17th Ottawa group meeting
Rome, 7 – 10 June 2022

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

Authors: Adrien Montbroussous

Abstract (maximum length of 400 words)

To improve the quality of hotel price index, a new method to collect data, web-scraping, could be useful. This paper presents an experiment using solely one website, booking.com. The website has been selected among several platforms and according to Eurostat recommendation. This platform has the highest number of monthly visits among the touristic hosting domain.

There are several axes on which new data could help improve the quality of the consumer price index. One of the principal target is to improve the coverage of the index, by including better touristic areas such as mountains and littorals. Another one is to enhance the sample size. To better seize the consumer behavior, it will also be interesting to take into account nights booked in advance. Indeed, the prices are currently only collected for the night of the survey: price collectors are going to the hotels and requesting the cost of a night for 2 persons with breakfast for the night of the collection.

The data collection has been realized using python. After a few tests, an automation of the scraping has been set in place using Kubernetes and GitLab CI on the platform Onyxia (platform developed by and for the french statistics services).

One of the difficulties is to have consistency in what we are collecting: there are various offers in the website and the description of each product is quite detailed. It is important as well to have a stable process because the platform booking.com is often the subject of technical changes such as HTML structure.

After collecting the data, an important aspect of the work was to clean it.

Along with the data collection part, the computation of an index using homogeneous classes has been studied. The idea is that the classes are homogeneous enough to consider that rooms are substitutable inside the classes and compute Jevon indexes for each one of these classes which will be aggregated with Laspeyres formulas in higher levels. A focus on the weight choice among the different available ones will also be presented.

Finally, the study will show the indexes obtained between December 2020 and May 2022 with this new data, and compare them with the current consumer prince index which is computed from data collected by price collectors.

REFERENCES