INSTITUTO NACIONAL DE ESTADISTICA

Medición del número de viviendas turísticas en España y su capacidad.

Proyecto técnico

Instituto Nacional de Estadística
Febrero de 2022

Experimental

Experimental

Índice

1 Introducción

Las Encuestas de Ocupación en Alojamientos Turísticos (establecimientos hoteleros, apartamentos turísticos, campings, alojamientos de turismo rural y albergues) dan respuesta al **Reglamento (CE) nº 692/2011** del Parlamento Europeo y del Consejo, de 6 de julio de 2011. Este reglamento exige remitir información mensual a EUROSTAT de las siguientes categorías de la CNAE:

- CNAE 55.1: Hoteles y alojamientos similares
- CNAE 55.2: Alojamientos turísticos y otros alojamientos de corta estancia
- CNAE 55.3: Campings y aparcamientos para caravanas

Para proporcionar la información referente a cada una de las CNAEs se utilizan las siguientes fuentes: información de la Encuesta de Ocupación Hotelera (EOH) para la 55.1, información de las Encuestas de Ocupación en Apartamentos Turísticos (EOAP), Albergues (EOAL) y Alojamientos de Turismo Rural (EOTR) para la 55.2 e información de la Encuesta de Ocupación en Camping (EOAC) para la 55.3.

Sin embargo, dentro de la CNAE 55.2 se encuentran incluidas también las denominadas viviendas de uso turístico; de las que hasta el desarrollo de esta estadística experimental no se disponía de ninguna información. Esta problemática era común en la mayoría de los países europeos y es por eso que Eurostat comenzó en 2020 un piloto con cuatro de las principales plataformas de alojamiento turístico para recuperar esta información.

Por otro lado, con el auge que han tenido en los últimos años estas plataformas de alojamiento turístico, la oferta y la demanda de este tipo de alojamientos han crecido de manera muy considerable en los últimos años. Muchas ciudades y barrios están cambiando su idiosincrasia debido al aumento de viviendas de este tipo. Es por esto que el análisis y estimación del número de este tipo de viviendas turísticas era fundamental y son muchos los usuarios que demandaban esta información desde hace tiempo.

Desde finales de 2019, la S.G. de Estadísticas de Turismo, Ciencia y Tecnología empezó a moverse para tratar de recoger información para poder elaborar una estimación de este tipo de viviendas. La primera medida fue contactar con los responsables en materia turística de cada comunidad autónoma, para recabar la información concerniente a este tipo de viviendas de cada una de ellas. La respuesta a este ejercicio no fue del todo satisfactoria, ya que algunas de las comunidades no disponían de un directorio de vivienda turística y otras lo tenían desactualizado.

Por otro lado, se empezaron a desarrollar en la subdirección programas de web scraping que extraían la información de alojamientos turísticos de las principales plataformas de alojamiento turístico. El buen resultado de este ejercicio conllevó el desarrollo de una estadística experimental basada en esta técnica.

2 Objetivos

Los principales objetivos del proyecto son:

- Estimar el número de alojamientos de vivienda turística que hay en España, así como su capacidad.
- Dar respuesta a la creciente demanda de información que hay sobre esta materia.
- Establecer una metodología que pueda servir para proporcionar la información de la vivienda turística en España de manera regular.
- Completar la información proporcionada a Eurostat referente a la CNAE 55.2: Alojamientos turísticos y otros alojamientos de corta estancia

3 Ámbito del Proyecto

3.1 ÁMBITO POBLACIONAL

El ámbito poblacional está constituido por el conjunto de las viviendas turísticas del territorio nacional. La delimitación de este tipo de viviendas se establece en la sección 6.

3.2 ÁMBITO GEOGRÁFICO O TERRITORIAL

Se estudiarán las viviendas turísticas del conjunto del territorio nacional. Para las 17 comunidades autónomas y las dos ciudades autónomas, Ceuta y Melilla, las 50 provincias, los 8116 municipios, 10517 distritos y 35960 secciones censales[1].

3.3 ÁMBITO TEMPORAL

La descarga de información de las 3 plataformas de alojamiento se hace en los meses de febrero y agosto de cada año.

Experimental

[1] Debido a que en la operación se usan los datos de viviendas del Censo de Población y Viviendas 2011 (último dato disponible al nivel de granularidad suficiente) para estimar el porcentaje de vivienda turística, se utilizan las desagregaciones geográficas definidas en este periodo.

3.4 VARIABLES DE ESTUDIO Y CLASIFICACIÓN

Las variables de estudio son el número de **viviendas turísticas**, número de **plazas** y las **plazas por vivienda turística**. Además, a partir del dato total de viviendas del censo se ha calculado el **porcentaje de vivienda turística** sobre el total de viviendas[2].

Las variables de clasificación utilizadas son:

– Geográficas: comunidad autónoma, provincia, municipio, distrito y sección censal

– Otras: tamaño del municipio, zona (costera o interior) y grado de urbanización.

4 Web scraping de plataformas de alojamiento turístico

4.1 INTRODUCCIÓN

El web scraping es una técnica que utiliza programas de software para extraer información de sitios web. Se basa en recorrer y captar la información de una página web, analizando la estructura particular propia del diseño de cada página y creando un conjunto de datos estructurados que puedan ser almacenados y analizados en una base de datos. La información que se puede extraer y utilizar de una web es muy variada: texto, cifras, fotos, videos integrados, mapas,…

Básicamente, el proceso de web scraping que se aplica es el siguiente:

– Cargar la URL de la que se quiere obtener información.

– Encontrar la información que se desea de esta página.

– Ejecutar el código de extracción de esta información.

– Almacenarla de forma estructurada.

Tras un análisis de las plataformas de alojamiento turístico en España, se decidió implementar este proceso para tres de las más utilizadas.

4.2 DESCRIPCIÓN DE LAS PLATAFORMAS

Las tres plataformas analizadas llevan operando desde hace tiempo en España con un modelo de negocio basado principalmente en las comisiones que cobra por cada reserva. El funcionamiento de las tres es muy parecido, disponen de un motor de búsqueda con las siguientes celdas a completar:

– Destino/Nombre del alojamiento

[2] Dato total de viviendas: Censo de Población y Viviendas 2011 publicado por el INE: (https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176992&menu=ultiDatos&idp=1254735572981).

– Fecha de entrada y de salida
– Número de huéspedes

Cuando se realiza la búsqueda aparece el listado de los alojamientos disponibles con esas características, con una información básica que depende de cada plataforma, como el nombre del alojamiento, su puntuación, la localización, el número de comentarios que tiene o la foto principal.

Además las páginas permiten:

– Volver a filtrar los resultados mediante otras variables como el tipo de alojamiento o sus estrellas.
– Reordenar los alojamientos de acuerdo a criterios como el precio o la ubicación.

Si se quiere obtener más información del alojamiento es necesario seleccionarlo. Una vez hecho esto, la información que aparece es mucho más amplia, como por ejemplo los comentarios de los usuarios, la descripción del alojamiento, su dirección, restricciones para reservar, puntuaciones categorizadas o las demás fotos del alojamiento.

4.3 EXTRACCIÓN DE DATOS

La extracción de los datos en las tres plataformas sigue un procedimiento muy similar que se divide en dos fases:

Fase 1

La primera de las fases consiste en listar todos los alojamientos presentes en el territorio nacional. Para ello se divide el territorio en zonas que se buscan consecutivamente en el motor de búsqueda de la página. El resultado de estas búsquedas es un listado de los alojamientos con la información básica de cada uno de ellos. La información extraída de esta primera fase depende de cada plataforma, aunque no suele cambiar mucho de una a otra. Algunas de las variables descargadas son las siguientes:

– Identificador del alojamiento
– Nombre del alojamiento
– Localización
– Capacidad
– Puntuación
– Tipo del alojamiento
– Fecha y hora de captación del alojamiento

La entrada a este proceso es un excel con la definición de las zonas sobre las que se va a realizar la búsqueda. La salida es un fichero csv por cada una de las zonas con la información básica de todos los alojamientos presentes en estas.

Fase 2

Una vez se ha obtenido la información básica para cada uno de los alojamientos de España, se carga el fichero csv de cada región de la primera fase y se accede al link de cada uno de los alojamientos de los que se ha obtenido la información básica.

Después, de cada uno de los alojamientos se recorre la página y se extrae información adicional de cada uno de ellos. Al igual que en la fase 1 la información extraída depende de cada plataforma. Algunas de las variables que se obtienen en esta fase son las siguientes:

- Subtipo de alojamiento
- Nombre del anfitrión
- Licencia
- Descripción del alojamiento, vecindario y anfitrión
- Número de comentarios
- Dirección
- Número de dormitorios, camas y baños
- Dimensión
- Servicios de internet y parking
- Disponibilidad de piscina
- Si permite fumar y llevar mascotas
- Empresa que gestiona el alojamiento

La salida de esta fase 2 es un archivo de tipo csv por cada una de las zonas, con la información recogida en esta fase para cada alojamiento, así como la información básica recogida en la fase1 concatenada.

Debido a la gran cantidad de alojamientos que hay en las plataformas, se ha implementado un algoritmo que optimiza la descarga. Este algoritmo permite reducir el tiempo de ejecución de los scrapers una vez que se haya hecho una descarga completa de las dos fases para toda España. En scrapings posteriores solo es necesario ejecutar completamente la fase 1, que es la más rápida. La fase 2 solo se ejecutará para aquellos alojamientos recogidos en la fase 1 que no se descargaron en la última descarga.

5 Directorios de vivienda turística de las CCAA

La S.G de Estadísticas de Turismo y Ciencia y Tecnología realiza peticiones periódicas de los directorios de vivienda turística a los responsables en materia turística de cada comunidad autónoma. Debido a que no se pueden recuperar los directorios de todas las comunidades, y a las diferencias en cuanto al grado de actualización y las variables proporcionadas, estos directorios se utilizan únicamente a modo de contraste con los resultados estimados vía plataformas.

6 Delimitación de la vivienda turística

6.1 VIVIENDA TURÍSTICA POR COMUNIDAD AUTÓNOMA

Las plataformas de alojamientos turísticos utilizadas en este proyecto ofertan alojamientos de diversos tipos como hoteles, campings, apartamentos turísticos, albergues, etc; sin embargo, el objetivo de este proyecto es medir la vivienda turística, ya sean ofertadas al completo o por habitaciones. Para ello de toda la información descargada de las plataformas es necesario hacer un filtrado que nos permita obtener únicamente la tipología deseada.

La vivienda turística no es un concepto claramente definido y del cual haya una definición única en España. Esta definición depende de las comunidades autónomas, y cada una clasifica los alojamientos turísticos siguiendo sus propios criterios. Después de un análisis de la legislación en cada una de ellas, se seleccionaron los tipos de alojamiento definidos como vivienda turística. El siguiente cuadro resume los tipos de alojamiento seleccionados en cada comunidad autónoma, así como la nomenclatura de sus licencias:

	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

A partir de la información presente en el cuadro, se estableció un algoritmo de selección de la vivienda turística que se presenta en la siguiente sección.

6.2 ALGORITMO DE SELECCIÓN DE VIVIENDA TURÍSTICA

El algoritmo de selección de la vivienda turística para los alojamientos extraídos de las plataformas es el siguiente:y

Paso 1:

Se separan los alojamientos que tienen licencia de los que no y se armoniza sobre aquellos que la tienen definida. Ejemplos de armonización:

vut 629 as -> VUT/AS/000000000000000629

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

Paso 2:

Sobre aquellos que tienen licencia, se extraen aquellos que hemos determinado como vivienda turística. Si se cumple cualquiera de estas condiciones el algoritmo seleccionará el alojamiento:

- Si CCAA = Andalucía y Tipo de Licencia = VFT o CTC
- Si CCAA = Aragón y Tipo de Licencia = VU
- Si CCAA = Principado de Asturias y Tipo de Licencia = VUT o VV
- Si CCAA = Islas Baleares y Tipo de Licencia = ETV o VTV
- Si CCAA = Canarias y Tipo de Licencia = VV
- Si CCAA = Cantabria y Tipo de Licencia = VUT
- Si CCAA = Castilla y León y Tipo de Licencia = VUT
- Si CCAA = Castilla – La Mancha y Tipo de Licencia = VUT
- Si CCAA = Cataluña y Tipo de Licencia = HUT
- Si CCAA = Extremadura y Tipo de Licencia = AT
- Si CCAA = Comunidad Valenciana y Tipo de Licencia = VT
- Si CCAA = Galicia y Tipo de Licencia = VT o VUT
- Si CCAA = Comunidad de Madrid y Tipo de Licencia = VT
- Si CCAA = Región de Murcia y Tipo de Licencia = VV
- Si CCAA = Comunidad Foral de Navarra y Tipo de Licencia = UVT o UAT
- Si CCAA = País Vasco y Tipo de Licencia = E o L
- Si CCAA = La Rioja y Tipo de Licencia = VT

Paso 3:

Aquellos alojamientos que no tienen licencia o en los que la licencia no venga definida correctamente[3], se seleccionarán o no en función de las variables subtipo de cada plataforma. La determinación de los subtipos considerados como vivienda turística se basa en un análisis de los cruces subtipo-licencia.

7 Algoritmo de desduplicado

7.1 OBJETIVO

La información de vivienda turística recuperada de las plataformas no puede ser sumada para dar un total de alojamientos en España; la razón de esto es que, en muchas ocasiones los propietarios registran los alojamientos en más de una plataforma para darle más visibilidad. Es por esto que se hace imprescindible implementar un algoritmo de desduplicado que elimine alojamientos que están presentes en más de una plataforma a la vez.

7.2 ALGORITMO

7.2.1 Esquema general

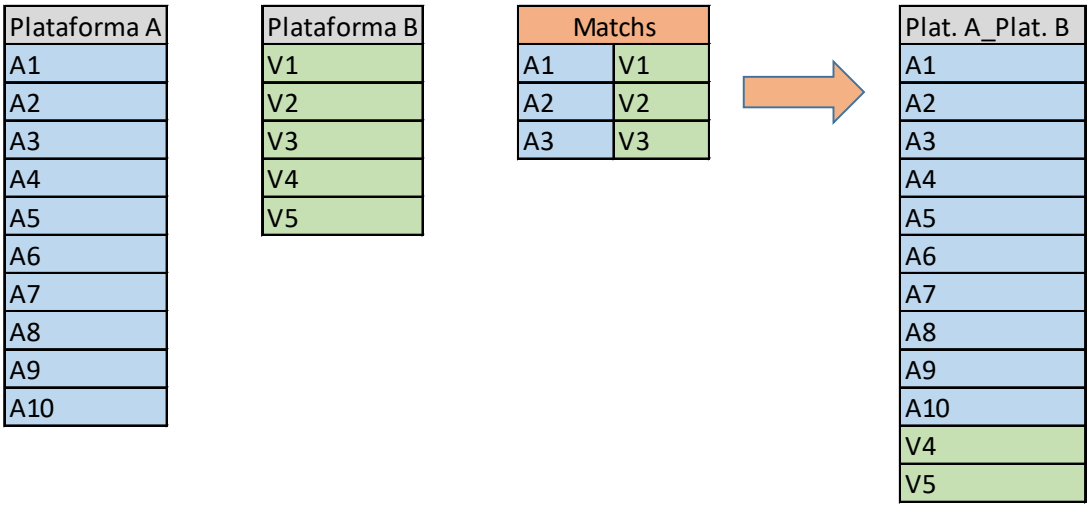
Las entradas para el algoritmo de desduplicación consisten en los alojamientos descargados mediante web scraping de las tres plataformas, y filtrados mediante el algoritmo definido previamente para disponer únicamente de las viviendas turísticas.

El procedimiento que implementa el algoritmo es el siguiente:

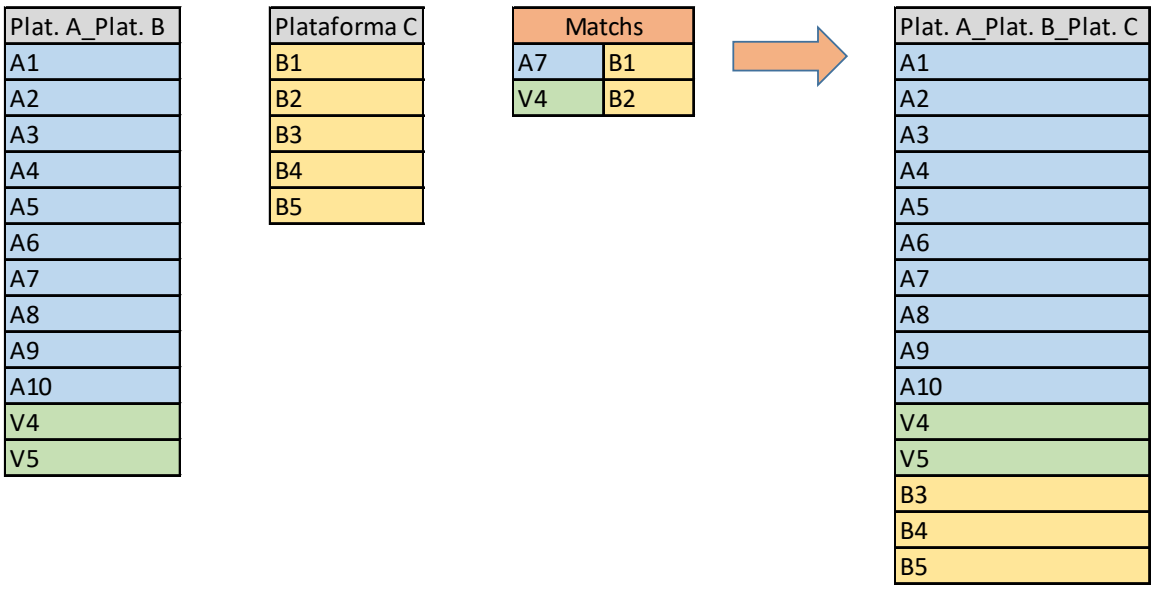
Paso 1:

En primer lugar, se toma como referencia el fichero de una de las plataformas, plataforma A. Después, con el fichero de la plataforma B, se compara cada alojamiento de este con todos los alojamientos de la plataforma A. Aquellos alojamientos que no se encuentren en la plataforma A se añaden a un fichero conjunto de ambas plataformas.

[3] Aproximadamente un 43% de los alojamientos.



Paso 2:



El proceso de comparación entre plataformas puede ser muy costoso si no se hace un filtrado previo de los alojamientos a comparar. Para ello, en el proceso explicado arriba los alojamientos de cada plataforma se comparan por municipio, y además, para aquellos

municipios con un número de alojamientos elevado, se comparan para cada alojamiento únicamente los 300 alojamientos más cercanos de la otra plataforma.

7.2.2 PROCESO DE DESDUPLICADO ENTRE PLATAFORMAS

Para los pasos del punto anterior es necesario definir un algoritmo que determine si los alojamientos de una de las plataformas se encuentran en la plataforma con la que se compara.

Suponiendo que estamos comparando las plataformas A y B el procedimiento sería el siguiente:

Paso 1:

Se realiza un proceso de armonizado previo de las variables para las dos plataformas.

Paso 2:

En el fichero de salida se añaden todos los alojamientos de la plataforma A.

Paso 3:

Se filtran los ficheros de las plataformas A y B para seleccionar únicamente los correspondientes al municipio a comparar.

Paso 4:

Se selecciona el primer alojamiento de la plataforma B y se compara con el primero de la plataforma A.

Paso 4.1:

Se comparan las licencias armonizadas de ambos alojamientos, resultando dos opciones:

- Si coinciden las licencias, no se añade el alojamiento de la plataforma B al fichero de salida ya que se considera un duplicado del de la plataforma A. Se volverá al punto 4 realizando el procedimiento para el siguiente alojamiento del fichero de la plataforma B.
- Si no coinciden las licencias, se continúa con el proceso.

Paso 4.2:

Se comparan variables comunes entre las dos plataformas para determinar si los alojamientos son el mismo. Algunas de las variables comparadas son el nombre del alojamiento, nombre del anfitrión, capacidad, número de dormitorios, subtipo, servicios de internet o parking, si permite mascotas, distancia entre alojamientos,…

A cada una de estas variables se les da un peso en función de la aptitud que tengan para desduplicar. Una vez ponderadas se compara el resultado con un valor límite previamente definido, resultando dos opciones:

– Si se supera este valor límite, no se añade el alojamiento de la plataforma B al fichero de salida ya que se considera un duplicado del de la plataforma A. Se volverá al punto 4 realizando el procedimiento para el siguiente alojamiento del fichero de la plataforma B.
– Si no se supera este valor límite, hay de nuevo dos opciones:
    1. Si hay más alojamientos de la plataforma A frente a los que comparar, se vuelve al paso 4.1 y se compara con el siguiente.
    2. Si no hay más alojamientos de la plataforma A frente a los que comparar, se añade el alojamiento de la plataforma B al fichero de salida y se considera como un nuevo alojamiento que no estaba registrado en la plataforma A. Se volverá al punto 4 realizando el procedimiento para el siguiente alojamiento del fichero de la plataforma B (siempre que existan más).

Una vez se haya finalizado con todos los alojamientos de la plataforma B del municipio se vuelve al paso 3 para seleccionar el siguiente municipio. Así, hasta analizar el total de municipios del territorio nacional.

El fichero de salida de este proceso es un fichero con la combinación de ambas plataformas: **A Y B**

Para los otros dos pares de plataformas, Plataformas C y B y plataformas C y A, el procedimiento es el mismo, la única diferencia son las variables utilizadas para realizar el desduplicado.

7.2.3 Determinación de las plazas

Si uno de los alojamientos está presente en más de una plataforma, se le asignan las plazas de la plataforma que se utiliza como referencia inicial; es decir, primero se utilizan las de la plataforma A, luego las de la B, y si únicamente está en la plataforma C, las de esta. Esta decisión se ha tomado en base a un análisis de los alojamientos presentes en más de una plataforma.

8 Difusión

Se publica información de la vivienda turística en España en forma de tablas y mediante mapas que reflejan la información de un modo más visual.

Las desagregaciones geográficas utilizadas son: por comunidad autónoma, provincia, municipio, distritos y secciones censales. Además se desagrega por tamaño del municipio, zona (costera o interior) y grado de urbanización.

Los resultados se publican en torno a cuatro meses después de la fecha de descarga de información de las tres plataformas; esto es, en junio y diciembre del año correspondiente.

9 Calendario

El calendario de las principales fases de la operación ha sido el siguiente:

– Desarrollo de los programas para el análisis de las tres plataformas: de noviembre de 2019 a julio de 2020.

– Desarrollo del algoritmo de desduplicado: abril de 2020 a septiembre de 2020.

– Integración de la información y elaboración de productos de difusión: octubre de 2020 a noviembre de 2020.

– Primera publicación: en diciembre de 2020 se publicaron los resultados de agosto del mismo año.

– Publicaciones sucesivas: en junio y diciembre de 2021 se publicaron los resultados obtenidos en febrero y en agosto de dicho año, respectivamente.