INSTITUTO NACIONAL DE ESTADISTICA

Measurement of the Number of Tourist Dwellings in Spain and their Capacity

Technical Project

National Statistics Institute
December 2020

experimental

INDEX

1. INTRODUCTION

The Non-Hotel Tourist Accommodation Occupancy (hotel establishments, tourist apartments, campsites, rural tourism accommodations and hostels) respond to **Regulation (EC) No. 692/2011** of the European Parliament and of the Council, of July 6, 2011. This regulation requires the monthly submission of information to EUROSTAT for the following CNAE categories:

- CNAE 55.1: Hotels and similar accommodation
- CNAE 55.2: Holiday and other short-stay accommodation
- CNAE 55.3: Camping grounds, recreational vehicle parks and trailer parks

The following sources are used to provide information regarding each of the CNAEs: information from the Hotel Occupancy Survey (HOS) for 55.1, information from the Occupancy Surveys for Holiday Dwellings (HDOS), Hostels (HOS) and Lodgings of Rural Tourism (RTAOS) for 55.2 and information from the Camping Occupancy Survey (COS) for 55.3.

However, the CNAE 55.2 also includes the so-called "vacation rentals" (reserved using internet platforms/apps); for the time being, there is not enough information available for to analyse their impact on the tourism sector and improve the quality of the data subject to the Regulation. This is a problem common to most European countries. Eurostat has begun a pilot with four of the main vacation rental platforms to collect this information.

With the boom experienced by these tourist accommodation platforms in recent years, supply and demand has grown considerably. The style of many cities and neighbourhoods is changing due to the increases in this type of housing. The analysis and estimation of this type of tourist housing has now become essential, and many users have been demanding this information for some time.

Starting several months ago, the G.S. of Statistics of Tourism, Science and Technology began to try to collect information in order to prepare an estimate of this type of rental housing. The first measure was to contact those responsible for tourism in each autonomous community to gather information regarding vacation rentals. The response to this exercise was not entirely satisfactory: some of the communities did not have a tourist housing directory, and in others, it was out of date.

Starting at the end of 2019, the G.S. began to develop web scraping programs to extract information on tourist accommodations from the main vacation rental platforms.

2. OBJECTIVES

The main objectives of the project are:

- Estimate the number of tourist housing accommodations in Spain, as well as their capacity.
- Respond to the growing demand for information on this matter.
- Establish a methodology that can be used to provide information on tourist housing in Spain on a regular basis.

- Complete the information provided to Eurostat regarding CNAE 55.2: Holiday and other short-stay accommodation

3. PROJECT SCOPE

3.1 POPULATION SCOPE

The population scope is made up of all the tourist dwellings in the national territory. The delimitation of this type of dwelling is established in section 6.

3.2 GEOGRAPHICAL OR TERRITORIAL SCOPE

The tourist dwellings for the entire national territory will be studied: for the 17 autonomous communities and the two autonomous cities, Ceuta and Melilla, the 50 provinces, the 8,116 municipalities, 10,517 districts and 35,960 census sections[1].

3.3 TEMPORAL SCOPE

The information was downloaded from the three vacation rental platforms in August 2020. The tourist housing directories used have the same time reference.

3.4 STUDY AND CLASSIFICATION VARIABLES

The study variables are the number of **tourist dwellings**, the number of **bed-places** and the **bed-places per tourist dwelling**. In addition, using the total number of dwellings from the census, we have calculated the **percentage of tourist dwellings** over the total dwellings[2].

The classification variables used are:

- Geographic: autonomous community, province, municipality, district, and census section
- Others: municipality size, area (coastal or inland) and degree of urbanization.

4. WEB SCRAPING OF VACATION RENTAL PLATFORMS

4.1 INTRODUCTION

Web scraping is a technique that uses software programs to extract information from websites. It works by sifting through and capturing web page information, analysing each page's design structure and creating a set of structured data that can be stored and analysed in a database. The information that can be extracted and used from a webpage is very varied: text, figures, photos, integrated videos, maps,...

[1] Since the operation uses the housing data from the Population and Housing Census 2011 (latest data available at the level of sufficient granularity) to estimate the percentage of tourist housing, the geographic breakdowns defined in this period are used.

[2] Total housing data: Population and Housing Census 2011 published by the INE: (https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176992&menu=ultiDatos&idp=1254735572981).

The web scraping process applied is essentially the following:

- Load the URL from which information is to be obtained.
- Find the desired information on the page.
- Run the code to extract this information.
- Store it in a structured way.

After an analysis of the vacation rental platforms in Spain, it was decided that this process would be implemented for the three most-used platforms.

4.2 DESCRIPTION OF THE PLATFORMS

The three platforms analysed have been operating in Spain for some time, with a business model based mainly on the commissions charged for each reservation. All three operate in a similar fashion - they have a search engine with the following cells to complete:

- Destination/Accommodation name
- Arrival and departure dates
- Number of guests

When the search is carried out, the list of accommodations available with these characteristics appears, with basic information that depends on each platform_ the name of the accommodation, its score, the location, the number of comments made, or the main photo.

In addition, the pages allow you to:

- Re-filter the results using other variables such as the type of accommodation or number of stars.
- Reorder the accommodations according to criteria such as price or location.

If you want to obtain more information about a specific accommodation, you must select it. The information that then appears is much more extensive, such as user comments, the description of the accommodation, its address, booking restrictions, categorized ratings, and additional photos of the accommodation.

4.3 DATA EXTRACTION

Data extraction on the three platforms follows a very similar procedure that is divided into two phases:

4.3.1 Phase 1

The first phase consists of listing all the lodgings present in the national territory. To do this, the territory is divided into zones that are searched consecutively in the page's search engine. The result of these searches is a list of the accommodations with the basic information for each of them. The information extracted from this first phase depends on each platform, although it does not usually change much from one to another. Some of the variables downloaded are the following:

- Property identifier
- Accommodation name
- Location
- Capacity
- Rating
- Accommodation type
- Date and time of accommodation capture

The entry format for this process is an Excel sheet with the definition of the areas for which the search will be carried out. The output is a csv file for each area, with the basic information on all the accommodations in each of them.

4.3.2 Phase 2

Once the basic information has been obtained for each of the accommodations in Spain, the csv file for each region of the first phase is loaded, and the link for each of the accommodations (for which the basic information has been obtained) is accessed.

For each of the accommodations, the page is then sifted through and additional information is extracted. As in phase 1, the information extracted depends on each platform. Some of the variables obtained in this phase are the following:

- Accommodation subtype
- Host name
- License
- Description of the accommodation, neighborhood and host
- Number of comments
- Address
- Number of bedrooms, beds and bathrooms
- Dimension
- Internet and parking services
- Pool availability
- If smoking and pets are allowed
- Company that manages the accommodation

The output of this phase 2 is a csv file for each of the areas, with the information collected in this phase for each accommodation, as well as the basic information collected in phase 1, concatenated.

Due to the large number of accommodations on the platforms, an algorithm has been implemented that optimizes the download. This algorithm allows scraper execution time to be reduced once a complete download of the two phases has been carried out for all of Spain. In subsequent scrapings, only phase 1 - which is the fastest - need be fully executed. Phase 2 will only be executed for accommodations collected in phase 1 that were not downloaded in the previous download.

5. TOURIST HOUSING DIRECTORIES OF THE AC

The G.S. of Tourism Statistics and Science and Technology makes periodic requests for the tourist housing directories from those responsible for tourism in each autonomous community. For this project, a request was made in August 2020 for those nearest to the time reference used.

Since it was not possible to obtain the directory for all the communities, and due to the differences in updating and the variables provided, it was decided that these directories would be used only to provide contrast with the results estimated via platforms.

6. DELIMITATION OF TOURIST HOUSING

6.1 TOURIST HOUSING BY AUTONOMOUS COMMUNITY

The vacation rental platforms used in this project offer various types of accommodations, such as hotels, campsites, tourist apartments, hostels, etc; However, the objective of this project is to measure tourist housing, whether it be for an entire accommodation or by rooms only. All the information downloaded from the platforms must thus be filtered, so that we can obtain only the desired type.

In Spain, tourist housing is not a clearly defined concept for which a unique definition exists. This definition depends on the autonomous communities, and each classifies “tourist accommodation” according to its own criteria. Following an analysis of the legislation in each AC, the types of accommodation defined as tourist housing were selected. The following table summarizes the types of accommodations selected in each autonomous community, as well as the nomenclature for their licenses:

	Denominación	Licencias
Andalucía	Vivienda con fines turísticos	VFT
	En trámites de conseguir una licencia de vivienda con fines turísticos	CTC
Aragón	Vivienda de uso turístico	VU
Principado de Asturias	Vivienda de uso turístico	VUT
	Vivienda vacacional	VV
Illes Balears	Estancia turística en vivienda	ETV
	Vivienda turística vacacional	VTV
Canarias	Vivienda vacacional	VV
Cantabria	Vivienda de uso turístico	VUT
Castilla y León	Vivienda de uso turístico	VUT
Castilla - La Mancha	Vivienda de uso turístico	VUT
Cataluña	Vivienda de uso turístico	HUT
Comunitat Valenciana	Vivienda turística	VT
Extremadura	Apartamento turístico	AT
Galicia	Vivienda turística	VT
	Vivienda de uso turístico	VUT
Comunidad de Madrid	Vivienda de uso turístico	VT
Región de Murcia	Vivienda de uso turístico	VV
Comunidad Foral de Navarra	Vivienda turística	UVT
	Apartamento turístico	UAT
País Vasco	Vivienda para uso turístico	E
	Alojamiento en habitación de vivienda particular	L
Rioja, La	Vivienda de uso turístico	VT

Based on the information present in the table, a tourist housing selection algorithm was established, which is presented in the following section.

6.2 TOURIST HOUSING SELECTION ALGORITHM

The tourist housing selection algorithm for the accommodations extracted from the platforms is as follows:

- **STEP 1:**

The accommodations that have a license are separated from those that do not and harmonized with those that have it defined. Examples of harmonization:

vut 629 as -> VUT/AS/000000000000000629

rta: vtar ca 01455 -> VTAR/CA/000000000000001455

- **STEP 2:**

For those that have a license, those that we have determined as tourist housing are extracted. If any of these conditions are met, the algorithm will select the accommodation:

- If AC = Andalucia and License Type = VFT or CTC
- If AC = Aragon and License Type = VFT or CTC
- If AC = Principado de Asturias and License Type = VUT or VV
- If AC = Islas Baleares and License Type = ETV or VTV
- If AC = Canarias and License Type = VV
- If AC = Cantabria and License Type = VUT
- If AC = Castilla y León and License Type = VUT
- If AC = Castilla - La Mancha and License Type = VUT
- If AC = Cataluña and License Type = HUT
- If AC = Extremadura and License Type = AT
- If AC = Comunitat Valenciana and License Type = VT
- If AC = Galicia and License Type = VT or VUT
- If AC = Comunidad de Madrid and License Type = VT
- If AC = Región de Murcia and License Type = VV
- If AC = Comunidad Foral de Navarra and License Type = UVT or UAT
- If AC = País Vasco and License Type = E or L
- If AC = La Rioja and License Type = VT

- **STEP 3:**

Accommodations that do not have a license, or where the license is not defined correctly[3], will be selected based on the subtype variables of each platform. The determination of the subtypes considered as tourist housing is based on an analysis of subtype-license cross-referencing.

7. DEDUPLICATION ALGORITHM

7.1 OBJECTIVE

The tourist housing information gathered from the platforms cannot be added to provide a total of accommodations in Spain; this is because, in many cases, the owners post the accommodations in more than one platform to give them more visibility. This is why it is essential that deduplication algorithm be implemented to eliminate accommodations present on more than one platform at the same time.

[3] Approximately 43% of the accommodations.

7.2 ALGORITHM

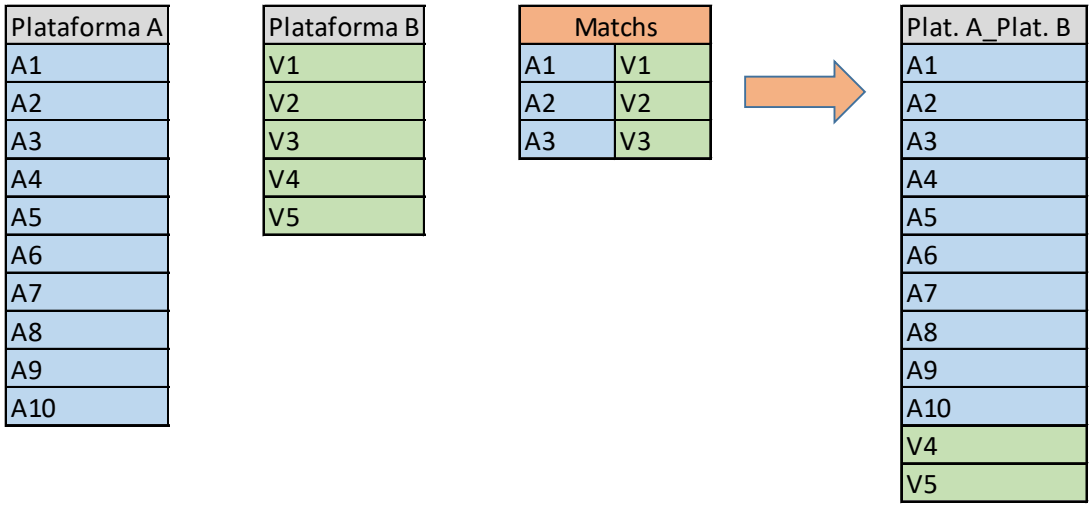
7.2.1 General scheme

The inputs for the deduplication algorithm consist of the accommodations downloaded through web scraping from the three platforms, and filtered using the previously defined algorithm to leave only tourist homes.

The procedure implemented by the algorithm is as follows:

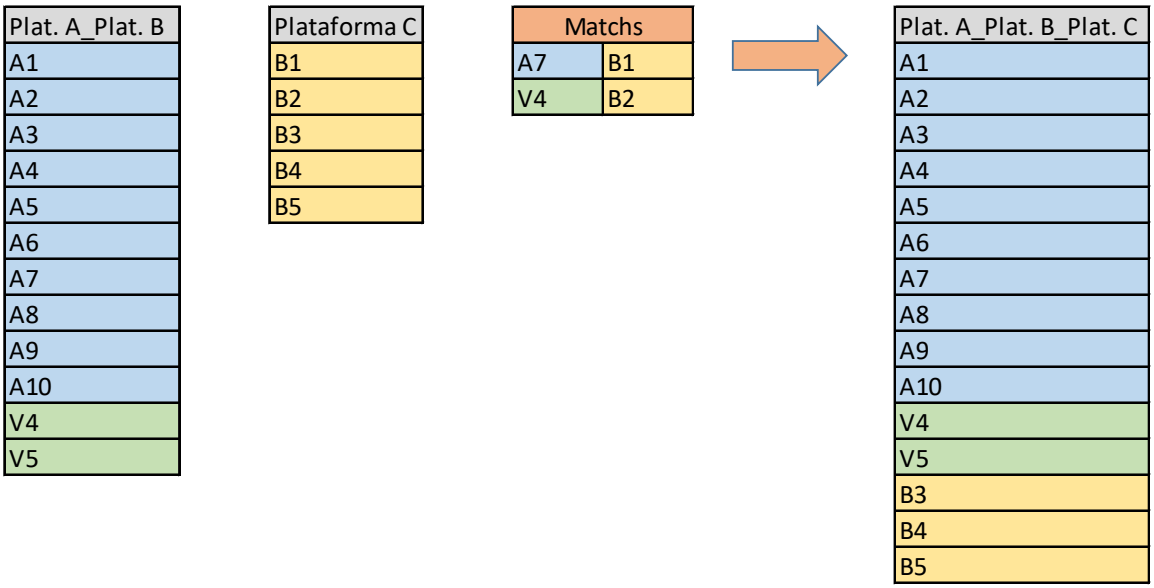
- **STEP 1:**

First, the file of one of the platforms, platform A, is taken as a reference. Then, each accommodation of platform B is compared with all the accommodation of platform A. Those accommodations that are not found on platform A are added to a joint file for both platforms.



- **STEP 2:**

Next, using the file that combines both platforms, the same procedure is carried out with the file for platform C. In this case, since the file for platforms A and B have different variables, the deduplication algorithm will be different for each platform.



The comparison process between platforms can be quite difficult if the accommodations to be compared are not first filtered. To this end, in the process explained above, the accommodations on each platform are compared by municipality. Furthermore, for municipalities with a high number of accommodations, only the 300 nearest accommodations on the other platform are compared for each accommodation.

7.2.2 Deduplication process between platforms

For the steps of the previous point, an algorithm must be defined to determine if the accommodations on one of the platforms are on the platform to which it is being compared.

Assuming that we are comparing platforms A and B, the procedure would be as follows:

- **STEP 1:**

An initial variable harmonization process is carried out for the two platforms.

- **STEP 2:**

All the accommodations for platform A are added to the output file.

- **STEP 3:**

The files from platforms A and B are filtered to select only those corresponding to the municipality to be compared.

- **STEP 4:**

The first accommodation on platform B is selected and compared with the first on platform A.

- **STEP 4.1:**

The harmonized licenses of both accommodations are compared, resulting in two options:

- If the licenses match, the platform B accommodation is not added to the output file, since it is a duplicate of the accommodation on platform A. We will return to point 4, carrying out the same procedure for the next accommodation in the platform B file.

- If the licenses do not match, the process continues.

- **STEP 4.2:**

Common variables between the two platforms are compared to determine if the accommodations are the same. Some of the variables compared are the accommodation name, name of the host, capacity, number of bedrooms, subtype, internet and parking services, if pets are allowed, distance between accommodations,...

Each of these variables is given a weight based on their ability to deduplicate. Once weighted, the result is compared with a previously defined limit value, resulting in two options:

- If this limit value is exceeded, the platform B accommodation is not added to the output file, since it is a duplicate of the accommodation on platform A. We will return to point 4, carrying out the same procedure for the next accommodation in the platform B file.

- If this limit value is not exceeded, there are again two options:

    1. If there are more lodgings on platform A to compare against, go back to step 4.1 and compare with the next one.

    2. If there are no more accommodations on platform A to compare against, the accommodation on platform B is added to the output file and it is considered a new accommodation was not registered on platform A. We will return to point 4, carrying out the procedure for the need accommodation in the platform B file (provided there are more).

Once all the accommodations on platform B of the municipality have been completed, go back to step 3 to select the next municipality. Continue as such, until all the municipalities in the national territory have been analysed

The output file for this process is a file with the combination of both platforms: **A AND B**

For the other two pairs of platforms, Platforms C and B and platforms C and A, the procedure is the same; the only difference is the variables used to perform the deduplication.

7.2.3 Determination of bed-places

If one of the accommodations is present on more than one platform, it is assigned the bed-places on the platform used as an initial reference: that is, first those of platform A are used, then those of B, and if the accommodation is only on platform C, then the bed-places shown on this platform are used. This decision has been made based on an analysis of the accommodations present on more than one platform.

8. DISSEMINATION

Information on tourist housing in Spain is published in the form of tables and using maps, to reflect the information in a more visual way.

The geographical breakdowns used are: by autonomous community, province, municipality, districts, and census sections. Information is also broken down by size of the municipality, area (coastal or inland), and degree of urbanization.

The feasibility of future publications will be analysed.

9. CALENDAR

The schedule for the main operational phases is as follows:

- Development of the programs for the analysis of the three platforms: from November 2019 to July 2020.
- Development of the deduplication algorithm: April 2020 to September 2020.
- Integration of information and production of dissemination products: October 2020 to November 2020.
- Publication: December 2020.