INTERNET COMO HERRAMIENTA DE PREDICCIÓN Y MEJORA DE ESTADISTICAS: EL CASO DEL TURISMO EN ESPAÑA

Lucía Inglada-Pérez[1], Pablo Coto-Millán[2], Pedro. Casares Hontañón[3], Vicente Inglada López de Sabando[4]

1 Departamento de Estadística e Investigación Operativa. Universidad Complutense. Correo-e: lucia.inglada.perez@ucm.es Departamento de Economía Aplicada y Estadística, UNED. Correo-e: lucia.inglada@madrid.uned.es

2 Departamento de Economía, Universidad de Cantabria, Cantabria, España. Correo-e: cotop@unican.es

3 Departamento de Economía, Universidad de Cantabria, Cantabria, España. Correo-e: casaresp@unican.es

4 Departamento de Economía Aplicada y Estadística. UNED. Correo-e: vinglada@cee.uned.es

Resumen

Durante las últimas décadas el uso acelerado de Internet ha generado un gran volumen de datos con numerosas aplicaciones en el ámbito económico y social. En esta investigación se estudian dos de estas aplicaciones para el sector turístico: (i) predicción de la actividad turística y (ii) mejora de estadísticas sobre el sector, es decir, facilitar el conocimiento de la evolución del sector antes de la publicación de los datos oficiales. La herramienta utilizada es Google Trends que incluye información desde enero de 2004 sobre las búsquedas realizadas en el motor de búsqueda más utilizado y que en nuestro caso se utiliza para medir la popularidad relativa de las palabras clave asociadas a viajar en España por los residentes. Para medir la actividad turística en España de los residentes se utilizan dos variables: número de viajeros y número de pernoctaciones. Se estiman varios modelos para predecir la evolución de cada una de las dos variables, introduciendo como variable explicativa adicional el índice del volumen de búsquedas de la palabra clave escogida. Los resultados obtenidos muestran la mayor precisión de las predicciones de los modelos estimados frente al mejor modelo univariante ARIMA y corroboran su utilidad como herramienta complementaria a las investigaciones muestrales.

Palabras clave: Turismo, Ciencia de datos, Google Trends, predicción, modelo ARIMA, estadísticas de turismo.
Área Temática: E1. Ciencia de Datos para la Economía Aplicada y Economía Cuantitativa.

INTERNET AS A TOOL FOR PREDICTION AND IMPROVEMENT OF STATISTICS: THE CASE OF TOURISM IN SPAIN

Abstract

During the last decades, the accelerated use of the Internet has generated a large volume of data with numerous applications in the economic and social sphere. This research studies two of these applications for the tourism sector: (i) prediction of tourism activity and (ii) improvement of statistics on the sector, i.e., facilitating knowledge of the evolution of the sector before the publication of official data. The tool used is Google Trends, which includes information since January 2004 on searches performed in the most widely used search engine and which in our case is used to measure the relative popularity of keywords associated with traveling in Spain by residents. Two variables are used to measure tourism activity in Spain by residents: number of guests and number of overnight stays. Several models are estimated to predict the evolution of each of the two variables, introducing as an additional explanatory variable the index of the volume of searches for the chosen keyword. The results obtained show the greater accuracy of the predictions of the estimated models compared to the best univariate ARIMA model and corroborate their usefulness as a complementary tool to surveys.

Key Words: Tourism, Data Science, Google Trends, forecasting, ARIMA model, tourism statistics.
Thematic Area: E1. Data Science for Applied Economics and Quantitative Economics.

1. INTRODUCCIÓN

Durante las últimas décadas el uso acelerado de Internet y los continuos avances en tecnología de la información y comunicaciones han generado una serie de nuevas actividades y ha cambiado la forma en que se realizan las actividades tradicionales. Asimismo, cabe destacar la emergencia de un gran volumen de datos con numerosas aplicaciones en el ámbito económico y social (Artola et al., 2015).

En esta investigación se estudian dos de estas aplicaciones para el sector turístico: (i) predicción de la actividad turística y (ii) mejora de estadísticas sobre el sector, es decir, facilitar el conocimiento de la evolución del sector antes de la publicación de los datos oficiales.

Los motores de búsqueda son las herramientas más útiles de Internet para adquirir las últimas noticias relevantes sobre un término (Yu et al., 2019). Google es el motor de búsqueda más utilizado en España. Las tendencias de Google reflejan la atención o el sentimiento del público hacia una determinada palabra clave de búsqueda. En concreto, una tendencia de Google es el volumen de búsqueda de una determinada consulta en relación con el número total de búsquedas en Google, en una escala de 0 a 100.

Desde el trabajo seminal de Choi y Varian (2009), han surgido numerosas investigaciones en muy diversos campos científicos basadas en los datos de búsqueda de Google. Por ejemplo, se ha utilizado Google Trends para predecir con éxito brotes de enfermedades (Carneiro y Mylonakis, 2009), flujos turísticos (Siliverstovs y Wochner, 2018) y el comportamiento de los mercados financieros (Preis y otros, 2013). Específicamente, en el ámbito del turismo, los investigadores han utilizado datos sobre las búsquedas en Internet y el volumen de tráfico en la red para predecir la llegada de turistas y la ocupación hotelera. Los resultados muestran la validez de los diferentes tipos de datos en línea (Pan y Yang, 2017).

La herramienta utilizada en esta investigación es Google Trends que incluye información desde enero de 2004 sobre las búsquedas realizadas en Google y que en nuestro caso se utiliza para medir la popularidad relativa de las palabras clave de búsqueda asociadas a viajar en España por los residentes.

Para medir la actividad turística en España de los residentes se utilizan dos variables: número de viajeros y número de pernoctaciones. Se estiman varios modelos para predecir la evolución de cada una de las dos variables, introduciendo como variable explicativa adicional el índice del volumen de búsquedas de la palabra clave escogida.

Los resultados obtenidos muestran la mayor precisión de las predicciones de los modelos estimados frente al mejor modelo ARIMA y corroboran su utilidad como herramienta complementaria a las investigaciones muestrales.

Con estos objetivos, la estructura de esta investigación es la siguiente. En la sección segunda se describen y analizan los datos utilizados. En la sección tercera se describe la metodología empleada. A continuación, en la sección cuarta se discuten los resultados obtenidos, y por último, en la sección quinta, se exponen las principales conclusiones.

2. ANÁLISIS DE DATOS

2.1. Variables utilizadas

Las variables utilizadas en este trabajo son las siguientes:

(i) Número de viajeros (VIA) residentes en España

(ii) (Número de pernoctaciones (PER) por residentes en España

(iii) Búsquedas en Google Trends de "Hoteles en España" (BUS)

Las tres series de datos son mensuales y abarcan desde enero de 2004 hasta abril de 2022. En particular se consideran dos periodos: enero de 2004 hasta marzo de 2014 con el fin de estudiar el comportamiento durante la recesión que se inicia en 2008 y el periodo comprendido entre enero de 2018 y abril de 2022 que cubre todo el periodo de la pandemia de COVID-19.

Los datos correspondientes a las dos primeras variables (VIA y PER) se han obtenido del INE, dentro de la Encuesta de Ocupación Hotelera. Dicha encuesta ha sustituido desde enero de 1999 a la antigua Encuesta de Movimiento de Viajeros en Establecimientos Hoteleros, ampliando la investigación a la categoría de una estrella y similares. Las unidades de análisis son todos los establecimientos hoteleros inscritos como tales en el correspondiente registro de las Consejerías de Turismo de cada Comunidad Autónoma. Son establecimientos hoteleros aquellos establecimientos que prestan servicios de alojamiento colectivo mediante precio con o sin otros servicios complementarios (hotel, hotel-apartamento o apartahotel, motel, hostal, pensión, ...).

Asimismo, los datos de la variable búsqueda (BUS) se han obtenido utilizando la herramienta Google Trends. Cebe citar que los datos de Google Trends consisten en un índice que refleja el número de búsquedas que se han realizado de términos específicos, en relación con el número máximo de búsquedas para el mismo término de Google en el período de tiempo que abarca. Google Trends elimina las búsquedas repetidas de un mismo usuario en un período corto de tiempo, por lo que el nivel de interés no se ve impactado artificialmente por estas búsquedas.

Existen una serie de términos de Google Trends relacionados con el comportamiento de las variables VIA y PER. Inicialmente se escogieron los siguientes términos: "hoteles en España"; "hotel España"; "hotel en España"; "alojamiento España". Se considera la categoría de Viajes para evitar búsquedas relacionadas con

otra actividad. Finalmente, se ha seleccionado el término "hotel España" después de analizar la correlación de los diferentes términos con las variables de estudio.

2.2. Análisis de datos

Al analizar la evolución de las series BUS y VIA en 2004-2022 que se muestran en la figura 1, se observan los siguientes rasgos:

- La evolución de las series es similar, aunque se observa una mayor variabilidad en la serie VIA.
- Ambas series son estacionales, con patrón estacional algo menos marcado en la serie BUS.
- Durante el periodo de recesión se observa un patrón similar en ambas series.
- Cabe señalar la tremenda caída en los valores de ambas series producida por la pandemia de COVID. El descenso es especialmente pronunciado en el número de viajeros que toma un valor cero en el periodo de confinamiento.
-

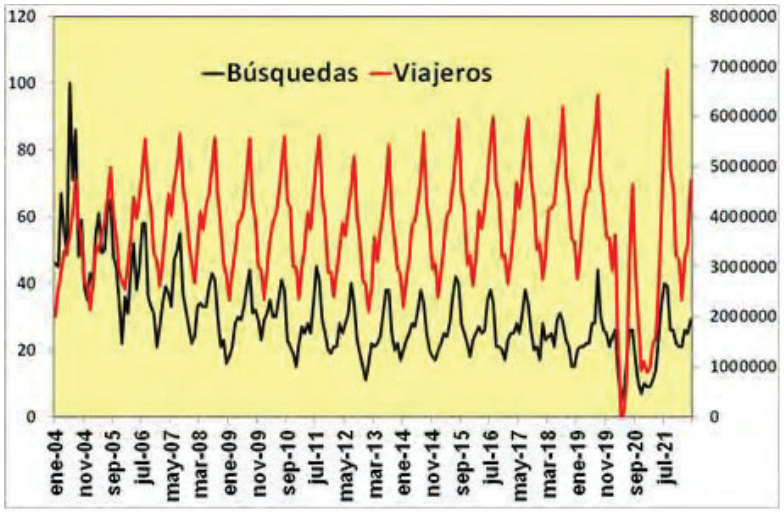

Figura 1. Evolución de las series BUS y VIA.

3. METODOLOGÍA

3.1. Modelos utilizados

El proceso metodológico adoptado en este trabajo consiste básicamente en definir un modelo inicial explicativo del comportamiento de cada una de las dos variables VIA y PER, objeto de nuestro interés, para posteriormente proceder a la estimación de dicho modelo en las dos alternativas correspondientes, respectivamente, a la

consideración o no de la variable búsquedas BUS como variable explicativa independiente. El análisis comparativo de los resultados de ambos modelos alternativos, con base en una serie de criterios, nos servirá para contrastar estadísticamente la relevancia de la variable BUS en la modelización y predicción de las dos variables, número de viajeros y de pernoctaciones, que constituye el objetivo principal de este trabajo.

Para el modelo inicial explicativo del comportamiento de las variables VIA y PER, en el que se incorporará BUS como variable explicativa, se ha elegido el modelo ARIMA. Las variables se expresan en forma logarítmica. Una ventaja de utilizar esta transformación es que los efectos estimados de las variables explicativas sobre la variable endógena están expresados en forma de tasa de variación y además se contribuye a la reducción de la heterocedasticidad en las estimaciones.

A su vez se estimarán los modelos en dos escenarios temporales diferentes, que abarcan respectivamente a la crisis económica de 2008 y a la pandemia del COVID-19, iniciada en marzo de 2020.

3.2. Modelos ARIMA

Un modelo ARIMA (p, d, q) surge a partir de un modelo ARMA (p, q) definido sobre una serie que exige *d* integraciones. Por lo tanto, se trata de un modelo autorregresivo de orden *p*, integrado *d* veces y de medias móviles de orden *q*. Todo ello se aplica de forma análoga al caso de existir estacionalidad, generándose el correspondiente modelo ARIMA(p,d,q)x(P,D,Q)$_{12}$ para el caso de estacionalidad mensual. En general, hay que modelizar la parte regular ARIMA(p,d,q) y también la parte estacional: ARIMA(P,D,Q)$_s$. El modelo conjunto se denomina ARIMA(p,d,q)x(P,D,Q)$_s$.

Esta nomenclatura nos indica lo siguiente:

- La serie ha necesitado *d* diferencias regulares para ser estacionaria.
- La serie ha necesitado *D* diferencias de orden *s* para ser estacionaria
- Por tanto, la serie $\nabla^d \nabla_s^D y_t$ es estacionaria.
- La parte regular de la serie estacionaria $\nabla^d \nabla_s^D y_t$ sigue un modelo ARMA(p,q)
- La parte estacional de la serie estacionaria $\nabla^d \nabla_s^D y_t$ sigue un modelo estacional ARMA(P,Q)$_s$

Por ejemplo, la forma de formular que una variable y_t sigue una estructura ARMA(1,1) X ARMA(0,1)$_{12}$ es la siguiente: $(1-\phi_1 B)y_t = (1-\theta_1 B)(1-\Theta_1 B^{12})\varepsilon_t$.

En el caso de estos modelos autorregresivos integrados de media móvil (ARIMA), introducidos por Box y Jenkins (1970), para evitar problemas como estacionalidad, existencia de varias tendencias o variabilidad proporcional a la media, se realizan

las siguientes transformaciones y operaciones: a) Utilización de los logaritmos de los valores de la serie primitiva con el fin de reducir la posible heterocedasticidad; b) Aplicación de una diferencia regular y otra estacional con el objetivo de eliminar, respectivamente, las tendencias en media y la estacionalidad. Consecuentemente, para cada serie de viajeros y pernoctaciones se procede a analizar: $\nabla\nabla_{12}\mathrm{Log}(y_t)$, donde y_t es respectivamente VIA_t o PER_t. En definitiva, una vez obtenida la estacionariedad en la serie mediante la aplicación de la transformación logarítmica y de los operadores diferencia regular y estacional, respectivamente, estamos en condiciones de identificar y estimar el modelo ARIMA final para cada una de las dos series. El mejor modelo ARIMA se selecciona de acuerdo con el criterio de Akaike.

3.3. Modelos ARIMAX

El modelo ARIMAX se obtiene mediante la incorporación de la variable BUS como variable explicativa en el modelo ARIMA anterior. Para calcular el retardo de esta variable, es decir el periodo medio que transcurre entre la búsqueda y el viaje, se han utilizado los correspondientes coeficientes de correlación entre esta variable y las variables VIA y PER. Si la variable BUS fuese significativa en la estimación, se concluiría que la aportación de los datos de Google Trends mejora el modelo y nos facilita la predicción del número de viajeros y pernoctaciones.

Además, con base en una serie de criterios, se comparan los resultados obtenidos en la estimación del modelo anterior con los correspondientes al modelo alternativo que no incluye la variable BUS como explicativa. Si el modelo que incluye a BUS se comportara de forma más satisfactoria, se corroboraría la hipótesis de partida consistente en la potencia predictiva de esta variable en relación con el número de viajeros y pernoctaciones en los hoteles españoles por residentes. Los criterios utilizados son: R^2 Ajustado, Criterio de Información de Akaike, Error estándar de la regresión, Suma de los cuadrados de residuos y Error absoluto medio (MAE).

Como se ha indicado anteriormente, se selecciona el término "hotel España" dentro de la categoría viajes a partir de cuatro términos basándonos en la correlación con las variables y el interés del término.

4. RESULTADOS

4.1. Periodo de recesión

En primer lugar se selecciona el modelo ARIMA que se va a ajustar: De acuerdo con el criterio de información de Akaike el modelo ARIMA seleccionado es el $(2,1,0)x(0,1,1)_{12}$. A continuación, se ha realizado un análisis comparativo mediante varios criterios de los modelos de viajeros (i) ARIMA y (ii) ARIMAX con variable BUS, durante el periodo que abarca desde enero de 2004 hasta marzo de 2014, es

decir un periodo que incluye la recesión acaecida en la economía española. Los criterios utilizados son: R^2 Ajustado, criterio de Información de Akaike, Error estándar de la regresión, suma de los cuadrados de residuos y error absoluto medio (MAE). Los principales resultados obtenidos en este periodo son los siguientes:

- El valor del estadístico R^2 ajustado mejora al incluir BUS como variable explicativa, de 0,719 a 0,722.
- El criterio de Akaike muestra que el modelo mejora con la inclusión de BUS ya que disminuye de -3,32 a -3,34.
- Otros estadísticos como el error estándar de regresión y la suma de los cuadrados de residuos también disminuyen con la inclusión de BUS.
- Finalmente, el error absoluto medio, que mide la capacidad predictiva del modelo, mejora con la inclusión de BUS ya que disminuye de 0,053 a 0,051.

4.2. Periodo que incluye la pandemia de COVID-19

En primer lugar se selecciona el modelo ARIMA que se va a ajustar: De acuerdo con el criterio de información de Akaike el modelo ARIMA seleccionado es el $ARIMA(0,1,1)(1,1,0)_{12}$. A continuación, se ha realizado un análisis comparativo mediante los criterios citados anteriormente de los modelos de viajeros (i) ARIMA y (ii) ARIMAX con variable BUS, durante el periodo que abarca desde enero de 2017 hasta abril de 2022, es decir un periodo que incluye la pandemia de COVID-19.

Los principales resultados obtenidos en este periodo son los siguientes:

- El valor del estadístico R^2 Ajustado mejora significativamente al incluir BUS como variable explicativa, de 0,573 a 0,751.
- El criterio de Akaike muestra que el modelo mejora con inclusión de BUS ya que disminuye de 0,842 a 0,330.
- Otros estadísticos como el error estándar de regresión y la suma de los cuadrados de residuos también disminuyen con la inclusión de BUS.
- El error absoluto medio, que mide la capacidad predictiva del modelo, mejora con la inclusión de BUS ya que disminuye de 0,281 a 0,244.

5. CONCLUSIONES

Los resultados obtenidos en esta investigación muestran que la información de Google Trends contribuye a mejorar la precisión de la predicción del número de viajeros y de pernoctaciones en los hoteles españoles por los residentes. Este comportamiento se ha contrastado en periodos de turbulencias: crisis económica y pandemia de COVID-19.

Asimismo, por su disponibilidad inmediata cabe considerar a los datos de búsquedas como un indicador adelantado y en tiempo real del número de viajeros y de pernoctaciones en los hoteles españoles. Todo ello resalta su interés para la producción de estadísticas oficiales en el sector turístico que se difunden con retraso temporal.

Los resultados obtenidos sobre el poder predictivo del indicador de búsquedas utilizado son particularmente relevantes para la industria turística, tan importante para la economía española, en relación con la gestión de las actividades turísticas y particularmente las hoteleras.

Entre las limitaciones de este estudio cabe indicar que Google cuantifica como una búsqueda, aunque el propósito del usuario sea buscar hotel también para otras personas. Otro problema es que Google Trends no suministra el número total de búsquedas sino un valor relativo en el intervalo de 0 a 100.

Hay varias direcciones para seguir investigando. Las investigaciones futuras deberían centrarse en la aplicación de la metodología empleada en esta investigación desagregando las búsquedas por Comunidad Autónoma o considerando otras variables como el número de turistas en España. Por último, también sería interesante perfeccionar los resultados considerando la existencia de cointegración entre las variables BUS con VIA y PER.

REFERENCIAS

ARTOLA, C.; FERNANDO PINTO, F; DE PEDRAZA GARCÍA, P. (2015): Can internet searches forecast tourism inflows? *International Journal of Manpower*, 36 (1). 103–116.

BOX G.E.P.; JENKINS G.M. (1970): *Time series analysis: forecasting and control*, Holden-Day, San Francisco.

CARNEIRO, H.A.; MYLONAKIS, E. (2009): Google Trends: A Web-Based Tool for Real-Time Surveillance of Disease Outbreaks. *Clinical Infectious Diseases*, 49, 1557-1564.

CHOI, H.; VARIAN, H. (2009): Predicting the Present with Google Trends, Technical report, Google. Available from: http://google.com/googleblogs/pdfs/google_predicting_the_present.pdf.

PREIS, T.; MOAT, H.; STANLEY, H. (2013): Quantifying Trading Behavior in Financial Markets Using Google Trends. *Sci Rep* 3, 1684 (2013). https://doi.org/10.1038/srep01684

SILIVERSTOVS, B.; WOCHNER, D.S. (2018): Google Trends and reality: Do the proportions match? *Journal of Economic Behavior & Organization*, 145, 1-23.

YU, L; ZHAO, Y; TANG, L.; YANG, Z. (2019): Online big data-driven oil consumption forecasting with Google trends. *International Journal of Forecasting*, 35 (1), 213-223.