Nueva fuente de datos para el cálculo del ADR (tarifa media diaria por habitación ocupada)

Introducción

EUSTAT, consciente de las enormes oportunidades de obtención de información mediante la utilización de distintas técnicas englobadas dentro del término de *Big Data,* está trabajando en diferentes proyectos para la incorporación de datos obtenidos mediante estas técnicas en sus procesos de estimación. Se están estudiando procedimientos de recogida de información, así como su utilización en la estadística oficial, bien como fuente única de datos bien en combinación con datos de origen muestral y/o administrativo.

En este contexto, en el año 2017 se inició un proyecto para la obtención de la tarifa media diaria por habitación ocupada (ADR) de la Encuesta de establecimientos turísticos receptores utilizando metodología *Big Data*.

A partir de octubre de 2020 los datos difundidos de esta variable (ADR) y de la variable derivada *Ingresos por habitación disponible* (RevPar), tienen su origen en la captura de información disponible en internet a través de las plataformas de reserva *online* de habitaciones en establecimientos hoteleros.

La captura de información de la web (*web scraping*) se realiza siguiendo las recomendaciones de Eurostat respecto a garantías legales, competencia, protección de datos y secreto estadístico.

Modelización del ADR

Para la estimación final del ADR de los establecimientos hoteleros de la C.A. de Euskadi se ha definido el siguiente modelo estadístico que combina la información obtenida de las plataformas de reserva online utilizando técnicas de *web scraping*, la presente en el directorio de la encuesta, así como la

obtenida por encuestación de los propios establecimientos:

$$ADR = \beta_0 + \beta_1 Mediana + \beta_2 Max + \beta_3 Min + \beta_4 Categoria$$

$$+\beta_5 Estrato + \beta_6 ReservaOnline + \beta_7 Ocupacion + \beta_8 Tamaño + \varepsilon$$

donde las variables explicativas vienen dadas por:

- ***Mediana***, ***Max*** y ***Min*** son la mediana, máximo y mínimo de precios diarios capturados en Internet durante los últimos de 4 meses
- ***Categoría***: Se consideran 5 niveles: 4 y 5 estrellas, 3 estrellas, 2 estrellas, 1 estrella, pensiones.
- ***Estrato***: zona geográfica en la que se ubica el establecimiento
- ***ReservaOnline:*** variable dicotómica que indica la presencia o no del establecimiento en plataformas de reserva *online* de habitaciones.
- ***Ocupación***: tasa de ocupación por habitaciones del establecimiento durante el mes de referencia
- ***Tamaño:*** bloque de tamaño según número de habitaciones del establecimiento (<9, 9-15, 16-39, > 40 habitaciones).

La obtención de la información en plataformas de reserva *online* se realiza para cada establecimiento hotelero mediante *web scraping* una vez al día para cada uno de los 120 días precedentes, de forma que para cada establecimiento y día se disponen de 120 precios por plataforma de reserva. Por consiguiente, en un mes de 31 días obtendría un total de 3.720 precios distintos para cada establecimiento y plataforma.

Durante un período de dos años se ha contrastado el ADR obtenido a partir de la información encuestada con el ADR estimado utilizando distintos modelos y procedimientos de captura de datos, siendo el modelo propuesto más arriba el más satisfactorio.

Tanto la captura de la información como el procedimiento de estimación están integrados en los procesos producción de la Encuesta de establecimientos turísticos receptores de Eustat.