Evaluating different index methods for short-term rentals

Ken Van Loon[1]

Ottawa Group on Price Indices

Ottawa, Canada: 13-15 May 2024

Abstract

This paper aims to establish a methodology for a price index for short-term rentals or short-stay accommodation offered via collaborative economy platforms (i.e. Airbnb). With the increasing prevalence of short-term rentals and the evolving nature of the "sharing economy", including this sector in the CPI basket is crucial. Using information from tourism statistics and the number of listings on the Airbnb website three cities in Belgium have been scraped regularly. A brief overview of the scraped data is given. The collected data is used to evaluate different index methods. We start with traditional matched model index where the same accommodation is matched over time. Compiling a matched index does have a downside, namely the price difference between new and disappeared rentals is ignored. We examine three methods that try to take (or quality adjust for) the price difference between new and disappeared rentals. We first apply hedonics using the scraped characteristics. We then try to combine a matched index with imputations for new and disappeared dwellings using an imputation multilateral method. Finally we then compare this to a basic stratification method which clusters similar dwellings. Each of the methods will be evaluated in terms of its advantages and drawbacks. To integrate Airbnb in the Belgian CPI a weight needed to be compiled. We try to estimate a weight using scraped price information and volume information from the platform data for tourism statistics.

[1] Statistics Belgium, email: ken.vanloon@economie.fgov.be

This paper continues the research that was started under Eurostat grant 2019-BE-PRICE, for which the author expresses his gratitude. The views expressed in this paper are those of the author and do not necessarily reflect the views of Statistics Belgium.

Introduction

Airbnb is in Belgium, as is the case in most countries, the largest platform for short-term rentals. This paper examines different methodologies for calculating a price index for short-term rentals (or short-stay accommodation) offered via Airbnb and to be able to include this in the CPI (and HICP) basket.

This paper covers all the necessary steps to compile an index for Airbnb: data collection, determining weights and finally index methodology.

Section 1 gives an overview how and what data is collected from the Airbnb website. We try to collect all available characteristics which can then be used in the index methods. Section 2 describes how the cities that are scraped were selected. Section 3 describes the problems we faced when determining a weight for Airbnb, in the same section we also try to compile a weight for it using scraped price information and information on the number of stays from the platform data for tourism statistics. In March 2020, an agreement between Eurostat and four large online collaborative economy platforms (Airbnb, Booking, Expedia Group and Tripadvisor) was signed. This agreement provides experimental data on among others guest nights spent in short-stay accommodation offered via those four collaborative economy platforms

Section 4 is the main part of the paper, in this section a price index for each city is compiled using different index methods. We start with a traditional matched model index where the same accommodation is matched over time, this done using a time product dummy method and a multilateral matching GEKS-Jevons index. A matched index ignores the price difference between new and disappeared accommodation. To accommodate for this, we then look at 3 methods which try to take this into account. First, we apply hedonics, specifically the multilateral time dummy hedonic method and chaining adjacent periods using bilateral time dummy hedonic indices. We then try to combine a matched index with imputations for new and disappeared dwellings with an imputation GEKS-Jevons index. Finally, we then compare this to a basic stratification method which clusters similar accommodations. Each of the methods will be evaluated in terms of its advantages and drawbacks. Section 5 concludes the paper, discussing the main findings.

1. Data collection

Using information from tourism statistics and the number of listings on the Airbnb website, three cities in Belgium have been scraped regularly since 2020. For one city (Brussels), web scraping has been carried out since 2018 as a test case, although with data gaps. The two other cities (Antwerp and Ghent) have been scraped regularly since 2020. In this paper we only limit our analysis to the data from November 2020 onwards for all three of the cities, because setting up the web scraping for the 3 cities took some time and obviously also the COVID-19 pandemic happened. This resulted in the tourism sector being one of the most hard-hit sectors during because of lockdowns and travel restrictions. Due to this, we are only going to use data from November 2020 onwards. Data before that period is not going to be representative given the impact of hard lockdowns and travel restrictions.

How did we carry out our scraping and what information is collection? The procedure to scrape was as follows. For each of the cities a search was carried out, this then gives an overview page of the accommodations with a brief description of the accommodation, an image and a rating (and price, if dates and number of guests are selected), as well as a map with the location of the accommodations. Every search result is always limited to a maximum of 270 accommodations, with 18 accommodations being listed on every page, which gives a total of 15 pages to scrape. This limit of 270 observations is obviously problematic since we are interested in all the listings in each city.

To get all the listings for a city, more detailed searches must be carried out using iterative procedures. For instance, an iterative procedure that was applied, was limiting the price range, and then incrementally increasing the price range or by searching for more detailed observations (e.g., a neighborhood within a city). The goal of the iterative procedure is to find a search query that gives a number of accommodations below 270 and then by changing the price range or other criteria, all the accommodations can be obtained for a city.

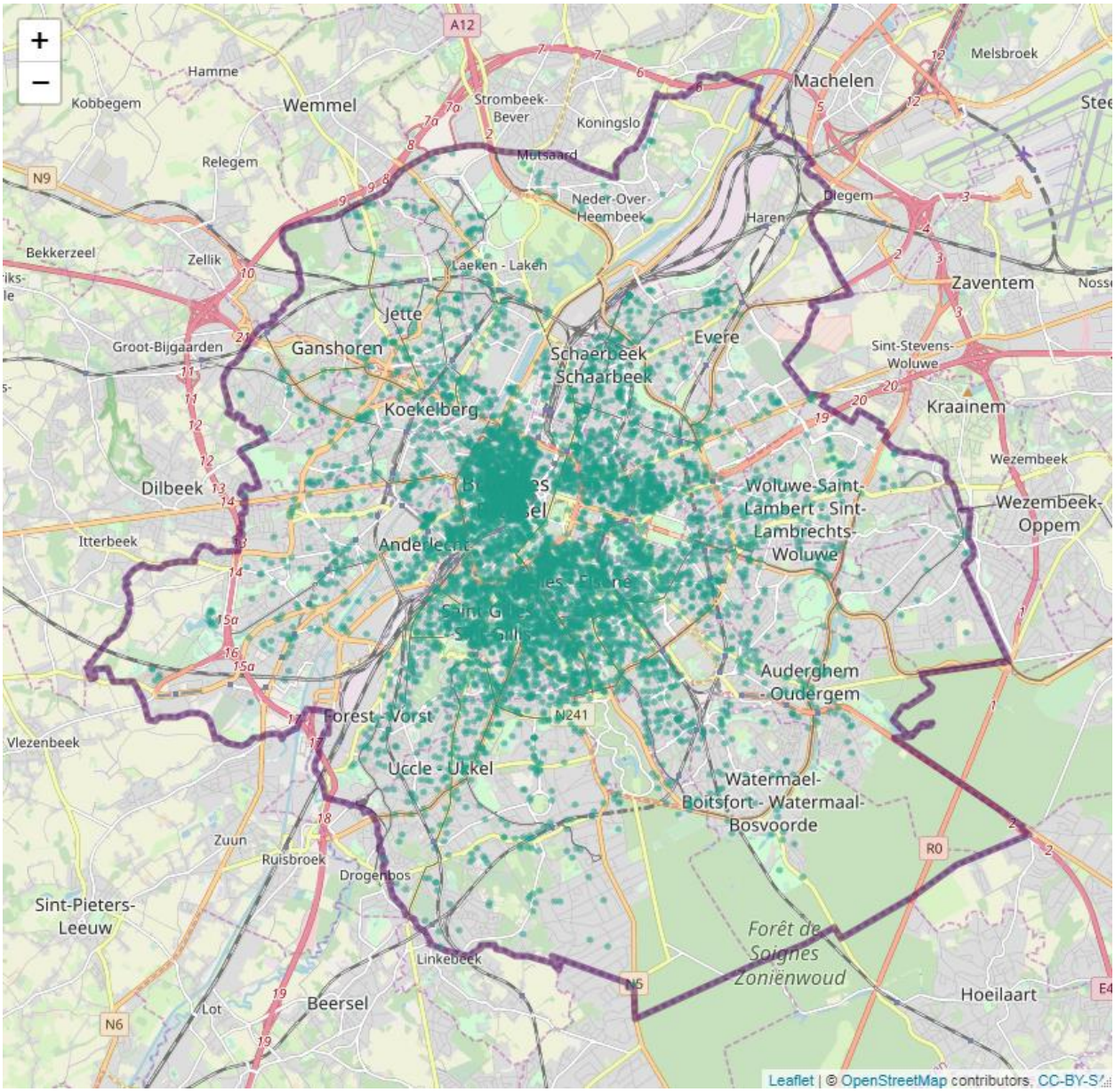
More detailed information for each listing or accommodation can then be found by clicking on it. Each accommodation also has a unique identifier, which can be used to go directly to the page of that accommodation at which detailed information and potential price determining characteristics can be found. This detailed information contains some typical characteristics of the accommodation (number of guests, bedrooms, bathrooms), as well as information about the amenities (Wi-Fi, dryer, free parking …). On the page of the accommodation also a calendar can be found with check-in dates, which gives the availability, the minimum number of days required for a stay as well as the corresponding price. Finally at the bottom of the page, a detailed rating of the accommodation can be found (cleanliness, location, accuracy, …) as well as an overall rating.

For the 3 cities, the scraping was implemented using the procedures outlined above as an input. All the data is collected using web scraping and carried out in R.

This results in over 100 potential variables that are available for each accommodation. In addition to the characteristics of the dwelling (appartement, paid parking, dryer, number of bedrooms, ...) there are also variables for the location (including. geo-coordinates).

In figure 1 below, the observations for Brussels are visualized. Each green dot represents one accommodation, especially in the center of Brussels there are many available accommodations.

Figure 1. Map of Airbnb accommodations in Brussels



The geo-coordinates of the listings are not 100% accurate, because Airbnb does not give the exact address of a property before completing the booking procedure for both privacy and security reasons. The location of the listing shown by Airbnb on a map is also not 100% accurate for the same reason. However, the proximity to the exact location is close. Consequently, the geo-coordinates do give a good indication of where the property is located, although with a small deviation.

2. Determining which cities to scrape

In the previous section we outlined our strategy for scraping, namely an iterative procedure to get all the accommodations and then collected detailed information for each individual accommodation. However, before we can do that, we first must determine which cities in Belgium we should scrape, due to resource and time constraints it is impossible to scrape everything.

In 2018 we carried out an analysis to determine which cities to scrape. The analysis used traditional tourism statistics, which at that time did not have information on short-term rentals (i.e., Airbnb), as well as the total number of listings per city found on Airbnb. We determined that the following 3 cities were the most important:

- Brussels with around 7500 accommodations every month
- Antwerp with around 2000 accommodations every month
- Ghent with around 1500 accommodations every month

All other cities had less than 1000 listings. To determine which cities to scrape, no administrative data could be used. Rental income is typically untaxed in Belgium, nor were (and are) accommodations on Airbnb required to have a license or register themselves with an administrative authority (as is the case in some other countries).

However, from a statistical point of view there have been improvements. In March 2020, an agreement between the European Commission (Eurostat) and four large online collaborative economy platforms (Airbnb, Booking, Expedia Group and Tripadvisor) was signed. This agreement provides experimental data on among others guest nights spent in short-stay accommodation offered via those four collaborative economy platforms from 2018 onwards. This data was first received by Eurostat and member states in 2021 and is then since updated yearly (although with a delay)

This data confirmed that the cities we chose were right ones, in the table below the number of stays for 2019 is given, this the most recent year with detailed information available before covid-19. The table shows that the 3 cities cover around 52.4% of the stays. With Brussels itself accounting for 35.4% of the stays. Now how important is this in terms of consumption expenditures?

Table 1. Number of short-term rentals in Belgium and reported cities

	2019	
	Total number of stays	**%**
Belgium	820.703	100,0%
Brussels	290.756	35,4%
Antwerp	84.996	10,4%
Ghent	53.952	6,6%
Charleroi	6.101	0,7%
Liège	24.826	3,0%
Bruges	48.600	5,9%
Ostend	24.478	3,0%

3. Estimating consumption expenditures

To integrate a potential Airbnb index into the Belgium CPI/HICP a weight for it needs to be compiled. Determining a weight for short-stay accommodation offered via collaborative economy platforms isn't easy. For fiscal reasons, Airbnb transactions for Belgium are carried out through their Luxembourg and Irish subsidiaries (as if the case for some other European Union countries). This means that Airbnb does not have a Belgian VAT number, nor does it have a tax representative in Belgium, neither is subsidiary of Airbnb is registered in Belgium and finally no annual accounts are being declared. Resulting in no centralized administrative data source being available which could potentially be used to determine a weight for Airbnb or other (smaller) platforms that provide accommodation services. Even if administrative data would be available, it might not be an easy task as shown by Rochlenge & Johannessen (2023) using data from the Norwegian Tax Authority.

We also cannot use national accounts information, because at the moment the consumption of accommodation services from the "sharing economy" is not captured under accommodation services in the Belgian national accounts (Basselier, Langenus and Walravens, 2018). The reasons for this are similar to the ones highlighted above, namely a lack of reliable data. In practice, (a part of the) expenditures are currently recorded as owner-occupied housing services, thus imputed rents for housing. However, this segment is neither captured in the Belgian CPI/HICP.

Luckily, the data sharing agreement with the four largest platform companies highlighted in the previous section does provide us with some information which could help us to determine a weight, especially if it is combined with other tourism statistics.

While the Eurostat data does not allow us to make a distinction between the four different companies, determining the total expenditure of short-stay accommodation offered via collaborative economy platforms would be a good start. Neither is it unusual in price statistics to only measure the price evolution of the total expenditure with a small sample of companies. And for this segment Airbnb is by far the most dominant player according to expert estimates, therefor we will extrapolate the information from Airbnb to the whole sector.

We scraped price information for other cities in 2020 to obtain an average price for each city, for everything else we applied spatial coefficients to determine a price. Now we need to determine what a stay is: how long and how many guests? From the Eurostat data we know that people on average stay 3 nights and the average guest size is 3 people. Using this information, we can calculate the average price for a stay on Airbnb. We also include the cleaning fee (on average around 25 euro at that time), the fee for additional guests (on average around 15 euro at that time), as well as the service charge of Airbnb which is around 14.5% of the total price.

This gives the following formula to obtain the average price for an Airbnb stay:

Airbnb price= ((price per night + fee for additional guest) x

number of nights + cleaning fee) x service charge.

This price per city can be multiplied with the number of stays per city to get estimated expenditure information. The results are shown in the table below.

Table 2. Estimated expenditure for short-term rentals

	2020		
	Price	Stays	Expenditure = price * stays
Brussels	338 €	97.653	33.013.608 €
Antwerp	427 €	32.165	13.723.995 €
Ghent	434 €	27.255	11.835.124 €
Charleroi	254 €	3.245	823.937 €
Liège	327 €	13.581	4.436.355 €
Bruges	491 €	19.095	9.383.845 €
Ostend	475 €	20.538	9.756.546 €
Total (excl. cities)	389 €	223.706	86.928.636 €
Total (incl. cities)			169.902.046 €

From traditional tourism statistics we know that around 25% of the stays are for business purposes, this means that around 75% of the estimate in the table is for consumption purposes, making the expenditure to be used in the CPI/HICP around € 127.358.500 in 2020. If we compare this to the total household financial monetary consumption expenditure (excluding for instance imputed rents) for 2020, the weight in the basket would be around 0,1 %. This would put it around the threshold of 1 part per thousand, this would require it to be covered in the HICP basket. Obviously, the calculation we carried out is only a rough estimate and the number of stays as well as total household consumption expenditure was heavily impact by the COVID-19 crisis. For instance, the number of stays in 2020 was around half of the number of stays in 2019. So, in "normal" years the weight in the basket would be around 0.15%.

Also, in the national accounts the estimation would be a lot more complex. For instance, a part of the consumption expenditure to accommodation services would need to be deducted from the imputed rents to avoid double counting.

Another aspect is intermediate consumption of Airbnb hosts, the goods and services they purchase to provide their rental services. This covers a wide range of goods and services such as consumption expenditures on toilet paper, food and drinks for breakfast, cleaning products, costs for water, heating and electricity, … as well as fees to Airbnb. This intermediate consumption should be deducted from their respective product categories.

We do not claim to be national accounts, therefore we leave this task up to them. The goal here is to make an estimate of what the weight for accommodation services from platforms could be, not to solve all the problems in the national accounts. This said, it should still be noted that deducting the intermediate consumption will normally have an impact on the weight of certain other product/service categories in the basket.

Beyond the challenges of avoiding double counting and determining how much intermediate consumption is, national accountants face many other challenges when trying to measure Airbnb and other platforms that provide accommodation services. An example for the Netherlands is given by (Hiemstra, 2017).

4. Index methods

While a weight is important to integrate Airbnb in the Belgian CPI/HICP, a proper price index is obvious essential. Using the scraped data of the 3 cities we will examine different index methods. Since it is scraped data, there is an absence of weights information. All the examined methods will therefore be unweighted price index formulas.

We start with traditional matched model index where the same accommodation is matched over time. We use two multilateral methods to do this, a time product dummy method and a multilateral matching GEKS-Jevons index. The results of these two methods are compared with a simple average price. Compiling a matched index does have a downside, namely the price difference between new and disappeared accommodation is ignored. We will then look at 3 methods which try to take into account (or quality adjust for) the price difference between new and disappeared accommodations.

We first apply hedonics using the scraped characteristics, we compile multilateral a Time Dummy Hedonic method. We will also look at chaining adjacent periods using bilateral time dummy hedonic indices, because chain drift is unlikely to be a problem with unweighted indexes. We then try to combine a matched index with imputations for new and disappeared dwellings using an imputation multilateral method, namely an imputation GEKS-Jevons index. Finally, we then compare this to a basic stratification method which clusters similar dwellings.

Each of the methods will be evaluated in terms of its advantages and drawbacks. In all of the calculations the index for November 2020 is equal to 100. All methods are compiled on the same dataset after data cleaning and outlier removal. Differences between the indices can thus only be attributed to methodological differences and not due to differences in the underlying data. Also, no splicing or extension methods are used which might cause differences. In practice however it is necessary to use a splicing or extension method when compiling non-revisable indices when using multilateral methods (Chessa, 2021).

4.1. Matched model indices

The first method we are going to examine, is a matched index. Recall that each accommodation has a unique identifier number. Using this variable, we can track the same accommodation over time. Calculating matched model indices is a traditional method for compiling a price index, namely get the price for the same "item" every month… but the same accommodation is not available every month. In other words, an accommodation might not be available in all months in the dataset.

One way to solve this problem of unavailability would be: making explicit imputation for the month when that dwelling is missing. However, here we have opted for a different strategy by compiling a multilateral matched model index, these types of indices use all possible matching items (or in this case accommodations) in the data (Ivancic, Diewert and Fox, 2011). To compile the index, we use two multilateral methods: a time product dummy index and a multilateral GEKS-Jevons index (de Haan and van de Laar, 2021).

Both methods have a particular interest to us, since Statistics Belgium currently uses variants of the GEKS-method for scanner data of supermarkets, as well as for consumer electronics and household appliances (Van Loon, Mierop and Roels, 2023). Likewise, the time product dummy is used for the "actual rentals for housing" index in the CPI/HICP, this approach is also used for the same market in New Zealand (Bentley and Krsinich, 2022). Therefore, these methods are of particular interest to us from a production point of view, since we want to limit methods across product segments.

A **time product dummy index** for a set of i accommodations for months 0 to T can be written as follows:

$$\ln p_i^t \;=\; \alpha + \sum_{t=1}^{T} \delta^t D_i^t + \sum_{i=1}^{N-1} \gamma_i D_i + \varepsilon_i^t$$

The parameters δ^t are the time dummy parameters, γ_i represent the accommodation fixed effects. The time dummy variable D_i^t has value 1 if accommodation i is available in period t and 0 if it is not available. The dummy variable D_i has value 1 if the observation relates to observation i and 0 otherwise. The price p_i^t is the average daily price for which the Airbnb unit can be rented in each month. We estimated the model using ordinary least squares (with a sparse matrix). By exponentiating the time dummy parameters, the index is obtained for all the periods in the pooled regression. The base period is by definition 100.

A multilateral **GEKS-Jevons index** for the same set of accommodations and periods can be written as:

$$P_{GEKS-J}^{0,t} = \prod_{l=0}^{T} \left(P_J^{0l} P_J^{lt}\right)^{(1/T+1)}$$

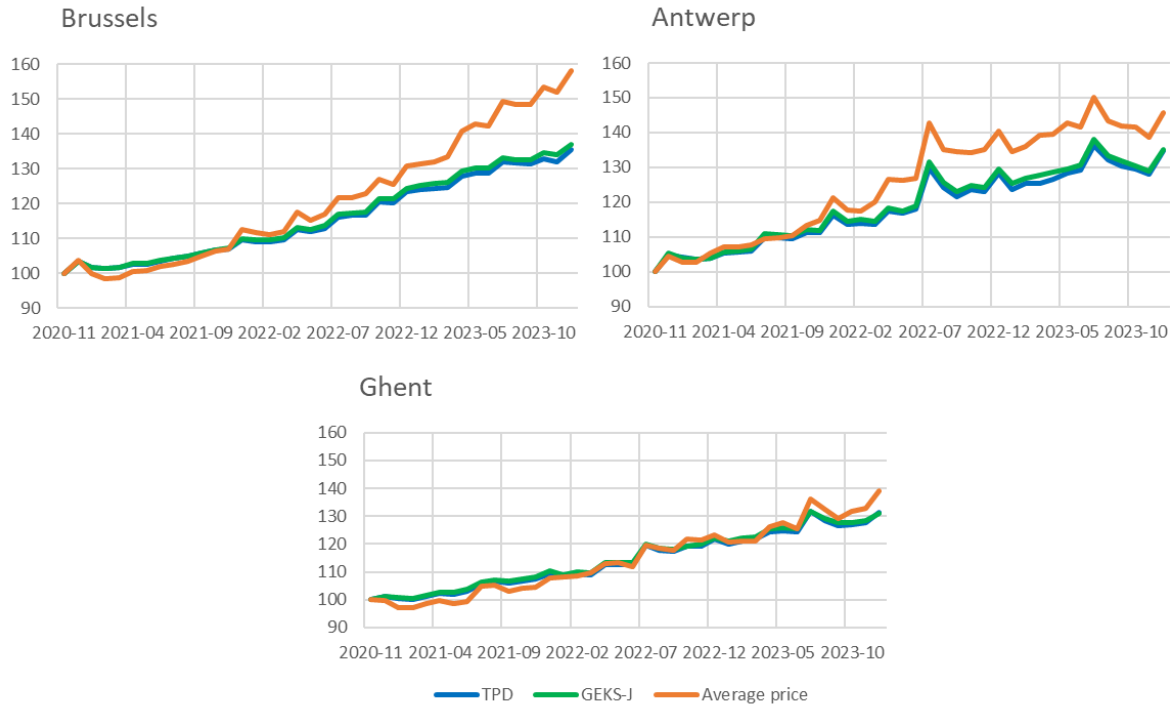
The indices P_J^{0l} and P_J^{lt} are the matching bilateral Jevons indices between periods 0 and l and period l and t respectively. For a set of $U_M^{0,t}$ matching accommodations between periods 0 and t, with $N_{0,t}$ being

the number of matching accommodations, the matching bilateral Jevons index between periods 0 and t is defined as:

$$P_J^{0,t} = \prod_{i \in U_M^{0,t}} \left(\frac{p_i^t}{p_i^0} \right)^{1/N_{0,t}}$$

The results for both methods for the 3 cities are shown in the figure 2. Also, an index compiled with a simple average price of all accommodations is shown.

Figure 2. Comparison of Time Product Dummy, GEKS-Jevons and a basic average price



For all 3 cities the trend is remarkably similar. An upward trend after the COVID-19 pandemic subdued, and the index tends to shift a bit upwards in the holiday periods (mostly December, July and August). The difference between the TPD and the GEKS-Jevons is limited, both give comparable results with minor short-term deviations. It is obvious that a simple average price across all accommodations cannot be used as a price index due to the heterogeneity, but it is merely shown for the sake of comparison. Using it would cause a significant upward bias of the index.

Compiling a matched index does have a downside. New accommodations that are posted online cannot be directly compared to existing (or disappeared) ones and that difference is thus not taken into. There are 2 ways to solve this problem. We can apply hedonics or stratification. We are first going to look at hedonics.

4.2. Hedonic methods

The hedonic method we are first going to examine is the multilateral **Time Dummy Hedonic** (TDH) method (Diewert, Heravi and Silver, 2009). This method is quite straightforward. We use a log-linear specification and estimate it using OLS.

$$\ln p_i^t = \alpha + \sum_{t=1}^{T} \delta^t D_i^t + \sum_{k=1}^{K} \beta_k z_{ik} + \varepsilon_i^t$$

With monthly average prices of all listings (i) for several periods (t) are pooled in the same regression, on their characteristics (z_{ik}) and on dummy variables for the periods (D_i^t).

The main advantage of the method is its simplicity since the index follows directly from the estimated time dummy parameters. After exponentiation of the time dummy parameters the index is obtained for all the periods in pooled regression. The base period equals by definition 100.

A disadvantage of the TDH is that it forces parameters to be fixed for the whole window. Another drawback is that, if there is no attrition of accommodations, there is no need to "quality adjust"; a matched index would be preferable in such cases (de Haan, 2010).

To try to accommodate for this we also compile a **Chained Bilateral Time Dummy Hedonic** method (TDH-Chained). In this method two adjacent periods are pooled together. This makes the model specification quite identical to the TDH, with the difference being that only 2 periods are pooled together and there is only 1 time dummy variable. This only forces the parameters to be fixed for adjacent months, instead of for all periods. The resulting index is again obtained after exponentiation of the time dummy parameter. However, since it now only gives the evolution between 2 consecutive periods, it is chained to the previous result to obtain the complete index series. A disadvantage of the bilateral chained TDH compared to the traditional TDH, is that fewer observations are used which might cause unstable coefficients with a more volatile index as a result.

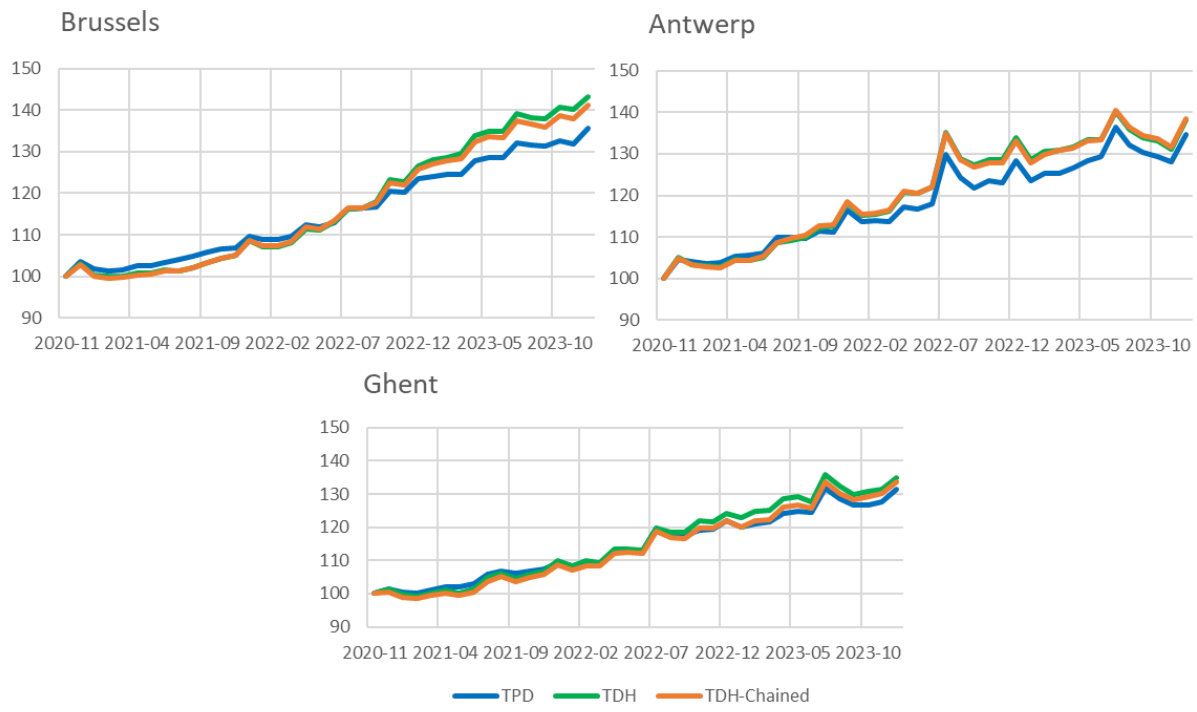
While monthly chaining can cause chain drift, we assume that chain drift is unlikely to be a problem for the TDH-Chained since we are using an unweighted index (de Haan, 2022). It is worth mentioning that in case there is no attrition of accommodations between two adjacent months, the bilateral time dummy hedonic indices between two months will be identical to a matched Jevons index.

The characteristic variables used in both methods are identical. While certain variables that traditionally have an impact on accommodation prices, such as living area, are not present in the dataset, these can be approximated by other variables such as the number of bedrooms, bathrooms and the number of guests that can stay. Also included in the model are the type of property, whether certain amenities are available (e.g., dryers, parking, garden, …), information related to host (e.g., response time, super host label, …), information related to the reviews (e.g., rating), as well as location variables (which area of the city). Using the geo-coordinates, we also compiled the distance to certain important landmarks. Using Cook's distance, outliers were determined. The price indices for all the methods (including the methods

of the previous section) are compiled using data without these outliers to guarantee the best comparability between methods.

Using all this information, resulted in a high "explanatory power" of the model. For instance, the R^2 was on average between 0.76 and 0.82 and did not differ that much from city to city. The results for both the hedonic methods for the 3 cities are shown in the figure 3 and compared with the TPD index of the previous section.

Figure 3. Comparison of Time Product Dummy, Time Dummy Hedonic method and Chained Bilateral Time Dummy Hedonic method



For all 3 cities the conclusion is quite similar. Taking into account the differences between and new and disappearing accommodation has an impact. For all 3 cities, the hedonic indices tend to be above the TPD index, this indicates that a matched pair method has a downward bias. The differences between the TDH and the Chained Bilateral TDH are limited, with some minor differences (the largest difference is noticeable for Brussels). This indicates that the fixity of the TDH is not that problematic or that the fewer observations used in the Chained Bilateral TDH is only a minor issue.

A disadvantage of the TDH is that, if there is no attrition of accommodations, there is no need to "quality adjust", since a matched index would be preferable in such cases. The Chained Bilateral TDH solves this for two adjacent months, but it does not solve it if an accommodation disappears for a few months and then returns. We will now look at a method that combines matching with hedonics.

4.3. Combining matching and hedonic methods

The previous section brings us to the imputation Jevons GEKS method with hedonic bilateral time dummy imputations as inputs. This method combines a matching index with explicit quality adjustments in the form of price imputations for new and disappeared products (or in our case accommodations). This imputation method with weighted hedonic bilateral time dummy indices was proposed by de Haan and Krsinich (2014) for the CCDI/GEKS Törnqvist. A method that we currently use for scanner data for consumer electronics and household appliances (Van Loon, 2021). Since no weighting information is available, we opt for the **imputation Jevons GEKS** index, which can be written as:

$$P_{GEKS-IJ}^{0,t} = \prod_{l=0}^{T} \left(P_{IJ}^{0l} P_{IJ}^{lt} \right)^{\left(1/T+1\right)}$$

This formula is equivalent to the matched GEKS Jevons index, with the bilateral matched Jevons indices P_J^{0l} and P_J^{lt} between periods 0 and l and period l and t being replaced by imputation Jevons indices P_{IJ}^{0l} and P_{IJ}^{lt}. The imputation Jevons index $P_{IJ}^{0,t}$ between period 0 and t, is defined as:

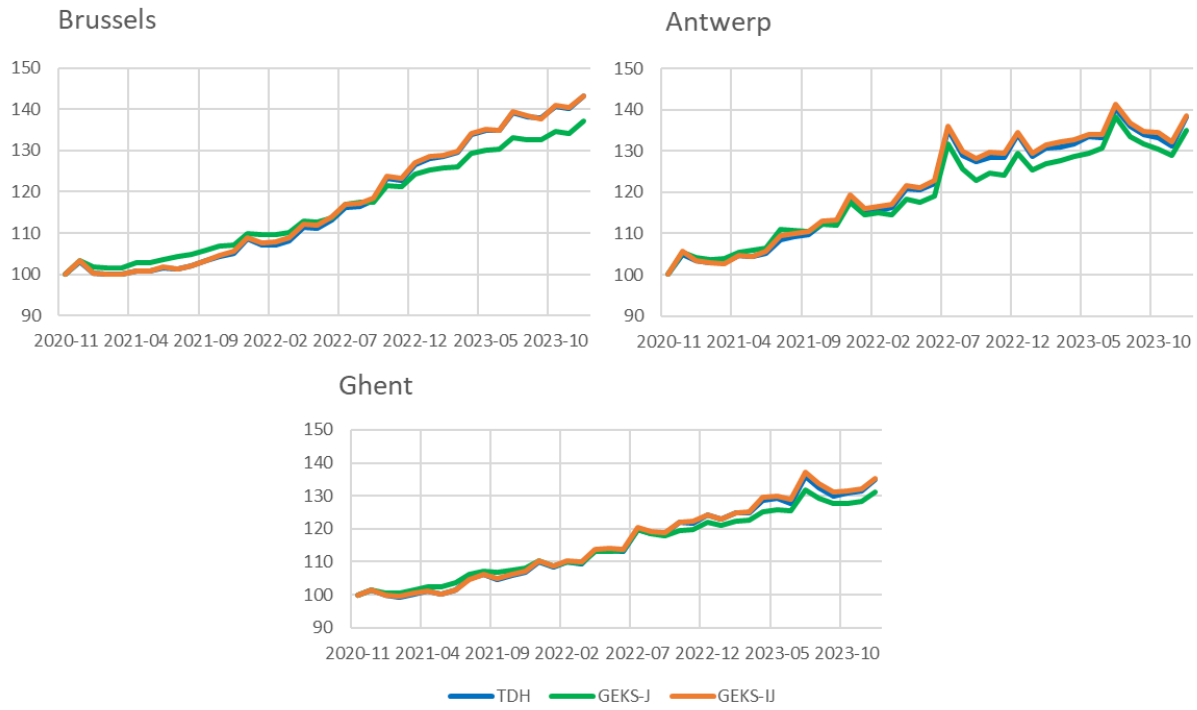
$$P_{IJ}^{0,t} = \prod_{i \in U_M^{0,t}} \left(\frac{p_i^t}{p_i^0} \right)^{0.5(N_0+N_t)} \prod_{i \in U_D^{0,t}} \left(\frac{\hat{p}_i^t}{p_i^0} \right)^{0.5(N_0)} \prod_{i \in U_N^{0,t}} \left(\frac{p_i^t}{\hat{p}_i^0} \right)^{0.5(N_t)}$$

With $U_M^{0,t}$ as the set of matched accommodations between the two periods, $U_D^{0,t}$ the set of disappeared accommodations which were available in period 0, but not in period t and U_N^{0t} the set of new accommodations which are available in period t, but not in 0. This results in $U_M^{0,t} \cup U_D^{0,t} \cup U_N^{0,t}$ being equal to $U^0 \cup U^t$. The number of accommodations of period 0 are denoted by the N_0 and those of period t by N_t. The imputed prices $\hat{p}$ are estimated by a bilateral time dummy hedonic method. If there is no attrition of accommodations (i.e. no new or disappearing dwellings), the above index is equal to a matched Jevons index and the resulting imputations Jevons GEKS is identical to the Jevons GEKS index. This is a desirable property, as a matched index would be preferable in such cases.

Using the bilateral time dummy hedonic method still forces fixity in the parameters, but now only between the two months compared, rather than over the whole window period as in the multilateral TDH. Other methods have been proposed to obtain the imputed prices $\hat{p}$ without parameter fixity, such as running a log-linear hedonic model for each period (De Haan and Daalmans, 2019). The advantage of running a log-linear hedonic model for each period is that the parameters are no longer fixed at all between periods, The downside of doing imputations for each period is that these cannot be carried out for new or disappeared characteristics. Put simply, new or disappeared characteristics have to be left out of the regression equation (Van Loon, 2021). Due to this, we do not investigate this method further and limit our research to using a bilateral time dummy hedonic imputation.

The results for the imputations Jevons GEKS index are shown in the figure 4 and compared with the TDH and GEKS-Jevons of the previous sections.

Figure 4. Comparison of Time Dummy Hedonic method, GEKS-Jevons and Imputation Jevons GEKS



The differences between the imputations Jevons GEKS and the TDH are hardly noticeable for all 3 cities. This confirms that the fixity of the coefficients in the TDH appears to be not much of a problem. The similarity of the TDH and the imputations Jevons GEKS can also be explained by the high explanatory power of the hedonic models. The imputations Jevons GEKS is more similar to the TDH than to the Chained Bilateral TDH. Indicating that it might be better to pool more months together than just 2 months as is the case of the Chained Bilateral TDH.

Besides hedonics, we can also use stratification to compare new accommodations directly with existing accommodations. If comparable listings are attributed to the same strata, then the price will be directly compared.

4.4. Stratification

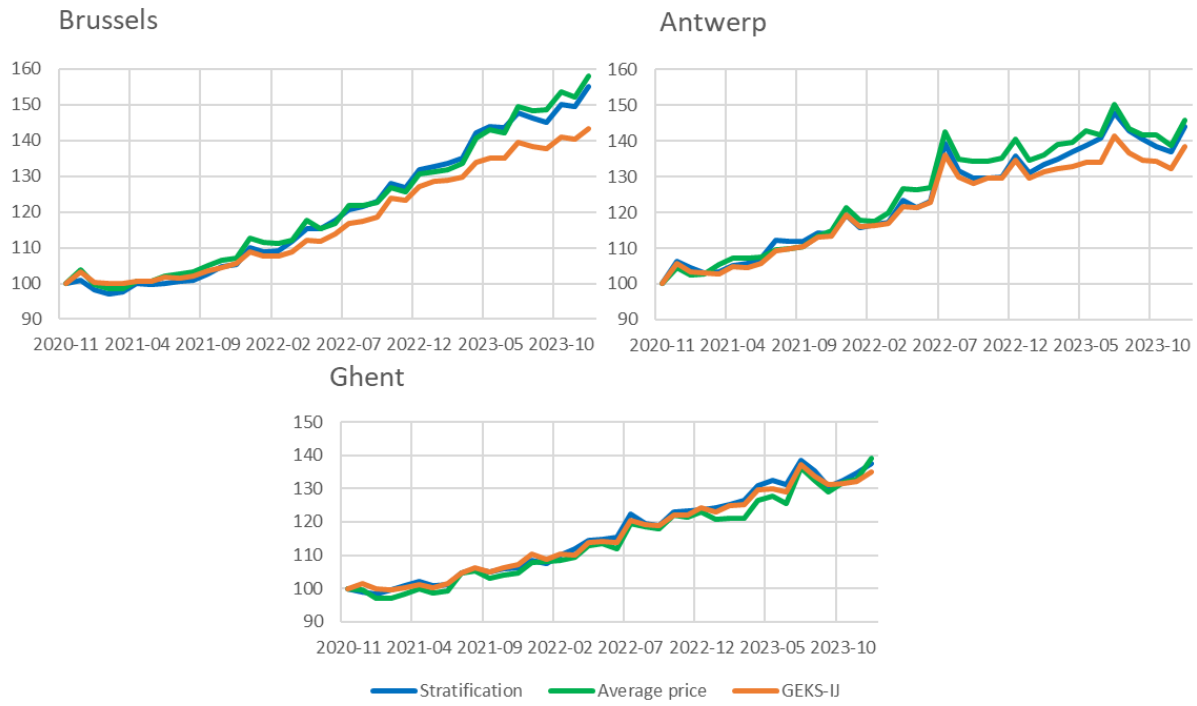
To stratify we used MARS (Chessa, 2019), this method combines the explained variance in prices with the item match over time. It thus makes a compromise between homogeneity and item match. It does this by calculating a "MARS score" which is derived by the multiplication of the two measures. The higher the MARS score the higher the combined value of the R^2 and the listing match. The highest MARS score is thus the optimal solution and will give you the best variables to stratify with (based on the MARS criteria).

A potential drawback of this is that while these methods try to make a compromise between homogeneity and product match, they might introduce a unit value bias if the strata end up being too heterogeneous (Daalmans, 2022).

Another drawback of stratification (and therefore also of MARS) is that only categorical variables can be used to stratify with in contrast to hedonics where this is not required, likewise geocoordinates cannot be used. So, to be able to use MARS, some variables needed to be transformed into categorical variables or be excluded. To be able compile stratification indices, we used the highest MARS-score to determine the variables that need to be used for stratification and then we compiled an index for each stratum. Then we aggregated those indices to obtain the global index per city.

The results are shown in figure 5 and compared with a simple average price across all accommodations and the imputations Jevons GEKS from the previous section (which was quite similar to the TDH).

Figure 5. Comparison of stratification, average price and Imputation Jevons GEKS



Only for Ghent is the difference between stratification and the imputations Jevons GEKS limited. This could be due to luck, since even a simple average price does not deviate much. For the other two cities the differences are significant, and the stratification results do not deviate much from the average price. Indicating that there is a unit value bias when using stratification. This bias is mostly caused by not being able to include certain variables or having to transform them which results in the strata being to heterogeneous.

5. Conclusion

Using information from tourism statistics and the number of listings on the Airbnb website, we determined which 3 cities we were going to scrape. The Eurostat platform data for Tourism Statistics received in 2021 confirmed we have chosen representative cities.

To integrate Airbnb in the Belgian CPI/HICP a weight needs to be compiled. We tried to estimate a weight. Using our procedures, we got a weight which was around the 1 in thousand threshold for 2020, meaning that consumption expenditures appear to be significant enough for it to be included in the basket. Obviously, further collaboration with our national accounts colleagues in the upcoming years will be necessary to solve some of the outstanding issues (mostly related to owner-occupied housing services and intermediate consumption).

Using the scraped data, we compiled indices using 4 methods: matched indices (TPD and GEKS-J), hedonic indices (TDH and bilateral chained TDH), an index that combines matching and hedonics (GEKS-IJ) and finally we also used stratification.

Our research indicates that it is possible to compile plausible Airbnb indices using scraped data. Since there is attrition of accommodations and a price and quality difference between new and disappeared accommodations, it is best to use methods that can capture this. In our datasets a matched index underestimated the price increase. While the difference between a TDH and bilateral chained TDH was small, the TDH appears to perform a bit better. This indicates that the fixity of the parameters is not that much of an issue. However, just to be sure it might be worthwhile to use an imputations Jevons GEKS. This index was almost identical to the TDH, but it has as an advantage that in case of no attrition it will be identical to a matched index. Finally, we also examined stratification, this appears to suffer from a unit value bias due to some variables which needed to be transformed into categorical variables or be excluded from the calculation.

References

Basselier R., Langenus G. and Walravens L. (2018), De opkomst van de deeleconomie – *Economisch Tijdschift, September 2018*.

Bentley A. and Krsinich, F. (2022), Timely Rental Price Indices for thin markets: Revisiting a chained property fixed-effects estimator - *Paper presented at the 17th Ottawa Group meeting, 7-10 June 2022, Rome, Italy.*

Chessa A. (2019), MARS: A Method for Defining Products and Linking Barcodes of Item Relaunches - *Paper presented at the 16th meeting of the Ottawa Group, 08-10 May 2019, Rio de Janiero, Brazil.*

Chessa A. (2021), Extension of multilateral index series over time: Analysis and comparison of methods - *Paper presented at the Meeting of the Group of Experts on Consumer Price Indices, 2-10 June 2021, Geneva, Switzerland.*

Daalmans J. (2022), Multilateral indices and the relaunch problem: product clustering and alternative solutions - *Paper presented at the 17th Ottawa Group meeting, 7-10 June 2022, Rome, Italy.*

Diewert E. W., Heravi S. and Silver M. (2009), Hedonic Imputation versus Time Dummy Hedonic Indexes - *NBER Chapters in Price Index Concepts and Measurement, 161-196, National Bureau of Economic Research.*

de Haan J. (2010), Hedonic Price Indexes: A Comparison of Imputation, Time Dummy and 'Re-Pricing' Methods - *Journal of Economics and Statistics 230, 772-791*.

de Haan J. and Krsinich, F. (2014), Scanner Data and the Treatment of Quality Change in Non-Revisable Price Indexes - *Journal of Business and Economic Statistics 32(3), 341-358*.

de Haan J. and Daalmans J. (2019), Scanner Data in the CPI: The Imputation CCDI Index Revisited - *Paper presented at the 16th meeting of the Ottawa Group, 08-10 May 2019, Rio de Janiero, Brazil.*

de Haan J. and van de Laar R. (2021), House Price Indexes: A Comparison of Repeat Sales and Other Multilateral Methods – *Room Paper at the 17th meeting of the Ottawa Group, 07-10 June 2022, Rome, Italy.*

de Haan J. (2022), Multilateral Hedonic House Price Indexes - *Room Paper at the 17th meeting of the Ottawa Group, 07-10 June 2022, Rome, Italy.*

Hiemstra L. (2017), Measuring challenges of the sharing economy: the case of Airbnb - *Paper presented at the OECD Working Party on National Accounts, 09-10 November 2017, Paris.*

Ivancic L., Diewert, E. W. and Fox, K.J. (2011), Scanner data, time aggregation and the construction of price indexes - *Journal of Econometrics 161, 24-35*.

Rochlenge C. & Johannessen R. (2023), Sharing economy or just utilization of new business models? - *Paper presented at the Meeting of the Group of Experts on Consumer Price Indices, 07-09 June 2023, Geneva, Switzerland.*

Van Loon K. (2021), Implementing transaction data for consumer electronics and household appliances - *Report written for Eurostat grant project.*

Van Loon K., Mierop A. and Roels D. (2023), Using multilateral hedonic methods to capture product relaunches - *Paper presented at the Meeting of the Group of Experts on Consumer Price Indices, 07-09 June 2023, Geneva, Switzerland.*