Discussion paper

Estimation of the number of guests and overnight stays in platform-related accommodations

S.T. Spinder
(supervised by N. Heerschap, D. Windmeijer, S. Ortega)

November 2019

Content

1. Introduction 3
   1.1 Background 3
   1.2 Other research 4
   1.3 Approach 5

2. Data 7

3. Methodological approach 9
   3.1 Review rate 10
   3.2 Extracting numbers of guests and nights 12
   3.3 Removing bias 13
   3.4 Background of guests 14

4. Results 15
   4.1 Number of listings (step 1) 15
   4.2 Number of bookings (step 2) 18
   4.3 Number of guests (step 3) 21
   4.4 Country of residence 23
   4.5 Number of nights (step 4) 24
   4.6 Comparison with TAS 26
   4.7 External validation 27
   4.8 Obtaining totals and removing duplicates 29

5. Conclusion and suggestions 31

References 32

Appendix 33

1. Introduction

1.1 Background

Statistics Netherlands is responsible for the production of the basic tourism statistics in the Netherlands. One of these statistics concerns the number of tourists[1] by country of residence, that stay in tourism accommodation establishments in the Netherlands by region and by month (Tourism Accommodation Statistics[2]; TAS). The TAS also includes the number of nights these tourists stay in such an accommodation. Tourists can be from abroad (inbound tourists), but also from the Netherlands (domestic tourism). The output of the TAS provide important indicators for the monitoring of the development of tourism in the Netherlands.
To produce the TAS Statistics Netherlands uses a sample survey among registered tourism accommodation establishments, such as hotels, bungalow parks and camping sites. Every month Statistics Netherlands asks these establishments how many guests they received and how many nights these guests stayed in their accommodation. To minimize the survey burden, smaller establishments are excluded from the sample.
Of course tourists also stay in other kinds of accommodation, like second homes, rented apartments and homes, small B&B's, boats and with family and friends. Until recent it was believed that this was a relatively small segment of tourism in the Netherlands.[3] In addition, in practice this segment was also difficult to identify and measure. Therefor this segment of tourism was not part of the regular monitoring. However, internet and especially the emergence of platforms, like Airbnb, Booking etc., has clearly changed this situation. These kinds of platforms have made it much easier for private individuals to rent rooms, apartments or even houses to tourists. This has led to the rapid growth of this segment of tourism, especially in those places where the demand is high and the supply of hotel beds is low or relatively expensive. Big cities, like Amsterdam, or coastal areas are good examples of such a situation. Platforms also offer people with unused space to earn some extra money, although there is a rapid commercialization in this domain as well. The growing attention for this phenomenon has led to questions, such as: where are these accommodations located; how big has this segment of tourism really become; how many and which tourists book through these platforms; and how many nights do they stay in a certain region? The attention for this phenomenon was further reinforced by the supposed negative effects of the growing use by tourists

[1] Besides a tourist, who stays – by definition - at least one night in an accommodation, tourism also covers same day visitors. Tourism also includes business travellers or travelling for other purposes than leisure, as long as people are outside their usual environment.

[2] De Statistiek Logiesaccommodaties.

[3] Around 20-30 percent of all nights spent by tourists, when looking at demand side statistics.

of these kinds of accommodation. For example, do they cause nuisance to neighborhoods, where tourists normally do not stay? Does it lead to higher prices of real estate? Is living space withdrawn from the market which could otherwise be used for local people?
So, it has become inevitable for Statistics Netherlands to have a more detailed look at this segment of the tourism market. The first question then is how big is the supply of these kind of accommodations and where are these accommodations located? And how can the number tourists, who uses these kinds of accommodation, be measured, including their country of residence and the number of nights they stay?
The first possibility which was investigated by Statistics Netherlands was to ask the needed information directly from the platforms. Due to competitive reasons platforms are not quite willing to disclose this information. There is also no law that forces platforms to supply this information to Statistics Netherlands. Besides this, especially the bigger international platforms often have no branch in the Netherlands. However, the fact that these platforms now are digital provides the possibility to web scrape information from the internet and also use this information to estimate the number of tourists using these kinds of accommodation and the number of nights tourists spend there.
So, the need for more information on the market share of this segment of tourism, difficulties to obtaining information directly from platforms themselves or the people who rent space through these platforms and possibilities to use information from the web, has led to the assignment to web scrape data from Airbnb and similar platforms and try to estimate on that basis the number of tourists and the number of overnight stays, distinguished by region and country of residence.

1.2 Other research

Statistics Netherlands is not the first organization to web scrape platforms like Airbnb and estimate the number of tourists and nights (see e.g. Colliers International, 2018 and Ecorys, 2018). The figures that are produced by these organizations, however, differ strongly. This is, partially, because of the many different methods used to estimate the total number of guests and nights. This paper presents an attempt to come at a more statistically sound estimate.
Information about Airbnb is known from other sources, such as AirDNA (AirDNA, 2018), InsideAirbnb (Inside Airbnb, 2019) and Beformation (Beformation, 2019).[4] These organizations scrape the website of Airbnb themselves. There after they use different methods to estimate the total number of guests and overnight stays. AirDNA (AirDNA, 2018), for example, uses the calendar on the Airbnb website to get to statistics. The calendar shows whether an Airbnb-accommodation is available on a certain date. However, it is not clear whether an Airbnb is not available, because it is rented or, because the host has decided not to rent on that particular day. This information was available before 2015 Q4, after which Airbnb decided

[4] Beside the research which is mentioned here, Airbnb is also web scraped by other organizations, like the cities of Amsterdam, The Hague and Utrecht. This information is often web scraped for the enforcement of local rules concerning the use of Airbnb.

not to publish this information anymore. To distinguish between the options 'already rented' and 'temporally not for rent', AirDNA uses a model based on the 18 months of data before 2015 Q4. This assumes the data before 2015 Q4 is representative for the period after that date, which can be doubted because of the growth of Airbnb and the increasing regulation in cities such as Amsterdam (Ecorys, 2018), which may or may not have caused a different composition of hosts.
Inside Airbnb (Inside Airbnb, 2019) uses the number of reviews to get to statistics. They convert the number of reviews to a number of bookings, where 1 review is equal to 2 bookings. The conversion rate of 50% (i.e., half of the bookings have a corresponding review) chosen by Inside Airbnb lays between two values 72% and 30.5%. The first comes from Airbnb founder Brian Chesky (Airbnb, 2012), and the second from a report by the San Francisco Board of Supervisors Budget and Legislative Analyst (San Francisco Board of Supervisors Budget and Legislative Analyst, 2015). To convert the number of bookings to a number of nights, Inside Airbnb uses either information from Airbnb itself (Airbnb, sd), that states guests stay an average 3.9 nights in Amsterdam, either 3 nights per booking or the minimum number of nights that can be stayed somewhere. The first value, 3.9, is from June 2013.
Research by Beformation (Beformation, 2019), for the province of Noord-Holland, uses similar methodology. They also scrape the number of reviews, but use a different review rate, namely 80%, i.e., 80% of bookings have corresponding reviews. To convert this to number of overnight stays, they use the minimum number of nights that can be stayed at an accommodation. This is only a lower boundary for the real number of nights. Besides that, Beformation (Beformation, 2019) defines an overnight stay as 'the stay of one or more guests for one night', while we are more interested in 'every night stayed by every guest'. The latter requires one to estimate the number of guest staying as well.
To get values that are more in line with the TAS, a method is needed that uses the fewest possible number of assumptions and can be applied to not only Airbnb, but also to other platforms, such as Booking.com, Micazu, HomeAway etc. Besides that, there is a need to get more insight into the platforms used in all of the Netherlands, rather than just focusing on large cities, as previous research has done.

1.3 Approach

The first step of the process was to identify the major booking platforms, that are active in the Netherlands. This was mainly done on the bases of the number of listings (accommodations for rent) on the platforms. The second step was to build web scrapers for these platforms, so that information could be collected on a regular basis. Information was, among others, collected on the listings, reviews and profiles. Because the duration of the whole project was rather short, the amount of data collected was also rather limited. The legal aspects of web scraping were also looked into. Because this is an experimental research for national statistics, web scraping is allowed. Data are also only published on an aggregated level. The way this first part of the project was carried out is described in Sluijpers, 2019, 'Gegevens van Airbnb-achtige platformen' (Data from Airbnb-like platforms;

Sluipers, P., 2019). The second part of the project mainly focused on the estimation process. This second part of the project is described in this paper.
Before ending the first part of the project and starting the second part of the project a brainstorm session was held to distinguish and evaluate different methods to estimate the number of tourists and the number of overnight stays on the basis of web scraped data. Three methods emerged:

- Using the occupancy rate of small tourism accommodation establishments of the TAS;
- Using calendar information from the platforms;
- Using review information from the platforms.

Given that not all platforms have a calendar, the decision was taken to research the review method, i.e. similar to the approach used by Inside Airbnb (Inside Airbnb, 2019) and Beformation (Beformation, 2019). Rather than using aggregates from Airbnb, or using the minimum number of stays, an attempt was made to extract the number of guests and the number of nights per booking from the review text. The research also looked into other ways to get to the conversion rate, i.e., the percentage of bookings with corresponding reviews, depending on the platform. Because of the time constraints of the project, the exercise was limited to data of Airbnb.com and Booking.com.
Finally, it must be clear that this concerns an experimental study. The figures mentioned in this paper must also be seen as such. They are not official statistics from Statistics Netherlands, they are experimental data.

2. Data

Data were obtained from the websites of Airbnb.com and Booking.com. These platforms were chosen as they comprise a large share of the market for accommodations. In the case of Airbnb, most of the accommodations are part of the population of interest, namely accommodations with less than 5 beds. For Booking.com, the share of such accommodations is unknown. Subsequent research could focus on obtaining data from other platforms, such as HomeAway, Micazu and Belvilla, i.e. incorporating the work of Sluijpers (2019). As time was limited, the decision was taken to limit the research to the two mentioned platforms. Nonetheless, the methods described in this paper should extend to those platforms as well.

Firstly, listings (accommodations for rent) were obtained from Airbnb.com and Booking.com at several evenly distributed points in time. This was done by using the region filter “the Netherlands”. For Airbnb, this was done only six times during the research period, as Airbnb makes it hard to obtain information by blocking the connection if too many requests are sent to the server. For Booking.com, the listings were taken from the website almost daily. The total number of listings was found to be almost 40,000 for Airbnb, at any point in time, and between 11,000 and 15,000 for Booking.com. Section 4.1 gives a more detailed overview of these values for every point in time.

For each listing, an attempt was made to scrape as much information as possible that could help with estimating the total number of guests and nights. Variables of interest for Airbnb are, among others, the total capacity, the latitude, the longitude and the total number of reviews. For Booking.com latitude, longitude and review counts were also found. Information about capacity is available on Booking.com, but is presented split by room. As it is unclear whether Booking.com shows all the rooms of every accommodation or merely a subset, this information was not used.

After obtaining the listings, reviews were scraped for every listing. For Airbnb.com there is one text field for the entire review, in which guests can freely provide the information they want to leave. In addition, information about the language of use is attached to every review. This information can be useful for languages different than English, but does not tell too much about the country of residence if they are in English, as many guests from countries outside the Anglosphere also leave reviews in English.

In total 1,002,368 were found for the 37,436 Airbnb accommodations that were obtained on the 8th of June 2019. As there was a considerable lag between scraping the accommodations on the 8th of June and scraping the reviews on the 27th, and as not all of the listings obtained were actually in the Netherlands, eventually 940,000 reviews were left that could be linked to an accommodation. The review dates range from the 2th of February 2010 to the 27th of June 2019.

For Booking.com every review has a title field, a positive review field and a negative review field. All these fields were merged into one review. Information can also be found on the country of residence of every guest.

A total of 3,020,665 reviews were found and retrieved for the 13,432 Booking.com accommodations that were obtained on the 4th of July. The earliest review is from the 5th of July 2017, while the last review can be found on the 11th of July 2019.

As only two years of reviews can be found, it is suspected that Booking.com removes older reviews to stay GDPR[5] compliant. For Booking.com eventually 2,800,140 reviews could be linked to the set of listings.
For both Airbnb.com and Booking.com, only the review date is listed and not the date when the guest left the accommodation. As the latter is of main interest, an adjustment is needed. For this purpose, peaks of review dates were compared to holidays. On average, they seemed to differ about 8 days. Therefore, the date 8 days before the review date was taken as the 'leaving date'. The leaving date is used by the TAS to count the tourist in that specific month. Subsequently all the nights spent by the tourist are also allocated to that month, even in the case that the tourist already arrived in the month before (see below).

[5] The European General Data Protection Regulation (GDPR) is applicable as of May 25th, 2018 in all member states to harmonize data privacy laws across Europe. More information at https://gdpr-info.eu/

3. Methodological approach

In this research the variables of primary interest are the number of guests and number of nights guests stay in Airbnb-like accommodations, in a certain period. The particular definition that is used for a guest (or a tourist), is in accordance with the TAS and is the following:

> *'A visitor that stays one or more nights at an accommodation. A guest who stays for more than two months is considered a regular user and does not count for this statistic. For each month, the guests are counted that leave (not arrive) in a particular month'*

The last sentence implies that a guest arriving at an accommodation on the 15th of December and leaving on the 2nd of January is counted for January and not for December. For nights, the definition is:

> *'All nights that guests spend in an accommodation'*

In this definition, each night by each guest is counted as a night.[6] E.g., if three guests stay at an accommodation for two nights, this one booking is counted as six nights. For every individual guest, this gives the following formula:

$$n_i = g_i * npg_i$$

where

n_i: The number of nights for booking i
g_i: The number of guests for booking i
npg_i: The (average) number of nights per guest for booking i

In an ideal situation (in which this information is known for every booking), a simple sum over all the bookings could be made to derive both the total number of guests and the total number of nights:

$N = \sum_{i=1}^{B} n_i = \sum_{i=1}^{B} g_i \, * \, npg_i$ where B is the total number of bookings.

However, platforms do not provide any information on g_i, npg_i nor B. To extract this information, the reviews of the accommodations on the platforms can be used. Reviews contain, besides a judgement on the quality of the corresponding accommodation, also information on the guest's experience. Aspects of this experience are, among others, the length of stay and the number of guests accompanying the main guest (who made the booking). Such information is considered valuable to make estimates of aggregates.
Section 3.1 will deal with the issue of turning reviews into bookings by means of a review rate, i.e. the rate with which guests leave reviews after they have stayed in an accommodation. Section 3.2 describes a way to extract information on numbers

[6] Also here all the nights spent are counted in the month that the guest leaves the accommodation.

of guests and numbers of nights from reviews. As not all reviews contain information on numbers of guests and number of nights, adjustments have to be made to obtain estimates for the complete set of reviews. Section 3.3 covers these adjustments.
An overview of the complete process is shown in Figure 1.

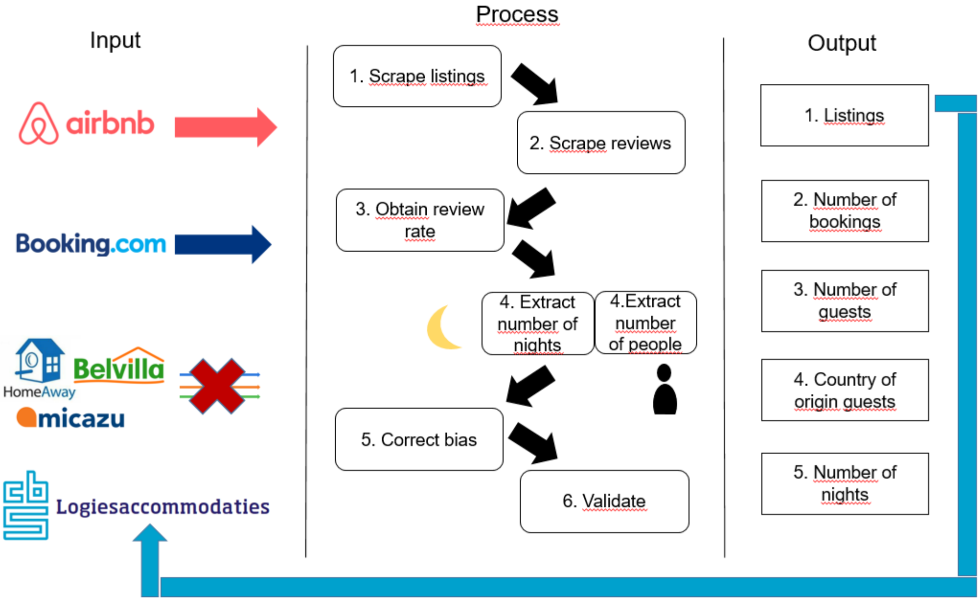

Figure 1: A flowchart showing the methodology used in this paper.

3.1 Review rate

As noted before, the reviews are basically a subset of the total set of bookings. The composition of the population of bookings is likely to differ from that based on the review sample. It is also not known in what way they differ, as there is no information on the guests who book, but also do not leave reviews. In other words, little is known about the population of guests whose reviews are not available, neither on possible differences in their length of stay nor on the number of guests that stay along with them, if any.
Hence, in this research and without loss of generality, the following assumptions are required to keep the calculations bounded.

> *Assumption 1*: Guest(s) who leave reviews do not differ significantly on relevant variables from those of the general population of bookings, i.e. on the length of stay or the number of guest staying along.
> *Assumption 2*: The review rate is consistent across time and across guests.

The latter makes it possible to divide the number of reviews (R) by the review rate (r) to get to a number of bookings. The review rate is not known exactly, but can be estimated based on publicly available data.
For Airbnb, several sources report on the review rate. According to Brian Chesky, founder of Airbnb and current CEO, the worldwide review rate was about 72% in

2012 (Airbnb, 2012). The (San Francisco Board of Supervisors Budget and Legislative Analyst, 2015) considers a high impact scenario in which the review rate is equal to 30.5%. (Inside Airbnb, 2019), a third party specialized in reporting on Airbnb, uses the average of these values as a review rate, about 50%. Finally, (Beformation, 2019), a consultancy company, mentions a review rate of 80%.
As these values have a large variance, in this project a different approach was taken to get to a more plausible range of possible values. Forum posts were looked at in which hosts explicitly mention their own review rate. These can be found on websites such as airhostsforum.com and ourbnb.com. Review rates were found in 56 posts. 29 of these posts also mentioned the exact number of bookings they had and the exact number of guests that had left reviews. Averaging the review rates in the 56 posts gives a review rate of 76.6%. Summing the number of reviews and the number of bookings for the 29 posts and dividing the former by the latter gives a rate of 67.3%. The former is based on more observations, but the latter is based on more reliable observations, i.e., hosts providing exact numbers. The middle of these values is exactly 72%, the review rate mentioned by Airbnb itself (Airbnb, 2012). So, a review rate of 72% is used for the estimations later in this paper.
For Booking.com there are no estimates from external sources. However, Booking.com does mention how many times listings have been booked in the last 24 hours, for a certain set of listings. These bookings can be aggregated to the total number of bookings and compared to the average number of reviews on a day, within a month, for this same set of listings. By doing this several times, a histogram of possible review rates can be made, as shown in Figure 2. The mean of this distribution is used (24%) as the review rate in subsequent calculations.

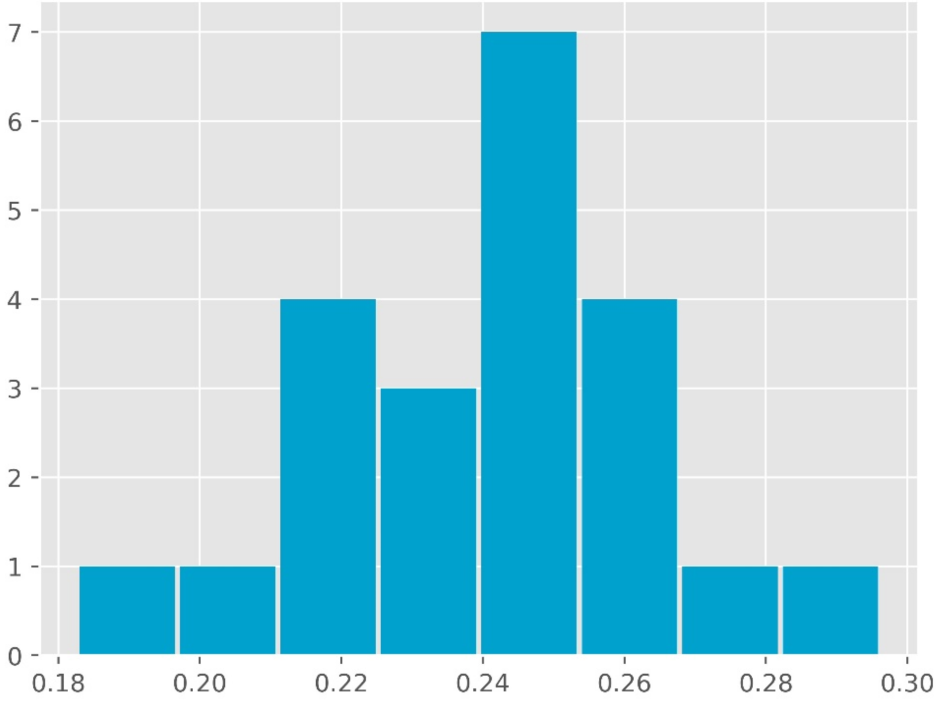

Figure 2: Histogram of review rates on Booking.com based on bookings in the period 17th July 2019 to 28th August 2019. For the average number of reviews on a day, data from April was used, as we did not have reviews for July and August. April was chosen as it has similar traffic to July and August in the Tourism Accommodations Statistics (TAS).

3.2 Extracting numbers of guests and nights

Figure 3 gives a schematic overview of the proportion of bookings that have corresponding reviews and the proportion of bookings for which reviews contain useful information on numbers of nights and numbers of guests. It is estimated that about 70% of guests leave reviews on Airbnb. Approximately 1% of these reviews were found to have information on either the number of nights or the number of guests staying. As the total number of reviews is over 1 million, this still offers a set of about 10,000 reviews with useful information.

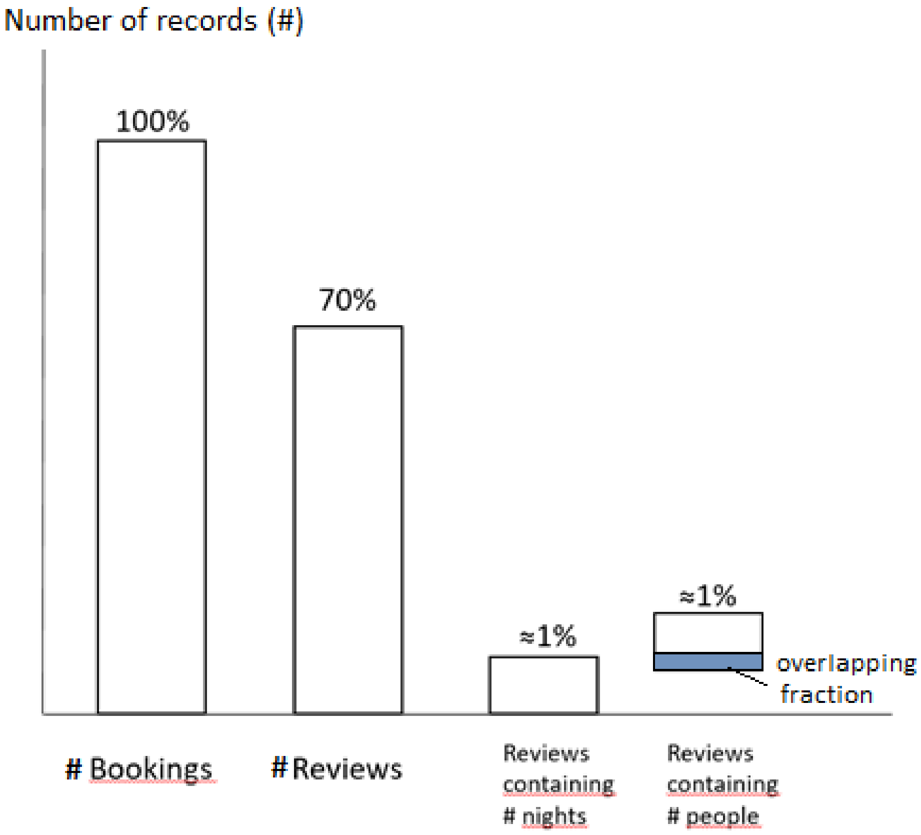

Figure 3: A schematic overview of the number of bookings, the number of reviews and the number of reviews containing useful information, i.e., the number of nights and the number of guests that have used an Airbnb/Booking.com accommodation

Estimates for the total number of guests are made by means of a filter that extracts reviews containing information on the number of guests. First, the review is split up into several sentences. Then for each sentence, it is checked whether there is a verb that expresses a guest has stayed at an accommodation and an unit that expresses a number of guests. If this is the case, the characters in front of the unit are extracted and checked for numbers. An example of a useful sentence is the following:

> *'I stayed at this Airbnb with three friends'*

In this case, 'stayed' is the verb and 'friends' is the unit. We then take, e.g., ten characters in front of 'friends' and obtain 'ith three ', clearly containing the number of three. This number is recorded as the number of friends and therefore also the total number of guests. For this example, we obtain a wrong number of guests, as the total number of guests should also include the word ‘I’ (i.e., referring to the term ‘me’). Therefore, a sample of reviews is annotated that the filter has marked. Then the filter’s precision is estimated in obtaining the number of guests. If the precision is higher than a certain threshold, the filter can be safely used in subsequent estimates. If it is lower, we annotate more and use the annotations for our estimate.

As reviews are written in different languages, they may contain different information. So this procedure was performed for six different languages, namely Dutch, English, German, French, Italian and Spanish. Theseare the most frequently used languages in reviews for accommodations in the Netherlands.

Instead of using annotations one could try to improve the filter by only allowing certain patterns or removing sentences that contain particular phrases. The example above could lead one to remove all sentences containing the word ‘with’. However, removing all such phrases or terms leads to a large reduction in reviews that are left after filtering and does not necessarily ensure a high precision on the remaining sample. By annotating, the annotated set will always be a correct sample. In addition, the annotated set can be used as a training set for machine algorithms. Such an algorithm will be able to extract useful patterns itself, without relying on a programmer to explicitly define sets of patterns.

3.3 Removing bias

After either applying a precise filter or using annotations, a small sample of reviews is left for which the number of guests and/or the number of nights is known. These reviews might differ in characteristics from reviews for which the number of guests or number of nights could not be found. Therefore, it is not simply possible to take an average of the number of guests or number of nights and use that for estimation. Instead, the small set of reviews was split into several strata and then an average for each strata was calculated. Using proportions of reviews on in the complete set, a weighted average can be derived.

To get to a correct estimate, the strata should be as homogeneous as possible on the target variable, which is either the number of nights or the number of guests. It then makes sense to split the set in strata based on the variables that are most determining for the target variable.

For the number of guests, a natural splitting variable is the total capacity of the accommodation, i.e., the larger the accommodation the more guest are likely to stay there. Given that the capacity boundaries of 1 guest or larger than 7 guests are relatively rare, the group with 1 guest is grouped together with 2 guests and one stratum is created for capacities larger than 7 guests. For each stratum, an occupancy rate is calculated, which is equal to calculating an average per stratum, and then dividing by the total capacity for the stratum.

For the length of stay, it is harder to determine natural splitting variables. Research (Barros, 2010) indicates, however, that most of the variation can be caught by looking at the country of residence of the guest and the destination. It would be preferable to create as many strata as possible. However, the use of every combination of country of residence and destination province would result in many un-

filled, or barely filled strata. So, the countries of residence are grouped in three regions and the destination in ‘Amsterdam’ or ‘Not-Amsterdam’. The latter is chosen as other sources (Colliers International, 2018; Airbnb, 2019) show that Amsterdam has a large number of guests and it is likely to differ in number of nights, because of it being the capital.
For the three regions of the country of residence, we choose regions based on proximity to the Netherlands. The first region is traditionally referred to as the Benelux and includes the Netherlands itself, Belgium and Luxembourg. The second region includes the large neighbors of the Netherlands, namely the United Kingdom, Germany and France. The third region includes all other countries. It can be discussed whether countries such as Switzerland, Austria, Spain and Sweden should be added to the second region, but guests coming from these countries seem to have lengths of stay that are more similar to those coming from the United States or China than to those coming from countries in the second group.

3.4 Background of guests

Both Airbnb and Booking.com provide information on the background of guests that write reviews. Booking.com does this by means of a flag and the name of the respective country above every review. Airbnb includes the language of the review in the information they provide through their API, but the specific country is also mentioned when navigating to the page of the user. So the country information was scraped from both websites. To stay GDPR compliant, usernames were not scraped. For Booking.com information was retrieved on the country of the reviewer for every review. For Airbnb this requires a large number of requests to the Airbnb servers. As similar information can be obtained by means of random sampling, a random sample was taken of the reviews. Information was then only extracted for the reviews in the sample.

4. Results

The sections 4.1 to 4.5 are dedicated to estimates of the total number of guests and nights on Airbnb and Booking.com. This includes hotels and similar accommodations, that might already be present in the TAS. In 4.6 the estimates are compared to the results of other sources and information from Airbnb itself. In 4.7 we discuss how the results of this research can be used to get to aggregate values of the complete set of small accommodations.

The steps that are made in this chapter:

- results of web scraping of the number of listings and output per province and major cities;
- conversion from reviews to bookings (review rate), including adjustment rates (for bias) and cancelation rates (for Booking);
- deriving the number of guests, including country of residence;
- deriving the number of nights.

For the extracting and processing of the data and all the experiments a laptop was used. This laptop, a Lenovo Thinkpad X1-Extreme, had 32 GB of DDR4 RAM and an Intel Core i7-8750H processor. All the code was written in Python.

4.1 Number of listings (step 1)

Active listings (accommodations for rent) were obtained from the websites of Airbnb and Booking at different points in time. This set of listings only includes those accommodations that were present on the website when web scraping was done.

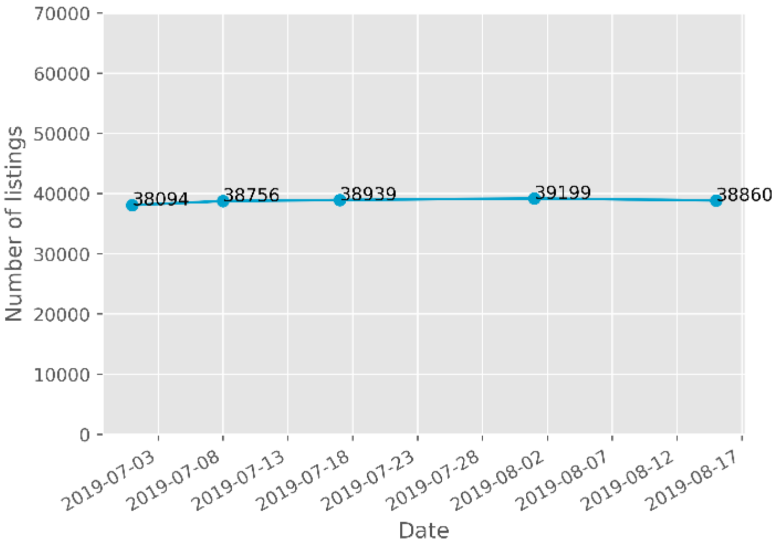

Figure 4: Number of listings on Airbnb at chosen dates, obtained by means of web scraping (See also Table 16 in Appendix)

For Airbnb, 6 measurements were taken from the 8th of June to the 15th of August. The first measurement is not listed in this section, as this measurement did not include the latitude and longitude of every accommodation, which means we were not able to verify whether accommodations were actually in the Netherlands. This first measurement, however, will be of importance in later sections, as the accommodations were used to obtain a set of reviews.

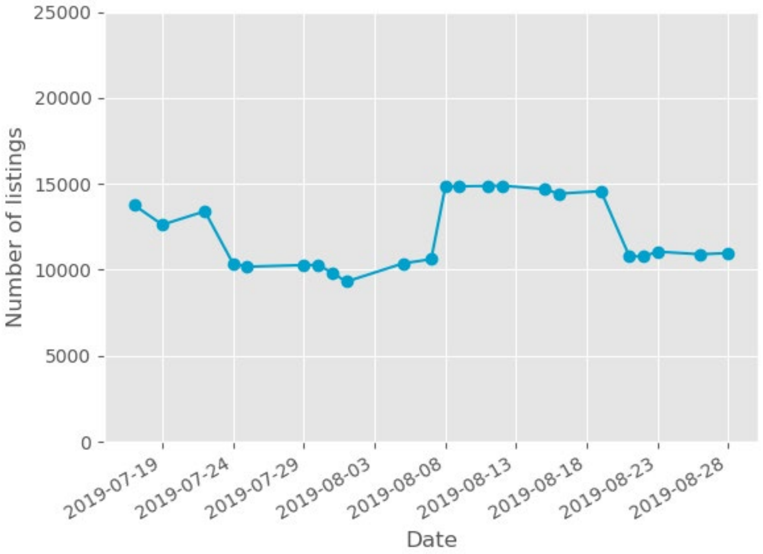

Figure 5: Number of listings on Booking.com at chosen dates, obtained by means of web scraping (See also Table 17 in Appendix)

The number of active listings are shown in Figure 4 for Airbnb and Figure 5 for Booking.com. For Airbnb, the total number of listings is fairly constant at around 38,500 accommodations, with a maximum peak of 39,199 at the beginning of August. For Booking.com, the number of accommodations is around 14,000 in the middle of July, then drops to 10,000 from the end of July to the beginning of August, then suddenly increases to around 14,700, after which it drops back to around 11,000 at the end of August. There was no clear explanation for this pattern, but it seems to apply to every province and municipality in a similar way, as shown by Table 1 and Table 2.
The same information is shown for Airbnb in Table 3 and Table 4, which, as expected, shows less variation. For both Airbnb and Booking.com, most listings are concentrated near the coast and in the big cities. This is in accordance with the TAS. Airbnb has a relatively large number of listings in cities, with the four largest cities making up the top 4, while Booking.com has more listings near the coast, in municipalities such as Bergen (NH), Zandvoort (NH) and Veere (ZE). In both cases, Amsterdam shows the largest share of listings. Notice that Amsterdam even has three times the number of listings of the second largest municipality found using Booking and more than six times the number of listings found on Airbnb.

Table 1: Development of the number of listings on Booking.com at chosen dates, in provinces in the Netherlands

	Date			
Province	2019-07-17	2019-07-30	2019-08-15	2019-08-28
Drenthe	574	370	568	398
Flevoland	123	107	122	83
Friesland	1188	939	1318	981
Gelderland	1791	1249	1774	1401
Groningen	371	263	372	280
Limburg	1127	812	1125	840
Noord-Brabant	846	767	1048	819
Noord-Holland	4176	2755	4059	2997
Overijssel	765	517	771	564
Utrecht	396	280	390	270
Zeeland	1327	1246	1788	1326
Zuid-Holland	1084	974	1359	1011
Total	13768	10279	14694	10970

Table 2: Development of the number of listings on Booking.com, at chosen dates, for the 13 municipalities with the most listings on the month of July

	Date			
Municipality	2019-07-17	2019-07-30	2019-08-15	2019-08-28
Amsterdam (NH)	1476	950	1389	1065
Bergen (NH)	545	324	525	387
Zandvoort (NH)	400	285	397	268
Veere (ZE)	366	351	494	379
Sluis (ZE)	239	245	356	255
Zuidwest-Friesland (FR)	219	182	247	168
Schagen (NH)	208	135	215	162
Texel (NH)	200	168	205	150
Apeldoorn (UT)	194	140	193	127
Nunspeet (GE)	184	142	183	168
Ameland (FR)	167	111	184	140
Schouwen-Duiveland (ZE)	167	140	217	160
‘s-Gravenhage (ZH)	164	146	206	157

Table 3: Number of active listings on Airbnb at chosen dates, in provinces in the Netherlands

	Date			
Province	2019-07-01	2019-07-17	2019-08-01	2019-08-15
Drenthe	745	741	766	772
Flevoland	360	373	399	405
Friesland	1898	1949	1963	1970
Gelderland	3491	3528	3616	3614
Groningen	972	993	999	998
Limburg	1624	1619	1649	1664
Noord-Brabant	2030	2070	2087	2111
Noord-Holland	16544	17024	16912	16530
Overijssel	1209	1210	1255	1270
Utrecht	1915	1963	1942	1919
Zeeland	1966	2014	2103	2105
Zuid-Holland	5340	5455	5508	5502
Total	38094	38939	39199	38860

Table 4: Number of active listings on Airbnb, at chosen dates, for the 13 municipalities with the most listings on the 17th of July

	Date			
Municipality	2019-07-01	2019-07-17	2019-08-01	2019-08-15
Amsterdam	9768	10162	9987	9574
‘s-Gravenhage	1492	1522	1480	1488
Rotterdam	1179	1209	1223	1198
Utrecht	995	1027	1018	994
Haarlem	806	804	768	767
Bergen (NH.)	783	785	813	823
Groningen	601	611	625	618
Noordwijk	514	518	584	583
Veere	506	529	554	551
Maastricht	454	444	415	407
Nijmegen	439	427	360	346
Ede	429	437	478	486
Zandvoort	383	385	378	394

4.2 Number of bookings (step 2)

After obtaining the listings, reviews were scraped of each listing. Listings that were used for the scraping of the reviews were obtained on the 8th of June for Airbnb and the 4th of July for Booking.com. The reviews were scraped from the 24th to the 28th of June for Airbnb and the 5th of July to the 11th of July for Booking.com. As scraping of reviews was slow, it was not possible to scrape everything in one day.
In total 1,002,368 reviews were scraped on Airbnb. As the accommodation file from the 8th of June did not include all necessary information, the reviews were

then linked to the accommodations that were scraped on the 17th of July. Therefore, only 940,000 reviews were usable. The date of the reviews range from March 2010 to June 2019. Airbnb reviews are plotted in Figure 6 and Figure 7. Figure 6 shows, in particular, the season-related penetration rate of Airbnb in the Netherlands since 2013.

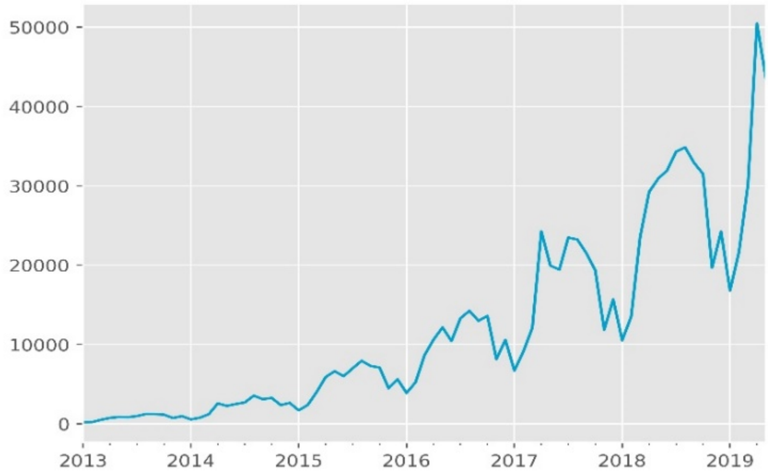

Figure 6: Number of reviews on Airbnb, per month, 2013-2019

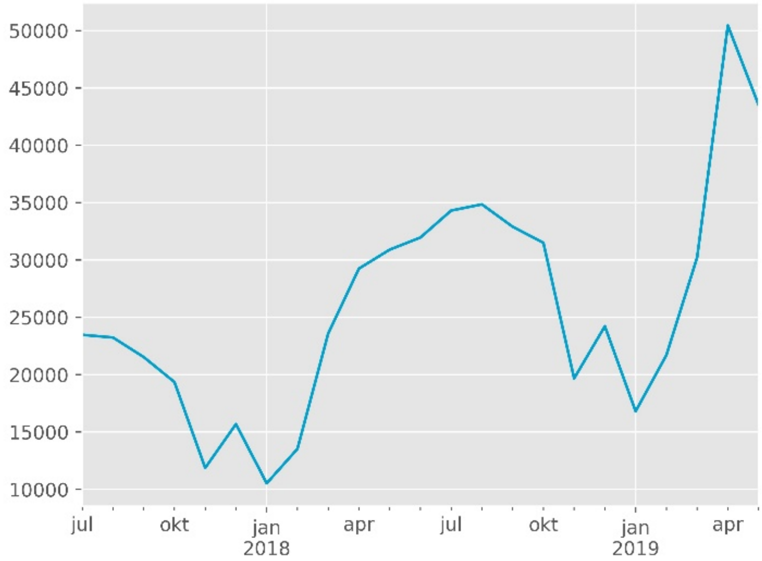

Figure 7: Number of reviews on Airbnb, per month, July 2017-May 2019

For Booking.com 3,020,665 reviews were scraped, of which 2,800,140 could be linked to the accommodation file from the 17th of July. The reviews are plotted in Figure 8.

The reviews of Airbnb show an increasing trend. If we were to take the shown trend and consider it the complete traffic on Airbnb, large growth rates would be implied. However, this set of reviews only includes those corresponding to the accommodations that were found on the 8th of June and that could be linked to those found on the 17th of July. As many accommodations have disappeared from Airbnb since the start of the time series, real growth rates are more flat. Optimally, accommodations and reviews are scraped every month, but we did not have enough time to do that during this research. Therefore we restrict ourselves to the last complete months, April and May of 2019.

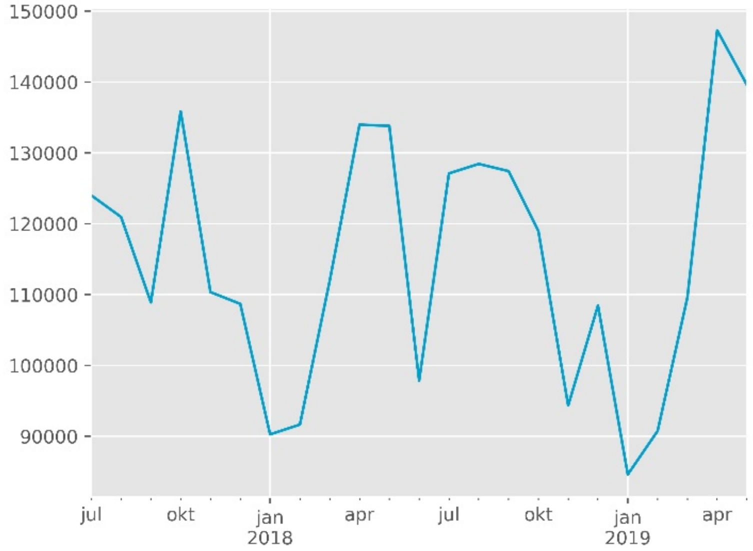

Figure 8: Number of reviews on Booking.com, per month, July 2017-May 2019

For those months, the number of reviews for both Airbnb and Booking can be found in Table 5. To go from reviews to numbers of bookings, two things have to be considered: (i) reviews were only scraped from one set of accommodations; those scraped on the 8th of June. This is just one point in time and not representative for the entire month, (ii) not all bookings have reviews. The latter is solved by dividing by the review rate, which we found to be equal to 72% for Airbnb and 24% for Booking.com. See paragraph 3.1.

The former consideration can be solved by calculating an adjustment rate by means of the capture-recapture technique. The capture-recapture technique is a method that is frequently used to estimate animal populations based on two measurements at different points in time. First, a set of $\boldsymbol{n}$ animals is caught and marked. Then, at a later point in time, one catches a number of animals, $\boldsymbol{K}$, again and counts the number of marked animals, $\boldsymbol{k}$. The population can then be estimated by

$$\boldsymbol{N} = \frac{\boldsymbol{Kn}}{\boldsymbol{k}}.$$

For accommodations the procedure is as follows: we take the number of listings at the 17th of July, multiply by the number of listings on the 15th of August and divide by the number of listings that they have in common. This gives us an estimate of the total population of that month, given that there has been no immigration or emigration from the population within that period. We then divide this total population by the number of listings on the 17th of July to get the adjustment rate. This is done separately per province. Notice that due to its Airbnb size, Amsterdam is presented separated from its province. On average the adjustment rate is 1.136 for Airbnb and 1.125 for Booking.com.

For Booking.com, an additional problem is that the calculated review rate is based on bookings from the last 24 hours. However, it is known a large number of bookings are cancelled before the date of arrival. (d-edge Hospitality Solutions, 2019) reports that this cancellation rate is approximately 50% for Booking.com. As such, the number of bookings is divided by 2 to get to the final number of bookings for Booking.com.

Results of these calculations are also shown in Table 5, together with intermediate values, which make it possible to calculate using different values of the review rate, adjustment rate and calculation rate.

Table 5: Number of reviews and bookings, after applying different multiplication rates, for both Airbnb and Booking.com, in April 2019 and May 2019

Platform	Month	Number of reviews	Number of bookings: after review rate (72% and 24%)	Number of bookings: after adjustment rate (1.136 and 1.125)	Number of bookings: after cancellation rate (50% for Booking)
Airbnb	April 2019	50,445	70,063	79,592	
	May 2019	43,513	60,435	68,654	
Booking	April 2019	147,274	613,642	690,347	345,174
	May 2019	139,540	581,417	654,094	327,047

4.3 Number of guests (step 3)

Different methods were used to calculate the number of guests for respectively Airbnb and Booking.com. For Airbnb 1,375 reviews were annotated with the number of guests and compared to the total capacity of the corresponding listings. An occupancy rate was calculated for every capacity. These occupancy rates are listed in Table 6. Note that this occupancy rate is given that the accommodation is occupied. The occupancy rate was subsequently multiplied by the capacity for every review.

For Booking.com the total capacity was unclear, so we resorted to information about the number of reviews with a certain type of guests, which are ‘solo guest’, ‘couple’, ‘family with children’ and ‘group of friends’. This information is listed on Booking.com for every listing with at least 5 reviews. Reviews were annotated with the types and number of guest to get to an average number of guests for every type. For solo guest and couple these were naturally equal to 1 and 2. For ‘family with children’ and ‘group of friends’ averages of 3.73 and 2.90 were found.

Then, a weighted average was taken to get to an average number of guests for every accommodation. As this average has a theoretical max of 3.73 for every accommodation, many large accommodations will have a lower estimate than is realistic. Similarly, accommodations with a capacity of two might get higher estimates than is possible. Aggregated per province, however, estimates should not be far from reality. Table 7 lists the average number of guests per booking for every province and Amsterdam.

In table 7 it can, firstly, be observed that guests stay, on average, with more guests in an Airbnb than in a Booking.com accommodation, with a mean of 3.06 for Airbnb and 2.13 for Booking.com. This can partially be attributed to the type of accommodations that are on Airbnb, which are often apartments or even homes, while Booking.com is dominated by hotel rooms. Secondly, guests in Airbnb tend

to stay with more guests in accommodations that are in the countryside, in provinces such as Friesland and Zeeland, than in large cities, such as Amsterdam. These former regions offer more spacious apartments than those in Amsterdam.
After imputation of the total set of listings, total number of guests can also be found. These are shown in Figure 8 and Figure 9. The total estimated number of guests in April 2019 is equal to 247,151 on Airbnb and 654,154 on Booking.com. The number of guests per region is also shown. The regions listed are provinces, with Amsterdam being listed separately from Noord-Holland.

Table 6: Occupancy rate per capacity on Airbnb, based on annotated reviews. The occupancy rate is given that the accommodation is occupied.

Capacity	Number of annotated reviews	Occupancy rate (given occupancy)
1, 2	379	104.4%
3	89	83.9%
4	381	91.5%
5	92	89.6%
6	165	87.7%
=> 7	269	82.7%
Total	1375	

Table 7: Average number guests per booking, based on annotated reviews for Airbnb and available review information for Booking.com

Region	Average number of guests per booking (Airbnb)	Average number of guests per booking (Booking.com)
Amsterdam	2.61	2.16
Drenthe	3.71	2.14
Flevoland	2.90	2.11
Friesland	3.81	2.24
Gelderland	3.34	2.05
Groningen	3.27	2.02
Limburg	3.42	2.15
Noord-Brabant	3.37	2.04
Noord-Holland (excl. Amsterdam)	3.24	2.16
Overijssel	3.54	2.09
Utrecht	3.15	1.93
Zeeland	3.51	2.24
Zuid-Holland	3.04	2.13
Netherlands	3.06	2.13

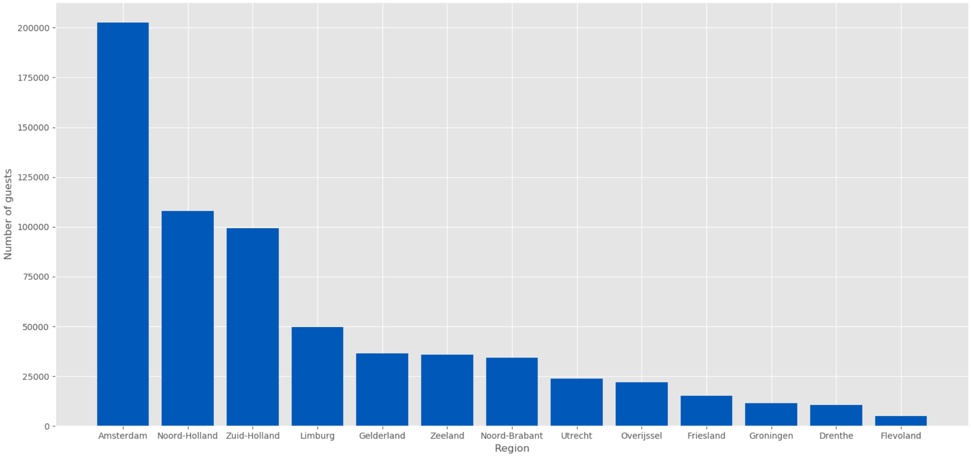

Figure 8: Number of guests staying in accommodations in April 2019 on Booking.com (See Table 18 in Appendix)

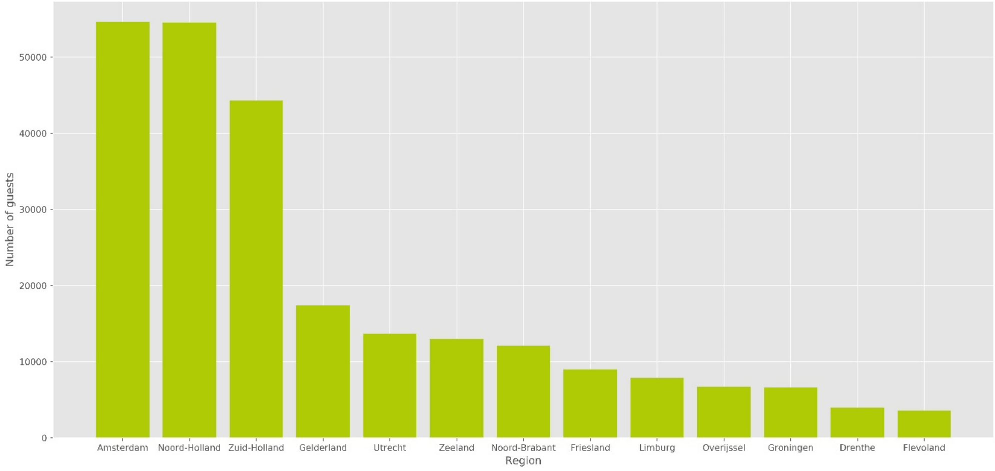

Figure 9: Number of guests staying in accommodations in April 2019 on Airbnb (See Table 18 in Appendix)

The regions with the most guests are Amsterdam, Noord-Holland (excl. Amsterdam) and Zuid-Holland, for both Airbnb and Booking.com. More guests visit Amsterdam through Airbnb and Booking than they do any complete province. For Booking.com the number of guests that visit Amsterdam is even twice as high as the largest province, which is Noord-Holland, excluding Amsterdam.

4.4 Country of residence

Booking.com provides information on the presupposed country of residence of the main guest[7] who made the booking. This can be found with every review. For Airbnb, the country of residence can be scraped from the user page of reviewers. To what extend this is correct cannot be checked.

[7] Primary guest is the guest that made the booking and whose name and profile is showed in the review extracted from either Airbnb- or Booking.com platforms.

Table 8 shows the percentage of reviews that were left by the main guests from a particular region. First, this does not take into account that guests from some regions may travel in larger groups than others. The assumption is also, that the guests accompanying the primary guest have the same country of residence.

Table 8: Percentage of bookings made by guests from the most commonly found home countries, for Airbnb and Booking.com. For Airbnb, a random set of approximately 140,000 reviews was taken, as it took long to scrape the home country from every user page. The scraping took place from the 22nd of July to the 28th of August.

Home country	% of guests (Airbnb), 2010-2019	% of guests (Airbnb), 2018	% of guests (Booking.com), 2017-2019
Netherlands	28.4%	29.1%	32.7%
Germany	16.5%	16.6%	15.3%
United States	10.3%	10.3%	1.7%
United Kingdom	8.3%	8.1%	7.4%
France	7.4%	6.9%	4.3%
Belgium	4.1%	4.2%	7.0%
Italy	2.2%	2.1%	4.1%
Canada	2.1%	2.1%	0.5%
Spain	1.7%	1.7%	2.6%
Australia	1.6%	1.6%	0.8%
Switzerland	1.5%	1.5%	1.4%
China	0.9%	0.9%	0.5%
Russia	0.8%	0.8%	1.7%
Austria	0.7%	0.8%	0.8%
Other	13.5%	13.3%	18.2%

For both Airbnb and Booking.com, most guests have a Dutch background (or give as country of residence the Netherlands). The percentage of guests that is Dutch is higher on Booking.com, at 32.7%, than on Airbnb, at 28.4%. For Airbnb, we see a relatively large number of guests from the United States, France, Canada and Australia, i.e., with the exception of France, Airbnb is used more by guests from the Anglosphere. For Booking.com, we see a relatively large number of guests from Belgium, Italy, Spain and Russia, which shows Booking.com is generally larger in Europe than it is in North-America.

4.5 Number of nights (step 4)

To determine the number of nights for each review, we use the review filter described in Section 3.2, which gives us information on 8,411 reviews for Airbnb and 859 reviews for Booking.com. As this is only a small, biased sample, we need some way to impute the rest of the reviews. We do this by calculating an average for a small set of strata and imputing the empty reviews with the relevant average. The average number of nights is shown per stratum in Table 9, together with the number of reviews that were found by the filter in each stratum.

Table 9: The average number of nights per stratum, for both Airbnb.com and Booking.com, based on accommodation data from the 17th of July and review data from the 24th of June to the 28th of June for Airbnb and the 5th of July to the 11th of July for Booking.com.

Destination	Country of residence	Average number of nights (Airbnb.com)	Average number of nights (Booking.com)
Amsterdam	NL, BE, LU	3.23 (137)	1.29 (24)
	UK, DE, FR	3.03 (1192)	2.56 (160)
	Other	3.98 (1615)	2.55 (152)
Not Amsterdam	NL, BE, LU	2.88 (1476)	2.04 (222)
	UK, DE, FR	3.89 (1966)	2.39 (180)
	Other	5.21 (1797)	2.60 (121)

NL = Netherlands; BE = Belgium; LU = Luxembourg (Benelux).
UK = United Kingdom; DE = Germany; FR = France.

Table 10: Total number of overnight stays in April 2019, estimated on Airbnb and Booking.com, by using accommodations scraped from the 17th of July and reviews scraped from the 24th of June to the 28th of June for Airbnb and the 5th of July to the 11th of July for Booking.com.

Region	Total number of nights Airbnb	Total number of nights Booking.com
Amsterdam	195,333	490,238
Drenthe	12,893	22,809
Flevoland	14,406	11,149
Friesland	29,979	32,959
Gelderland	58,273	79,026
Groningen	22,932	25,190
Limburg	26,720	107,379
Noord-Brabant	41,438	76,622
Noord-Holland (excl. Amsterdam)	220,213	250,313
Overijssel	22,640	47,959
Utrecht	53,899	53,738
Zeeland	43,874	78,488
Zuid-Holland	172,462	225,729
The Netherlands	915,061	1,501,619

The results of Table 9 show that guests from Airbnb in April 2019, when not visiting Amsterdam, tend to stay longer as they come from a destination farther away. That is less evident for guests staying in Amsterdam. A similar pattern can be observed for Booking.com, albeit with lower numbers of nights.
After imputing, the total number of nights per region in April 2019 are shown in Table 10. For Airbnb the bigger number of nights can be seen in Noord-Holland (excluding Amsterdam). Amsterdam shows the second largest number of nights. Also Zuid-Holland show a relatively large number of nights. The other provinces do not come close to the numbers mentioned. For Booking the pattern is a little bit different. Concerning the number of nights, Amsterdam comes clearly first, followed by Noord-Holland (excluding Amsterdam) and Zuid-Holland. Also the provinces of Limburg, Zeeland and Gelderland score relatively high.

4.6 Comparison with TAS

To validate the results we can compare our results to those from the Tourism Accommodation Statistics (TAS). As these statistics concern a different set of accommodations, comparisons of totals do not make sense. Therefore, the average number of nights are compared. Results are shown in Table 11 and Table 12.

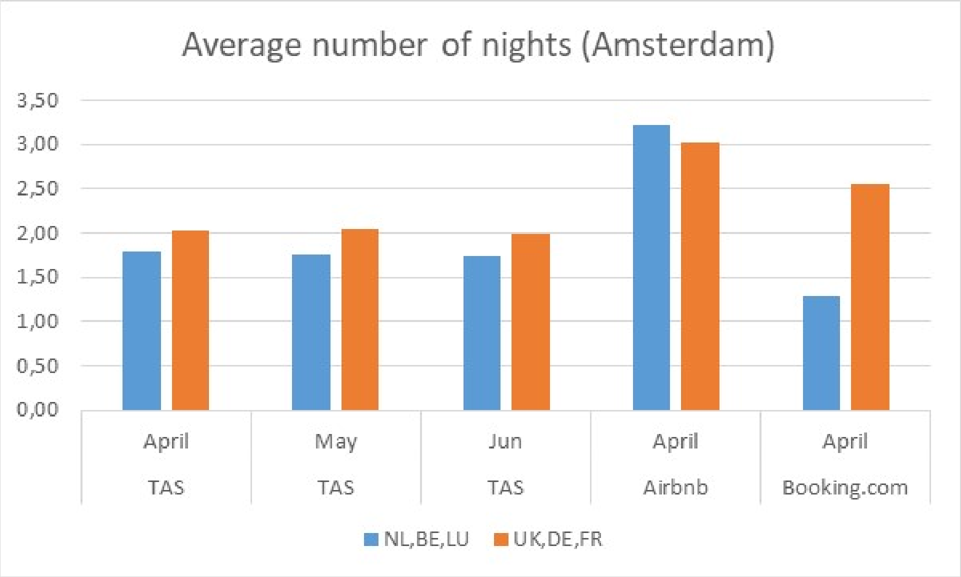

Figure 10: The average number of nights that guests stay at accommodations in Amsterdam, for TAS (denoted by CBS) and average number of nights we found for Airbnb and Booking.com (April 2019)

NL = Netherlands; BE = Belgium; LU = Luxembourg (Benelux).
UK = United Kingdom; DE = Germany; FR = France.

Figure 11 shows that the average number of nights of guests who stay in accommodations booked through Airbnb is considerably higher than the average number of nights for accommodations booked through Booking and the average number of nights from the TAS. The average number of nights from the TAS and Booking are more similar, although those of Booking are higher. The reason could be that the last comparison involves more comparable types of accommodation (e.g. mostly hotels). For average number of nights outside Amsterdam the same reasoning applies. Only for Airbnb, people coming from the United Kingdom, Germany and France tend to stay shorter than those coming from the Netherlands, Belgium and Luxembourg.

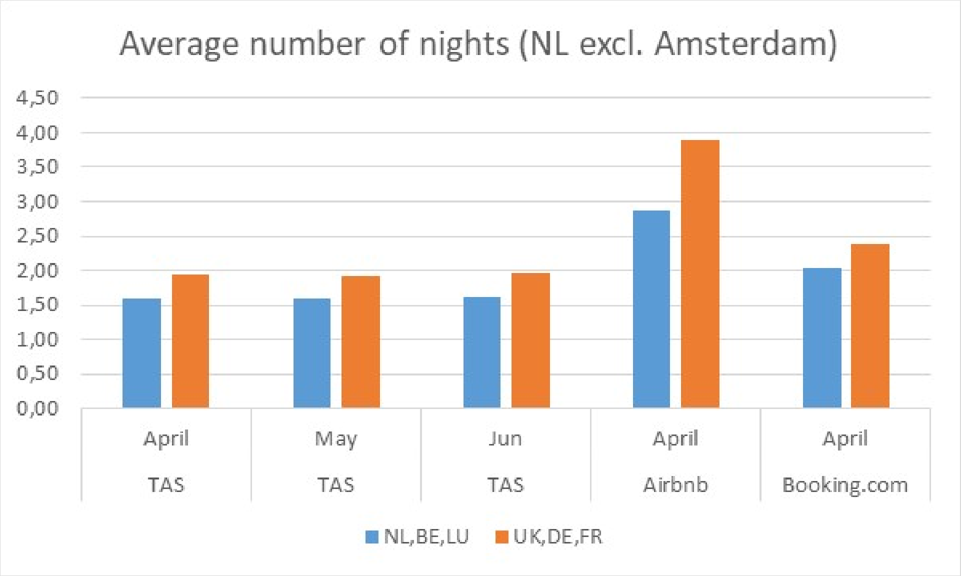

Figure 12: The average number of nights that guests stay at accommodations in the Netherlands (ex. Amsterdam), for TAS (denoted by CBS) and average number of nights we found for Airbnb and Booking.com

NL = Netherlands; BE = Belgium; LU = Luxembourg (Benelux).
UK = United Kingdom; DE = Germany; FR = France.

4.7 External validation

For Airbnb and Booking.com several sources have already reported on number of nights and number of bookings (Airbnb, 2019; Colliers International, 2018). In this section we compare the results from those sources to those estimated here.

Colliers International (Colliers International, 2018) gives the number of accommodations and nights for Airbnb accommodations in selected cities (

Table 11). These are numbers over a complete year (2018), while our statistics mostly focus on the month April in 2019.

Table 11: Number of guests and overnight stays on Airbnb, reported by Colliers International (Colliers International, 2018)

City	Number of Accommodations	Number of overnight stays
Amsterdam	25,380	1,976,000
‘s-Gravenhage	2,709	226,000
Utrecht	2,143	160,000
Rotterdam	2,501	188,000
Eindhoven	561	46,000

Airbnb (Airbnb, 2019) itself also posts statistics on their own accommodations in selected cities. These values are shown in

Table 12. The numbers of accommodations are for the 1st of January 2019, while the number of nights and guests are over 2018. As Colliers (Colliers International, 2018) reports over a complete year and Airbnb only reports at one point in time,

Colliers' numbers are generally higher by a factor 1.1-1.2. An exception to this is Utrecht, where Airbnb reports 2600 accommodations, but Colliers only reports 2150. We suspect this might be a typo on Airbnb's part.
The 800,000 and 2,500,000 are from an external report by Ecorys (Ecorys, 2018), that is cited by Airbnb itself. If we were to apply the 30 nights per accommodation to 21,000 accommodations, a total of 630,000 nights is obtained. This would imply an average number of 2,500,000/630,000 = 3.97 guests per night, far from the 2.61 guests per night we obtained. As such, assuming the number of total accommodations in one year is equal to the number on the 1st of January is likely to be wrong.

Table 12: Number of accommodations, nights per accommodation and guests reported by Airbnb (Airbnb, 2019) and Ecorys (Ecorys, 2018). The number of accommodations are from the 1st of January 2019, nights per accommodation and number of guests are over 2018. Ecorys (Ecorys, 2018) values are denoted by *.

Place	Number of Accommodations	Number of nights per accommodation	Number of guests	Number of overnight stays
Amsterdam	21,000	30	800,000*	2,500,000*
's-Gravenhage	2,000	39		
Utrecht	2,600	30		
Rotterdam	2,000	39		
Eindhoven	500	50		
The Netherlands	55,000	33	1,900,000	

If we apply the 30 nights per accommodation to the 25,380 accommodations from Colliers (Colliers International, 2018), we get a total of 761,400 nights. This would imply 2,500,000/761,400 = 3.28 guests per night with the number of overnight stays from Ecorys (Ecorys, 2018) or 1,976,000/761,400 = 2.60 guests per night with the overnight stays from Colliers (Colliers International, 2018). The latter is more in line with our results. Using similar reasoning, we can come to an implied number of guests per night, which is shown in Table 13 .

Table 13: Implied number of guests per night using values from Airbnb (Airbnb, 2019) and Colliers International (Colliers International, 2018), compared to guests per night using review filters and annotations.

City	Implied guests per night	Our guests per night	Ratio
Amsterdam	2.60	2.61	1.00
's-Gravenhage	2.14	2.77	1.29
Utrecht	2.49	2.89	1.16
Rotterdam	1.93	3.02	1.56
Eindhoven	1.64	2.78	1.70

This implied number of guests is lower than the number of nights found by the review filter. A possible explanation is that the number of guests is overestimated as we did not have sufficient solo guests in our biased sample of annotated reviews.

A future filter should take this into account and look more for terms such as ‘alone’ or ‘on my own’.

To make a valid comparison we assume the number of nights to follow the same pattern as the number of nights for the TAS and extrapolate, i.e., if the month March has 80% of the number of nights as the month April for TAS , we multiply our number of nights by 0.8. We then get the following numbers for the above municipalities:

Table 14: Number of accommodations, overnight stays and guests on Airbnb. Accommodation numbers were obtained on the 17th of July 2019. Overnight stays and number of guests are estimated using filters and annotations.

Place	Number of accommodations	Number of guests	Number of overnight stays
Amsterdam	10,162	520,567	1,808,646
‘s-Gravenhage	1,522	95,304	361,517
Utrecht	1,027	61,106	231,034
Rotterdam	1,209	88,689	350,639
Eindhoven	294	19,525	75,597
The Netherlands	38,939	2,452,185	8,863,687

These numbers are most likely an overestimate, given the reasoning above and the fact that the total number, 2.45 million, is above the 1.9 million from Airbnb. If we divide the numbers by the ratios given in Table 13, the following values are obtained.

Table 15: Number of accommodations, overnight stays and guests on Airbnb. Accommodation numbers were obtained on the 17th of July 2019. Overnight stays and number of guests are estimated using filters and annotations, but adjusted by using values from Airbnb (Airbnb, 2019) and Colliers International (Colliers International, 2018)

City	Number of guests	Number of overnight stays
Amsterdam	520,567	1,808,646
‘s-Gravenhage	73,628	279,295
Utrecht	52,648	199,057
Rotterdam	56,679	224,084
Eindhoven	11,518	44,597

This is more in line with the values from Colliers (Colliers International, 2018), albeit still a bit higher. Possible reason for this is that the TAS assumes almost no growth, while Airbnb most likely has had more growth, especially outside Amsterdam, where regulation has not been as strict.

4.8 Obtaining totals and removing duplicates

The goal of this research was to get to the total number of guests and overnight stays in platform-based accommodations. To obtain those totals, two steps still have to be made. Those are: (i) Filtering out the larger accommodations, that are already present in the TAS and (ii) linking accommodations between platforms, so duplicates are removed. Attempts were made to do this, but failed, as the quality of the data was often not sufficient to properly link. Nonetheless, we will discuss some issues that occurred and possible solutions for both filtering and linking.

For filtering, the most important variable is the total capacity of the accommodation. Accommodations with more than 5 sleeping places should already be in the TAS and can therefore safely be removed. For Airbnb, this is relatively easy, as they give the total capacity for every accommodation. For Booking.com, it is not clear what the capacity of every accommodation is. Attempts should be made to scrape this information and impute the capacity for accommodations where it is not completely clear.

To link accommodations, exact linkage will not work on variables such as postal code, as these are not listed on Airbnb. Airbnb also does not give exact latitudes and longitudes but approximations. To link, a first step can be to create a list of possible options for every accommodation, that are geographically close to it. For rural areas, these lists will not be long as the density of accommodations is low and therefore fuzzy linkage based on the names of accommodations should often be sufficient. For urban areas, especially Amsterdam, this can be more problematic, as many accommodations are in the same region and often have similar names, based on the region where they are, e.g., hotels called ‘Hotel Rembrandtplein’ and ‘B&B Rembrandtplein’.

Linkage should then be done based on other variables, such as capacity, normalized numbers of reviews and for the best linkage, the pictures on the platforms, as these were often found to be the same over all platforms. That also accounts for the text describing the accommodation.

5. Conclusion and suggestions

Estimating the number of guests and overnight stays at accommodations offered on platforms such as Airbnb, Booking etc. is a task that has been of interest to many organizations. A main reason for this interest is that this segment of tourism has grown considerably the last ten years. In addition, there is also many interest in the presupposed negative effects of these kinds of accommodations.
So, Statistics Netherlands has web scraped the data from the two biggest platforms, to get insight in the number of accommodations (listings) on offer and in their location (provinces and municipalities). In addition, an attempted was made to estimate the number of guests and the number of nights by using the reviews that are posted on such platforms. The reviews were first taken from the platforms, after which it was tried to find information on guest's length of stay and the size of the groups they travelled in.
The values were unbiased by calculating a weighted average and imputing with it. Resulting values were not in line with other sources for Airbnb, in terms of absolute values, as our estimates were significantly higher than those from other sources (Colliers International, 2018; Airbnb, 2019). We suspect it can be attributed to the fact that the filter used to extract the size of groups does not take into account solo guests well. Subsequent research should focus on fixing this and extending the filter in such a way that more reviews with useful information are left. In addition, research can be done into replacing the deterministic filter by a machine learning approach. As the current way of estimating already requires one to annotate results from the filter to get an estimation of its accuracy, the annotations provide a ground to do machine learning on.
To get proper estimates, it is also needed to scrape more consistently at certain intervals, preferably by means of a job scheduler, instead of manually turning on the scraper. By scraping at least monthly, time series can be built of the number of accommodations, instead of relying on extrapolation using the Tourism Accommodations Statistics (TAS).

In the meantime a discussion has taken place on the European level (Eurostat) to see if data can be directly provided by the bigger platforms, like Airbnb, Expedia, Trip Advisor and Booking. The consultation with these platforms is at an advanced stage. That would mean that data will be provided per month, in the same way data is available for the TAS. This does not mean that web scraping and estimations are not necessary anymore. First, data on a micro-level are needed to tackle the problems with overlap of listings between platforms and the TAS and between platforms themselves. Also estimates are still required for platforms that operate locally.

References

Airbnb. (2012, September 8). What percent of Airbnb hosts leave reviews for their guests? Retrieved from Quora.com: https://www.quora.com/What-percent-of-Airbnb-hosts-leave-reviews-for-their-guests

Airbnb. (2019, May 2). Nieuwe Cijfers 2018: Airbnb Groeit Verantwoord en Verspreidt Toerisme Over Het Hele Land. Retrieved from Airbnb Press Website: https://press.airbnb.com/nl/nieuwe-cijfers-2018-airbnb-groeit-verantwoord-en-verspreidt-toerisme-over-het-hele-land/

Airbnb. (n.d.). Airbnb Economic Impact. Retrieved from blog.atAirbnb.com: https://blog.atairbnb.com/economic-impact-airbnb/#san-francisco

AirDNA. (2018, October 26). The AI that Fuels AirDNA. Retrieved from AirDNA.co: https://www.airdna.co/blog/short-term-rental-data-methodology

Barros, C. P. (2010). The length of stay in tourism. Annals of Tourism Research, 692-706.

Beformation. (2019, Maart 6). Airbnb-onderzoek Noord-Holland. Retrieved from Noord-Holland.nl: https://www.noord-holland.nl/Actueel/Archief/2019/Maart_2019/Airbnb_onderzoek_Noord_Holland

Colliers International. (2019, April 30). Airbnb in Nederland - de belangrijkste cijfers over 2018. Retrieved from Colliers.com: https://www2.colliers.com/nl-NL/Research/20190501Airbnb2018

d-edge Hospitality Solutions. (2019). How Online Hotel Distribution is Changing in Europe. Retrieved from d-edge.com: https://www.d-edge.com/how-online-hotel-distribution-is-changing-in-europe/

Ecorys. (2018, October 9). Tourism in Amsterdam - Today and Tomorrow. Retrieved from AirbnbCitizen.com: https://www.airbnbcitizen.com/new-report-on-tourism-in-amsterdam/

Inside Airbnb. (2019, Mei 22). About Inside Airbnb. Retrieved from InsideAirbnb.com: http://insideairbnb.com/about.html

San Francisco Board of Supervisors Budget and Legislative Analyst. (2015). Analysis of the impact of short-term rentals on housing. San Francisco.

Sluijpers, P. (2019). Gegevens Airbnb-achtige platformen. (Projectverslag) Den Haag.

Appendix

Table 16: Number of listings on Airbnb at chosen dates, obtained by means of web scraping

Date	Number of unique listings (Airbnb)
2019-07-01	38094
2019-07-08	38756
2019-07-17	38939
2019-08-01	39199
2019-08-15	38860

Table 17: Number of listings on Booking.com at chosen dates, obtained by means of web scraping

Date	Number of unique listings (Booking.com)
2019-07-17	13768
2019-07-19	12620
2019-07-22	13403
2019-07-24	10359
2019-07-25	10177
2019-07-29	10276
2019-07-30	10279
2019-07-31	9835
2019-08-01	9307
2019-08-05	10368
2019-08-07	10614
2019-08-08	14840
2019-08-09	14860
2019-08-11	14883
2019-08-12	14895
2019-08-15	14694
2019-08-16	14426
2019-08-19	14576
2019-08-21	10768
2019-08-22	10782
2019-08-23	11054
2019-08-26	10904
2019-08-28	10970

Table 18: Number of guests staying in accommodations in April 2019 on Airbnb and Booking.com.

Region	Total number of guests (Airbnb)	Total number of guests (Booking.com)
Amsterdam	54,589	202,716
Drenthe	3,939	10,706
Flevoland	3,592	4,951
Friesland	9,002	15,274
Gelderland	17,382	36,599
Groningen	6,593	11,366
Limburg	7,836	49,661
Noord-Brabant	12,061	34,309
Noord-Holland (ex Ams)	54,550	107,816
Overijssel	6,684	21,857
Utrecht	13,682	23,954
Zeeland	12,967	35,734
Zuid-Holland	44,272	99,212
Netherlands	247,151	654,154

Verklaring van tekens

Niets (blanco)	Een cijfer kan op logische gronden niet voorkomen
.	Het cijfer is onbekend, onvoldoende betrouwbaar of geheim
*	Voorlopige cijfers
**	Nader voorlopige cijfers
2018–2019	2018 tot en met 2019
2018/2019	Het gemiddelde over de jaren 2018 tot en met 2019
2018/’19	Oogstjaar, boekjaar, schooljaar enz., beginnend in 2018 en eindigend in 2019
2016/’17–2018/’19	Oogstjaar, boekjaar, enz., 2016/’17 tot en met 2018/’19

In geval van afronding kan het voorkomen dat het weergegeven totaal niet overeenstemt met de som van de getallen.

Colofon

Uitgever
Centraal Bureau voor de Statistiek
Henri Faasdreef 312, 2492 JP Den Haag
www.cbs.nl

Prepress
Centraal Bureau voor de Statistiek

Ontwerp
Edenspiekermann

Inlichtingen
Tel. 088 570 70 70
Via contactformulier: www.cbs.nl/infoservice