# **Method for compiling statistics on rented holiday homes**

Katja Löytynoja, Statistics Finland, katja.loytynoja@stat.fi

Johannes Kolu, Statistics Finland, johannes.kolu@stat.fi

Matti Kokkonen, Statistics Finland, matti.kokkonen@stat.fi

Ossi Nurmi, Statistics Finland, ossi.nurmi@stat.fi

**Abstract**

*The short term renting of holiday homes is a major business, rarely covered by existing statistics. In Finland, there are some half a million buildings classified as holiday homes. Owners use intermediate parties or rent them directly to end users via multiple internet platforms.*

*In this paper we present a preliminary study on estimating the occupancy and capacity of rented holiday homes in Finland. The objectives were to:*

1. *Identify the target population and capacity of holiday homes available for rent*
2. *Develop a method for estimating their occupancy and sales*
3. *Design a production process for regularly published statistics*

*We utilized web scraping in collecting data from all major Finnish web platforms offering holiday home rentals. We obtained structured data of over 20,000 holiday home listings, including location, number of beds, building square area, and weekly rates of the holiday homes. One of the challenges was to identify unique buildings. The current method is a combination of text analysis and comparison of building characteristics: size, location and year of construction.*

*Estimating the precise number of holiday homes is also challenging as not all web sites can be scraped, and not all cottages are advertised. However, with the help of web scraped data and Finnish Building and Dwelling Register we hope to estimate the rented holiday home population for each of the 19 regions in Finland.*

*To estimate occupancy and sales figures we conducted a survey aimed at regional booking agencies as well as the sole largest national booking agency. According to a survey made in spring 2018, these agencies represent roughly half of the total target population. The study proposes to publish first results from 2018 by April 2019.*

***Keywords:***  *Web scraping, holiday homes, accommodation, tourism*

The short term renting of holiday homes is a major business, rarely covered by existing statistics. In Finland, there are some half a million buildings classified as holiday homes. Owners use intermediate parties or rent them directly to end users via multiple internet platforms.

In this paper we present a preliminary study on estimating the occupancy and capacity of rented holiday homes in Finland. The objectives were to:

1. Identify the target population and capacity of holiday homes available for rent
2. Develop a method for estimating their occupancy and sales
3. Design a production process for regularly published statistics

# **Identifying target population**

It’s not easy to get information of which privately owned holiday homes or other buildings are rented out. It would not be very cost effective to sample the owners and estimate number of rented out holiday homes and their parameters, when there is information available online. For this reason, we utilized web scraping in data collection.

## *Scraping websites*

First, we searched online for rental platforms and agencies. Then we selected those that had the greatest number of holiday homes for rent and that could be easily scraped with available open source software. The sites selected include both national and regional platforms within Finland.

In all, we identified 34 sites of which 12 were initially scraped. These are very large sites with good data structure and coverage. Data scraped included most of the information available, excluding pictures, and reservation calendars that are compute-intensive to loop through, which would possibly burden the targeted servers. In most cases we scraped the address, map coordinates, free text field describing the holiday home, additional information concerning nearby activities, and the information of owner. The latter is used in identifying duplicate holiday homes available on multiple platforms, and any personal information such as names and phone numbers are to be eradicated in pseudonymization process used by Statistics Finland.

The number of scrapers dropped finally to four in production as some sites merged, some did not have reliable location information, and some became obsolete as their holiday homes were also found to be advertised in the largest platforms. We also had to exclude Åland Islands.

Currently production scrapers run once a month. Most likely quarterly or biyearly run could also be sufficient as listing contracts usually last for about half a year, and most holiday homes are rented out throughout the year.

## *Removing duplicates*

As the same rental holiday home is often advertised on multiple websites it is possible to get duplicates of the same listing. To get the number of physical rental holiday homes instead of number of listings, these duplicates must be removed. There are several ways this can be achieved, but the content available on different web sites restrict the possibilities. For example, we could do this by using the advertiser’s identity, building size, and location, but the exact information for the same rental holiday home varies by website or all required data may not be available on all websites. However, the free text field describing the rental is often quite long and detailed and seem to be very similar from one site to another.

In our method we first break the free text field into single words. From these the 500 most frequent are filtered out. This removes for example conjunctives and other words that have no individual descriptive information. After that, listings are matched for the remaining words and exact building floor area (m2). Matching is run for 10 times and if more than one match is found after all the runs, the one that has most similarities is picked. After identifying the matching listings, they are allocated an id that will identify the individual rental holiday home in the final data.

In the future we are also considering performing picture analysis in the process of removing duplicates, as the pictures of the same rental holiday homes are often identical on different web sites. There is also a need to methodologically test how accurate our removal method is.

## *Estimating the number of rental holiday homes*

The production scrapers provide structured data of over 20,000 holiday home listings. Of these, some 12 000 are unique buildings. However, as not all web sites can be scraped, and not all rental holiday homes are listed, we need a method to estimate the actual number of holiday homes.

As the sites we scrape are selected based on a number of criteria (number of listings, availability of relevant information such as location information, technology used in creating the site, permission from site owners, spatial coverage of the entire set of scraped web sites), the data gathered is not a representative sample of the population but a non-probability sample. However, we consider our sample a purposive sample as opposed to other non-probability sample types because although we restricted scraping to larger sites, we aimed to cover the whole of Finland (excluding Åland islands), we aimed especially to scrape privately owned holiday home listings that are not identified in administrative data, and if a site was considered important enough, technical difficulty of scraping was second to being able to obtain the data. In some cases, we also checked that the smaller site listings were also found from a larger site before excluding it from the sample.

We tried linking the scraped data with Building Registry to get an idea of the representativeness of our scraped data, and possibly estimate the number of rental holiday homes. The distribution of registered use of the private owned rental holiday homes was also of interest. However, linking by address or coordinates proved to be very difficult as scraped coordinates did not point to the exact location of the building. Address data also was not always accurate. We tried cross checking addresses to coordinates but the inaccuracies still were a problem.

There are known problems with the Building Registry: Not all buildings are registered in Finland. Not all changes in usage are reported to the registry. Not all buildings marked as rental holiday houses are currently in use. All of these may be small variations in the data, but still cause error. Tax administrative data of rental income could be more usable for us. It would also provide information on occupancy and sales. However, linking the scraped data with Tax administrative data produced similarly poor results as linking with Building Registry. In Tax administrative data, the buildings were identified in a free text field that could include anything from house name to several taxation related building id’s that in some cases were wrongly typed. In addition, the electronic data only cover about 40 % of observations, and buildings cannot be identified in paper return data as they are aggregate level. However, Tax administration is strongly investing in the digitalization of tax returns and we hope in future to be able to use that for identifying the buildings and rental price estimations.

Not being able to use administrative data to estimate total number of rentals, we checked the regional distribution of scraped holiday home listings on each scraped site. These were very consistent across all the four sites and thus could be thought to represent the actual distribution of rentals in Finland. In addition, we believe we have very close to total amount of rentals from Northern Savo in our data and thus it could be used as reference region to estimate rentals across the country. Hence, we calculated regional averages across scraped sites and used Northern Savo as reference region to estimate the number of rented holiday homes per region. However, we don’t know yet how to validate the method, so we currently use the scraped data without duplicates as it is.

## *Ethical and legal aspects*

During the process of creating the scrapers, it became clear that we should consider the ethical and legal aspects of gathering data online. To avoid possible pitfalls, we have negotiated an agreement with all the sites that we regularly scrape.

Data ownership is not a straightforward issue as the sites have varying practices concerning user agreements and immaterial rights. There are cases where data ownership is clear, and site provides immaterial rights notice in user agreement whereas in other cases there is no user agreement, or any information related to site usage. In one case, the data on the website originates mostly from servers of their clients and the site owners have only a small amount of data on their own server. Nevertheless, the site owner has an agreement to use all data available on their site, and we are allowed to scrape the site in exchange for statistical feedback such as municipal level data. We consider this a form of informant feedback that is in line with Finnish statistical law and the practices of Statistics Finland.

All in all, despite the difficulties and on the contrary to our expectations, scraping has been well received by site owners. We’ve created generic feedback for all the site owners which positions the site against the whole market of rental holiday homes. We do not pay for the privilege of scraping.

# **Estimating occupancy and sales for the target population**

To estimate occupancy and sales we conduct a survey aimed at regional booking agencies as well as the sole largest national booking agency three times a year. According to a survey made in spring 2018, these agencies represent roughly half of the total target population of rental holiday homes.

In the Survey we inquire the number of holiday homes, beds, sold days, domestic and international overnight stays, and sales revenue. Based on this, we calculate monthly indicators for sold days per holiday home, overnight stays per bed, and occupancy rates per holiday home and bed. Also, daily rates for holiday homes and beds are calculated. Using these estimates we calculate volumes for the target population identified through web scraping. Information on volume is published on national level, and capacity on regional level (NUTS 3 level). The final published statistics on the national level are domestic and foreign overnight stays, utilization rate of holiday homes and beds, rental day average price, and overnight average price. Regional published information consists of the number of rentals and beds.

Statistics are published three times per year in Statistics Finland’s web pages, and as database tables in Statistic Finland’s and Visit Finland’s statistical databases[[1]](#footnote-1).

# **Designing the production process**

The process flow for scraped data is presented in figure 1. The scrapers reside on a computer inside Statistics Finland. Web scraping is done once a month using Anaconda Python 3.7 version with packages Scrapy 1.6.0 and Selenium 3.141.0. Scraping is logged and data is written to XML file during scraping. We use SAS 9.4 for statistics production.

Data are mostly parsed straight from source code and not further processed in scrapers. If there are nodes or other source code related text items in the data, they are cleaned in the scrapers. Ensuring correct encoding is also done in scrapers as SAS is particular in that XML processing uses UTF-8 encoding.

Further parsing is done in SAS if needed. This has a practical reason behind it: most statisticians can work with SAS but Python is relatively unknown among the staff. This ensures flexibility and minimum data loss in case new variables are needed. After parsing, the data goes through an automated reporting procedure where key values are presented as time series. Data is further processed by enriching location information using QGIS. The aim is to have regional information on all the rentals. If coordinates are missing, region is deduced from municipality or postal code. After removing rentals that are already present in Accommodation statistics, and removal of duplicate listings per rental, the data set is considered to represent the monthly target population of the statistics.

Respondent feedback is produced after all the steps and it consists of web site specific information on total number and share of rentals and beds per region.



Figure 1. Web scraped data process flow.

The process flow for Survey data is presented in figure 2. Survey is conducted three times a year using a web based questionnaire. After receiving the data it is checked for logical inaccuracies and returned for corrections if needed. Missing data is imputed, finalized data weighed and key figures estimated. In respondent feedback the values from individual respondent are compared to the total capacity published in the statistics.



Figure 2. Survey data process flow.

# **Statistics on rented holiday homes based on the method**

The first experimental publication on rented holiday homes took place in April 2019. According to the web scraped data there are around 11 000 holiday homes for rent on the websites of the biggest finnish market places and booking agencies offering rental holiday homes. A big share of these are located in Lapland.

Figure 3. Rental holiday homes by region in April 2019

Based on the web scraped capacity and on the survey for the booking agencies, there were around 2.5 million overnight stays in the year 2018. Around 1 945 000 were nights stayed by finnish residents and 539 000 by foreigners. The most popular months among the Finns was March as the most popular among the foreigners was January.

Figure 4. Overnight stays by month in 2018

Prices for holiday homes per day vary depending on the month. The cheapest days were in September as the average price for per day was 82.79 € whereas the most expensive month to rent a cottage was December when the average price for per day was 140.47 €.

Figure 5. Average prices for a holiday home per day by month in 2018

1. http://visitfinland.stat.fi/PXWeb/pxweb/en/VisitFinland/ [↑](#footnote-ref-1)