Big Data y la Encuesta de Establecimientos Turísticos Receptores

Jorge Aramendi Rique, Javier San Vicente Alfambra
EUSTAT - Euskal Estatistika Erakundea / Instituto Vasco de Estadística

El proyecto

El *Big Data* se define de manera generalizada como un conjunto de datos de volumen, variedad y velocidad creciente, con frecuencia no estructurado o que no se ajusta bien al modelo de bases de datos relacionales con las que las oficinas estadísticas vienen trabajando.

El *Big Data* es una interesante fuente de información para las oficinas estadísticas, bien para su aprovechamiento independiente bien en combinación con otras fuentes de datos más tradicionales como las encuestas por muestreo o los registros administrativos.

Una de sus mayores ventajas es que posibilita el acceso a un ingente volumen de información a un costo menor y más rápido que los datos estadísticos tradicionales. Pero también abre una serie de retos e incógnitas. Obviamente, se está ante un nuevo paradigma en lo que al uso de tecnología de información se refiere. Por otra parte, es indudable el impacto que el uso del *Big Data* tendrá en los métodos de recogida, explotación y difusión de las estadísticas.

Ante los retos y oportunidades que plantea el *Big Data* Eurostat en su Memorando Scheveningen (2013) propone:

- Promover el estudio del potencial de *Big Data* para la estadística oficial y exploración de su uso en coordinación con otros organismos gubernamentales.
- Dar prioridad a la identificación de riesgos del *Big Data* (protección de datos, derechos de las personas, etc.)
- Entrenar al personal en la cuestión. Colaborar con otros para alcanzar el objetivo del aprovechamiento del *Big Data*.
- Crear nuevas metodologías de trabajo y adaptar las tradicionales.

En el marco del este memorando se han llevado a cabo diversos estudios sobre, entre otros, estadísticas de sociedad de la información, estadísticas de turismo, precios, estadísticas de actividad económica o estadísticos de confianza de consumidor.

EUSTAT, consciente de que el *Big Data* es una interesante fuente de información para las oficinas estadísticas, bien para su aprovechamiento independiente, bien en combinación con otras fuentes de datos más tradicionales como las encuestas por muestreo o los registros administrativos, ha llevado a cabo un estudio piloto para el estudio de las series diarias de precios de habitaciones de hotel y su posible aprovechamiento en la Encuesta de Establecimientos Turísticos Receptores (ETR). En esta ponencia EUSTAT presenta los resultados del que es su primer estudio de caso en el ámbito del *Big Data*.

La ETR de la C.A. de Euskadi es una operación de carácter mensual cuyo objetivo es medir la evolución del movimiento turístico (entradas y pernoctaciones) registrado en los establecimientos turísticos de la C.A. de Euskadi. Hasta el momento se recoge y difunde información de hoteles, pensiones y agroturismos.

Por lo que respecta a los hoteles y pensiones se trata de una encuesta censal, es decir, recoge información de todos los establecimientos turísticos de la C.A. de Euskadi, aunque no con el mismo tratamiento para todos ellos. Para los hoteles de 3, 4 y 5 estrellas se recogen los movimientos registrados en el establecimiento todos los días del mes de referencia. En el caso de los hoteles y pensiones de 1 y 2 estrellas se recoge los movimientos acaecidos durante 7 días consecutivos –asignados al azar- del mes de referencia.

Los datos obtenidos se publican atendiendo a una estratificación geográfica y por categoría de hotel difundiéndose además de las entradas y pernoctaciones de viajeros, según su origen y destino, otros indicadores como el número de establecimientos hoteleros de la C.A. de

Euskadi, las plazas ofertadas, los grados de ocupación, la estancia media o el personal empleado en el sector.

Entre los objetivos de la operación para este año 2016 se ha señalado la difusión de tablas estadísticas para el ADR y el REVPAR de los establecimientos hoteleros con el mismo nivel de desagregación que para el esto de variables. Se define el ADR como la tarifa media diaria aplicada a una habitación doble con baño por alojamiento descontado cualquier tipo de otro servicio y/o impuestos, siendo el REVPAR el ingreso por habitación disponible. Los dos tipos de cuestionario, el mensual y el semanal, incorporan una pregunta relativa al ADR y al REVPAR.

El ADR se obtiene como promedio ponderado por el porcentaje de habitaciones ocupadas de nueve tipos diferentes de clientes cada uno de ellos con su ADR: tour operador tradicional, empresas, agencia de viajes tradicional, particulares, grupos, internet, agencia de viajes on-line, **tour operador on-line** y otros.

Web scraping o escrapeado

El objetivo es la puesta en marcha de un proyecto piloto para obtener los precios medios diarios de los establecimientos hoteleros, hoteles y pensiones, que conforman el Directorio de Establecimientos Turísticos Receptores (DIRTU). La unidad de análisis es la habitación doble estándar con baño, sin desayuno y sin IVA.

Además de las propias páginas webs de los establecimientos hoteleros, existen numerosas páginas web que ofertan la reserva de habitaciones. Por citar las más relevantes: Booking, Atrápalo, Trivago, Rastreator… Considerando los objetivos del proyecto se decidió seleccionar únicamente una de ellas, Booking, estimando que sería suficiente para poner de relevancia la casuística derivada tanto de la herramienta informática como del método para obtener los precios medios diarios.

El Web scraping o escrapeado es una técnica utilizada mediante programas de software para extraer información de sitios web.

Para el desarrollo del prototipo se hizo uso de la tecnología de *import.io*. Consiste en un programa que permite crear plantillas de escrapeado adaptado a un tipo de estructura de página web concreto, la página web de Booking en este caso. Para su desarrollo no son precisos profundos conocimientos de programación, ya que todo se realiza mediante la interfaz gráfica del programa. La plantilla diseñada se aplica a la página web que, siempre que mantenga la estructura para la cual se generó la plantilla, será escrapeada y guardada en una tabla. La tabla generada tendrá tantas columnas como campos y tantas filas como repeticiones de la estructura detectadas (en nuestro caso, tantas filas como hoteles). Añadidamente, Import.io genera una API independiente para cada plantilla, de forma que el programa recibe como respuesta el conjunto de datos escrapeado en formato XML.

En este caso, la página web de Booking es fácilmente escrapeable ya que muestra los resultados de la búsqueda en un listado bien estructurado fácilmente detectable por Import.io. No obstante, los resultados de la consulta se muestran en su página en un listado de 15 hoteles máximo, con lo que es necesario lanzar más de 50 consultas cada día para capturar la información de todos los hoteles disponibles en Booking. En el caso de Euskadi, Booking oferta información sobre más de 80intenr0 establecimientos hoteleros de distinto tipo: hoteles, pensiones, apartamentos y agroturismos.

El propósito del proyecto se centró inicialmente en obtener precios únicamente para los hoteles y pensiones. Eustat está ampliando la difusión de la estadística de establecimientos turísticos receptores con vistas a difundir resultados de ADR y REVPAR. En este sentido, tal y como se ha señalado, uno de los objetivos fundamentales del proyecto es implementar en sus estadísticas información procedente de *Big Data*. Actualmente el directorio de Eustat existen 556 establecimientos de estos tipos, hoteles y pensiones.

Por otra parte, un requisito del escrapeado era obtener los precios ofertados a siete y a catorce días. Inicialmente se han ensayado estos plazos de manera tentativa. En el análisis de resultados se hablará más extensamente sobre la conveniencia de determinar adecuadamente los plazos para garantizar resultados óptimos.

Para los objetivos del proyecto se consideró suficiente escrapear la oferta hotelera de Booking durante 45 días. Todos los días desde el 27 de enero al 11 de marzo a las 9:00 de la mañana se ejecutaba un script con la consulta de precios a 7 y 14 días vista.

El resultado de ejecutar la consulta diariamente se muestra en la Cuadro 1. En esta primera aproximación se incluyen todos los establecimientos que resultaron escrapeados al menos una vez durante el período de referencia.

El universo de Booking para la C.A. Autónoma de Euskadi durante estos 45 días contemplaba 955 establecimientos diferentes, dos menos tanto en el caso de a siete días vista como en el de a catorce días vista. Sin embargo, para más de 300 establecimientos no se consiguió un precio de reserva a semana o dos semanas vista.

Los motivos de ausencia de oferta pueden ser variados. La fiabilidad de import.io no es alta, ocurren numerosos errores con los que la respuesta no es la esperada. Entre ellos, se han detectado errores de conexión, errores de escrapeo (el número de hoteles es 0 o menor del esperado), tiempo de respuesta demasiado grande… Se trata de errores que es necesario identificar, controlar y corregir para obtener una tabla depurada. Por supuesto, también es posible que el establecimiento esté completo y no permita reservas en las fechas determinadas o, simplemente, esté temporalmente cerrado.

Cuadro 1. Booking. establecimientos, escrapeos totales y escrapeos medios por establecimiento según plazo de la consulta

	7 días	14 días	Total	7 días	14 días	Total
Establecimientos	953	953	955	616	618	628
Escrapeos totales	48.102	47.882	95.984	14.296	14.610	28.906
Escrapeos medios por establecimiento	50,40	50,10	100,5	23,2	23,6	46,0

Filtrando los establecimientos para los que sí se obtuvo algún precio en la consulta, el número de establecimientos se reduce hasta el 65,8% de los iniciales del universo Booking y, más aún, el total de escrapeos con precio, 28.906, que se queda en un tercio de los de partida. En consecuencia, el número de escrapeos medios por establecimiento se limita a algo más de 23 tanto si se trata de precio a semana vista como si es a dos semanas vista.

Antes de delimitar correctamente las unidades coincidentes del Directorio ETR de Eustat y el universo Booking, para poder analizar los precios, hubo de solucionarse otros problemas. A los ya mencionados de errores en la respuesta esperada, hay que añadir que el universo Booking incluye todo tipo de establecimientos: alojamientos rurales, apartamentos individuales, conjunto de apartamentos, residencias, hoteles etc. Por su parte, el Directorio de Eustat contempla todos los tipos definidos por la Ley de Turismo del País Vasco de 13, de 2016 de 18 de julio si bien ETR, de momento, solo difunde resultados para hoteles y agroturismos. Por lo tanto, para establecer una correcta comparación hubo de relacionar el universo Booking con las unidades del Directorio de Eustat pero únicamente con los establecimientos hoteleros en que se concretó esta prueba piloto.

Otro de los problemas detectados durante la fase del proyecto fue que los responsables de import.io comenzaron a penalizar la realización de consultas repetidas en un corto espacio de tiempo, por lo que fue necesario modificar el script para incorporar un tiempo de retraso.

Igualmente, en un momento dado, Booking.com modificó la estructura de sus páginas devolviendo un número no constante de hoteles por página, frente a los 15 por página que venía devolviendo. Tratándose de un problema de difícil tratamiento la solución se limitó a

forzar un número mínimo de establecimientos por consulta, los 15 iniciales, estimando que de este modo se garantizaba una alta cobertura. Añadidamente, Booking también alteró la estructura de la página con lo que el escrapeado dejó de funcionar correctamente y solo se pudo solucionar a posteriori, re-entrenando la plantilla.

También se tuvo que controlar la existencia de *outliers*. Varios de los precios devueltos eran manifiestamente erróneos – por ejemplo, 999 euros para un hotel de una estrella –. Este problema es de fácil solución para los *outliers* más evidentes, se suprimieron directamente, pero, a futuro, requiere de un análisis más profundo para su correcto tratamiento.

Con todo, los tratamientos más significativos para el proyecto fueron los derivados de la necesidad de enlazar el universo Booking con el Directorio de ETR, exclusivamente para los hoteles, propósito del prototipo. Inicialmente se consideró el nombre del hotel como identificador único, pero quedó patente que no era una buena aproximación. Por un lado, no siempre coincide el nombre de Booking con el del directorio. Por este motivo, se cotejó la dirección disponible en el directorio con la que aparece en la página web específica del establecimiento en Booking para establecer el emparejamiento de establecimientos con nombres similares. Además, se encontraron varios establecimientos que compartían el mismo nombre.

Por otro lado, Booking permite cambiar el nombre al hotel. Así, se han observado cambios parciales de nombres a lo largo de estos 45 días lo que invalida el uso del nombre como identificador único.

Considerando el plazo limitado de tiempo en que el prototipo iba a estar en funcionamiento, así como el relativo escaso número de establecimientos afectados los tratamientos de enlace de establecimientos Booking con establecimientos del Directorio de Eustat fueron llevados a cabo a posteriori y de forma manual, asignando a los coincidentes el identificador único del Directorio de Eustat.

Una vez finalizadas las labores de tratamiento y depuración, se constituyó una tabla con todos aquellos establecimientos del Directorio de Eustat para los que la consulta sobre Booking había devuelto al menos un precio válido a 7 ó 14 días durante los 45 días. Esta tabla consta de 381 hoteles de los 556 que componen DIRTU y 15. 209 registros o precios. A la tabla, se incorporaron el resto de variables de DIRTU que se consideraron de interés, como categoría, estrato, empleo o municipio.

Cobertura de Booking sobre el directorio de EUSTAT

Por lo tanto, un elemento clave para determinar la bondad de la utilización de BigData en la posible determinación de los precios de oferta para la Estadística de Establecimientos turísticos Receptores es el grado de cobertura obtenido de la información proveniente de Booking.

En el cuadro 2 se aprecia el grado de cobertura de los establecimientos ofertados por Booking es del 70% para el total con respecto al directorio de Eustat. Como se ha mencionado, la consulta se diseñó para escrapear los precios a una y dos semanas vista a partir de la fecha de consulta. Se ha considerado que una única coincidencia es suficiente para formar parte del cálculo de la cobertura.

Cuadro 2. Ratio de cobertura de establecimientos de Booking sobre el directorio ETR

	Gasteiz	Resto Alava	Rioja Alavesa	Bilbao	Area Metro. Bilbao	Bizkaia interior	Bizkaia costa	Donostia-San Sebastián	Area Metro.DSS	Gipuzkoa interior	Gipuzkoa costa	EUSKADI
H1	60%	86%	100%	50%	100%	56%	80%	100%	80%	77%	85%	78,8%
H2	100%	33%	67%	100%	60%	73%	91%	100%	100%	73%	100%	82,4%
H3	100%	100%	100%	84%	100%	100%	78%	100%	0%	78%	100%	90,2%
H4	100%	0%	100%	100%	100%	100%	100%	100%	100%	100%	80%	96,1%
H5	100%	–	100%	100%	–	–	–	100%	–	–	–	100%
P1	0%	14%	–	58%	44%	13%	28%	63%	57%	25%	54%	42,9%
P2	57%	0%	–	75%	60%	33%	44%	86%	86%	11%	69%	67,3%
TOTAL	66,7%	45,8%	83,3%	78,6%	71,0%	53,1%	63,6%	80,8%	77,8%	55,4%	79,0%	70,1%

Como cabía esperar, el grado de cobertura es directamente proporcional a la categoría del hotel, de forma que es del 100% para los hoteles de 5 estrellas, descendiendo a medida que lo hace la categoría, limitándose al 79% en los hoteles de una estrella. De igual forma sucede en el caso de las pensiones: 67% de cobertura en las de dos estrellas frente al 43% en las de una estrella. Hay que recordar que la disponibilidad de página web por parte del establecimiento es requisito necesario para ofertarse en Booking y este hecho es, lógicamente, menos frecuente entre los establecimientos de categoría inferior.

Además de por la categoría del hotel, Eustat difunde resultados conforme a una zonificación significativa que distingue capitales, zonas metropolitanas (excepto en Álava), costa e interior para Bizkaia y Gipuzkoa y Rioja Alavesa y Resto de Álava en el caso de este último Territorio. Atendiendo a esta zonificación, las coberturas más amplias se obtienen en las capitales de Bizkaia y Gipuzkoa y en sus áreas metropolitanas, además de en Gipuzkoa Costa, zonas que concentran buena parte de la oferta hotelera de la Comunidad (casi el 70%). En el caso de Rioja Alavesa, aún obteniendo la cobertura más amplia, su oferta hotelera es muy reducida.

Es de destacar igualmente el hecho de que en 28 de los 77 estratos conformados por la combinación de categoría y la zona geográfica, la cobertura es total.

Se puede obtener más información atendiendo al ratio de cobertura por plazas. Para ello, se establece la relación entre las plazas de los hoteles de DIRTU ofertados en Booking sobre el total de plazas de DIRTU. Siguiendo la estratificación anterior los resultados se muestran en el cuadro 3.

Cuadro 3. Ratio de cobertura de plazas de Booking sobre el directorio ETR

	Gasteiz	Resto Alava	Rioja Alavesa	Bilbao	Area Metro. Bilbao	Bizkaia interior	Bizkaia costa	Donostia-San Sebastián	Area Metro.DSS	Gipuzkoa interior	Gipuzkoa costa	EUSKADI
H1	79%	77%	100%	36%	100%	49%	88%	83%	95%	78%	90%	81,2%
H2	100%	33%	85%	100%	61%	68%	81%	100%	100%	84%	100%	88,8%
H3	100%	77%	100%	92%	100%	100%	97%	100%	-	100%	100%	97,4%
H4	100%	0%	100%	100%	100%	85%	100%	100%	100%	100%	89%	98,3%
H5	100%	-	100%	100%	-	-	-	100%	-	-	-	100%
P1	0%	25%	-	79%	74%	14%	28%	59%	19%	34%	73%	52,1%
P2	41%	0%	-	64%	84%	68%	49%	86%	83%	6%	45%	63,8%
TOTAL	93,0%	48,7%	97,0%	94,5%	92,7%	64,6%	78,0%	93,7%	80,2%	76,1%	87,6%	88,3%

Para todas las zonas geográficas y categorías de hotel contempladas el ratio de cobertura en plazas es superior al de establecimientos. Para el conjunto de la C.A. de Euskadi la cobertura de plazas es casi 20 puntos superior a la de establecimientos, aproximándose al 90%.

La cobertura por zonas geográficas, lógicamente, se ve condicionada por la distribución de los diferentes tipos de establecimiento. Los establecimientos de mayor categoría ofertan un mayor número de plazas y se concentran en las capitales y zonas en las que existe oferta turística. Así, en las tres capitales, el Área Metropolitana de Bilbao, Gipuzkoa Costa y, especialmente, en Rioja Alavesa la cobertura de plazas de los establecimientos coincidentes en Booking se sitúa en porcentajes próximos o que superas el 90%. El resto de zonas observa ratios de cobertura en plazas entre el 60% y el 80%. Solo en la zona Resto de Álava no se alcanza el 50% de cobertura.

Algunas categorías de hotel, las más altas, aproximan la cobertura en plazas de Booking prácticamente al 100% de las plazas de DIRTU. Es el caso de los hoteles de 3, 4 y 5 estrellas (en los de 5 estrellas ya se alcanzaba un 100% de cobertura de establecimientos). Únicamente en las pensiones de dos estrellas coincidentes en Booking señalan un mayor porcentaje de establecimientos en el DIRTU que la proporción sobre las plazas que representan.

Por lo tanto, se puede concluir que la cobertura de Booking sobre DIRTU es positiva; muy elevada para todas las categorías de hotel y razonablemente alta en el caso de las pensiones. Más considerando que la información obtenida en ETR de las pensiones se basa en un diseño muestral – se obtiene información de una de cada cuatro semanas - y que son los establecimientos que presentan una mayor tasa de no respuesta (entre el 8% y el 10% en el año 2015).

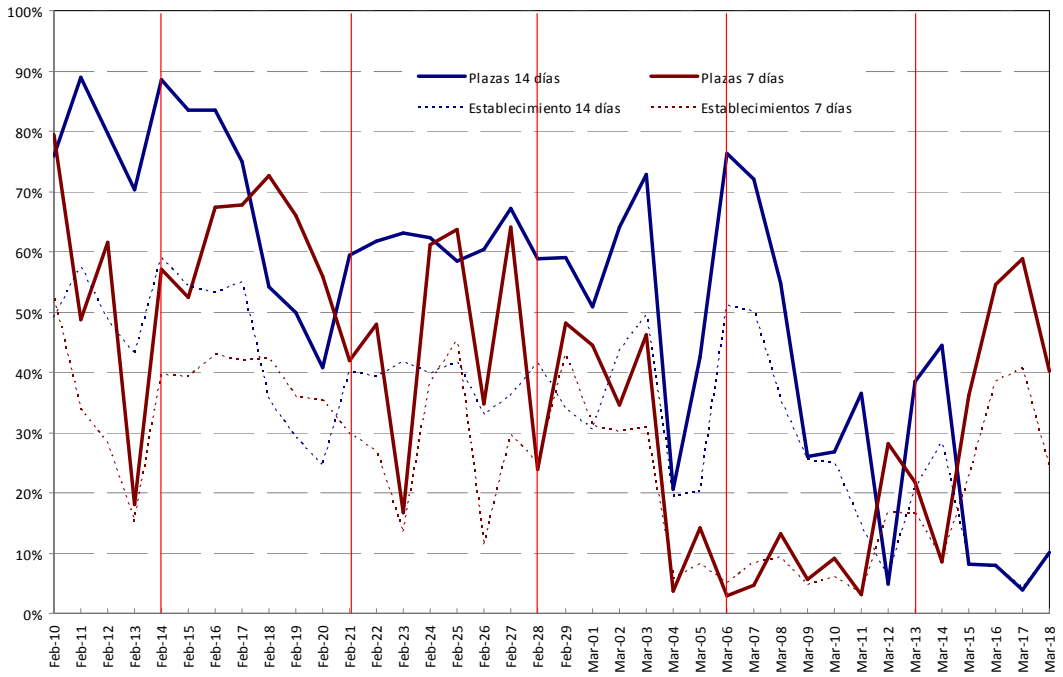
Uno de los aspectos que se ha mostrado decisivo durante el análisis para la obtención de los precios de Booking es el establecer correctamente los plazos de consulta sobre la reserva. En este prototipo se escogieron dos plazos: siete y catorce días.

El gráfico 1 muestra la evolución día a día del ratio de cobertura de plazas total para las fechas coincidentes en ambos plazos. Para un día concreto se obtienen dos coberturas: la correspondiente a la consulta generada la semana previa, cobertura a siete días, y la de la consulta generada dos semanas antes, cobertura a 14 días. La cobertura calculada para cada día considera los hoteles para los que Booking devuelve algún precio ese día y es, por tanto,

inferior a la cobertura para todo el período en que el prototipo estuvo operativo mostrada en los cuadros uno y dos.

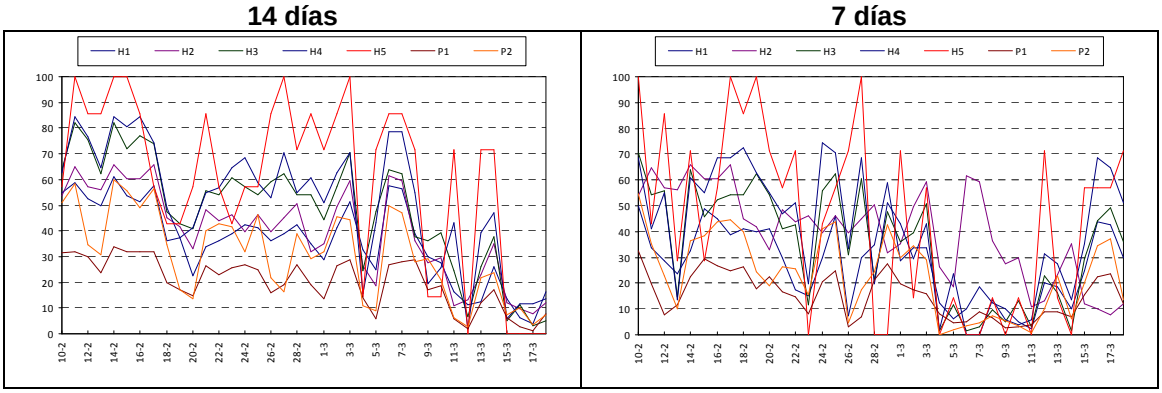
Salvo en días concretos, el ratio de cobertura de plazas a 14 días es siempre significativamente superior al de 7 días consecuentemente con una previsible mayor disponibilidad en la medida en que el plazo para la reserva es superior. Las excepciones a esta regla derivan del ratio de cobertura por establecimiento en las fechas en que se han capturado un número mayor de establecimientos.

Gráfico 1. Ratio de cobertura en plazas y establecimiento según plazo de la consulta



Sin embargo, no parece existir una correlación clara entre los dos plazos de cobertura. No se observa un patrón común a ambos plazos cuyas líneas muestran perfiles diferentes a lo largo de los 45 días, tanto en lo que se refiere a cobertura de plazas como de establecimientos. De hecho, entre el 18 y el 20 de febrero y los últimos días del proyecto la cobertura a siete días es claramente superior a la de catorce días. Lo mismo cabe decir de los perfiles del ratio de cobertura de plazas. Cabría esperar que las desigualdades en el nivel de cobertura para ambos plazos fueran aproximadamente constantes durante el período considerado, particularmente en unos meses como febrero y marzo cuando la ocupación no presenta picos estacionales, de forma que las diferencias se debieran en buena medida a un patrón de comportamiento diferenciado de la demanda en función del plazo de reserva. Pero no es así, los márgenes entre las líneas de los dos plazos se estrechan y se ensanchan a lo largo de los 38 días en que ambos plazos de consulta coinciden.

Gráfico 2. Ratio de cobertura en establecimiento según categoría del establecimiento



Analizando la cobertura por establecimiento según su categoría se observa que a catorce días todas las categorías de hotel siguen un patrón muy similar (gráfico 2). Si bien los niveles de cobertura son claramente diferentes, los picos y valles se producen, de manera general, en los mismos días independientemente de la categoría, observándose una evolución muy parecida para todos ellos. Mientras, en el caso de la cobertura a siete días, para una misma fecha se alternan coberturas relativamente altas en unas categorías y relativamente bajas en otras categorías. Cabe deducir, por tanto, una mayor estabilidad en la cobertura a catorce días, quedando más influenciada por factores circunstanciales la cobertura a plazos más reducidos.

Cuadro 4. Ratio de cobertura de plazas de Booking por categoría del hotel según día de la semana y plazo

	LABORABLE		FESTIVO	
	7 días	14 días	7 días	14 días
H1	30,2%	38,4%	22,0%	30,8%
H2	32,7%	43,3%	22,4%	32,2%
H3	36,2%	51,2%	32,2%	41,0%
H4	41,4%	55,4%	36,5%	41,9%
H5	39,7%	61,9%	55,8%	55,8%
P1	17,5%	22,8%	9,8%	14,2%
P2	27,8%	37,0%	11,3%	15,5%
TOTAL	28,4%	37,8%	19,4%	25,8%

En cuanto al día de la semana, cuadro 4, los festivos (viernes y sábados) Booking constata una menor cobertura en los días festivos frente a los laborables tanto a siete días como a catorce días vista e independientemente de la categoría del hotel. Además, la diferencia entre laborables y festivos se incrementa en el plazo de a catorce días. Si en promedio la cobertura a siete días es 9 puntos superior a la de catorce días, favorable a los días laborables, en el plazo a catorce días se eleva hasta los 12 puntos, con un máximo de casi 22 puntos de diferencia para las pensiones de categoría dos. La cobertura a siete días únicamente es superior para los hoteles de cinco estrellas, pero el escaso número de hoteles de esta categoría quizás esté introduciendo variaciones aleatorias.

ADR y precios de Booking

Para este año 2016 EUSTAT tiene previsto difundir mensualmente tablas estadísticas para ADR y REVPAR por zona geográfica y categoría de hotel. En el cuestionario incluye la correspondiente pregunta sobre la tarifa media diaria aplicada a una habitación doble, distinguiendo nueve tipos diferentes de clientes, entre ellos clientes que han reservado la habitación por medio de tour operador on-line. Esta categoría de cliente, y su ADR, sería, por lo tanto, equiparable al precio medio obtenido en el escrapeado de Booking.

Los datos que aquí se presentan están en fase de depuración inicial, no habiéndose llevado a cabo tratamiento de *outliers* ni de establecimientos particulares en cuanto a su ADR. El objetivo al presentarlos en esta ponencia, aunque a título provisional, es ilustrar la posible bondad de la utilización de información proveniente de *Big Data* en la determinación del ADR.

En primer lugar, según ETR, en promedio uno de cada cuatro de los viajeros del año 2015 realizaron sus reservas por medio de tour operador on-line (cuadro 5). Los hoteles de 1 y 3 estrellas se sitúan en el promedio general, y ligeramente por encima los de 2 estrellas, pero los porcentajes son más reducidos en los hoteles de 4 y 5 estrellas para los que este tipo de cliente apenas alcanza el 10%, mientras que en el caso de las pensiones se eleva hasta corresponder a uno de cada dos clientes. En los meses de febrero y marzo de 2016, la proporción de clientes de tour operador on-line es muy similar a la del año 2015 completo; únicamente las pensiones de 2 estrellas y, en menor medida, los hoteles de 2 estrellas aumentan la proporción de clientes de este tipo (5 y 9 puntos porcentuales respecitvamente).

Cuadro 5. ADR, ADR de T.O. On-line y proporción de clientes de T.O. On-line por categoría del hotel. Año 2015 y febrero de 2016. Datos provisionales

	2015			2016 febrero y marzo		
	ADR	ADR T.O. ON LINE	% T.O. ON LINE	ADR	ADR T.O. ON LINE	% T.O. ON LINE
H1	61	71	24%	59	65	24%
H2	67	73	32%	63	63	37%
H3	64	67	20%	61	64	21%
H4	97	83	11%	89	74	12%
H5	157	157	8%	149	119	6%
P1	42	57	55%	39	49	58%
P2	52	68	44%	50	61	53%
TOTAL	61	76	26%	57	66	30%

Dependiendo de la categoría del hotel, las diferencias entre ADR tour operador on-line y el ADR total pueden ser considerables. En promedio, el ADR del tipo de cliente tour operador on-line es un 20% superior al ADR total en el año 2015. En el caso de las pensiones la reserva de tour operador on-line es un 25% más cara que el promedio. Al contrario, los hoteles de 4 estrellas presentan un ADR próximo al 20% más barato si la reserva se efectúa por medio de un tour operador on-line. A este respecto, hay que considerar que los tipos de clientes particulares y de empresa cubren en torno al 35% del total de reservas de los hoteles de la comunidad y son los tipos de cliente que presentan las tarifas más ventajosas, mientras el ADR del cliente de tour operador on-line suele situarse entre los precios de tipo de cliente más caros en el caso de las pensiones y hoteles hasta 3 estrellas.

Cuadro 6. Booking. Precio medio y nº de observaciones por categoría de establecimiento según plazo de la consulta

	7 días		14 días		Total	
	Media	N precios	Media	N precios	Media	N precios
H1	66	1.149	66	1.184	66	2.333
H2	76	1.403	76	1.497	76	2.900
H3	77	1.119	77	1.203	77	2.322
H4	102	1.036	102	1.099	102	2.135
H5	177	163	165	174	171	337
P1	49	1.233	49	1.270	49	2.503
P2	61	1.325	62	1.354	62	2.679
TOTAL	73 €	7.428 €	73 €	7.781 €	73	15.209

Por su parte, los precios analizados de Booking, cuadro 6, corresponden a los meses de febrero y marzo de 2016 y prácticamente no presentan diferencias en función del plazo de la consulta. Excepto en los hoteles de 5 estrellas, en el que el precio a 14 días se reduce 12 euros (6,8% menos), para las demás categorías no se constata variación alguna. Para cada plazo figura el número de observaciones con la que se ha calculado la media.

En el cuadro 5, ADR obtenido por la encuesta ETR, se observa que los hoteles de 3 estrellas marcaban un ADR medio inferior al ADR de los hoteles de 2 estrellas tanto para 2015 como para febrero y marzo de 2016, peculiaridad que se incrementa cuando el ADR es para el tipo de cliente tour operador on-line y que, en cierta medida, se reproduce en los precios de Booking: entre los hoteles de 2 y 3 estrellas solo hay un euro de diferencia.

Antes de comparar los precios de una y otra fuente, cuadro 7, hay que indicar que el ADR se calcula con los precios de reserva realmente aplicados por los hoteles en el momento de la reserva. En cambio, los precios escrapeados en Booking son precios de consulta a los días que marca el plazo de la propia consulta (7 o 14 días).

Cuadro 7. ADR de T.O. On-line y precios Booking por tipo de cliente. 2016 febrero y marzo. Datos provisionales.

	ADR T.O. ON LINE	BOOKING	% ADR T.O. ON LINE / BOOKING
H1	65	66	-1,7
H2	63	76	-17,0
H3	64	77	-16,8
H4	74	102	-27,8
H5	119	171	-30,2
P1	49	49	-0,6
P2	61	62	-2,4
TOTAL	66	73	-9,6

En promedio, el ADR de la encuesta ETR para el tipo de cliente tour operador on-line es un 9,6% inferior al resultante del escrapeado en Booking para las mismas fechas con un rango de diferencias que va del -1,7% para los hoteles de 1 estrella, apenas un euro, al -30,2% para los de 5 estrellas, variación esta última muy próxima a la registrada en los hoteles de 4 estrellas (-27,8%). En las pensiones las diferencias de precios entre encuesta y Booking son prácticamente idénticas, de un euro en el peor de los casos. Por lo tanto, las discrepancias de precios en los hoteles y pensiones de categorías más bajas son insginifantes pero se hacen máximas y significativas en los de categorías altas.

Conclusiones

El objetivo de EUSTAT con el proyecto de escrapeado de precios de Booking era poner de relevancia las ventajas y dificultades que desde el punto de vista de un instituto de estadística oficial conlleva uso de información proveniente de *Big Data* en las estadísticas

El prototipo de escrapeado creado al efecto no ha estado exento de dificultades. Las más destacables han sido la necesidad de unir correctamente los establecimientos de DIRTU con los hoteles de Booking y los cambios en la estructura de la página web, si bien es cierto que estos problemas se han solventado sin un gran esfuerzo técnico ni en tiempo. En la parte positiva, la tecnología empleada es de manejo asequible, sin requerir profundos conocimientos informáticos, además de fácilmente reprogramable en el caso de cambios en la página web.

En cuanto a la cobertura sobre DIRTU, se puede calificar de muy elevada en los hoteles de categorías superiores y de razonable en el caso de los de menos estrellas, considerando que ETR tiene parte de diseño muestral para el caso de estos últimos establecimientos.

Quizás la parte menos positiva en lo que se refiere al proyecto ha sido la relativa a la comparación de precios de una y otra fuente, con diferencias que obligan a profundizar en sus causas. La definición del precio en una y otra fuente, el modo en que se obtiene el ADR en la encuesta de ETR - con una pregunta sobre precios medios - la tipología de los hoteles no capturados en el escrapeado y sus precios, el funcionamiento de Booking en la determinación de la oferta de plazas y sus precios, son algunas de las cuestiones en las que el proyecto debe de buscar su continuidad.