexe).

L'analyse des coefficients des régressions montre que :

- les hôtels 1 étoile se distinguent avec des prix plus bas que les hôtels non classés, toutes choses égales par ailleurs ; les hôtels 3, 4 et 5 étoiles se distinguent avec des prix plus élevés ;
- les hôtels indépendants pratiquent des prix plus élevés que les chaînes, toutes choses égales par ailleurs ;
- les nuitées sont moins chères en période de vacances scolaires ou en jour férié, toutes choses égales par ailleurs. Les prix de ces nuitées sont également plus bas le vendredi, samedi par rapport au dimanche qu'en semaine. Ces effets semblent montrer une différenciation de prix possible selon la clientèle ciblée (professionnelle ou pour raison personnelle) ;
- les prix des nuitées sont plus élevés pour les réservations à 60 jours d'antériorité par rapport à 30 jours d'antériorité ; ils sont plus bas pour les réservations à 30 jours d'antériorité par rapport au prix des réservations pour le jour-même et les prix des réservations à 60 jours d'antériorité sont légèrement plus élevés que ceux des réservations pour le jour-même, toutes choses égales par ailleurs ;
- les prix des nuitées d'avril à août sont plus élevés que les prix de décembre toutes choses égales par ailleurs ;
- les prix des nuitées dans les communes centre, isolées et hors unités urbaines sont plus élevés qu'en banlieue, toutes choses égales par ailleurs. Les prix des nuitées dans les massifs de montagnes sont plus élevés qu'en Île-de-France, toutes choses égales par ailleurs, alors que les prix sur le littoral, sur les espaces urbains de province et les autres espaces sont plus bas qu'en Île-de-France. Enfin, les prix sont plus élevés en Provence-Alpes-Côte d'Azur (PACA) par rapport à la région Auvergne Rhône-Alpes (ARA), toutes choses égales par ailleurs. À l'inverse, les prix dans les Pays de la Loire, Centre Val de Loire, Hauts de France, Occitanie, Grand Est, Bretagne, Bourgogne Franche-Comté sont plus bas que la région ARA [23].

23. L'analyse avec la variable département au lieu de la variable région montre que les prix dans les départements de Paris, des Hauts de Seine, des Yvelines, de Seine et Marne, du Vaucluse, de Haute-Saône, du Var sont plus élevés que dans l'Aube, toutes choses égales par ailleurs. À l'inverse, les prix dans les départements du Territoire de Belfort, de la Creuse et du Gers sont plus bas. Les exemples donnés sont ceux qui ont les coefficients les plus extrêmes en valeur absolue.

FIGURE 8 – Principaux résultats des régressions testées (R^2 ajusté, test d'hétéroscédasticité)

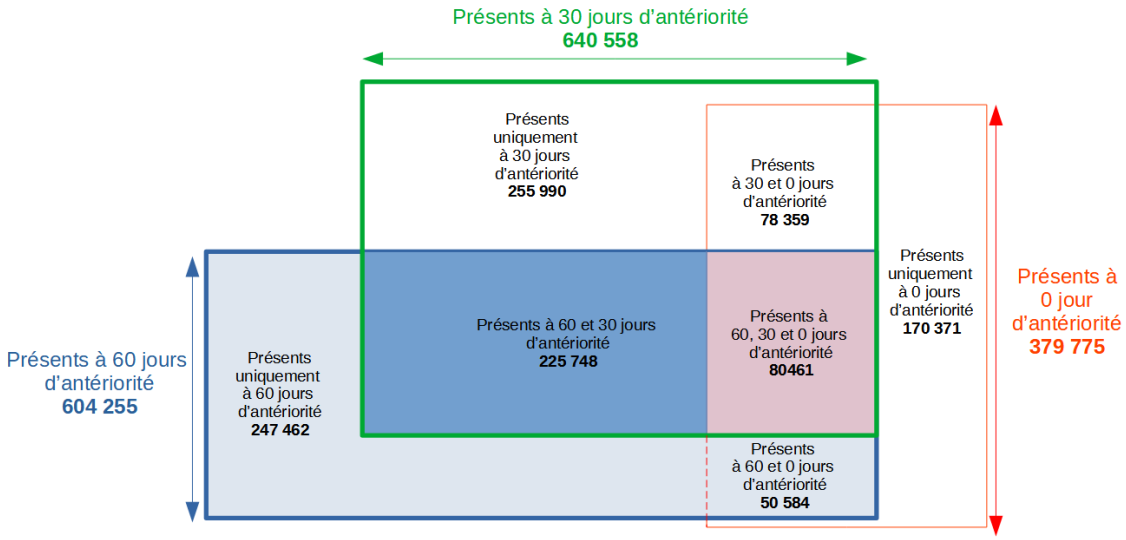
	Base avec les antériorités 30 et 60 jours		Base avec les antériorités 0, 30 et 60 jours	
	Ensemble de la base	Base échantillonnée 1000 fois	Ensemble de la base	Base échantillonnée 1000 fois
Nombre d'observations	792 289	10 000	897 896	10000
Variable : département				
R² ajusté	0,6823		0,6891	
Test de Breusch-Pagan (H0 : homoscédasticité des erreurs)	Rejet de H0 Présence d'hétéroscédasticité	Rejet de H0 Présence d'hétéroscédasticité pour tous les échantillons	Rejet de H0 Présence d'hétéroscédasticité	Rejet de H0 Présence d'hétéroscédasticité pour tous les échantillons
Test de Goldfeld et Quandt (H0 : homoscédasticité des erreurs)	Rejet de H0 Présence d'hétéroscédasticité	Rejet de H0 Présence d'hétéroscédasticité pour tous les échantillons	Rejet de H0 Présence d'hétéroscédasticité	Rejet de H0 Présence d'hétéroscédasticité pour tous les échantillons
Test de White (H0 : homoscédasticité des erreurs)	Rejet de H0 Présence d'hétéroscédasticité	Rejet de H0 Présence d'hétéroscédasticité pour tous les échantillons	Rejet de H0 Présence d'hétéroscédasticité	Rejet de H0 Présence d'hétéroscédasticité pour tous les échantillons
Variable : région				
R² ajusté	0,6985		0,7047	
Test de Breusch-Pagan (H0 : homoscédasticité des erreurs)	Rejet de H0 Présence d'hétéroscédasticité	Rejet de H0 Présence d'hétéroscédasticité pour tous les échantillons	Rejet de H0 Présence d'hétéroscédasticité	Rejet de H0 Présence d'hétéroscédasticité pour tous les échantillons
Test de Goldfeld et Quandt (H0 : homoscédasticité des erreurs)	Rejet de H0 Présence d'hétéroscédasticité	Rejet de H0 Présence d'hétéroscédasticité pour tous les échantillons	Rejet de H0 Présence d'hétéroscédasticité	Rejet de H0 Présence d'hétéroscédasticité pour tous les échantillons
Test de White (H0 : homoscédasticité des erreurs)	Rejet de H0 Présence d'hétéroscédasticité	Rejet de H0 Présence d'hétéroscédasticité pour tous les échantillons	Rejet de H0 Présence d'hétéroscédasticité	Rejet de H0 Présence d'hétéroscédasticité pour tous les échantillons

Source : Base avec filtres issue du webscraping de la plateforme de réservation en ligne, à la date du 30 juillet 2021.
Champ : France métropolitaine.

3.3 Analyse des antériorités choisies

Dans cette partie, seules les observations (hôtels x chambres) dont les prix sont collectés à 0, 30 et 60 jours d'antériorité pour un même jour sont étudiées, cela concerne 80 461 observations (cf. illustration n°9).

FIGURE 9 – Répartition des hôtels x chambres selon l'antériorité de réservation



Source : Base avec filtres issue du webscraping, à la date du 31 décembre 2021.
Champ : France entière.

66 % des hôtels x chambres ont un prix identique que la réservation ait été effectuée à 60 ou à 30 jours d'antériorité, alors que seulement 22 % ont un prix identique pour une réservation à 30 jours et à 0 jours. 16 % des hôtels x chambres sont à l'intersection de ces deux groupes, c'est-à-dire que ces hôtels pratiquent un prix stable quelle que soit l'antériorité de réservation de la nuitée (cf. tableau n°9). À priori l'antériorité de réservation à 60 jours semble moins pertinente mais demandera une analyse hors période de crise sanitaire.

TABLE 9 – Répartition des observations selon leur profil de tarification

	Répartition des observations
Prix identiques à 60 jours, à 30 jours et à 0 jour	16 %
Prix identiques à 60 et 30 jours puis à la baisse à 0 jour	10 %
Prix identiques à 60 et 30 jours puis à la hausse à 0 jour	40 %
Prix à la baisse à 30 jours puis à la baisse à 0 jour	4 %
Prix à la baisse à 30 jours puis à la hausse à 0 jour	12 %
Prix à la baisse à 30 jours puis identiques à 0 jour	3 %
Prix à la hausse à 30 jours puis à la hausse à 0 jour	8 %
Prix à la hausse à 30 jours puis à la baisse à 0 jour	3 %
Prix à la hausse à 30 jours puis identiques à 0 jour	3 %

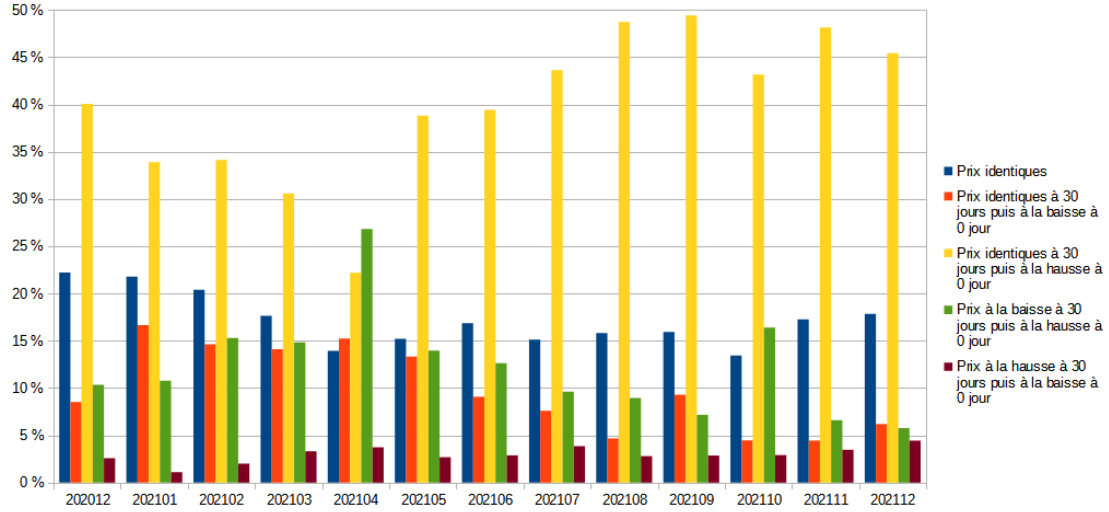
Source : Base avec filtres issue du webscraping, à la date du 31 décembre 2021.
Champ : France

Cette étude montre que les prix ont tendance à augmenter pour une réservation pour le jour-même par rapport à celle effectuée à 30 jours (60 % des observations), mais peu sont en

hausse à la fois entre 60 et 30 jours puis entre 30 et 0 jours d'antériorité (8 %). À l'inverse, 17 % des observations présentent une baisse de prix entre une réservation effectuée à 30 jours et une réservation effectuée le jour-même.

Par ailleurs, cette hausse des prix entre 30 et 0 jours a tendance à se vérifier plus amplement à partir de mai, notamment pour le profil des hôtels x chambres qui pratiquaient un prix stable entre une réservation à 60 et 30 jours d'antériorité puis une hausse pour une réservation le jour-même (32 % en moyenne chaque mois entre décembre à avril, 40 % entre mai et juillet et 47 % entre août et décembre, cf. graphique n°10). Pour rappel, le 3^{eme} confinement a eu lieu du 3 avril au 3 mai 2021 en France, ce qui peut expliquer le caractère atypique du mois d'avril, qui mériterait d'être analysé sur une autre année. À l'inverse le profil des hôtels x chambres pratiquant un prix stable lors d'une réservation à 60 et à 30 jours d'antériorité puis une baisse pour une réservation le jour-même semble être moins présent dernièrement (14 % en moyenne chaque mois entre décembre à avril, 10 % entre mai et juillet, et seulement 6 % entre août et décembre). Enfin le profil des hôtels x chambres pratiquant une stabilité des prix quelle que soit l'antériorité est à peu près stable au cours du temps (19 % en moyenne chaque mois entre décembre et avril, 15 % entre mai et juillet et 16 % entre août et décembre).

FIGURE 10 – Évolution de la part de certains profils de tarification au cours des mois civils de la nuitée



Source : Base avec filtres issue du webscraping, à la date du 31 décembre 2021.
Champ : France entière, uniquement les observations (hôtels x chambres) présentes aux trois antériorités de réservation.

Cette approche sur les seules observations (hôtels x chambres) présentes aux trois antériorités de réservation conduit à un résultat en opposition avec le panorama des prix moyens (partie 1 de ce chapitre). En effet, le panorama montrait une baisse des prix pour les réservations à 0 jour. Cette baisse est donc portée par les nouveaux entrants.

4 Discussion autour de la construction d'un nouvel indice des prix

Cette partie proposera la construction d'un indice des prix selon la méthode des classes homogènes. Cet indice sera ensuite comparé à l'IPC actuel. Deux tests seront ensuite menés pour mesurer (i) l'impact de la prise en compte du calendrier civil par rapport à l'utilisation du calendrier IPC ; (ii) l'impact de la prise en compte des pondérations de consommation 2020 impactées par la crise sanitaire par rapport aux données de 2019.

Dans tout ce chapitre, l'analyse portera uniquement sur la France métropolitaine.

4.1 L'approche des classes homogènes : des moyennes géométriques non pondérées des chambres au sein des classes agrégées par une formule de Laspeyres arithmétique

L'approche de la construction d'un indice à panier fixe a été écartée du fait d'un taux d'imputation important au cours du temps [24]. Il serait en moyenne de 45 % sur la période de janvier à juillet 2021.

Taux d'imputation dans le cadre d'une approche à panier fixe

L'approche à panier fixe consiste à suivre un échantillon d'hôtels x chambres défini en décembre 2020 un jour donné tout au long de l'année 2021. Dans le cadre de la collecte terrain de l'IPC, le produit est suivi pour un jour donné avec une tolérance à plus ou moins trois jours du fait de l'organisation pratique des tournées de prix des enquêteurs. Cette façon de procéder permet de s'assurer que l'on mesure bien des évolutions en moyenne sur un mois, de garantir l'ouverture du point de vente ce jour-là et de neutraliser d'éventuels effets « jour de la semaine » sur les prix. Concrètement si le produit doit être collecté tous les jeudis de semaine 1 du calendrier IPC, une collecte le mercredi de semaine 1 du calendrier IPC pour un mois de l'année est autorisée.
À partir de l'ensemble des hôtels x chambres collectés en décembre 2020, il est proposé de calculer un taux d'imputation pour les mois de janvier à juillet 2021, date de la première analyse sur le sujet qui n'a pas été réactualisée. Pour cela, le calendrier de collecte IPC est retenu, les produits devront toujours être collectés soit le week-end, soit en semaine pour la bonne semaine du calendrier IPC (semaine 1, 2 3 ou 4) et la bonne antériorité de réservation (0, 30, 60 jours).

TABLE 10 – Taux d'imputation du nombre d'hôtels x chambres dans le cadre d'une approche à panier fixe selon les mois du calendrier IPC

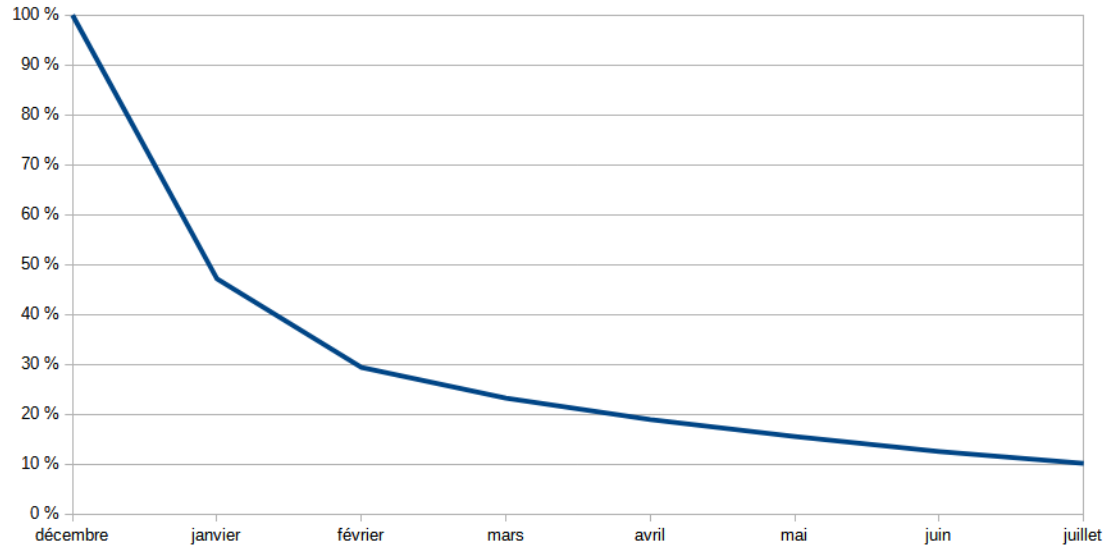
	Janvier	Février	Mars	Avril	Mai	Juin	Juillet
Taux d'imputation	53 %	49 %	43 %	43 %	41 %	43 %	46 %

Source : Base avec filtres issue du webscraping, à la date du 30 juillet 2021.
Champ : France entière.

24. Cette approche pourrait néanmoins être investiguée plus amplement en sélectionnant un échantillon de chambres « bien suivies, bien vendues », ce qui limiterait le taux d'imputation. Par ailleurs, cette approche nécessiterait de recoder la chaîne IPC actuelle en R (cas des remplacements, du recalcul du prix de base etc), ce qui demanderait un investissement informatique important.

En janvier ce sont donc 53 % produits (hôtels x chambres) qui sont à imputer ou à remplacer. Il est intéressant de compléter cette donnée par le taux des présents tout au long de l'année par rapport à l'échantillon défini en décembre 2020, ce qui permettrait de choisir entre l'imputation ou le remplacement.

FIGURE 11 – Évolution du taux de présence des hôtels x chambres présents dans le cadre d'une approche à panier fixe selon les mois du calendrier IPC



Source : Base avec filtres issue du webscraping, à la date du 30 juillet.
Champ : France entière.
*Note de lecture : 30 % des hôtels * chambres de l'échantillon présents en décembre sont encore présents en février (i.e. présents en décembre, janvier et février).*

Une autre approche a été retenue consistant à construire des classes suffisamment fines et homogènes pour considérer que les chambres d'hôtels sont substituables pour le consommateur à l'intérieur de ces strates. [25]. Cette approche a un double avantage : être plus simple à mettre en place et nécessiter peu d'imputations.

4.1.1 Agrégation et stratification

Au sein de ces classes homogènes, on suppose que le consommateur a la possibilité de substituer des produits entre eux car les différentes chambres lui permettent de satisfaire les mêmes besoins tout en ayant connaissance des différents prix pratiqués pour l'ensemble des hôtels [26] (toutes les chambres disponibles s'affichent en ligne sur une ou plusieurs pages lors d'une requête sur la plateforme). **Cette hypothèse de substituabilité des chambres au sein de ces strates homogènes conduit à retenir une moyenne géométrique des prix non pondérée (indice de Jevons)**. Au sein d'une strate s, le micro-indice est donc :

25. Cette approche est dénommée « monthly chaining and replenishment means » (MCR) dans le manuel d'Eurostat ou « class mean imputation » dans le manuel du FMI. Elle permet de prendre en compte la forte rotation de produits (notamment dans le secteur des produits électroniques) en autorisant un renouvellement mensuel complet de l'échantillon. Les manuels préconisent d'utiliser cette méthode sur des échantillons suffisamment grands, et de veiller à ce que les produits ne subissent pas des réductions trop fréquentes ni qu'ils sortent systématiquement à un prix réduit (à la fin du cycle de vie).

26. En réalité, uniquement pour les hôtels présents sur la plateforme de réservation en ligne.

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

n_0 (resp. n_m) est le nombre d'observations de la strate s au mois 0 (resp. au mois m)

Ce niveau de granularité choisi pour le micro-indice reflète l'hypothèse que certains paramètres dont dépend le prix sont importants pour le consommateur. Ils ont été sélectionnés notamment à partir de l'analyse des déterminants de prix [27] :

- la géographie appréhendée par le croisement de la région (hors Île-de-France) et des aires touristiques et par le croisement de l'Île-de-France et le statut des communes d'Île-de-France (centre, banlieue, isolée, hors unité urbaine) ;
- le confort de l'hôtel mesuré par son classement : non classé, 1, 2, 3, 4 ou 5 étoiles ;
- le modèle d'exploitation de l'hôtel : chaîne ou indépendant ;
- le confort de la chambre (classique / supérieur) ;
- les différentes antériorités 0, 30 et 60 jours car réserver une nuitée avec deux mois d'avance comporte plus d'incertitudes ou de contraintes qu'en dernière minute pour le consommateur. Les consommateurs recourant à une de ces trois antériorités ont donc des profils différents ;
- les deux périodes week-end / semaine permettent de contrôler les effets de calendrier pour être à utilité constante.

Pour la suite, la strate s est définie comme le croisement suivant : **zone géographique x confort hôtel x modèle d'exploitation x confort chambre x antériorité x période**. Les micro-indices sont alors agrégés par un **indice de type Laspeyres arithmétique**. C'est en effet l'indice classique utilisé dans l'IPC au-dessus d'un certain niveau d'agrégation. Il reflète une approche par panier-type qui élimine les effets de structure de consommation : on calcule une moyenne arithmétique des indices élémentaires de prix en fixant les poids au cours d'une année.

$$I^m_{hotels} = \sum_{g}^{\substack{zone \\ géographique}} \sum_{s}^{\substack{nombre \\ d'étoiles}} \sum_{e}^{\substack{modèle \\ d'exploitation}} \sum_{c}^{confort} \sum_{a}^{antériorité} \sum_{p}^{période} w_g * w_s * w_e * w_c * w_a * w_p * I^m_{s(g,s,e,c,a,p)}$$

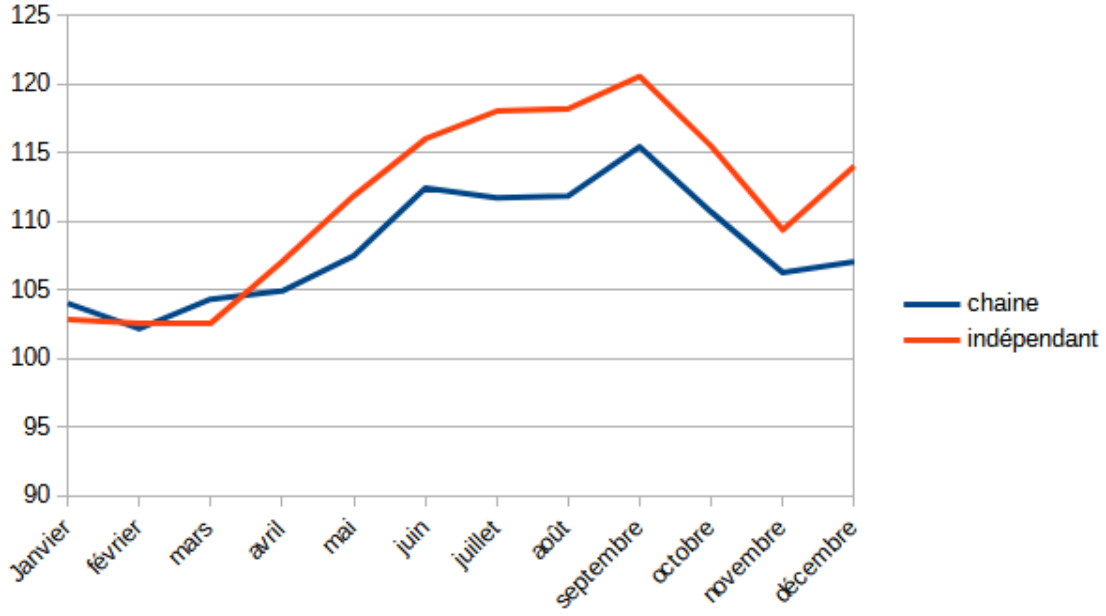
w_g est le poids de la zone touristique g, w_s du nombre d'étoiles s, w_e du modèle d'exploitation e, w_c du confort de la chambre c , w_a de l'antériorité a et w_p de la période p

4.1.2 Indices de prix selon les différentes variables

Lorsque l'ensemble des prix d'une même strate (*zone géographique * confort hôtel * modèle d'exploitation * confort chambre * antériorité * période*) sont absents ou lorsqu'un seul prix au sein d'une même strate est disponible, l'imputation est réalisée au niveau du micro-indice. Afin d'analyser les différentes dynamiques des prix selon les caractéristiques des hôtels et des chambres, nous avons calculé des indices par variable.

27. (i) cette analyse des déterminants portait sur les niveaux de prix et non les évolutions de prix ; (ii) tous les déterminants n'ont pas été retenus, un arbitrage opérationnel a lieu entre la taille des classes (plus elles sont fines, plus elles permettent de retracer au plus près le comportement du consommateur), et la volumétrie des imputations (plus les classes sont fines, plus le nombre de classes vides est potentiellement important).

FIGURE 12 – Indices de prix des nuitées hôtelières sans imputation selon la modalité d'exploitation entre janvier et août 2021 (base 100 = décembre 2020)



Source : Base avec filtres issue du webscraping, à la date du 31 décembre.
Champ : France métropolitaine, uniquement les observations (hôtels x chambres) présentes aux antériorités de réservation 30 et 60 jours.

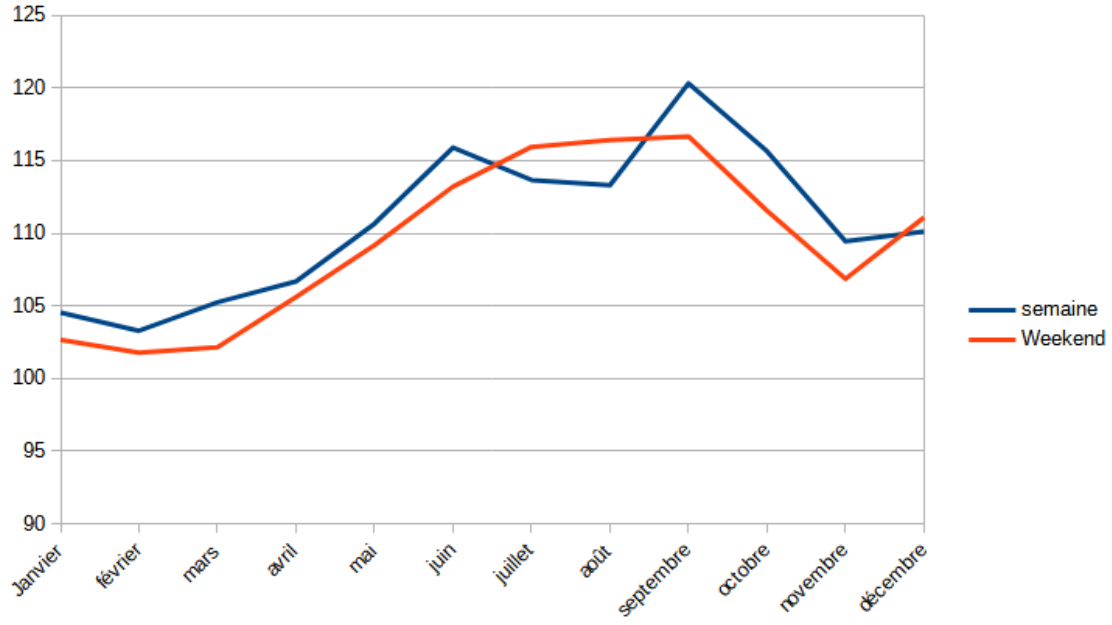
Les indices différenciés selon le confort de la chambre (cf figure n°13 ont un dynamisme très proche en début d'année, la hausse plus élevée des chambres de confort supérieur sur la deuxième partie de l'année (pic en septembre) est peut être due à une reprise plus importante des déplacements professionnels.

FIGURE 13 – Indices de prix des nuitées hôtelières sans imputation selon le confort de la chambre entre janvier et décembre 2021 (base 100 = décembre 2020)

Source : Base avec filtres issue du webscraping, à la date du 31 décembre.
Champ : France métropolitaine, uniquement les observations (hôtels x chambres) présentes aux antériorités de réservation 30 et 60 jours.

L'analyse du comportement des prix selon la période de la nuitée (cf figure n°14) a la particularité d'analyser un champ qui n'est actuellement pas pris en compte dans l'indice des prix : les réservations le weekend. On constate que les prix sont stables le weekend entre février et mars alors qu'en semaine il y a une hausse assez importante, une piste explicative peut être la période de vacances scolaires. La période estivale présente également des comportements différenciés dans l'évolution des prix entre la semaine et le weekend : les prix le weekend sont à la hausse entre juin et juillet tandis que les prix à la semaine baissent. On observe pour le weekend par la suite une stabilité jusqu'en septembre alors que pour les prix à la semaine, les prix évoluent fortement à la hausse entre août et septembre, ce qui peut s'expliquer par une reprise des nuitées pour raison professionelle.

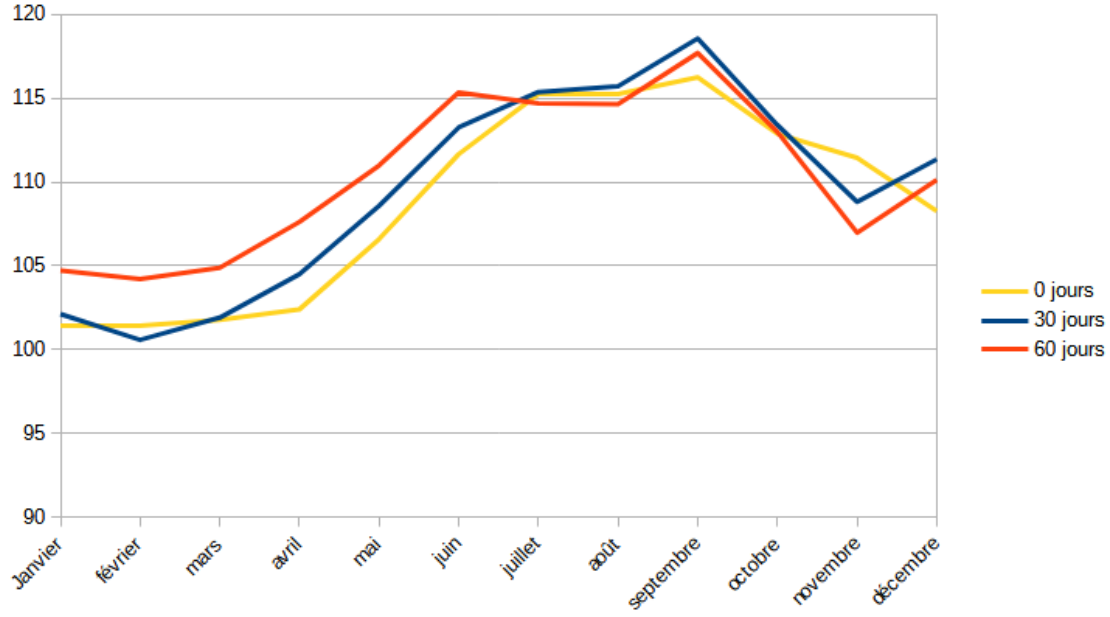
FIGURE 14 – Indices de prix des nuitées hôtelières sans imputation selon la période de la nuitée entre janvier et décembre 2021 (base 100 = décembre 2020)



Source : Base avec filtres issue du webscraping, à la date du 31 décembre.
Champ : France métropolitaine, uniquement les observations (hôtels x chambres) présentes aux antériorités de réservation 30 et 60 jours.

Hormis pour le mois de novembre, les évolutions de prix suivant l'antériorité sont assez proches (cf. figure n°15. On remarque néanmoins que les prix des nuitées achetées avec beaucoup d'anticipation (60 jours d'antériorité) augmentent un mois plus tôt sur le premier semestre.

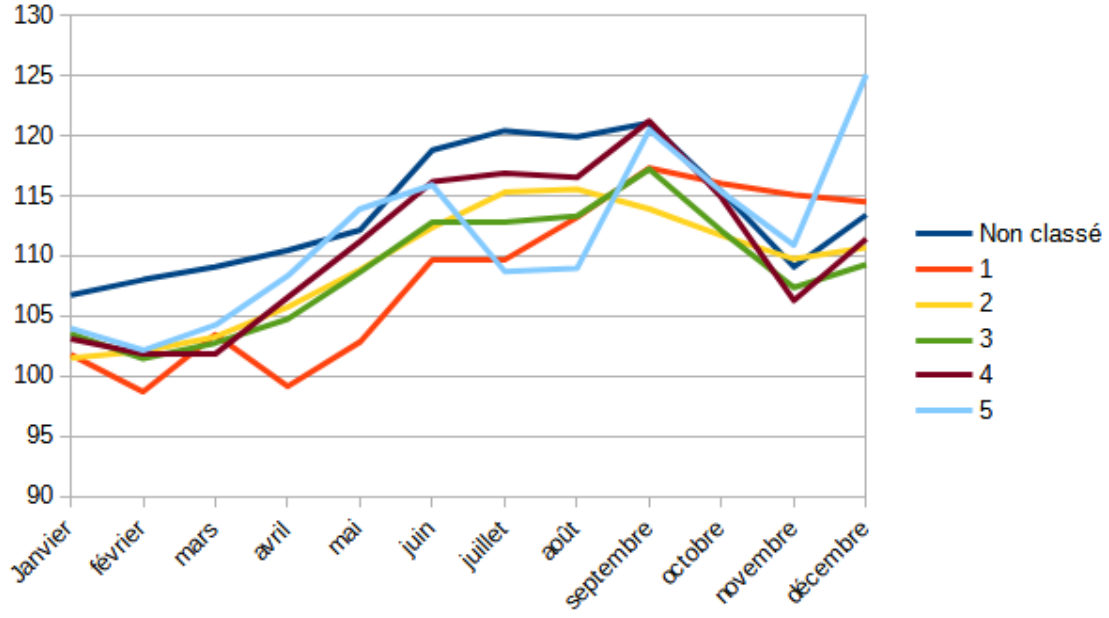
FIGURE 15 – Indices de prix des nuitées hôtelières sans imputation selon l'antériorité de réservation (base 100 = décembre 2020)



Source : Base avec filtres issue du webscraping, à la date du 31 décembre.
Champ : France métropolitaine, uniquement les observations (hôtels x chambres) présentes aux antériorités de réservation 30 et 60 jours.

Si les prix moyens des hôtels non classés et des hôtels 2 étoiles sont relativements proches, on constate que les dynamismes de prix eux ne le sont pas, ils sont peut être plus proches des 3 étoiles(cf. figure n°16. On peut noter le profil atypique des 5 étoiles dont les prix baissent sur l'été et ont une hausse prononcée en septembre et en décembre (un facteur explicatif peut être le fait que beaucoup d'hôtels 5 étoiles se situent en région Parisienne).

FIGURE 16 – Indices de prix des nuitées hôtelières sans imputation selon le nombre d'étoiles de l'hôtel entre janvier et décembre 2021 (base 100 = décembre 2020)

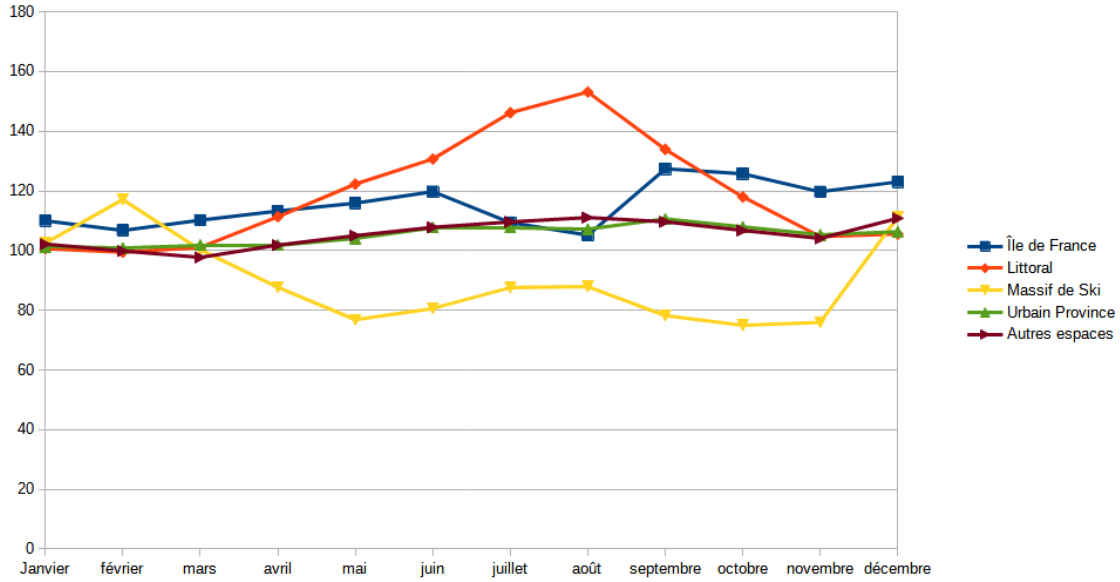


Source : Base avec filtres issue du webscraping, à la date du 31 décembre 2021.
Champ : France métropolitaine, uniquement les observations (hôtels x chambres) présentes aux antériorités de réservation 30 et 60 jours.

Pour les indices selon les zones touristiques[28] (cf. figure n°17), on constate que les prix ont des évolutions saisonnières distinctes selon la zone touristique : une hausse des prix dans les périodes de fin de printemps/été pour le littoral, une dynamique à la hausse pour les stations de ski en début et en fin d'année. La particularité pour les stations de ski est que le mois de base(décembre) est un des mois où les prix sont les plus elevés, ce qui explique que l'indice descende en dessous des 80 au cours de l'année. En ce qui concerne l'Île de France, l'indice baisse de juin à août et repart à la hausse en septembre, un facteur explicatif est la baisse de nuitées pour raisons professionnelles durant la période estivale.

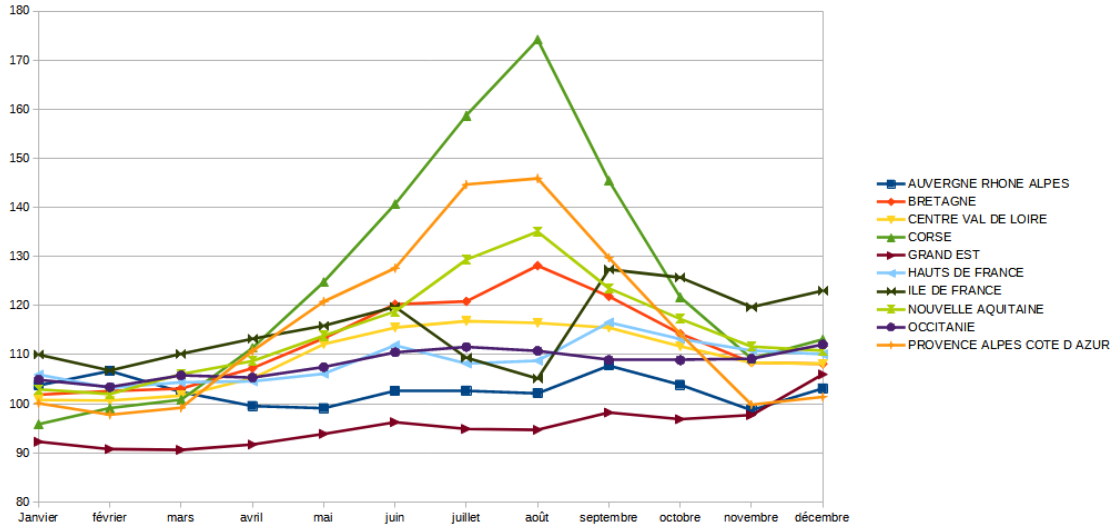
28. À partir de fin octobre la perte de collecte d'une variable due à un changement de la plateforme a amené des problèmes pour l'appariement des nouveaux hôtels au référentiel géographique (cela réprésente 5213 observations sur 427 243).

FIGURE 17 – Indices de prix des nuitées hôtelières sans imputation selon la zone touristique entre janvier et décembre 2021 (base 100 = décembre 2020)



L'analyse des indices par région (cf. figure n°18) présente certaines limites car toute les régions ne sont pas de la même taille (en terme de population et de superficie) et recouvrent parfois plusieurs zones touristiques de natures différentes ou non. On constate tout de même que la corse possède un dynamisme de prix proche de celui des zones littorales vu en figure n°17, avec une hausse importante en été.
Source : Base avec filtres issue du webscraping, à la date du 31 décembre 2021.
Champ : France métropolitaine, uniquement les observations (hôtels x chambres) présentes aux antériorités de réservation 30 et 60 jours.

FIGURE 18 – Indices de prix des nuitées hôtelières sans imputation selon la région entre janvier et décembre 2021 (base 100 = décembre 2020)



Source : Base avec filtres issue du webscraping de la plateforme de réservation en ligne, à la date du 31 décembre 2021.
Champ : France métropolitaine, uniquement les observations (hôtels x chambres) présentes aux antériorités de réservation 30 et 60 jours.

Lorsqu'aucun micro-indice par zone géographique x confort hôtel x modèle d'exploitation x antériorité x période n'est disponible sur un mois, on calcule l'évolution moyenne des chambres pour l'ensemble des critères confondus. Pour calculer un micro-indice, le nombre de chambres peut varier d'un jour à l'autre et au cours du mois soit du fait d'une indisponibilité, soit du fait de problème informatique (le robot n'a pu collecter correctement l'information). D'une certaine manière, cela revient à dire que les prix absents sont imputés implicitement au sein du micro-indice par la moyenne des autres prix de la même strate.

4.1.3 Pondérations

Les pondérations ont été déterminées à partir de données transmises par le pôle tourisme à partir de l'enquête mensuelle de fréquentation dans les hébergements touristiques [29]. Trois jeux de données sur les nombres de chambres occupées ont été transmis concernant les années 2019 et 2020 (année impactée par la crise sanitaire).

Jeu de pondérations n°1 – données 2019 brutes
Les données sur l'année 2019 permettent d'obtenir des poids sur les nombres de chambres occupées pour le croisement région x aire touristique x nombre d'étoiles (1-2 étoiles, 3 étoiles, 4-5 étoiles, non classés). Plusieurs hypothèses ont été ensuite ajoutées :

— afin de distinguer les poids pour les hôtels 1 et 2 étoiles et pour les hôtels 4 et 5 étoiles au sein de chaque région, une clé de répartition a été définie au niveau de chaque région

29. Une déformation de la répartition des hôtels webscrapés par rapport au parc hôtelier du référentiel tourisme avait été mis en évidence. Retenir les poids issus de l'enquête de fréquentation touristique (basé sur le parc hôtelier du référentiel) permet de se recaler sur cette enquête. On suppose que les chambres avec petit-déjeuner et annulable gratuitement sont un bon représentant de l'ensemble des chambres. La Belgique utilise également cette méthode de pondération se basant sur la fréquentation touristique malgré le biais de couverture identifié.

à partir du référentiel d'hôtels tenu par le pôle tourisme de Montpellier (et non en fonction d'une consommation) ;
— afin de répartir le poids de l'Île-de-France selon les différents statuts des communes (ville-centre, banlieue, ville isolée, hors unité urbaine), une clé de répartition a été définie à partir du référentiel d'hôtels (et non en fonction d'une consommation) ;
— le poids des chaînes et des hôtels indépendants est calculé au niveau régional et s'applique donc uniformément quel que soit l'aire touristique ou le classement des hôtels selon le nombre d'étoiles.

Jeu de pondérations n°2 – données 2019 pour raisons personnelles
Le jeu de pondérations n°1 a pu être amélioré en ne tenant compte que des chambres occupées par une clientèle pour raisons personnelles (ie hors clientèle professionnelle). Les chambres occupées uniquement pour des raisons personnelles (cf. tableau n°11) sont moins nombreuses en proportion en Île-de-France, dans les Hauts-de-France et dans les Pays de la Loire par rapport à l'ensemble des chambres occupées. Au contraire, les chambres occupées uniquement pour des raisons personnelles sont plus nombreuses pour la Corse et la région Provence Alpes Côte d'Azur.

Jeu de pondérations n°3 – données 2020 brutes
Les données sur l'année 2020 ne présentent pas le croisement fin région x aire touristique x nombre d'étoiles mais le croisement région x aire touristique. Une hypothèse supplémentaire a donc été ajoutée en appliquant le poids du classement hôtelier (1-2 étoiles, 3 étoiles, 4-5 étoiles, non classés) de chaque région quelle que soit l'aire touristique. La crise sanitaire débutée en 2020 a eu un impact particulièrement important sur la fréquentation touristique en Île-de-France (forte proportion d'hébergements fréquentés par les touristes non résidents, dans le haut de gamme en particulier [30]) . Cette déformation s'observe à partir de la répartition du nombre de chambres occupées en France métropolitaine (cf. tableau n°11) :
— poids plus faible pour l'Île-de-France car de nombreux hôtels sont restés fermés notamment durant l'été ;
— poids plus important pour les régions moins urbaines ou littorales (Bourgogne Franche-Comté, Normandie, Pays de la Loire, Bretagne, Nouvelle Aquitaine, Auvergne Rhône-Alpes). Ces régions ont retrouvé un meilleur niveau de fréquentation durant l'été 2020.

30. Par ailleurs, chute également du tourisme d'affaires à cause de l'annulation des réunions en présentiel et de l'annulation des grands évènements (hors périmètre IPC).

TABLE 11 – Répartition des chambres occupées en 2019 et 2020 selon les régions

Région	2019 (1)	2019 (2)	2020 (1)
Île-de-France	32 %	31 %	22 %
Centre	3 %	3 %	3 %
Bourgogne – Franche-Comté	3 %	3 %	4 %
Normandie	4 %	4 %	5 %
Hauts de France	5 %	3 %	5 %
Grand Est	7 %	7 %	7 %
Pays de la Loire	4 %	3 %	5 %
Bretagne	4 %	4 %	5 %
Nouvelle-Aquitaine	8 %	8 %	9 %
Occitanie	8 %	8 %	8 %
Auvergne-Rhône Alpes	11 %	11 %	14 %
Provence Alpes Côte d'Azur	11 %	13 %	11 %
Corse	1 %	2 %	1 %

Source : Enquête mensuelle de fréquentation dans les hébergements touristiques.
Note : (1) données brutes, (2) données pour les chambres occupées pour raison personnelle.
Champ : France métropolitaine.

Les jeux de pondérations n°1 et n°3 posent problème en raison de l'inclusion de nuitées pour motif professionnel.
Enfin d'autres pondérations ne sont pas disponibles dans l'enquête mensuelle de fréquentation dans les hébergements touristiques, il a donc fallu opter pour des conventions faute d'informations :

- le poids du confort de la chambre : la répartition est calculée à partir des observations webscrapées de la plateforme de réservation en ligne (France entière, sans utilisation de filtres). Il sera ainsi retenu un poids de 85 % pour les chambres classiques et de 15 % pour les chambres confort ;
- le poids de l'antériorité : équirépartition des poids ;
- le poids de la période : 40 % pour la semaine, 60 % pour le week-end afin de sur-pondérer le week-end.

4.2 Comparaison des indices actuels et des indices par classes homogènes

4.2.1 Comparaison des indices hôteliers proposés

Trois indices sont donc obtenus : un indice des prix pour les réservations le jour-même (indice n°1), un indice des prix pour les réservations à 30 ou 60 jours (indice n°2) et un indice des prix pour les réservations toutes antériorités confondues (indice n°3). Du fait de l'élargissement progressif du périmètre géographique dans le cadre du webscraping de la plateforme de réservation en ligne, toutes les régions ne sont pas présentes dans toutes les strates en décembre (mois de base). Les antériorités 30 et 60 jours ne permettent pas de couvrir les régions Bourgogne-Franche Comté, Normandie et Pays de la Loire. Les hôtels de ces régions sont néanmoins couverts par les réservations pour le jour-même. Les trois indices présentent un dynamisme à la hausse équivalent sur la période de mai à juillet par rapport à l'indice actuel mais diffèrent sur le début d'année 2021 (cf. graphique n°19). En effet, l'évolution des prix mesurée par ces trois indices semble plus dynamique en janvier que l'indice actuel, les trois indices diminuent en février (contre une hausse de l'indice actuel), augmentent dans une moindre mesure que l'indice actuel (à l'exception de l'indice pour des réservations pour le jour-même), puis augmentent plus sensiblement en avril alors que l'indice actuel baisse. Par la suite, seul l'indice des prix pour les réservations à 60, 30 jours d'avance et les réservations pour le jour-même sera étudié car il correspond à l'indice qui permet de prendre en compte la diversité des profils de tarifications mis en évidence précédemment [31]. Pour la fin d'année, on constate un écart de dynamique entre les indices avec les données webscrapées qui augmentent et l'indice IPC qui diminue entre novembre et décembre. Cela peut s'expliquer par le fait que l'indice IPC ne comprends pas les dernières semaines du mois de décembre avec les fêtes de fin d'année et les vacances scolaires. La prochaine partie va s'intéresser à analyser plus en détail l'impact du calendrier retenu.

31. Il faudrait idéalement collecter les prix à quelques jours seulement de la date de la nuitée (J-2 par exemple). Les réservations pour le jour-même sont ici un proxy de cette date de réservation.

FIGURE 19 – Comparaison de différents indices de prix des nuitées hôtelières entre décembre 2020 et décembre 2021

Source : Base avec filtres issue du webscraping, à la date du 31 décembre.
Champ : France métropolitaine.
Note : Le calendrier ici est celui du mois civil (sauf pour l'IPC), les pondérations utilisées ici sont celles de l'année 2019 pour la consommation de chambre pour raisons personnelles (sauf pour l'IPC, pondérations 2019 corrigées par un traitement spécifique tenant compte de l'impact de la crise sanitaire). L'écart est calculé comme la différence entre l'indice 0/30/60 jours – l'indice IPC

4.2.2 Impact du calendrier retenu

Un test est effectué pour estimer l'impact d'une modification du calendrier retenu : mois civil (calendrier non fixe) versus mois IPC (calendrier fixe mais ne tenant pas compte des week-ends et de manière imparfaite des vacances scolaires).

TABLE 12 – Calendrier de collecte IPC

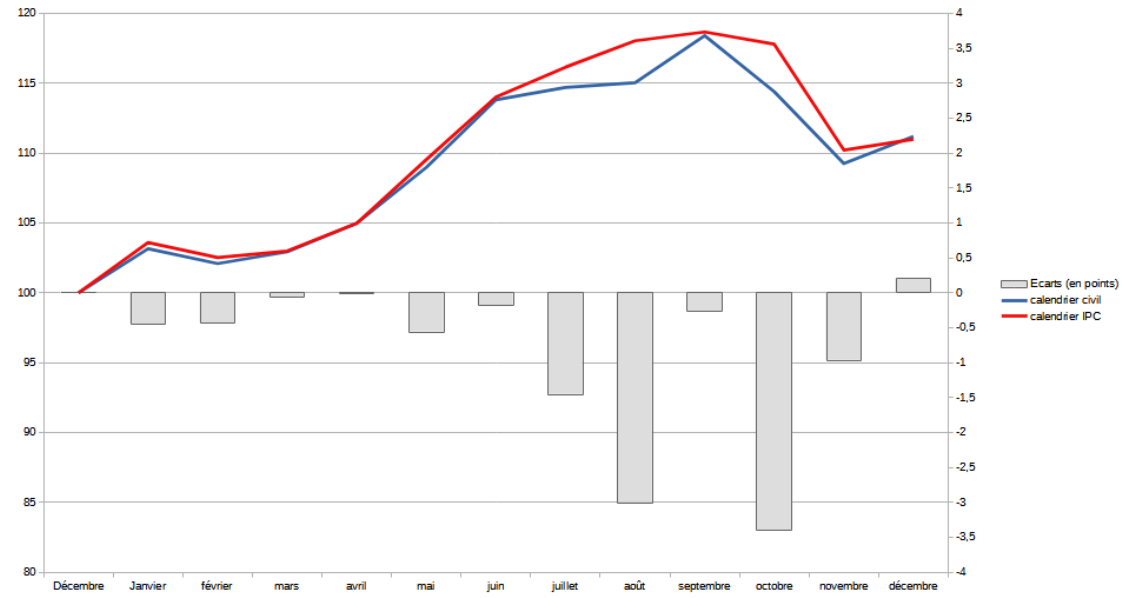
Mois	Date de début	Date de fin
Décembre 2020 (mois de base)	23-11-2020	20-12-2020
Janvier 2021	04-01-2021	29-01-2021
Février 2021	01-02-2021	26-02-2021
Mars 2021	01-03-2021	26-03-2021
Avril 2021	29-03-2021	23-04-2021
Mai 2021	03-05-2021	28-05-2021
Juin 2021	31-05-2021	25-06-2021
Juillet 2021	28-06-2021	23-07-2021
Août 2021	02-08-2021	27-08-2021
Septembre 2021	30-08-2021	24-09-2021
Octobre 2021	27-09-2021	22-10-2021
Novembre 2021	25-10-2021	19-11-2021
Décembre 2021	22-11-2021	17-12-2021

Regardons dans un premier les indices calculés avec les pondérations 2019 et 2020 pour les 3 antériorités confondues (cf. figure n°20). L'écart moyen en valeur absolue est de 0,92 point entre les deux indices. L'indice des prix des nuitées hôtelières calculé à l'aide du calendrier IPC est notamment :

- Janvier : l'indice avec le calendrier IPC est plus dynamique, il n'intègre pas les prix des premiers jours de janvier ni les derniers (qui sont un vendredi, un samedi, deux dimanches, un lundi et un mardi). De plus, le mois de base pour l'indice avec le calendrier civil intègre la dernière semaine de décembre avec les fêtes de fin d'année et les vacances scolaires avec des prix moyens plus élevés.
- Juillet : hausse un peu plus dynamique de l'indice avec le calendrier IPC que celui avec le mois civil. Le calendrier IPC de juillet intègre 1 semaine de prix de juin et a 1 semaine sans prix en juillet.
- Août : hausse très dynamique de l'indice avec le calendrier IPC, il n'intègre pas les prix du premier jour d'août ni les trois derniers (qui sont un vendredi, deux samedis et deux dimanches).
- Septembre : L'indice avec le calendrier IPC est peu dynamique alors que celui se servant du mois civil est très dynamique. Le calendrier IPC de septembre a 1 semaine de prix en moins en septembre (dernière semaine de septembre).
- Octobre : baisse moins marquée de l'IPC par rapport au mois civil : L'IPC intègre les prix de la dernière semaine de septembre mais pas la dernière semaine d'octobre.
- Novembre : baisse très marquée de l'indice avec le calendrier IPC, baisse aussi marquée mais un peu moins de l'indice avec le calendrier civil. L'indice avec le calendrier IPC intègre les prix de la dernière semaine d'octobre mais pas les 2 semaines de fin novembre.
- Décembre : hausse marquée de l'indice avec le calendrier civil, hausse moins marquée de l'indice avec le calendrier IPC qui intègre 2 semaine de novembre et 2 semaine de décembre alors que le mois civil prend la hausse des 15 jours de vacances de Noël.

Il est à noter que la collecte du robot un jour sur deux selon les deux approches (sans filtre et toutes les chambres sur une zone géographie restreinte, avec filtres avec la chambre mise en avant par la plateforme et sur tout le territoire) permet de couvrir les différents jours en deux semaines et permet donc d'avoir une bonne approximation de l'indice. Cependant, ce mode opératoire peut avoir des impacts pour les week-ends particulièrement réservés pendant les vacances scolaires (comme le dernier de juillet). Il a été arrêté en septembre où le choix a été fait de ne maintenir que la collecte avec filtres et ce tous les jours.

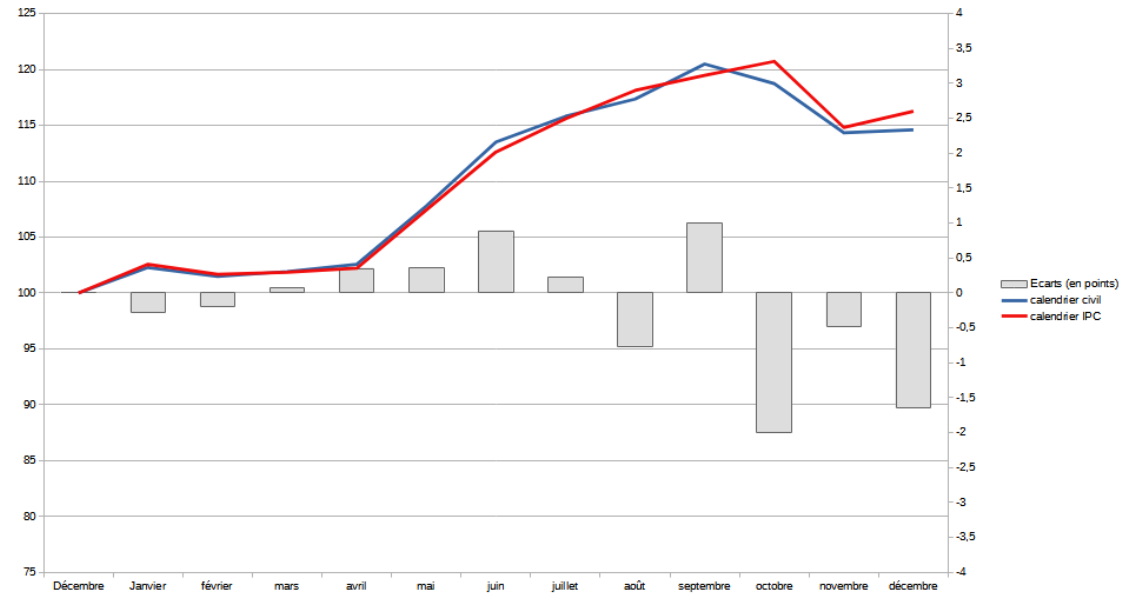
FIGURE 20 – Comparaison de l'évolution des prix des nuitées hôtelières en tenant compte des antériorités 0, 30 et 60 jours selon le calendrier civil et selon le calendrier IPC



Sources : Calculs auteur, base avec filtres issue du webscraping, à la date du 31 décembre 2021, enquête mensuelle de fréquentation dans les hébergements touristiques.
Champ : France métropolitaine.
Note : les pondérations utilisées ici sont celles de l'année 2019 pour la consommation de chambre pour raisons personnelles. L'écart est calculé comme la différence entre l'indice calendrier civil – l'indice calendrier IPC.

Le calendrier utilisé actuellement pour l'IPC est utilisé avec une collecte sur le terrain pour le jour même, un focus sur l'indice calculé seulement avec les données à 0 jour d'anteriorité (cf. figure n°21) va rendre plus aisé l'analyse de cet "effet calendrier". Le calendrier IPC se détache du calendrier civil principalement vers la fin d'année. Si on regade l'évolution des prix, l'indice avec le calendrier IPC augmente d'avril à octobre tandis que l'indice calendrier civil augmente d'avril à septembre et commence à diminuer dès octobre. Ce décalage peut s'expliquer par le fait que la dernière semaine du mois d'Octobre civil n'est pas comprise dans le mois IPC (cf. tableau n°12). Entre novembre et décembre l'indice avec le calendrier IPC augmente alors que l'indice civil reste stable.

FIGURE 21 – Comparaison de l'évolution des prix des nuitées hôtelières en tenant compte de l'antériorité 0 jour selon le calendrier civil et selon le calendrier IPC

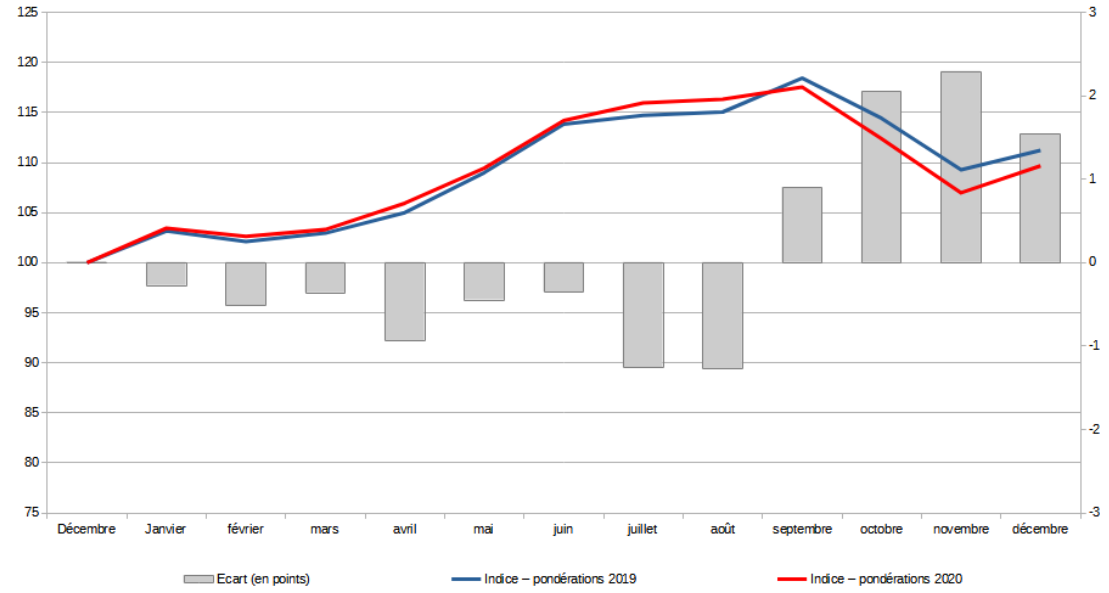


Sources : Calculs auteur, base avec filtres issue du webscraping, à la date du 31 décembre 2021, enquête mensuelle de fréquentation dans les hébergements touristiques.
Champ : France métropolitaine.
Note : les pondérations utilisées ici sont celles de l'année 2019 pour la consommation de chambre pour raisons personnelles. L'écart est calculé comme la différence entre l'indice calendrier civil – l'indice calendrier IPC.

4.2.3 Impact des pondérations

Un test est effectué pour estimer l'effet de la crise sanitaire sur les pondérations utilisées dans le calcul de l'indice.
Une première comparaison est faite entre les données de fréquentation de l'année 2019 et les données de fréquentations de l'année 2020(cf. graphique n°22). L'écart moyen en valeur est de 0,1 point entre les deux indices et l'écart moyen en valeur absolue est de 1,01 point. Il y a une compensation importante entre les périodes de janvier à août et de septembre à décembre. L'écart le plus important se mesure en octobre (+ 2,1 points). La principale différence entre les données de fréquentation 2019 et 2020 - le poids plus faible pour l'Île-de-France (cf. 4.1.3) - et le creux en été avec un rebond en septembre pour cette région (cf. figure n°17) expliquent l'écart.

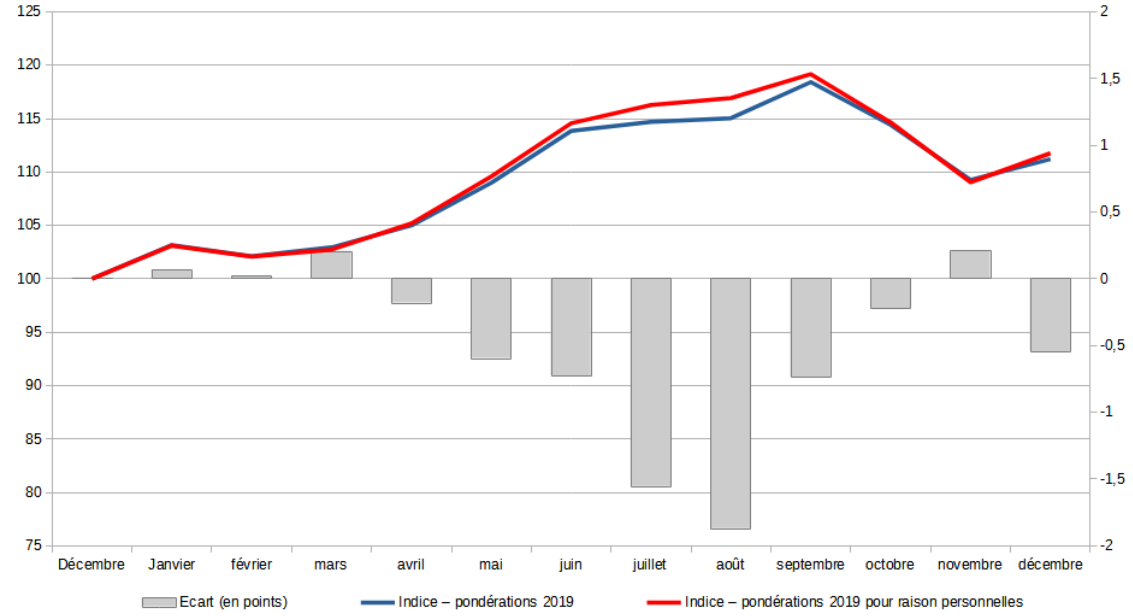
FIGURE 22 – Comparaison de l'évolution des prix des nuitées hôtelières en tenant compte des antériorités 0, 30 et 60 jours selon l'utilisation des pondérations 2019 et 2020



Sources : Calculs auteur, base avec filtres issue du webscraping, à la date du 31 décembre 2021, enquête mensuelle de fréquentation dans les hébergements touristiques.
Champ : France métropolitaine.
Note : le calendrier utilisé ici est le calendrier civil. L'écart est calculé comme la différence entre l'indice pondérations 2019 – indice pondérations 2020.

Une deuxième comparaison est faite entre les données de fréquentation de l'année 2019 et les données de fréquentations de l'année 2019 en prenant seulement en compte les chambres réservées pour raisons personelles(cf. graphique n°23). On constate que les deux indices sont très proches à l'exception de la période estivale (juin, juillet, août et septembre) où l'indice avec les pondérations pour raisons personnelles est supérieur de 1,2 points en moyenne.

FIGURE 23 – Comparaison de l'évolution des prix des nuitées hôtelières en tenant compte des antériorités 0, 30 et 60 jours selon l'utilisation des pondérations 2019 et 2019 uniquement avec les chambres réservées pour raisons personnelles



Sources : Calculs auteur, base avec filtres issue du webscraping, à la date du 31 décembre 2021, enquête mensuelle de fréquentation dans les hébergements touristiques.
Champ : France métropolitaine.
Note : le calendrier utilisé ici est le calendrier civil. L'écart est calculé comme la différence entre l'indice pondérations 2019 – indice pondérations 2019 pour raisons personnelles.

5 Conclusion

Cette étude porte sur la construction d'un indice de prix des nuitées hôtelières tenant compte de la tarification en temps réel à l'aide du webscraping d'une grande plateforme de réservation. L'indice retenu repose sur une agrégation d'indices élémentaires de classes suffisamment homogènes par une formule de Laspeyres. Le principal enjeu est donc de définir ces classes homogènes et de s'assurer de leur pertinence pour appréhender les évolutions de prix et identifier des effets de calendrier sur le niveau de prix. Pour cela, un prérequis indispensable est l'analyse des déterminants de prix mais elle reste à compléter à l'aide d'autres considérations :

- la disponibilité de données de consommation au niveau de ces classes afin de pouvoir les agréger ;
- l'intégration des dimensions temporelles (période [32] et antériorité) et leur signification :
- intégrer la période de consommation, visant à refléter l'utilité du consommateur selon sa date de départ, dans la construction des classes ou pondérer au sein d'une classe les prix selon la période ;
- intégrer la dimension de classe d'antériorité dans la construction des classes ou pondérer au sein d'une classe les prix selon la classe d'antériorité ;
- le choix de la maille d'agrégation peut également reposer sur un arbitrage opérationnel en fonction du nombre de micro-indices manquants à imputer au final. Plus cette maille est fine, plus cette méthode présente l'inconvénient de générer un nombre important de micro-indices manquants et d'être dépendant de classes avec très peu de prix [33].

Cette expérience à la fois de webscraping et de constitution d'un indice tenant compte du yield management pourra être mobilisée pour d'autres services concernés par ce type de tarification en temps réel.

32. Jour de la semaine, semaine/week-end, jour férié, vacances scolaires.

33. Ces classes requièrent une vigilance plus importante notamment dans le nettoyage et la validation des prix.

Bibliographie

[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 et 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 », *Présentation dans le cadre des 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 », *Présentation dans le cadre des NTTS 2021 (New techniques and technologies for statistics)*, 2021.

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

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

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

[15] P. Sillard et 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.

6 Annexes

FIGURE 24 – Analyse des déterminants des prix des nuitées hôtelières en métropole (coefficients des régressions)

	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 : Base avec filtres issue du webscraping, à la date du 30 juillet 2021.
Champ : France métropolitaine.

FIGURE 25 – Analyse des déterminants des prix des nuitées hôtelières en métropole (coefficients de régression)

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.
R^2 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 : Base avec filtres issue du webscraping, à la date du 30 juillet 2021.
Champ : France métropolitaine.