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, division Prix à la consommation
(**) Insee, Direction des statistiques démographiques et sociales
(***) Insee, Direction des statistiques démographiques et sociales

Mots-clés : indices, webscraping, prix à la consommation, tourisme, classes homogènes

Domaines concernés : Collecte, Analyse des données et data science

Résumé

L'indice des prix à la consommation pour les nuitées hôtelières est constitué à partir de relevés effectués sur le terrain par des enquêtrices et enquêteurs qui relèvent le prix d'une chambre pour une nuitée le soir de la collecte pour 2 personnes avec petit déjeuner. 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 constituer 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, 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 sont par ailleurs brutes à la collecte par le robot et nécessitent un traitement nettoyage, par exemple la valeur pour une caractéristique n'est pas forcément décrite de la même manière entre deux observations.

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). 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). De nouvelles antériorités vont également être rajoutées pour le mois de décembre 2021 et l'année 2022.

Après une analyse des déterminants du prix des nuitées hôtelières, une comparaison de plusieurs méthodes pour élaborer un indice de prix à partir des données webscrapées a été faite. 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 savoir quelle méthode choisir et quels sont les différents niveaux d'agrégation.

Nous proposerons une comparaison des résultats de la méthode proposée de calcul d'indice avec les données webscrapées depuis 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. L'étude sera poursuivie avec les résultats en 2022, qui permettront d'approfondir la méthodologie et de faire une année de double calcul d'indice. Une évaluation de la nouvelle méthode de collecte et ses résultats sera faite à la suite de cette année. Elle permettra notamment de juger s'il est pertinent d'intégrer ces données dans le calcul de l'Indice.