Indices des prix à la consommation des nuitées hôtelières :

L'expérience du webscraping d'une plateforme de réservation en ligne

Adrien Montbroussous, Camille Freppel et Ombéline Guillon

Département des statistiques démographiques et sociales
Institut National des Statistiques Économiques et Sociales

30 mars 2022

Plan

Plan

Contexte : Qu'est-ce que l'IPC ?

- C'est un indicateur statistique qui rend compte de l'évolution moyenne des prix des biens et services consommés par les ménages
- En pratique, on suit l'évolution d'un panier fixe de produits mois après mois
- Pour mesurer une évolution des prix :
  - **à qualité constante**
  - **à structure de consommation constante**

Contexte : Indice des services hôteliers

- Le poste location de chambres représente 0,8 % de la consommation du panier de l'IPC en 2021
- Les prix sont collectés sur le terrain par des enquêteurs dans les agglomérations ($>$ 2000 habitants) définies dans le cadre de l'IPC
- Les prix sont collectés du lundi au vendredi, une fois par mois (lundi de semaine n°1 de janvier $=>$ lundi de semaine n°1 de février) pour le jour-même (évaluation de l'hôtelier si l'hôtel est plein)
- Pour s'assurer de la qualité constante, le produit suivi est une nuitée **pour 2 personnes avec 2 petits-déjeuners compris**

Contexte : pistes d'amélioration, yield management

Méthodologie pour l'indice des prix des nuitées hôtelières.
Collecte sur le terrain d'un échantillon d'hôtels : le prix pour une nuitée pour une chambre pour deux personnes avec petit déjeuner pour le jour même est relevé.

- Pas de prise en compte des réservations en avance.
- Pas de prise en compte de la hausse de la consommation sur internet.
- Certaines zones touristiques ne sont pas bien représentées.
- Aucun prix pour des nuitées le samedi ou le dimanche soir ne sont collectés.

Le yield management et volume de données

Services à prix volatils ou dont le prix dépend de l'antériorité de réservation

Étudier les stratégies tarifaires dans le cadre du yield management nécessite de disposer d'un volume conséquent de données.

Comment disposer de telles données ?

- données de caisses
- Interface de programmation applicative (API)
- webscraping

Webscraping : définir un protocole de collecte

Le webscraping est reconnu comme une méthode de collecte possible par le comité du label.
Webscraping quotidien de la plateforme de réservation en ligne développé en Python.
Lancement automatisé depuis fin août via la plateforme du SSPCloud (docker + gitlab.ci).

- Requête envoyé pour 0, 30 et 60 jours d'avance.
- Webscraping brut : on parcourt le résultat HTML des requêtes.
- Récolte de prix pour des hôtels sur toute la France : filtres activés de petit-déjeuner compris et annulation gratuite.

Webscraping : les limites

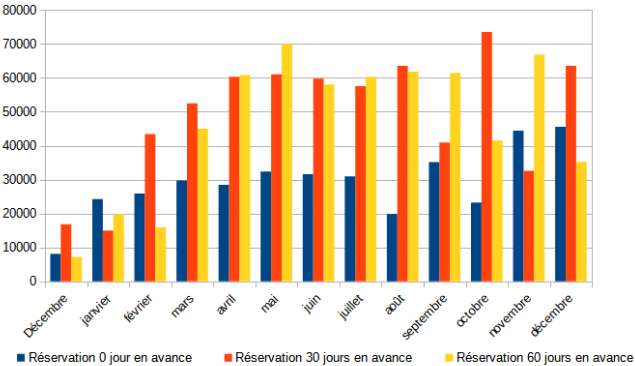
Dépendance au site internet

- Des interruptions de collecte dûes à des changements de la plateforme (interruption pendant une semaine en octobre 2021 par exemple).
- Des codes de nettoyage des variables à maintenir face aux changements de la plateforme, par exemple disparition/mauvais remplissage de certaines variables.

Plan

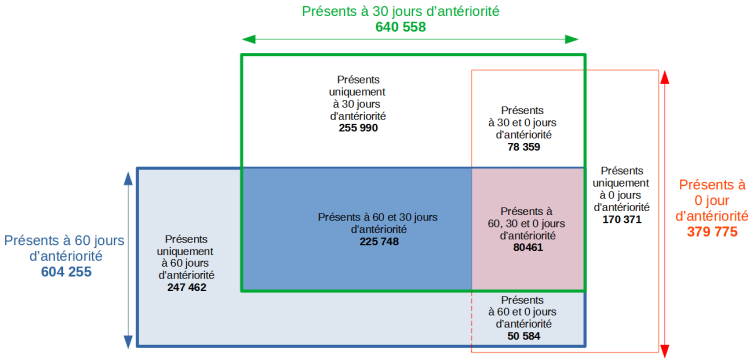
Données collectées

Figure – Répartition des observations selon le mois et l'antériorité de la réservation



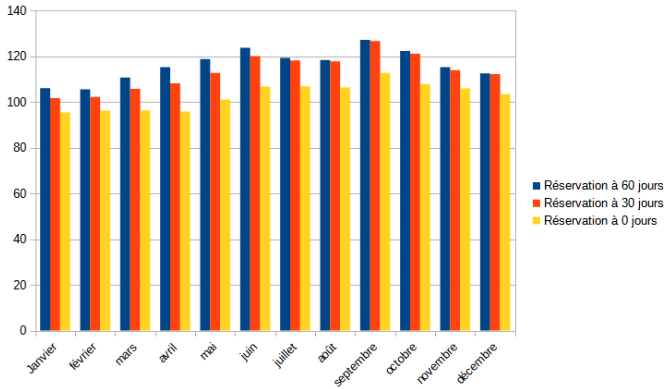
Note : Les données en décembre 2020, pour une réservation à 30 et 60 jours en avance en janvier 2021 et 60 jours en avance en février 2021 ne sont pas sur le champ géographique complet. Une interruption durant une semaine du robot de collecte en octobre explique des baisses à 0 jour en octobre 2021, 30 jours en novembre 2021 et 60 jours en décembre 2021

Une politique d'offre difficile à appréhender



Prix selon l'antériorité de réservation

Figure – Évolution du prix moyen en euros selon l'antériorité de réservation

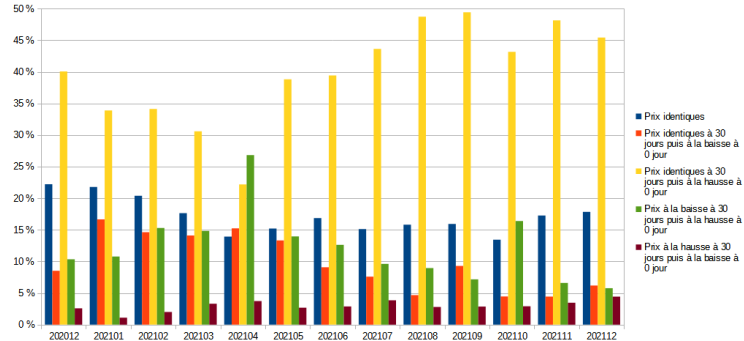


Source : Base avec filtres issue du webscraping, à la date du 31 décembre 2021. Champ : France entière.

Note : les prix moyens sont calculés à partir d'une moyenne géométrique.

Les différents profils de tarification

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

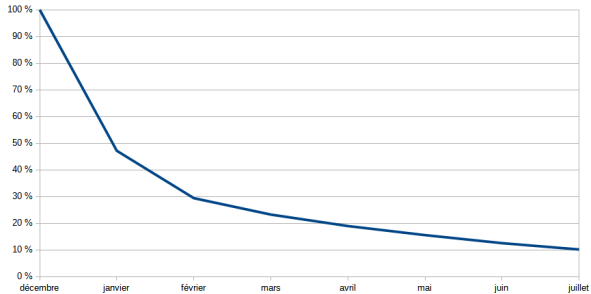


Source : Base avec filtres issue du webscraping, à la date du 31 décembre 2021. Champ : France entière.

Une baisse portée par les nouveaux entrants

Le problème d'un indice avec panier fixe

Figure – Évolution du taux de présence des hôtels x chambres selon le mois IPC



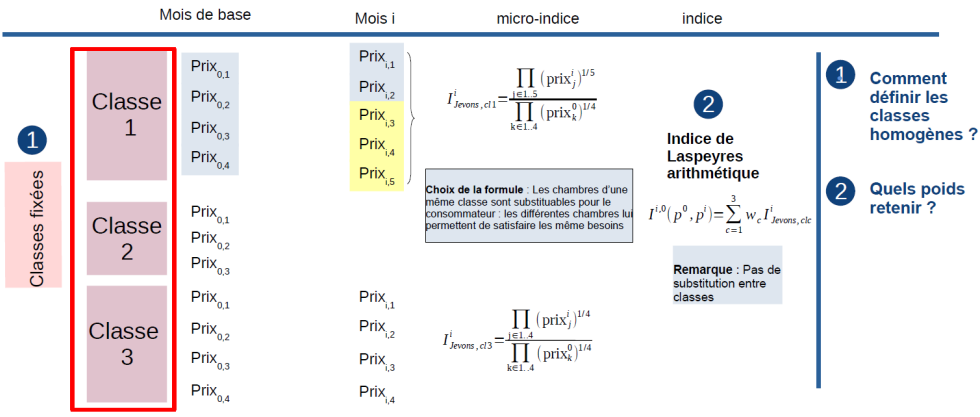
Source : Base avec filtres issue du webscraping, à la date du 30 juillet 2021. Champ : France entière.

- produit suivi : hôtel x chambre 1 jour donné d'une semaine donnée chaque mois
- potentiellement beaucoup de remplacements/imputations

Plan

Principe de la méthode

Champ : Pour 2 personnes, petits-déjeuners inclus, annulation gratuite, avec réservation à 0, 30 et 60 jours



Analyse des déterminants des prix des chambres

- Nettoyage des données et imputation
  - Étape conséquente pour des données webscrapées
  - Nettoyage des noms de ville pour appariement avec un référentiel (zone touristique et région/département)
  - Création de variable par analyse textuelle : type de chambre (supérieure, classique), chaîne d'hôtel, etc.
- Analyse de données
  - Utilisation d'arbres de regression et de modèles linéaires pour trouver les caractéristiques explicatives des prix : nombre d'étoiles, indépendant, jour, mois, type de chambre, information géographique...
  - Certains résultats sont inattendus : des prix plus bas durant les vacances scolaires et les jours de la semaine.

Classes retenues

1 Géographie

- Croisement de la région (hors IDF) et des aires touristique et de l'IDF avec le statut des communes d'IDF
- Limite : peu crédible, à moyen terme recourir au croisement département x aire touristique
- ***Dans toute la suite : France métropolitaine uniquement***

2 Type d'hôtel

- Nombre d'étoiles
- Chaîne/indépendant

3 Type de chambre

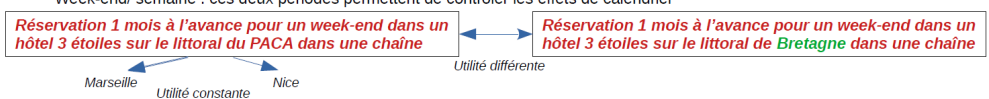
- Confort de la chambre : classique / supérieure

4 Antériorité

- Réserver une nuitée avec deux mois d'antériorité comporte plus d'incertitudes ou de contrainte qu'en dernière minute
- Les consommateurs recourant à une de ces trois antériorités ont des profils différents

5 Période

- Week-end/ semaine : ces deux périodes permettent de contrôler les effets de calendrier



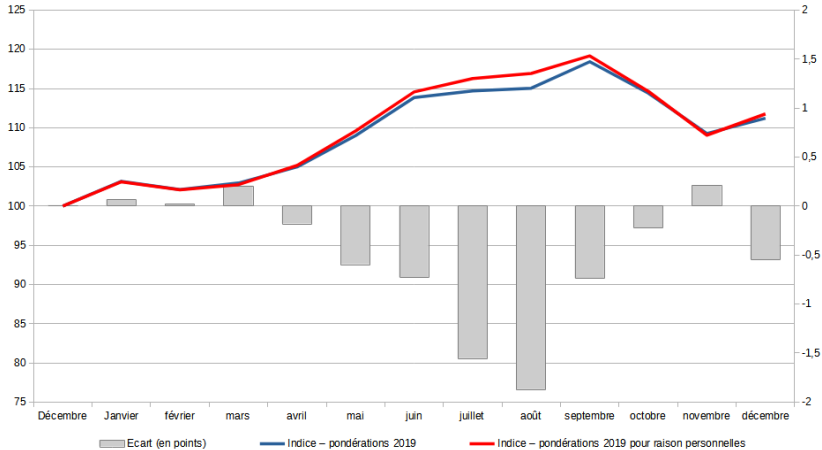
Le choix des pondérations

Pour pondérer les micros indices, trois jeux de données partielles issus du registre des établissements touristiques étaient envisageables :

- Jeu de pondérations n°1 – données 2019 brutes (ex : 32 % des chambres occupées en IDF)
- Jeu de pondérations n°2 – données 2019 pour raisons personnelles (ex : 31 % des chambres occupées en IDF)
- Jeu de pondérations n°3 – données 2020 brutes (ex : 22 % des chambres occupées en IDF)

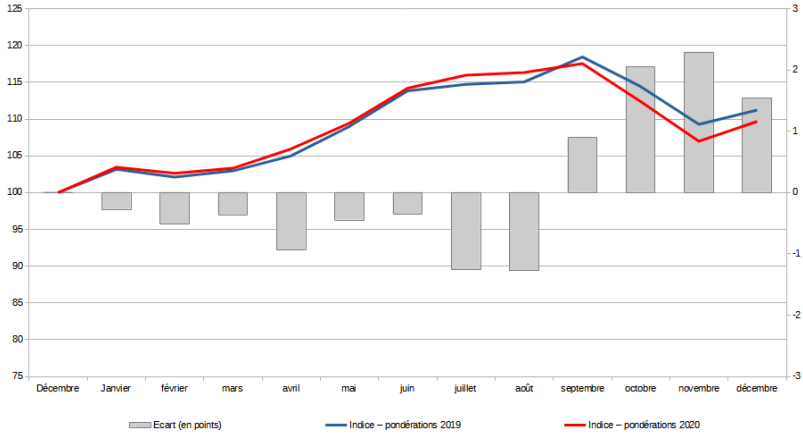
Le choix des pondérations : comparatif 1

Figure – Comparatif des pondérations 2019 brutes et pour raisons personnelles



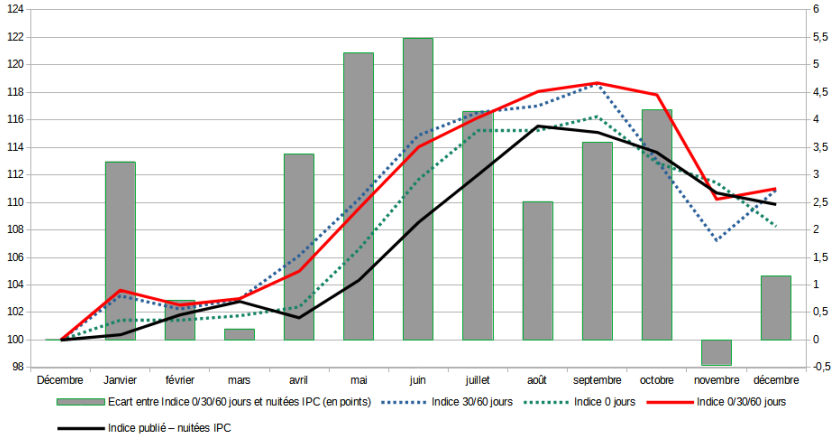
Le choix des pondérations : comparatif 2

Figure – Comparatif des pondérations 2019 et 2020 brutes



Indices vs indice publié

Figure – Comparatif des différents indices calculés



Plan

Conclusions

- Pour la suite de l'étude
  - Analyser les données pour les anteriorités 10 et 20 jours (collectées depuis décembre 2021).
  - Étudier les résultats avec un mois de base plus robuste (confinement entre le 30 octobre et le 15 décembre 2020)
  - Étudier s'il existe une substitution entre les antériorités
  - Effectuer une nouvelle année de tests et de réflexion sur une éventuelle utilisation pour l'IPC
- D'autres applications possibles
  Transport par autocar, festivals, évenements sportifs, location de véhicules, tourisme