UNIVERSITAT POLITÈCNICA DE VALÈNCIA

UNIVERSIDAD POLITÉCNICA DE VALENCIA

FACULTAD DE ADMINISTRACIÓN Y DIRECCIÓN DE EMPRESAS

TRABAJO FIN DE CARRERA

Licenciatura en Administración y Dirección de Empresas

"USO DE GOOGLE TRENDS PARA PREDECIR EL NIVEL Y LA ESTRUCTURA DEL DESEMPLEO EN ESPAÑA"

PRESENTADO POR:

Jorge Redondo Caballero

DIRIGIDO POR:

Josep Domènech i de Soria

Valencia, Julio de 2013

ÍNDICE GENERAL

ÍNDICE DE FIGURAS Y GRÁFICOS

ÍNDICE DE TABLAS

CAPÍTULO 1

INTRODUCCIÓN

CAPÍTULO 1. INTRODUCCIÓN

1.1. Resumen

Google Trends es una herramienta online elaborada por Google que proporciona la evolución de la popularidad de las búsquedas. El uso de dichos datos para realizar predicciones a corto plazo, sobre una gran variedad de series económicas y sociales, se ha visto incrementado de manera sustancial en los últimos años. Las investigaciones en esta materia han demostrado satisfactoriamente que los datos procedentes de Google Trends mejoran las predicciones sobre el nivel de desempleo en países donde el uso de Internet está ampliamente extendido entre la sociedad (por ejemplo, EE.UU y Alemania). (D'Amuri y Marcucci, 2009; Askitas y Zimmermann, 2009)

En este contexto, el presente Trabajo Final de Carrera (TFC) valida el uso de Google Trends para predecir el nivel agregado del desempleo en España, un país donde el porcentaje de hogares con acceso a Internet se encuentra por debajo de la media europea (Eurostat, 2012). Además, se lleva a cabo un análisis de la estructura del desempleo, examinando el tiempo que una persona lleva desempleada. De este modo, se pretende investigar si existe un determinado tipo de términos de búsqueda que ayuden a predecir mejor dichas partes de manera separada. Por ejemplo, una persona que acaba de ser despedida es más propensa a buscar en Internet información sobre prestaciones económicas que una persona que lleva desempleada más de un año. Así pues, el carácter innovador de este Trabajo Final de Carrera yace en la obtención de una serie de términos de búsqueda que ayuden a monitorizar de manera más eficiente las distintas partes que componen el desempleo en España.

La metodología de este TFC se basa en una gran variedad de artículos científicos que tratan sobre el uso de modelos de *nowcast* y, más concretamente, sobre el uso de Google Trends. Los datos numéricos han sido obtenidos a través del Instituto Nacional de Estadística y del Servicio Público de Empleo Estatal, así como de la propia herramienta online Google Trends.

Como se podrá apreciar a lo largo de este trabajo, el término de búsqueda *Paro* presenta una similitud notable con la evolución del nivel de desempleo en España. Asimismo, el término de búsqueda *Cobrar Paro* se encuentra muy ligado al número de personas que han trabajado con anterioridad y llevan desempleadas menos de 3 meses.

1.2. Introducción

Internet se ha convertido en una segunda realidad para la humanidad, en la cual podemos ver reflejados nuestros deseos y nuestras preocupaciones. Cada día, la actividad *online* genera una innumerable cantidad de información, dejando una huella digital a su paso. Aunque tal cantidad de información pueda parecer *a priori* abrumadora, existen herramientas que permiten a los investigadores analizar y comprender nuestro comportamiento desde un punto de vista tanto económico como social.

Los indicadores económicos basados en la información *online* poseen numerosas ventajas con respecto a los indicadores tradicionales. A saber, la información requerida para obtenerlos no presenta costo alguno y, por tanto, se pueden llevar a cabo en el intervalo de tiempo deseado. Ello permite un análisis en tiempo real y un margen de maniobra más amplio en materia intervencionista.

En este contexto, existe una literatura creciente que trata sobre la extracción sistemática de indicadores de la economía a través de Internet. Una de las maneras más popularizadas ha sido el uso de Google Trends; una herramienta online elaborada por Google que proporciona la evolución de la popularidad de las búsquedas. Dicha herramienta ha concedido a los investigadores la oportunidad de generar indicadores en tiempo real para una gran variedad de situaciones, como puede ser la evolución de la gripe (Ginsberg et al., 2009), predecir las ventas al por menor (Choi y Varian, 2009) o predecir el volumen de transacciones en la bolsa (Preis et al., 2010).

Con respecto a la predicción del nivel de desempleo, investigaciones previas (D'Amuri y Marcucci, 2009; Askitas y Zimmermann, 2009) han demostrado la pertinencia del uso de Google Trends para mejorar la predicción del nivel de desempleo en países donde el uso de Internet está ampliamente extendido entre la sociedad, como EE.UU y Alemania. (Corrocher y Ordanini, 2002)

Asimismo, esta investigación ha dado como resultado la elaboración de un artículo científico que ha sido aprobado por la Asociación Internacional de Economía Aplicada (ASEPELT). (Anexo 1)

En cuanto a la estructura organizativa, el presente Trabajo Final de Carrera se encuentra estructurado en 4 bloques principalmente, excluyendo la Introducción. El Capítulo 2 versa sobre el marco teórico que concierne a los modelos de *nowcast* y, más concretamente, a los modelos que emplean datos procedentes de Google Trends. El Capítulo 3 trata sobre la metodología empleada, analizando las variables, la muestra y los modelos econométricos propuestos. El Capítulo 4 muestra las estimaciones de dichos modelos, así como su validación econométrica y sus aplicaciones económicas. Finalmente, el Capítulo 5 plantea las conclusiones pertinentes y expone las limitaciones de este TFC.

1.3. Objeto del TFC y justificación de las asignaturas relacionadas

El objeto de este TFC consiste en el uso de Google Trends para predecir el nivel y la estructura del desempleo en España.

En cuanto a la justificación de las asignaturas relacionadas, para la redacción del Capítulo 2 han sido de ayuda los conocimientos sobre el desempleo que se imparten en las asignaturas de Microeconomía y Macroeconomía. Asimismo, los contenidos de la asignatura de Economía Española y Mundial son útiles para entender los mercados internacionales en los que están basados diversos estudios.

Con respecto al Capítulo 3, la asignatura de Introducción a la Informática facilita el manejo de bases de datos y hojas de cálculo. Por otra parte, las asignaturas de Introducción a la Estadística y Métodos Estadísticos proporcionan una visión descriptiva y aplicada de los datos. Finalmente, la asignatura de Econometría ofrece los conocimientos necesarios para el planteamiento de los modelos econométricos.

En lo referente al Capítulo 4, los conocimientos estudiados en las asignaturas de Introducción a la Estadística y Métodos Estadísticos ayudan a entender las estimaciones obtenidas en los modelos. Por otro lado, la validación de los modelos requiere de los conocimientos aprendidos en la asignatura de Econometría y, la interpretación económica de éstos, de la asignatura de Macroeconomía.

1.4. Objetivos del TFC

El objetivo principal de este TFC es validar el uso de Google Trends para predecir el nivel de desempleo en España. Asimismo, se pretende predecir la estructura del desempleo, ligando ciertos términos de búsqueda en Google con los diferentes estratos que conforman el nivel total de desempleo.

De este modo, se pueden enumeran los siguientes objetivos:

1. Validar el uso de Google Trends para predecir el nivel de desempleo en España.

2. Justificar que los datos obtenidos a través de Google Trends mejoran la predicción del nivel de desempleo en España basada en los datos procedentes del Servicio Público de Empleo Estatal (SEPE).

3. Averiguar qué tipo de términos de búsquedas están más ligados a la evolución del desempleo en España.

4. Predecir la estructura del desempleo.

CAPÍTULO 2

MARCO TEÓRICO

CAPÍTULO 2. MARCO TEÓRICO: SISTEMAS DE *NOWCAST* Y GOOGLE TRENDS

2.1. Introducción

El presente capítulo versa sobre los modelos de *nowcast* y, más concretamente, los modelos econométricos que emplean los datos facilitados por Google Trends. Se encuentra estructurado en tres secciones. En primer lugar, se explica el significado y funcionamiento de los modelos de *nowcast*, así como la literatura más importante en materia económica. Seguidamente, se presenta Google Trends, explicando su funcionamiento y su historia. Finalmente, se exponen las investigaciones más relevantes que emplean datos obtenidos a través de Google Trends.

2.2. Sistemas de *nowcast*

En este apartado se lleva a cabo una introducción a las técnicas del *nowcasting*, así como un análisis de los estudios más representativos.

2.2.1. ¿Qué es el *nowcasting*?

El término *nowcasting* proviene de las palabras inglesas "now" (ahora) y "forecasting" (predecir el futuro) y hace referencia a una serie de técnicas empleadas para realizar predicciones del presente, del pasado reciente y del futuro cercano. Los sistemas de *nowcast* son utilizados principalmente en ámbitos económicos y meteorológicos, siendo únicamente el primero de ellos de interés para este análisis. (Giannone, Reichlin y Small, 2008)

De acuerdo con Bańbura et al. (2012), el término *nowcasting* se puede definir como el ejercicio de observar, desde la perspectiva de un modelo, las publicaciones de corrientes de información en tiempo real.

Los modelos de *nowcast* han adquirido gran relevancia en el ámbito económico debido a que la mayoría de indicadores del estado actual de la economía son publicados con cierto retraso. Además, también resulta de gran utilidad para otro tipo de variables que

revelan el estado actual de la economía. El principio básico del *nowcasting* es la explotación de información publicada con anterioridad con el objetivo de obtener una estimación antes de que la variable en cuestión sea oficialmente publicada. A modo de ejemplo, si se pretende realizar una estimación del PIB de una nación determinada, se debe llevar a cabo un análisis de las variables que lo componen. Además, se pueden consultar encuestas o variables financieras. (Andreou, 2008)

La idea, por lo tanto, es que tanto la información "hard" (dura), como son los informes de producción industrial, y la información "soft" (blanda), como son las encuestas, pueden proporcionar una estimación previa de los acontecimientos actuales de la economía. (Bańbura et al. 2012)

2.2.2. Aplicaciones de los modelos de *nowcast*

A continuación se exponen las aplicaciones más relevantes, en materia económica, que ha suscitado el uso de los modelos de *nowcast,* como son: la predicción del PIB, el nivel de inflación y el desempleo.

2.2.2.1. Modelos de Nowcast para la predicción del PIB

El uso de sistemas de *nowcast* para predecir el nivel del PIB ha sido explotado por diversos autores a lo largo de los años. En el caso de los Estados Unidos (Giannone, Reichlin y Small, 2008) por parte de la Reserva Federal y, de forma independiente (Lahiri y Monokroussos, 2011; Aastveit et al., 2011). Para el área económica europea en su conjunto (Angelini, Bańbura y Rünstler, 2010; Angelini, Camba-Méndez, Giannone, Reichlin y Rünstler, 2011; Bańbura y Modugno, 2010; Bańbura y Rünstler, 2011; Camacho y Pérez-Quirós, 2010). Centrándose en concreto en un país determinado se pueden citar las investigaciones de (Barhoumi, Darn y Ferrara, 2010) para Francia, (Liebermann, 2012b) para Irlanda, (de Winter, 2011) para Holanda y (Aastveit y Trovik, 2012) para Noruega. Para el caso de España, (Antonio-Liedo y Fernandez, 2010).

Los resultados obtenidos en los estudios previos han proporcionado las siguientes conclusiones al respecto. En primer lugar, las ganancias institucionales de dichas predicciones son solamente sustanciales en un corto periodo de tiempo. Ello implica que

la predicción del nivel de PIB se ve afectada, principalmente, por el trimestre actual (y el anterior). En segundo lugar, los procedimientos estadísticos automáticos son equiparables a las predicciones institucionales, las cuales son el resultado de un proceso de modelos y valoraciones. Estos resultados sugieren que el *nowcasting* tendrá un papel importante en la literatura. En tercer lugar, los modelos de *nowcast* mejoran sus predicciones conforme la información relevante se acumula, lo cual sugiere la necesidad de incorporar nuevos datos al modelo tan pronto como sean publicados. (Bańbura, Giannone y Reichlin, 2011)

Los datos empleados para realizar las predicciones convergen en una serie de categorías, a saber:

- Información dura: prestaciones de desempleo, nivel de desempleo, venta de coches, mercado inmobiliario, producción industrial y ventas al por menor.
- Información blanda: indicadores de la confianza de los consumidores y empresarios.
- Información financiera: tipos de cambio, agregados monetarios y crédito del sector privado.

2.2.2.2. Modelos de Nowcast para la predicción del nivel de inflación

La literatura relacionada con los modelos de *nowcast* para la predicción del nivel de inflación no es tan abundante como para la predicción del nivel de PIB, debido a que se trata de un ámbito relativamente nuevo. Aún así, se pueden destacar los estudios propuestos por (Lenza y Qarmedinger; 2010; Monteforte y Moretti, 2010; Modugno, 2011) con respecto al Área Económica Europea.

2.2.2.3. Modelos de Nowcast para la predicción del nivel de desempleo

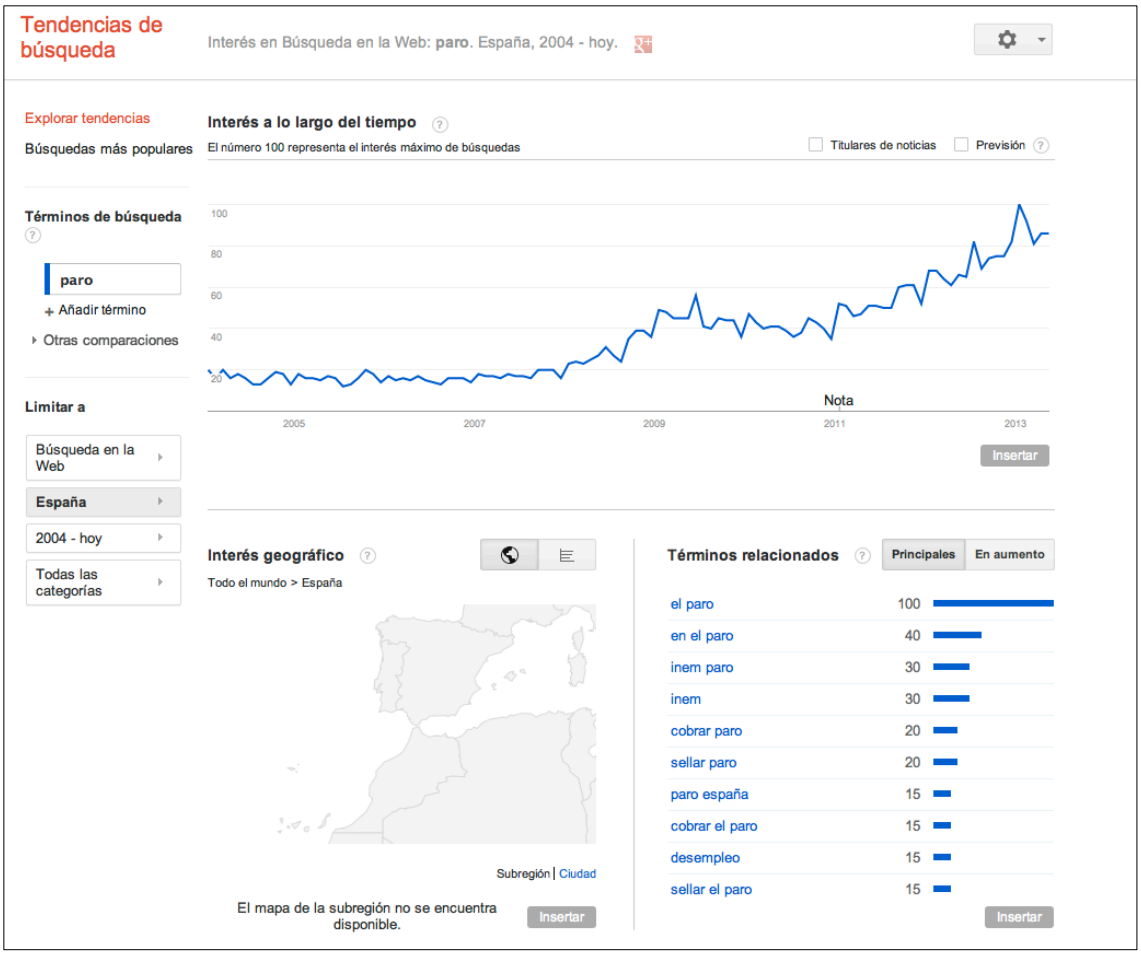
Ettredge et al. (2005) fueron los pioneros en proponer un modelo de *nowcast* que ayudara a predecir el número de personas que acababa de ser desempleada, basándose en los historiales de búsqueda de diversos buscadores. Dicho enfoque fue el que seguirían varios autores una vez que Google decidió hacer pública la popularidad de los términos de búsquedas para su pertinente análisis. (Véase epígrafe 2.3.2)

2.3. Google Trends

2.3.1. ¿Qué es Google Trends?

Google Trends proporciona series temporales relacionando el volumen de búsquedas de una determinada consulta en un área determinada. Los datos con los que se trabaja se encuentran estandarizados mediante un índice de 0 a 100 (denominado Google Index), en lugar de trabajar con los datos absolutos. La frecuencia de dichas series temporales es semanal, y tiene su punto de origen en 2004. La Figura 1 muestra el funcionamiento de Google Trends para el término de búsqueda *Paro* en España, desde 2004 hasta mayo de 2013.

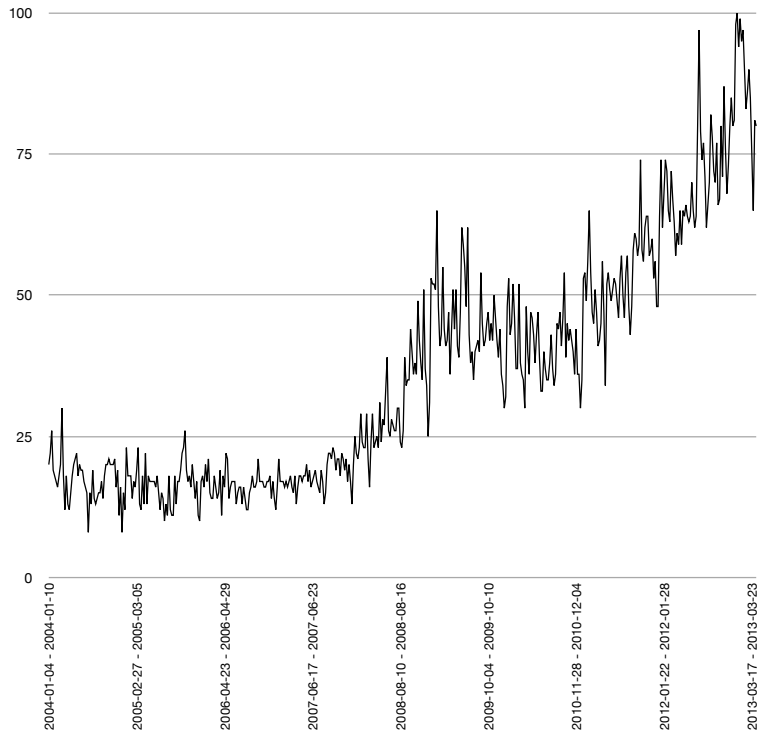
Figura 1 Captura de pantalla de Google Trends



Fuente: Elaboración propia

Los usuarios pueden especificar la palabra cuya popularidad deseen conocer o bien pueden examinar dicha popularidad a través de categorías que Google previamente ha implementado. Por ejemplo, la categoría "viajes" incluye todo tipo de consultas referidas a viajar. Asimismo, se puede seleccionar de forma manual el ámbito geográfico deseado, por lo que se puede limitar el análisis a 1 o más países, o regiones. A modo de ejemplo, el Gráfico 1 refleja la evolución de la popularidad del término *Paro* en España, desde enero de 2004 hasta marzo de 2013.

Gráfico 1. Evolución de la popularidad del término de búsqueda *Paro* en España (Google Index)



Fuente: Elaboración propia a partir de los datos de Google Trends

2.3.2. Historia de Google Trends

El 10 de mayo de 2006 fue publicada la primera versión de Google Trends. Desde entonces, se ha visto inmersa en una serie de cambios estructurales, de entre los que cabe mencionar la publicación de una herramienta online adjunta, denominada Google Insight, que permitía la exportación de los datos en formato CSV. Dicha herramienta supuso un punto de inflexión en el análisis estadístico de los datos y, en 2012, pasó a formar parte de Google Trends.

2.3.3. Aplicaciones de Google Trends

En este apartado se muestran los estudios más influyentes basados en el uso de Google Trends, tanto en materia económica como en el ámbito general.

2.3.3.1. Ámbito general

La literatura referente a los volúmenes de búsquedas ha sido relativamente abundante, teniendo en cuenta que los datos estadísticos sólo han estado disponibles desde 2008. El primer estudio en utilizar Google Trends fue Ginsberg et al.(2009), en el campo de la medicina. Descubrieron que las búsquedas de 45 términos asociados con la gripe permiten predecir brotes de gripe 2 semanas antes que los informes del CDC (Centro para el Control y Prevención de Enfermedades). Se suele recurrir a este estudio en la literatura para verificar la naturaleza predictiva de Google Trends y justificar su uso en la investigación académica.

Choi y Varian (2009) demostraron que los datos procedentes de Google Trends pueden ser utilizados para predecir las ventas de inmuebles, de automóviles y el turismo. Para este último caso, se partió de la suposición de que Google es usado para planear viajes y, por lo tanto, un incremento en un destino determinado supondría un incremento en la cantidad de turistas. El estudio se centra en el keyword (término de búsqueda) "Hong Kong" desde 9 países diferentes, y lo compara con los informes del Departamento de Turismo de Hong Kong. Los autores llegaron a la conclusión de que existe una alta correlación entre los dos parámetros (con la exclusión de Japón).

Goel et al. (2010), por otra parte, examinaron cómo se pueden utilizar las búsquedas en Google para predecir los ingresos en taquilla de la industria del cine, las ventas de videojuegos en su primer mes y el ranking de canciones en la lista Billboard Hot 100. Concluyeron que Google Trends ofrece la posibilidad de predecir los días de más ventas con bastante antelación. La fiabilidad de la predicción varía dependiendo del tema en cuestión, siendo más precisa con las películas y menos con las canciones.

2.3.3.2. Ámbito económico

Uno de los ámbitos en los que el uso de Google Trends ha sido más receptivo es el ámbito económico. McLaren y Shanbhogue (2011) analizaron el mercado de trabajo y viviendas de Reino Unido, mientras que Gill et al. (2012) se centraron en una serie de aspectos de la economía Australiana, ambos utilizando información *online* proporcionada por Google Trends.

Guzmán (2011) demostró que Google Trends puede ser utilizado para predecir el nivel de inflación. Para ello, comparó los datos obtenidos con 36 encuestas y llegó a la conclusión de que los datos obtenidos con Google Trends presentaban el menor error de predicción de entre todos los indicadores ya validados.

Da, Engelberg y Gao (2010), analizaron la situación económica basándose en un grupo de términos de búsqueda negativos, a saber: recesión, bancarrota, etc. El estudio demuestra que un incremento del conjunto de *keywords* conlleva una volatilidad extrema en el flujo de los fondos de inversión. De una manera similar, Dzielinksi (2012) utiliza el *keyword* "economic" para analizar la incertidumbre económica. Su hipótesis se basa en que un nivel alto de incertidumbre económica incrementa la demanda de información, lo cual sería rastreable en Internet a través del término de búsqueda "economic".

2.3.3.3. Ámbito del desempleo

El uso de Google Trends para predecir el nivel de desempleo ha sido un tema recurrente en la literatura desde su publicación, especialmente aplicado a países donde el uso de internet está completamente extendido. Una serie moderada de investigaciones obtuvieron datos favorables para países como Alemania (Askitas y Zimmermann, 2009), los Estados Unidos (D'Amuri y Marcucci, 2009) y el Reino Unido (McLaren y Shanbhogue, 2011). De forma similar, diversos investigadores han demostrado el poder predictivo de Google Trends para países que no se encuentran en la vanguardia tecnológica, como son: Israel (Suhoy, 2009) y Turquía (Chadwick et al., 2012).

Se ha empleado una gran variedad de términos y categorías con el objetivo de encontrar aquellos que mejor predigan el nivel de desempleo. Por un lado, Choi y Varian (2009) utilizaron las categorías "Jobs" (trabajos) "Welfare" (bienestar) para su estudio sobre prestaciones económicas en los Estados Unidos. Paralelamente, D'Amuri (2009) empleó la categoría "Job Offers" (ofertas de trabajo), en la cual quedan recogidas todas las búsquedas relativas a ofertas de empleo. Por otra parte, D'Amuri y Marcucci (2009) utilizaron el término "Jobs" (trabajos), al igual que McLaren y Sanbhongue (2011), que utilizaron los términos "Jobs" y "Unemployment" (desempleo).

Un enfoque diferente fue el propuesto por Askitas y Zimmermann (2009), los cuales propusieron cuatro grupos de términos con el objetivo de predecir el nivel agregado de desempleo en Alemania. La naturaleza innovadora de este estudio se basa en que cada grupo de términos está orientado a los flujos de entrada y salida de demandantes de empleo. Sin embargo, el poder predictivo de este grupo de términos sólo está validado contra el nivel agregado de desempleo, y no con cada uno de los distintos segmentos a los que hacían referencia.

El presente trabajo tiene como objetivo confirmar que diferentes términos de búsquedas pueden estar relacionados con diferentes partes de la estructura del desempleo, permitiendo una formulación de modelos más eficiente.

Así pues, de acuerdo con la revisión bibliográfica realizada, se exponen las siguientes hipótesis:

- H_1: La popularidad del término de búsqueda *Paro* ayuda a predecir el número total de desempleados en España.

- H_2: La popularidad del término de búsqueda *Trabajo* ayuda a predecir el número total de desempleados en España.

- H_3: La popularidad del término de búsqueda *Paro* mejora la predicción del número total de desempleados en España basada en los datos del desempleo disponibles.

- H_4: La popularidad del término de búsqueda *Cobrar Paro* está relacionada con aquellas personas que han trabajado anteriormente y llevan desempleadas poco tiempo.

- H_5: La popularidad del término de búsqueda *Cobrar Paro* no está relacionada con aquellas personas que han trabajado anteriormente y llevan desempleadas mucho tiempo.

2.3.4 Tabla resumen de la bibliografía relacionada con el desempleo

La Tabla 1 muestra, a modo de resumen, la bibliografía relacionada con el desempleo que ha sido expuesta en el epígrafe 2.3.3.3.

Tabla 1 Resumen de la bibliografía relacionada con la predicción del desempleo

Investigadores	Nombre del estudio	Revista	Año	País	Frecuencia	Término/Categoría
Askitas y Zimmermann	Google econometrics and unemployment forecasting	Applied Economics Quarterly	2009	Alemania	Mensual	Términos: "unemployment office", "unemployment rate"y "personal consultant"
Choi y Varian	Predicting the Present with Google Trends	Economic Record	2009	EEUU	Semanal	Categorías: "jobs" y "welfare"
	Predicting initial claims for unemployment insurance using Google Trends	-	2009	EEUU	Semanal	Categorías: "jobs" y "welfare"
D'Amuri	Predicting unemployment in short samples with internet job search query data	MPRA	2009	Italia	Cuatrimestral	Categoría: "job offers"
D'Amuri y Marcucci	Google it! Forecasting the US unemployment rate with a google job search index	FEEM, trabajo en curso	2009	EEUU	Mensual	Término: "jobs"
McLaren y Sanbhongue	Using internet search data as economic indicators	Bank of England Quarterly Bulletin	2011	Reino Unido	Mensual	Términos: "jobs", "unemployment" y "jobseeker allowance"

Fuente: Elaboración propia

CAPÍTULO 3

METODOLOGÍA

CAPÍTULO 3. METODOLOGÍA

3.1. Variables

En este apartado se presentan las variables que han sido utilizadas para la realización del presente TFC. En primer lugar, se expone la Encuesta de Población Activa (EPA), explicando su funcionamiento y delimitando su margen de acción. Seguidamente, se efectúa el mismo procedimiento para los datos del desempleo ofrecidos por el Servicio Público de Empleo Estatal (SEPE). Finalmente, se explica en más detalle la herramienta *online* Google Trends.

3.1.1. EPA

A continuación, se explica el funcionamiento e historia de la EPA, así como sus objetivos, ventajas, ámbito geográfico, etc. La información ha sido obtenida del Informe Técnico publicado por el INE en 2009, con referencia a la Encuesta de Población Activa.[1]

Historia

La primera publicación de la Encuesta de Población Activa data del 1964. En 1987 se modificó el cuestionario empleado, adaptándose a los estándares de la Encuesta de Fuerza de Trabajo de la Comunidad Económica Europea del 1986. De este modo, se recalcularon series retrospectivas de acuerdo a la nueva metodología desde el tercer trimestre de 1976.

En 1999, la periodicidad de las entrevistas pasa a ser de 13 semanas, en lugar de 12 como se venía haciendo anteriormente. Finalmente, en 2005 se llevó a cabo el último cambio metodológico, implantando un nuevo cuestionario y control centralizado del sistema de recogida mediante encuesta telefónica asistida por ordenador. Las cifras actuales de la encuesta se basan en la metodología introducida en 2005.

[1] http://www.ine.es/docutrab/epa05_disenc/epa05_disenc.pdf

Definición

La EPA es una investigación por muestreo de periodicidad trimestral, dirigida a la población que reside en viviendas familiares del territorio español y cuya finalidad es averiguar las características de dicha población en relación con el mercado de trabajo.

Los entrevistadores del INE (Instituto Nacional de Estadística) se ponen en contacto, personal o telefónicamente, con las viviendas seleccionadas para formar parte de la muestra y recogen la información de las personas que residen en ellas.

Está considerada como el mejor indicador del empleo y desempleo en España.

Objetivos

El objetivo principal de la EPA es conocer la actividad económica desde un punto de vista social. Los datos obtenidos representan categorías poblacionales relacionadas con el mercado de trabajo (ocupados, parados, activos e inactivos) y clasificaciones de éstas basadas en diversas características. Desde un punto de vista internacional, los datos obtenidos son comparables con el resto de países de la Comunidad Económica Europea debido a que están basados en los mismos principios.

Ventajas

Existen diversas fuentes que versan sobre estos temas, sin embargo, todas ellas presentan una serie de desventajas que hacen necesaria una encuesta específica. A continuación se citan las fuentes más relevantes y sus respectivos inconvenientes:

- Los Censos de Población: son fuentes que permiten obtener información sobre la fuerza de trabajo. No obstante, presentan un serie de desventajas, a saber: el distanciamiento en el tiempo, la recogida de datos se realiza mediante auto inscripción (esto es, el entrevistado cumplimenta por sí mismo el cuestionario), su elevado coste y la tardanza en la obtención del resultado.

- Las Encuestas de Salarios y las Encuestas Industriales: proporcionan información sólo sobre una parte de los ocupados: los asalariados, sin incluir todas las ramas de la actividad.

- El Paro Registrado y la afiliación a la Seguridad Social: debido a que deben realizar sus procedimientos en base a normas legales variables, la información ofrecida sólo representa una parte del colectivo, no permitiendo la obtención de series homogéneas.

Las ventajas de una Encuesta de Población Activa se pueden resumir en las siguientes categorías:

- Se puede realizar con la periodicidad que se desee.

- Mejora la profundización en los aspectos que más interesen con respecto a la fuerza de trabajo.

- La cumplimentación de los cuestionarios se lleva a cabo mediante entrevistadores especializados.

- Los resultados son obtenidos de forma rápida, debido al hecho de que se trata de una encuesta por muestreo.

- Las series obtenidas son homogéneas, puesto que las definiciones y el tratamiento de la información son uniformes a lo largo de las encuestas.

- Los resultados están enfocados tanto al conjunto nacional como a los subconjuntos territoriales (comunidades autónomas y provincias)

La mayor desventaja que presenta la Encuesta de Población Activa es, en esencia, su condición de encuesta. A modo de ejemplo, el número de personas activas de cada una de las sesenta divisiones de la Clasificación Nacional de Actividades Económicas en cada provincia resulta poco fiable, ya que un mayor análisis en la información conlleva un mayor error de muestreo.

Ámbito geográfico y poblacional

La Encuesta de Población Activa cubre todo el territorio nacional desde el segundo trimestre del 1988, momento en el que se incluyeron Ceuta y Melilla. Se encuentra dirigida a la población residente en viviendas familiares utilizadas todo el año como vivienda habitual.

Periodo de referencia

El periodo de referencia de los resultados de la EPA es trimestral, mientras que el periodo de referencia de la información es la semana inmediatamente anterior (de lunes a domingo) a la de la entrevista según el calendario.

Población desempleada o parada

La población parada o desempleada engloba a todas las personas de 16 o más años que reúnan simultáneamente las siguientes condiciones:

- No tener trabajo: no haber tenido un empleo por cuenta ajena ni por cuenta propia durante la semana de referencia

- En busca de trabajo: que hayan tomado medidas concretas para buscar un trabajo por cuenta ajena o hayan hecho gestiones para establecerse por su cuenta durante el mes posterior.

- Disponibilidad para trabajar: se deben hallar en condiciones de comenzar a hacerlo en un plazo de dos semanas a partir del domingo de la semana de referencia.

Asimismo, también se consideran desempleadas las personas de 16 o más años que durante la semana de referencia han estado sin trabajo, disponibles para trabajar y que no buscan empleo porque ya han encontrado uno al que se incorporarán dentro de los tres meses posteriores a la semana de referencia. Por consiguiente, en este caso no se exige el criterio de búsqueda efectiva de empleo.

3.1.2. Paro registrado

En este apartado se explica el funcionamiento del Paro Registrado, así como la categorización del término "desempleado" perteneciente al Servicio Público de Empleo Estatal.

¿Cómo se mide el paro registrado?

El Paro Registrado está constituido por el total de demandas de empleo en alta, registradas por el SEPE, existentes el último día de cada mes, excluyendo las que correspondan a situaciones laborales descritas en la Orden Ministerial del 11 de marzo de 1985 (B.O.E de 14/3/85) por la que se establecen los criterios estadísticos para la medición del Paro Registrado.

De acuerdo con la citada Orden Ministerial, se excluyen todas aquellas demandas que al final del mes de referencia se encuentran en alguna de las siguientes situaciones:

- Pluriempleo: demandantes que solicitan otro puesto de trabajo compatible con el que ejercen actualmente.

- Mejor empleo: demandantes que, estando ocupados, solicitan un trabajo para cambiarlo por el que realizan actualmente.

- Colaboración social: demandantes perceptores de prestaciones por desempleo que participan en trabajos de colaboración social.

- Jubilados: demandantes que son pensionistas de jubilación, por gran invalidez o invalidez absoluta y demandantes de edad igual o superior a 65 años.

- Empleo coyuntural: demandantes que solicitan un empleo para un período inferior a 3 meses.

- Jornada menor a 20 horas: demandantes que solicitan un trabajo con una jornada inferior a 20 horas semanales.

- Estudiantes: demandantes que están cursando estudios de enseñanza oficial reglada siempre que sean menores de 25 años o que superando esta edad sean demandantes de primer empleo. También hay que tener en cuenta los demandantes asistentes a cursos de Formación Profesional Ocupacional, cuando sus horas lectivas superen las 20 a la semana, dispongan de una beca al menos de manutención y sean demandantes de primer empleo.

- Demandas suspendidas: demandantes con demanda suspendida que permanezcan en esta situación ya que la suspensión de la demanda, que generalmente se tramita a petición del demandante y por causa que lo justifique, interrumpe la búsqueda de empleo.

- Compatibilidad de prestaciones: demandantes beneficiaros de prestaciones por desempleo en situación de compatibilidad de empleo por realizar un trabajo a tiempo parcial.

- Trabajadores Eventuales Agrícolas Subsidiados: demandantes que están percibiendo el subsidio agrario o que, habiéndolo agotado, no haya transcurrido un periodo superior a un año desde el día del nacimiento del derecho.

- Rechazo de acciones de inserción laboral: demandantes que rechacen acciones de inserción laboral adecuadas a sus características, según se establece en el Art. 17 apartado 2 del Real Decreto Legislativo 5/2000 de 4 de agosto (Rechazo de acciones de inserción laboral).

- Otras causas: demandantes sin disponibilidad inmediata para el trabajo o en situación incompatible con el mismo como demandantes inscritos para participar en un proceso de selección para un puesto de trabajo determinado, solicitantes de un empleo exclusivamente para el extranjero, demandantes de un empleo solo a domicilio, demandantes de servicios previos al empleo, demandantes que, en virtud de un expediente de regulación de empleo, están en situación de suspensión o reducción de jornada o modificación de las condiciones de trabajo, etc.

3.1.3. Google Trends

Puesto que en el Capítulo 2 se ha llevado a cabo una breve descripción del funcionamiento de Google Trends, a continuación se presentan las características más importantes a modo de conceptualización.[2]

- La muestra analizada por Google Trends es aleatoria, por lo que no tiene en cuenta el total de búsquedas.

- El índice mediante el cual se estandarizan las series temporales de Google Trends se denomina Google Index. Dicho índice representa la popularidad de los términos de búsquedas en relación con el total de búsquedas realizadas en una determinada área geográfica y en un intervalo de tiempo especificado. Las series temporales generadas no proporcionan, sin embargo, valores absolutos del volumen de búsquedas, sino la frecuencia relativa de las mismas.

- Google Index está normalizado en una escala de 0 a 100, dividiendo la popularidad relativa en cada momento *t* por el máximo valor del periodo. Si el volumen de búsquedas se encuentra por debajo del umbral de tráfico mínimo, Google Trends asignará un valor de 0.

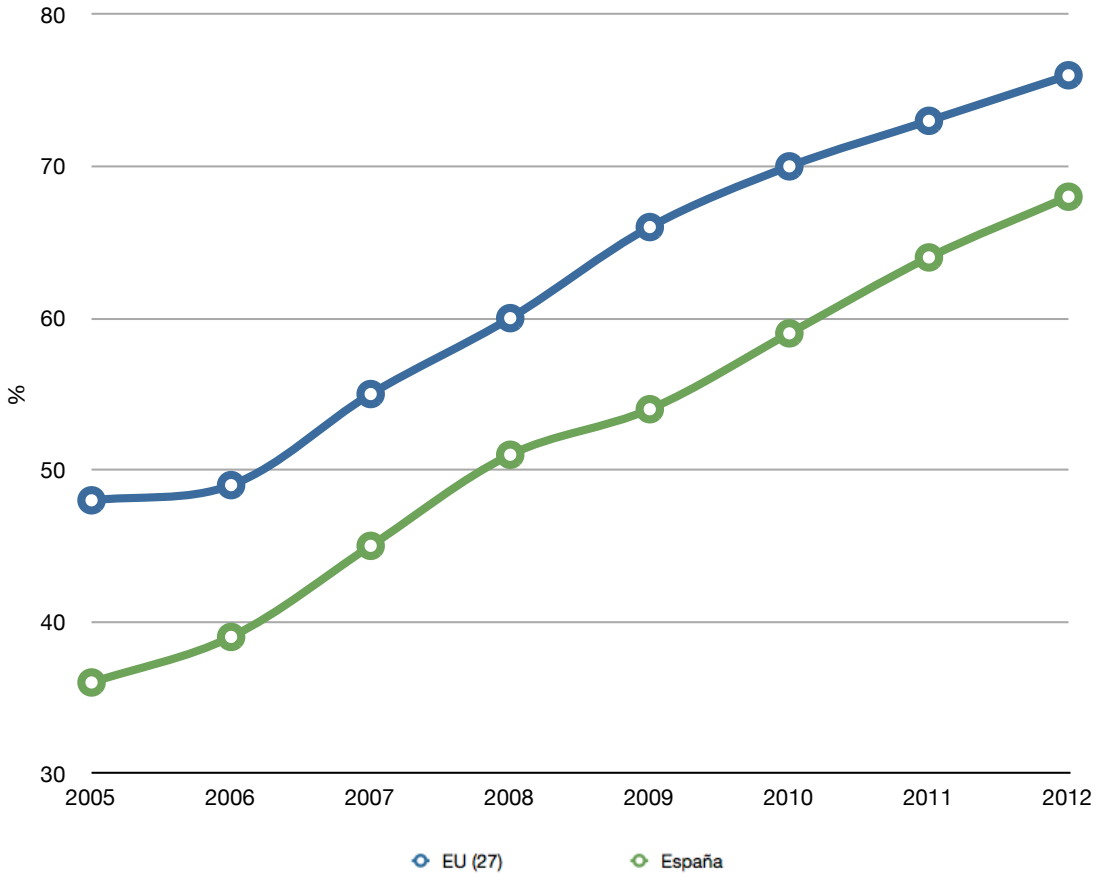
- El ámbito geográfico de búsqueda se puede delimitar en función del país o región. El periodo temporal, con origen en 2004, también se puede modificar (siendo la frecuencia siempre semanal).

- Las búsquedas se clasifican en 27 categorías de primer nivel y 241 subcategorías. Dicha clasificación se lleva a cabo de modo automático a través de un procesador de lenguaje natural.

[2] https://support.google.com/trends/

Justificación del uso de Google Trends en España

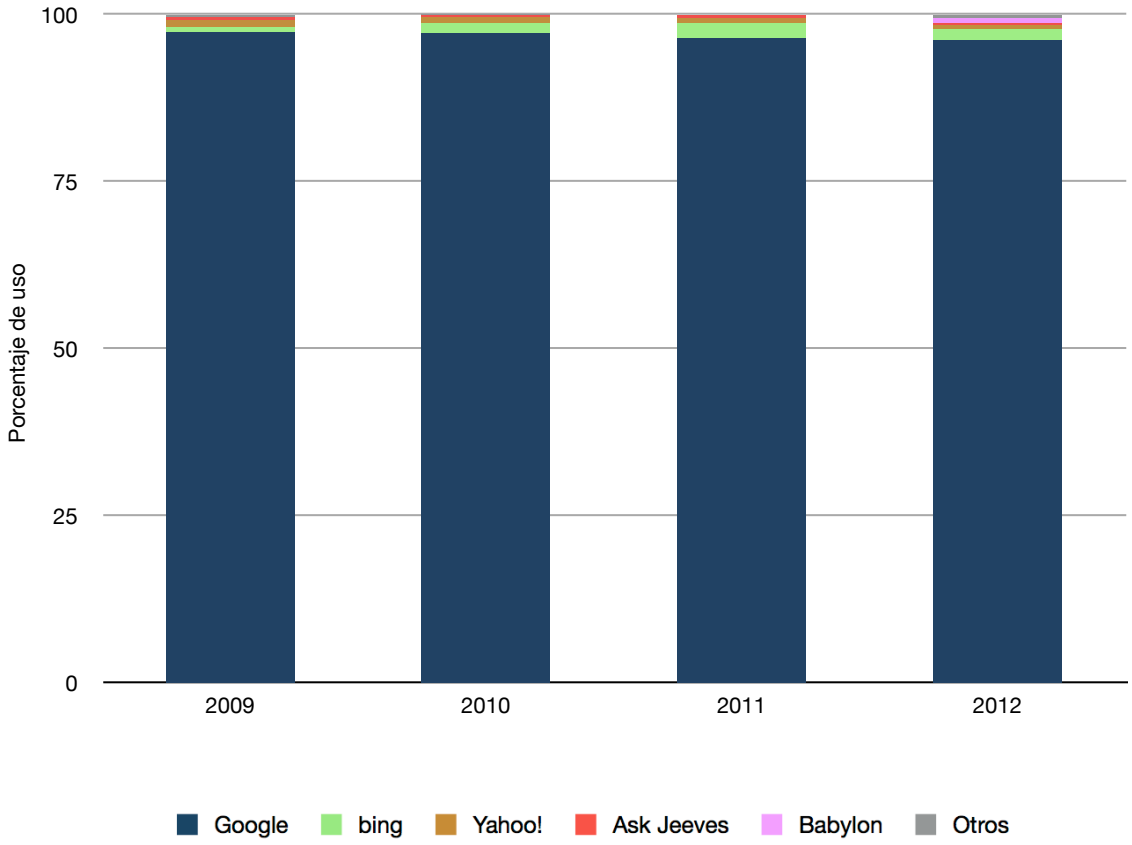
Gráfico 2. Hogares con acceso a Internet en la Unión Europa (27) y en España



Fuente: Elaboración propia a partir de los datos de Eurostat

El Gráfico 2 muestra la evolución del porcentaje de hogares con acceso a Internet. Como se puede apreciar, el porcentaje de hogares con acceso a Internet en España es menor que la media de la Unión Europea (27), sin embargo, se aprecia una convergencia en los últimos años. Así las cosas, dicho porcentaje resulta suficiente para llevar a cabo la presente investigación.

Gráfico 3. Evolución del uso de buscadores web en España (2009-2012)



Fuente: Elaboración propia a través de los datos de StatCounter

Con respecto al uso de Google en España, el Gráfico 3 muestra la evolución del uso de los buscadores web en España desde 2009 hasta 2012. Como se puede apreciar, el uso de Google representa más del 95%, por lo que es razonable pensar que los datos obtenidos a través de Google Trends son representativos de la sociedad española.

3.2. Muestra

En este apartado se expone, para cada variable previamente analizada, la muestra que ha sido utilizada para la realización del presente Trabajo Final de Carrera.

3.2.1. EPA

Las series temporales facilitadas por la Encuesta de Población Activa tienen periodicidad trimestral y comienzan en enero de 2005. Así pues, las series que han sido empleadas son las siguientes:

- Resultados Nacionales

    1.1. Parados por grupo de edad, sexo y sector económico.[3]

    1.2. Parados que han trabajado anteriormente por tiempo desde que dejaron el último empleo, sexo y grupo de edad.[4]

Como se puede apreciar, cada serie temporal está ligada a una o varias de las hipótesis expuestas en el Capítulo 2. De este modo, la serie temporal 1.1 está relacionada con las hipótesis 1, 2 y 3. Por otra parte, la serie temporal 1.2 está relacionada con las hipótesis 4 y 5.

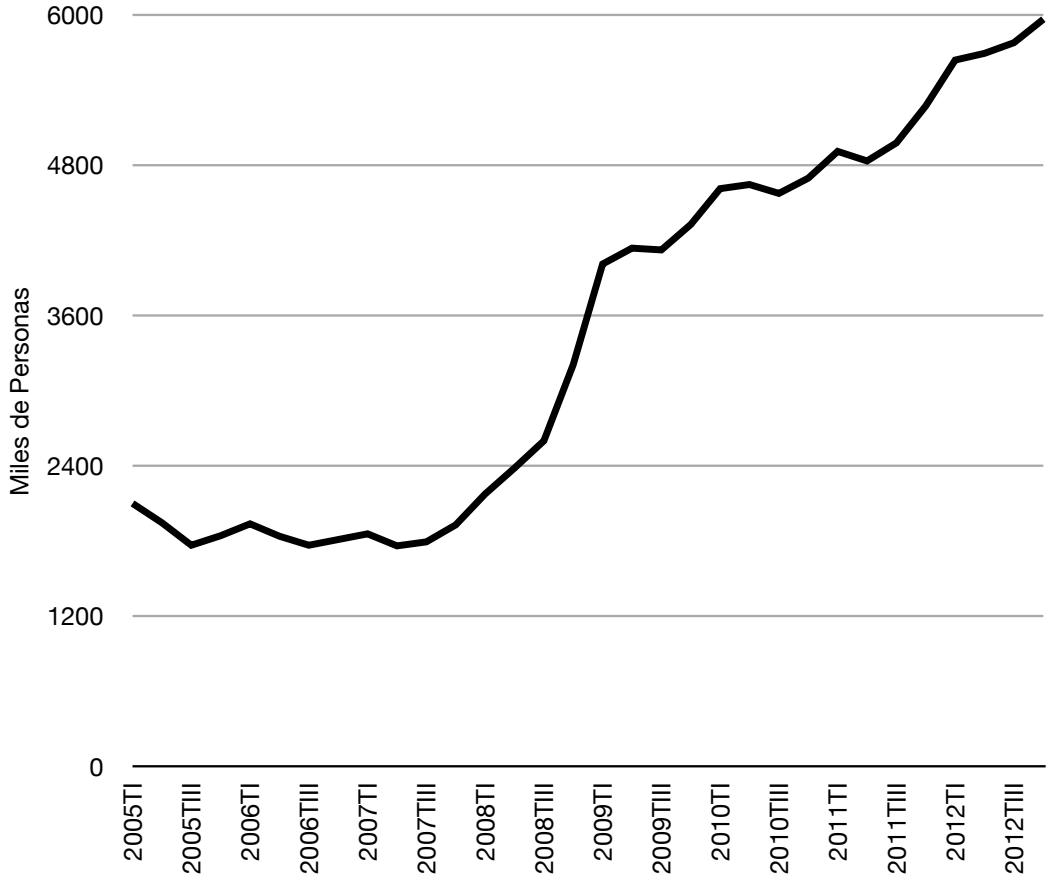
A continuación se representan gráficamente dichas series temporales, especificando qué parámetros se han tomado en consideración para su elaboración.

[3] http://www.ine.es/jaxiBD/tabla.do?per=03&type=db&divi=EPA&idtab=8

[4] http://www.ine.es/jaxiBD/tabla.do?per=03&type=db&divi=EPA&idtab=9

3.2.1.1. Total de desempleados en España

Gráfico 4. Total de desempleados en España (2005-2012)



Fuente: Elaboración propia a partir de los datos del INE.

La serie temporal del número total de desempleados en España ha sido obtenida a través de los resultados nacionales, en función del grupo de edad, sexo y sector económico (Serie 1.1.). En cada uno de los parámetros se seleccionó el valor total.

Como se puede observar, presenta una pendiente positiva que se ve incrementada de manera sustancial en el transcurso de los años 2008 y 2009, debido al impacto de la crisis económica en la economía española. La Tabla 2 muestra los estadísticos más relevantes para dicha serie temporal.

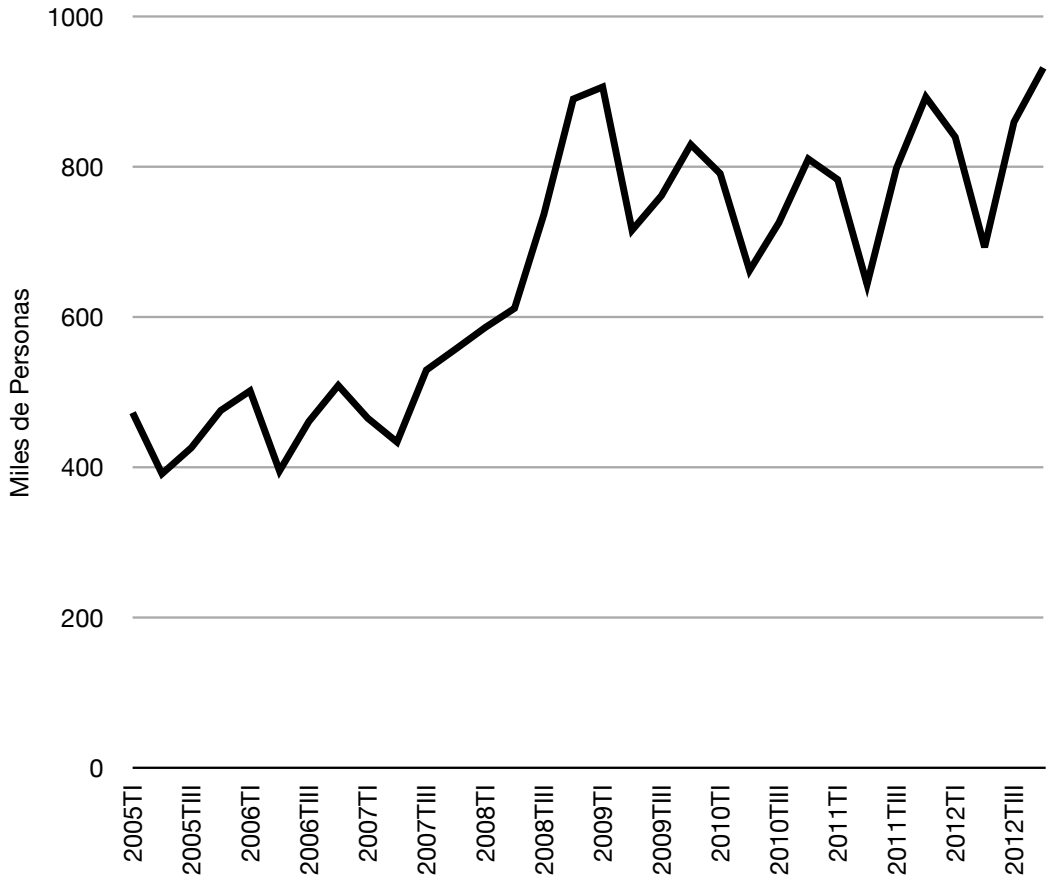
Tabla 2. Estadísticos más relevantes de la serie temporal: Total de desempleados en España (2005-12).

Media	3384,85
Desviación Típica	1463,39
Máximo	5778,1
Mínimo	1760
n	32

Fuente: Elaboración propia a partir de los datos del INE.

3.2.1.2. Parados que han trabajado anteriormente y llevan 3 meses en el Paro.

Gráfico 5. Parados que han trabajado anteriormente y llevan 3 meses en el Paro (2005-2012)



Fuente: Elaboración propia a partir de los datos del INE.

Para la elaboración de esta serie temporal se usaron los datos de los resultados nacionales, en función del tiempo transcurrido desde el anterior empleo, del sexo y del grupo de edad (Serie temporal 1.2). Se seleccionó el intervalo de *menos de 3 meses* para la variable tiempo, utilizando para el resto el volumen total.

Como se puede apreciar, la presente serie temporal presenta una forma de sierra, poniendo de manifiesto la estacionalidad. Asimismo, también se aprecia un claro punto de inflexión en los años 2008 y 2009, cuya explicación fue comentada previamente. La Tabla 3 muestra los estadísticos más relevantes para esta serie temporal.

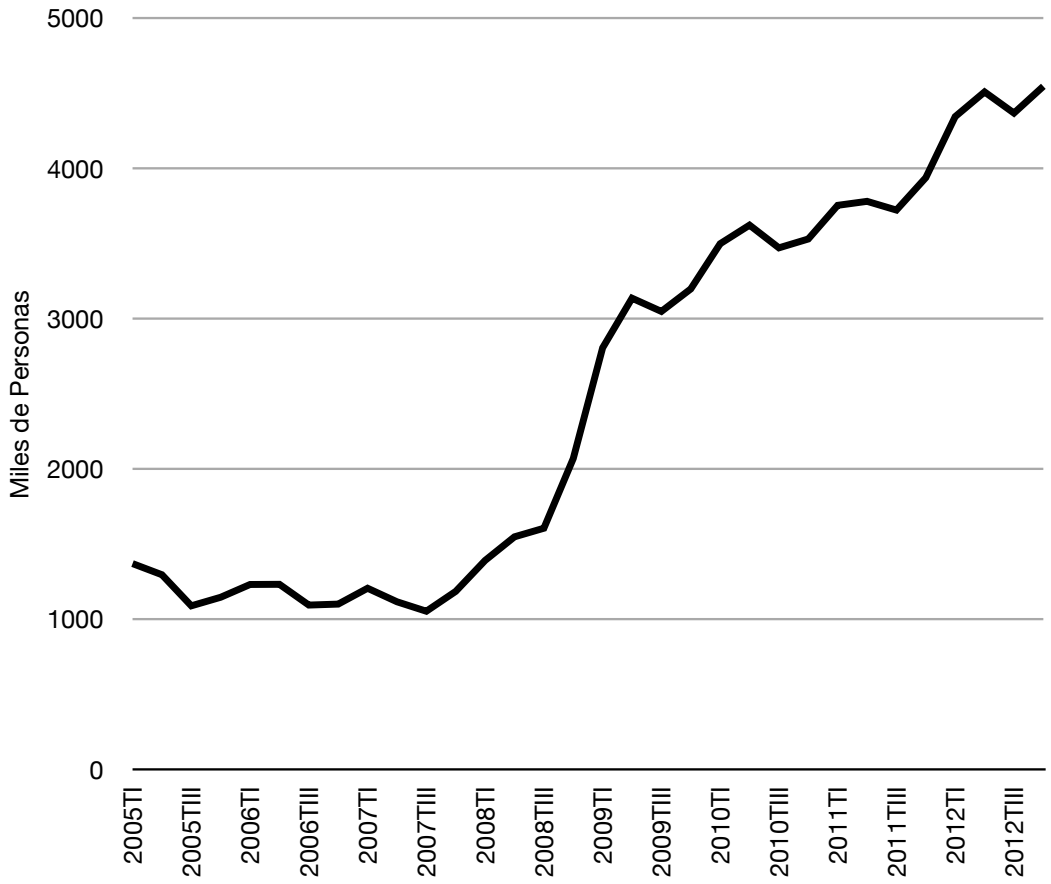
Tabla 3. Estadísticos más relevantes de la serie temporal: Parados que han trabajado anteriormente y llevan 3 meses en el Paro (2005-12)

Media	650,06
Desviación Típica	164,86
Máximo	906,1
Mínimo	391
n	32

Fuente: Elaboración propia a partir de los datos del INE.

3.2.1.3. Parados que han trabajado anteriormente y llevan más de 3 meses en el Paro.

Gráfico 6. Parados que han trabajado anteriormente y llevan más de 3 meses en el Paro (2005-2012)



Fuente: Elaboración propia a partir de los datos del INE.

La serie temporal "Parados que han trabajado anteriormente y llevan más de 3 meses en el Paro" fue generada con los resultados nacionales, de acuerdo con la serie 1.2, seleccionando, de forma agregada, en la variable tiempo: de 3 a 5 meses, de 6 a 12 meses y más de 1 año. Para el resto de las variables se tomó el valor total. Asimismo, se puede apreciar que guarda una gran similitud con el Gráfico 4, debido a que ambas series trabajan con gran cantidad de datos similares.

La única diferencia reside en que esta serie no tiene en cuenta aquellas personas que no han trabajado con anterioridad, mientras que en el Gráfico 4 sí que están recogidas. Asimismo, la Tabla 4 muestra los estadísticos más relevantes.

Tabla 4. Estadísticos más relevantes de la serie temporal: Parados que han trabajado anteriormente y llevan más 3 meses en el Paro (2005-12)

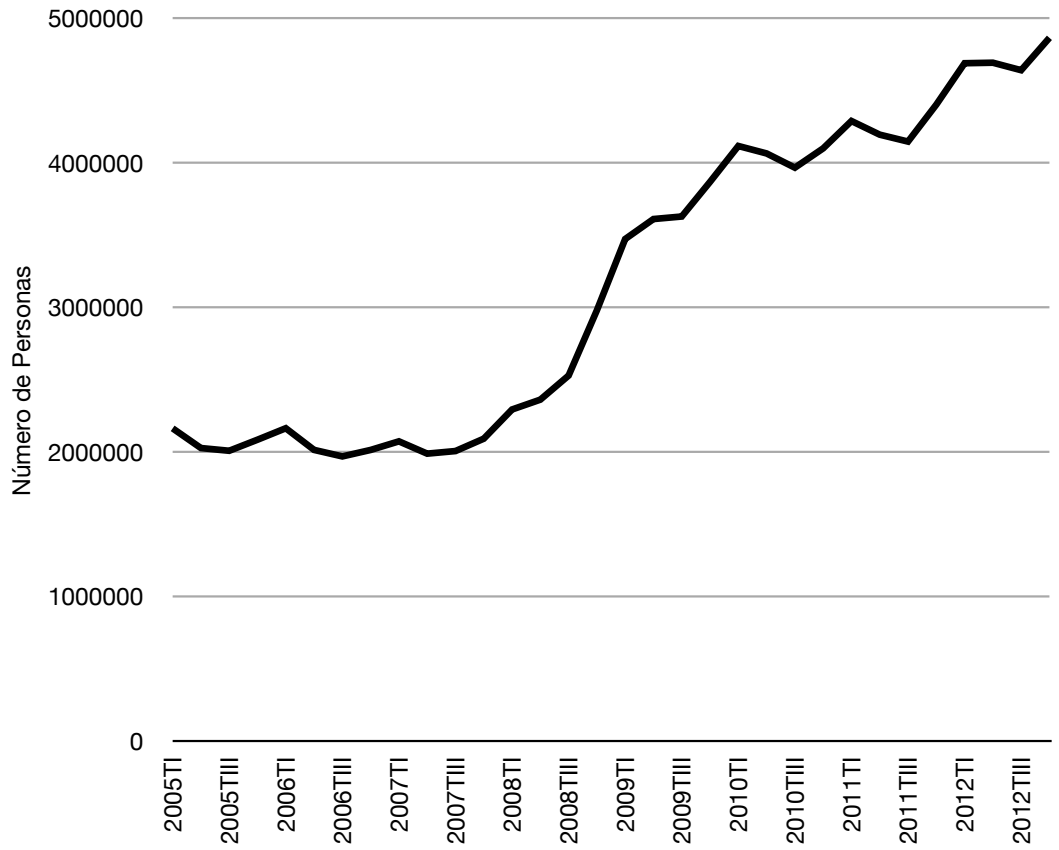
Media	2499,56
Desviación Típica	1270,78
Máximo	4545,5
Mínimo	1052,4
n	32

Fuente: Elaboración propia a partir de los datos del INE.

3.2.2. Paro registrado

La serie temporal empleada se trata de la serie *Paro Registrado por Sectores, Sexo y Edad*[5], facilitada por el Servicio Público de Empelo Estatal. Se trata de una serie mensual, por lo que se realizó una media aritmética para ajustarla a la serie trimestral de la Encuesta de Población Activa. Para todos las variables de la serie se seleccionó el valor total.

Gráfico 7 .Total de desempleados en España (2005 – 2012)



Fuente: Elaboración propia a partir de los datos del INE.

Se puede apreciar que guarda una gran similitud con los datos extraídos de la Encuesta de Población Activa recogidos en el Gráfico 4, aunque esta serie presenta más oscilaciones.

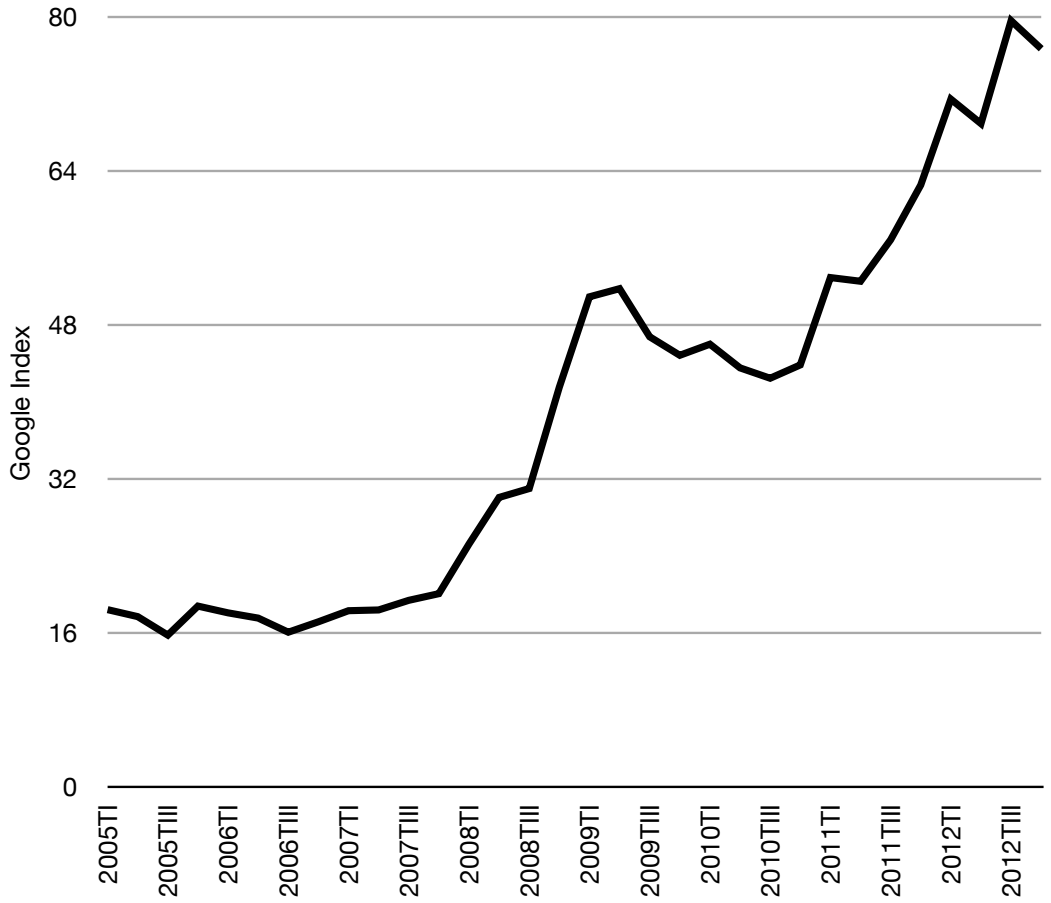
[5] http://www.sepe.es/contenido/estadisticas/datos_avance/pdf/empleo/evolparoseries.pdf

3.2.3. Google Trends

Las series temporales obtenidas con Google Trends se basan en los distintos términos de búsqueda que han sido empleados. Aunque los datos se encuentran disponibles desde enero de 2004, se comenzó en 2005 para poder ajustarlos con los datos de la EPA. Con el mismo objetivo, la periodicidad de las series de Google Trends (semanal) se ajustó a trimestral realizando una media aritmética (D'Amuri, 2009). A continuación, se representan los términos de búsqueda utilizados.

3.2.3.1. Término de búsqueda: Paro

Gráfico 8. Evolución del término de búsqueda *Paro* en España (2005 – 2012)



Fuente: Elaboración propia a partir de los datos de Google Trends

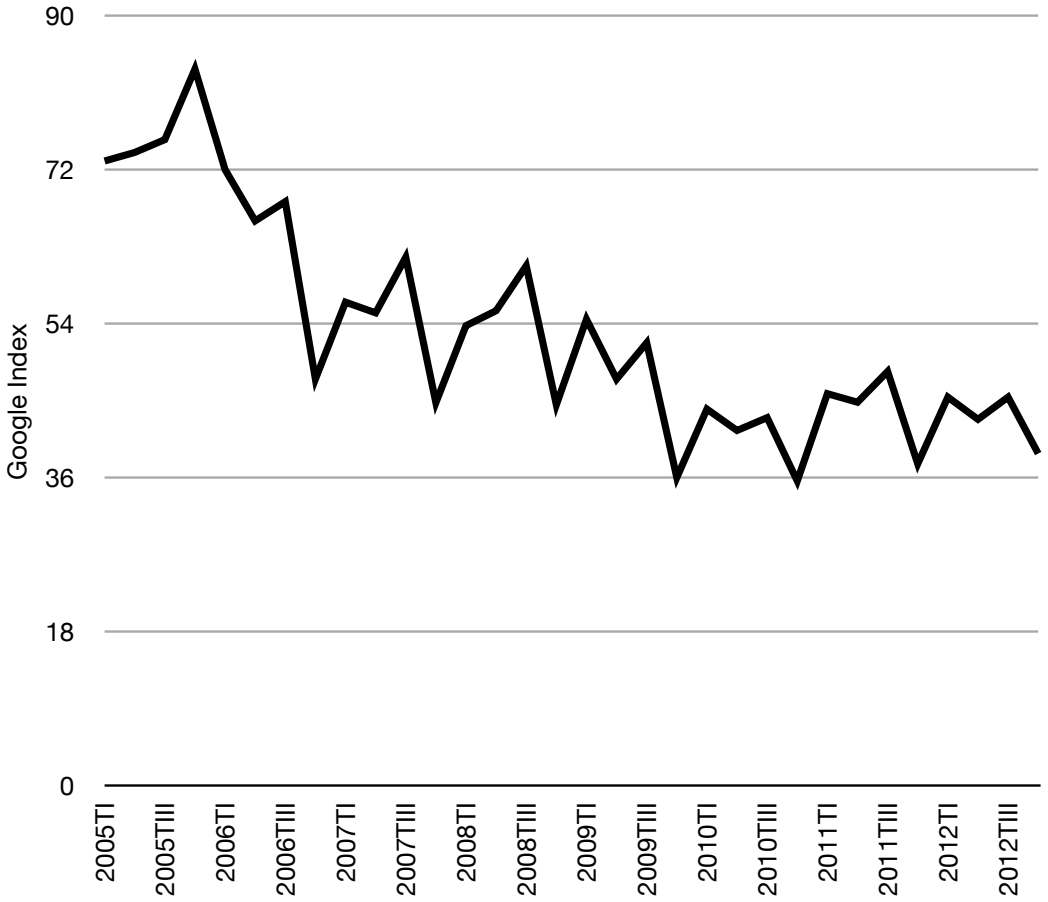
El Gráfico 8 representa el término de búsqueda *Paro*, desde el primer trimestre de 2005 hasta el cuarto trimestre de 2012. Al haberse realizado medias aritméticas para ajustar los datos semanales de Google Trends con los datos trimestrales de la EPA, ningún trimestre tiene el valor 100. El ámbito geográfico seleccionado fue España.

Se espera que el término *Paro* sea un buen predictor del número total de desempleados en España por varios motivos. En primer lugar, cualquier consulta que tiene que ver con el desempleo suele ir acompañada de la palabra *Paro*. En segundo lugar, un aumento en el nivel del paro implica una mayor relevancia social, y por lo tanto, un mayor número de búsquedas.

Se puede observar que la evolución en la popularidad del término *Paro* presenta una gran similitud con la evolución del número total de desempleados en España, representada en el Gráfico 4 y el Gráfico 7. Este hecho enfatiza la correlación existente entre la sociedad "real" y la sociedad "virtual", y pone de manifiesto que un aumento en la tasa real de desempleo implica un mayor volumen de búsquedas referidas al Paro.

3.2.3.2. Término de búsqueda: Trabajo

Gráfico 9 Evolución del término de búsqueda *Trabajo* en España (2005 – 2012)



Fuente: Elaboración propia a partir de los datos de Google Trends

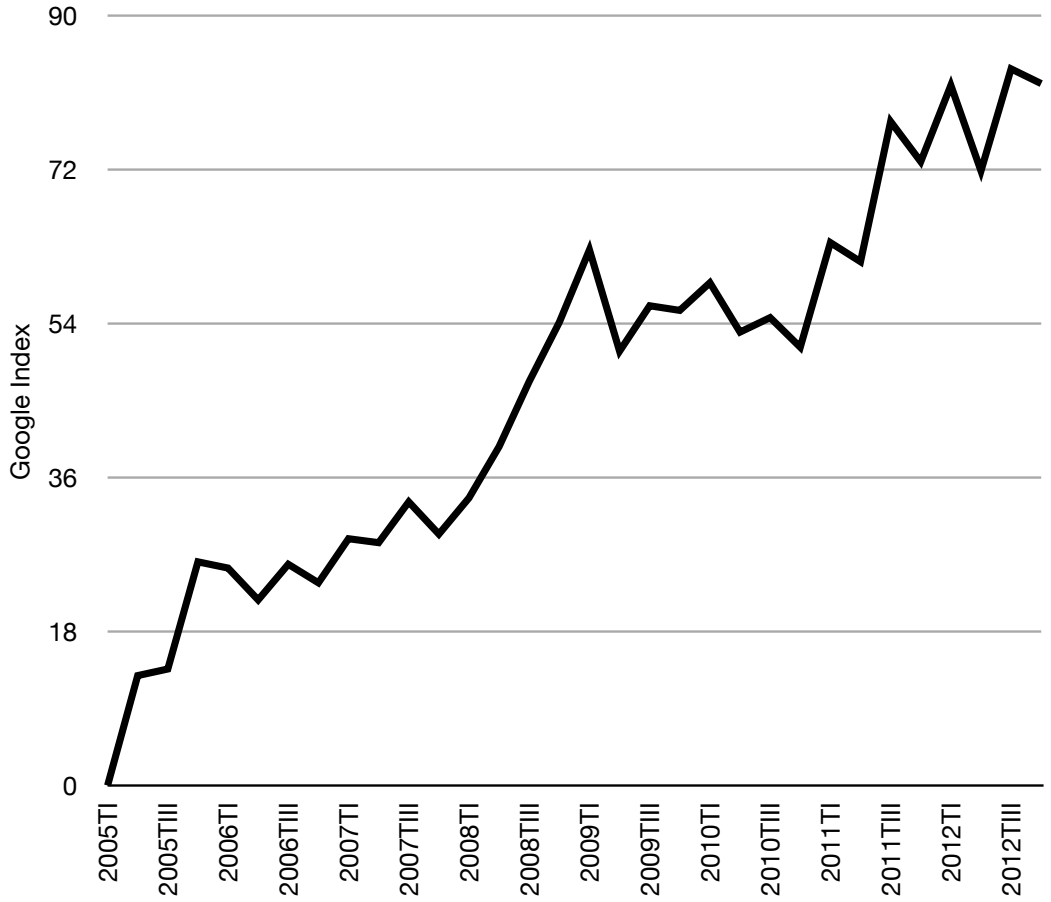
El Gráfico 9 representa el término de búsqueda *Trabajo*, desde el primer trimestre de 2005 hasta el cuarto trimestre de 2012. El ámbito geográfico seleccionado fue España.

Se espera que el término *Trabajo* esté relacionado con el número total de desempleados en España, al igual que sugieren diversos estudios realizados en países como Alemania (Askitas y Zimmermann, 2009) o Estados Unidos (D'Amuri y Marcucci, 2009), en los que los autores emplearon el término de búsqueda Jobs (trabajos).

Sin embargo, *a priori* se puede observar que la pendiente que presenta el Gráfico 9 es negativa, y no parece que esté relacionado con el Gráfico 4.

3.2.3.3. Término de búsqueda: Cobrar Paro

Gráfico 10. Evolución del término de búsqueda *Cobrar Paro* en España (2005 – 2012)



Fuente: Elaboración propia a partir de los datos de Google Trends

El Gráfico 10 representa el término de búsqueda *Cobrar Paro*, desde enero de 2005 hasta diciembre de 2012. El ámbito geográfico seleccionado fue España.

El término *Cobrar Paro* puede estar relacionado con aquellas personas que hayan trabajado con anterioridad y que se encuentren desempleadas en un intervalo de tiempo no muy espacioso. Así pues, las oscilaciones representadas por el Gráfico 10 guardan una similitud moderada con las representadas en el Gráfico 5.

3.3. Modelos propuestos

A continuación se detallan los modelos econométricos relacionados con las hipótesis postuladas en el Capítulo 2.

3.3.1 Modelos para predecir el número total de desempleados en España mediante Google Trends

Los modelos 1.a y 1.b tratan de verificar si los términos Paro y Trabajo ayudan a predecir el número total de desempleados en España respectivamente. Para ello, se proponen 2 modelos econométricos autorregresivos, basándose en los trabajos realizados por Choi y Varian (2012). Nótese que la variable dependiente (EPA) está basada en los datos de desempleo de la Encuesta de Población Activa, debido a que se trata del mejor indicador del desempleo en España.

$$\textbf{Modelo 1.a}.\ EPA_t = \beta_0 + \beta_1\, GT1_t + \sum_{j=1}^{4} \alpha_j\, EPA_{t-j} + \varepsilon_t$$

$$\textbf{Modelo 1.b}.\ EPA_t = \beta_0 + \beta_1\, GT2_t + \sum_{j=1}^{4} \alpha_j\, EPA_{t-j} + \varepsilon_t$$

EPA = número de desempleados en España de acuerdo con la EPA

GT1 = serie temporal basada en los datos de Google Trends para el término de búsqueda *Paro*

GT2 = serie temporal basada en los datos de Google Trends para el término de búsqueda *Trabajo*

$\sum_{j=1}^{4} \alpha_j$= sumatorio de los términos autorregresivos

ε = perturbación aleatoria

3.3.2 Modelos para predecir el número total de desempleados en España mediante Google Trends y el Paro Registrado

Se parte de la suposición de que es posible elaborar un modelo econométrico que trate de predecir el número de desempleados en España (con respecto a los datos oficiales de la EPA), basándose en los datos disponibles del Paro Registrado, elaborados por el SEPE. Así pues, el poder predictivo de los datos de Google Trends se demostrará si se mejora la predicción al incluirlos en el modelo. De manera similar al Modelo 1, se proponen dos modelo econométricos autorregresivos, basándose en los trabajos realizados por Choi y Varian (2012). Dicho esto, si el modelo 2.a funciona mejor que el modelo 2.b se demostrará el poder predictivo de Google Trends.

Modelo 2.a. $EPA_t = \beta_0 + \beta_1 GT1_t + \beta_2 PR + \sum_{j=1}^{4} \alpha_j EPA_{t-j} + \varepsilon_t$

Modelo 2.b. $EPA_t = \beta_0 + \beta_1 PR_t + \sum_{j=1}^{4} \alpha_j EPA_{t-j} + \varepsilon_t$

EPA = número de desempleados en España de acuerdo con la EPA

PR = número de desempleados en España de acuerdo con el SEPE

GT1 = serie temporal basada en los datos de Google Trends para el término de búsqueda Paro

$\sum_{j=1}^{4} \alpha_j$= sumatorio de los términos autorregresivos

ε = perturbación aleatoria

3.3.3 Modelos para predecir el número de personas que han trabajado anteriormente y llevan menos de 3 meses en el Paro

Los modelos 3.a y 3.b tienen como objetivo delimitar qué tipo de término de búsqueda se ajusta mejor al estrato del paro en el que las personas han trabajado anteriormente y llevan desempleadas menos de 3 meses. Así pues, si el modelo 3.b obtiene mejores resultados, implicará que el término de búsqueda *Cobrar Paro* ayuda a predecir mejor dicho estrato del desempleo que el término general *Paro*.

Modelo 3.a. $EPA\_L3M_t = \beta_0 + \beta_1 GT1_t + \sum_{j=1}^{4} \alpha_j EPA\_L3M_{t-j} + \varepsilon_t$

Modelo 3.b. $EPA\_L3M_t = \beta_0 + \beta_1 GT3_t + \sum_{j=1}^{4} \alpha_j EPA\_L3M_{t-j} + \varepsilon_t$

EPA_L3M = número de Parados que han trabajado anteriormente y llevan menos 3 meses en el Paro.

GT1 = serie temporal basada en los datos de Google Trends para el término de búsqueda *Paro*

GT3 = serie temporal basada en los datos de Google Trends para el término de búsqueda *Cobrar Paro*

$\sum_{j=1}^{4} \alpha_j$= sumatorio de los términos autorregresivos

ε = perturbación aleatoria

3.3.4 Modelos para predecir el número de personas que han trabajado anteriormente y llevan más de 3 meses en el Paro

Los modelos 4.a y 4.b funcionan como contrapartida de los modelos 3.a y 3.b. Han sido formulados para comprobar que el término de búsqueda *Cobrar Paro* no está relacionado con aquellas personas que han trabajado anteriormente y llevan desempleadas más de 3 meses. Por tanto, se espera que el modelo 4.b no funcione correctamente.

Modelo 4.a. $EPA\_M3M_t = \beta_0 + \beta_1 GT1_t + \sum_{j=1}^{4} \alpha_j EPA\_M3M_{t-j} + \varepsilon_t$

Modelo 4.b. $EPA\_M3M_t = \beta_0 + \beta_1 GT3_t + \sum_{j=1}^{4} \alpha_j EPA\_M3M_{t-j} + \varepsilon_t$

EPA_M3M = número de Parados que han trabajado anteriormente y llevan más de 3 meses en el Paro.

GT1 = serie temporal basada en los datos de Google Trends para el término de búsqueda *Paro*

GT3 = serie temporal basada en los datos de Google Trends para el término de búsqueda *Cobrar Paro*

$\sum_{j=1}^{4} \alpha_j$= sumatorio de los términos autorregresivos

ε = perturbación aleatoria

3.3.5 Tabla resumen: hipótesis y modelos propuestos

Tabla 5. Resumen: Hipótesis y modelos propuestos

Hipótesis	Modelo
▪ H_1: La popularidad del término de búsqueda *Paro* ayuda a predecir el número total de desempleados en España.	1.a
▪ H_2: La popularidad del término de búsqueda *Trabajo* ayuda a predecir el número total de desempleados en España.	1.b
▪ H_3: La popularidad del término de búsqueda *Paro* mejora la predicción del número total de desempleados en España basada en los datos del desempleo disponibles.	2.a y 2.b
▪ H_4: La popularidad del término de búsqueda *Cobrar Paro* está relacionada con aquellas personas que han trabajado anteriormente y llevan desempleadas poco tiempo.	3.a y 3.b
▪ H_5: La popularidad del término de búsqueda *Cobrar Paro* no está relacionada con aquellas personas que han trabajado anteriormente y llevan desempleadas mucho tiempo.	4.a y 4.b

Fuente: Elaboración propia.

CAPÍTULO 4

RESULTADOS

CAPÍTULO 4. RESULTADOS

4.1 Introducción

El presente capítulo se encuentra estructurado de la siguiente manera. En primer lugar se realiza una exposición de las estimaciones obtenidas para los modelos presentas en el capítulo previo. Seguidamente, se lleva a cabo una validación de dichas estimaciones, comprobando así que los modelos econométricos son consistentes. Finalmente, se exponen las conclusiones pertinentes respecto a cada modelo.

4.2 Estimaciones

En este apartado se presentan los resultados de los modelos previamente estipulados. El programa utilizado para obtener los datos ha sido EViews. Para cada uno de los modelos se han tenido en cuenta cuatro especificaciones:

- Especificación 1: no se consideran los términos autorregresivos.
- Especificación 2: se considera el término autorregresivo de primer orden.
- Especificación 3: se consideran los términos autorregresivos de primer y cuarto orden.
- Especificación 4: se consideran los términos autorregresivos de primer, segundo, tercer y cuarto orden.

La razón por la cual se han llevado a cabo las citadas especificaciones se debe a la estacionalidad de la serie temporal. Así pues, puede ser que los datos del paro del trimestre n no dependan de ningún trimestre anterior (especificación 1), o puede que dependan del trimestre n_{-1} (especificación 2). Sin embargo, también puede ser que dependan de los trimestres n_{-1} y n_{-4} (especificación 3), por ejemplo, que los datos del tercer trimestre dependan del segundo trimestre y del tercer trimestre del año anterior. Finalmente, podría darse el caso de que los datos del trimestre n dependieran de los trimestres n_{-1} , n_{-2} , n_{-3} y n_{-4} (especificación 4).

Para seleccionar la especificación del modelo que mejor describe los datos se debe buscar aquella forma que presente un mayor coeficiente de determinación ajustado (R^2 ajustado) y, menor valor de Akaike Info Criterion (AIC) y de Mean Absolute Error (MAE) respectivamente. Asimismo, se debe validar que las variables explicativas sean significativas (su P-valor debe ser inferior a 5%), descartando el término constante.

- El R^2 es un coeficiente que indica el porcentaje del ajuste que se ha conseguido con el modelo lineal. A mayor porcentaje mejor es el modelo para predecir el comportamiento de la variable explicada.
- El AIC mide de forma relativa la calidad de un modelo estadístico. Se basa en la complejidad del modelo y en la bondad de ajuste. Cuanto menor sea el nivel de AIC mejor será el modelo.
- El MAE es el error medio absoluto de un modelo. Cuanto menor sea el nivel de MAE más ajustada estará la predicción del modelo.

A pesar de todo lo expuesto, la elección de cada modelo no implica su veracidad, puesto que posteriormente se llevará a cabo su pertinente validación econométrica para comprobar su funcionabilidad.

4.2.1 Modelo 1

Tabla 6. Estimaciones del modelo 1.a

				Modelo 1.a
	Coeficientes			
	Especificación 1	Especificación 2	Especificación 3	Especificación 4
Variables				
Constante	605,6*	28560,86	4.128,66	1875,80*
GT1	74,09*	33,71*	31,20*	47,96
EPA_{t-1}		0,99*	1,13*	0,74*
EPA_{t-2}				0,90*
EPA_{t-3}				-0,63
EPA_{t-4}			-0,17	-0,09
N	32	32	32	32
R^2 ajustado	0,93	0,99	0,99	0,99
AIC	14,83	13,05	12,98	13,04
MAE	8,14	3,82	3,27	3,52

Nota: * p<0,05

Fuente: Elaboración propia

Para el **modelo 1.a** se puede apreciar en la Tabla 6 que las especificaciones 3 y 4 quedan descartadas al obtener términos autorregresivos que no son significativos (su P-valor es mayor que 5%). Por lo tanto, entre las especificaciones 1 y 2 se puede observar que:

- La especificación 2 dispone de un R^2 ajustado de 0,99, mientras que la especificación 1 dispone de un R^2 ajustado de 0,93. Por lo tanto, la especificación 2 tiene mejor valor de R^2 ajustado.

- Con respecto al valor del AIC, la especificación 2 obtiene un valor de 13,05, mientras que el valor de AIC para la especificación 1 es de 14.83. Por consiguiente, la especificación 2 obtiene un mejor valor de AIC.

- En lo referente al MAE, la especificación 1 obtiene un valor de 8,14, mientras que la especificación 2 obtiene un valor de 3,82. Así pues, la especificación 2 obtiene un mejor valor de MAE.

Por todo lo expuesto se puede concluir que la mejor especificación es la **número 2.**

Con respecto al **modelo 1.b** se puede observar en la Tabla 7 que las especificaciones 3 y 4 presentan términos autorregresivos no significativos, por lo que quedan descartadas. Por lo tanto, entre las especificaciones 1 y 2 se puede observar que:

- La especificación 1 dispone de un R^2 ajustado de 0,56, mientras que la especificación 2 dispone de un R^2 ajustado de 0,98. Por lo tanto, la especificación 2 tiene mejor valor de R^2 ajustado.

- Con respecto al valor del AIC, la especificación 1 obtiene un valor de 16,77, mientras que el valor de AIC para la especificación 2 es de 13,60. Por consiguiente, la especificación 2 obtiene un mejor valor de AIC.

- En lo referente al MAE, la especificación 1 obtiene un valor de 30,4, mientras que la especificación 2 obtiene un valor de 5,29. Así pues, la especificación 2 obtiene un mejor valor de MAE.

Por todo lo expuesto se debe seleccionar la especificación **número 2**. Sin embargo, como se puede apreciar, el valor de GT2 no resulta significativo; el modelo 1.b (especificación 2) se trata de un modelo válido *a priori*, pero la variable GT2 no está relacionada con el nivel de paro estimado por la EPA.

Tabla 7. Estimaciones del modelo 1.b

				Modelo 1.b
	Coeficientes			
	Especificación 1	Especificación 2	Especificación 3	Especificación 4
Variables				
Constante	8197,83*	49993,82	17.414,88	60.409,00
GT2	-83,31*	0,6	0,72	3,28
EPA_{t-1}		0,99*	1,11*	1,86*
EPA_{t-2}				-1,62
EPA_{t-3}				1,39*
EPA_{t-4}			-0,13	0,62*
N	32	32	32	32
R^2 ajustado	0,56	0,98	0,98	0,99
AIC	16,77	13,60	13,63	13,73
MAE	30,40	5,29	4,46	5,21

Nota: * p<0,05

Fuente: Elaboración propia

4.2.2 Modelo 2

Tabla 8 Estimaciones del modelo 2.a

				Modelo 2.a
	Coeficientes			
	Especificación 1	Especificación 2	Especificación 3	Especificación 4
Variables				
Constante	-852,86*	-856,86*	-883,08*	-889,95
GT1	11,73*	12,16*	11,72*	11,39*
PR	0,001*	0,001*	0,001*	0,001*
EPA_{t-1}		0,19	0,14	-0,10
EPA_{t-2}			0,27	-0,69*
EPA_{t-3}				-0,22
EPA_{t-4}				-0,09
N	32	32	32	32
R^2 ajustado	0,998	0,998	0,998	0,999
AIC	11,05	10,99	10,86	10,49
MAE	1,51	1,38	1,16	0,86

Nota: * p<0,05

Fuente: Elaboración propia

Para el **modelo 2.a** se puede apreciar que las especificaciones 2, 3 y 4 obtienen algunos términos autorregresivos no significativos, por lo que la especificación elegida es la **número 1**.

Tabla 9. Estimaciones del modelo 2.b

				Modelo 2.b
	Coeficientes			
	Especificación 1	Especificación 2	Especificación 3	Especificación 4
Variables				
Constante	-1065,36*	-1058,95*	5.568,78	-1052,39
PR	0,001*	0,001*	0,001*	0,001*
EPA_{t-1}		0,53*	0,79*	0,69*
EPA_{t-2}				-0,29
EPA_{t-3}				0,05
EPA_{t-4}			0,21	0,17
N	32,00	32,00	32,00	32,00
R^2 ajustado	0,996	0,997	0,997	0,997
AIC	11,90	11,65	11,64	11,77
MAE	2,00	1,68	1,54	1,39

Nota: * p<0,05

Fuente: Elaboración propia

Para el **modelo 2.b** se puede observar en la Tabla 9 que las especificaciones 3 y 4 disponen de términos autorregresivos que no son significativos, por lo que quedan descartadas. Así pues, entre las especificaciones 1 y 2 se puede observar que:

- La especificación 2 dispone de un R^2 ajustado de 0,997, mientras que la especificación 1 dispone de un R^2 ajustado de 0,996. La diferencia, por lo tanto, no es comparable.

- Con respecto al valor del AIC, la especificación 2 obtiene un valor de 11,65, mientras que el valor de AIC para la especificación 1 es de 11,90. Por consiguiente, la especificación 2 obtiene un mejor valor de AIC.

- En lo referente al MAE, la especificación 1 obtiene un valor de 2, mientras que la especificación 2 obtiene un valor de 1,68. Así pues, la especificación 2 obtiene un mejor valor de MAE.

De acuerdo a todo lo expuesto se puede concluir que la mejor especificación para el modelo 2.b es la **número 2**.

4.2.3 Modelo 3

Tabla 10.Estimaciones del modelo 3.a

				Modelo 3.a
	Coeficientes			
	Especificación 1	Especificación 2	Especificación 3	Especificación 4
Variables				
Constante	372,91*	382,12*	442,78*	441,43
GT1	7,41*	7,24*	6,40*	6,24
EPA_{t-1}		0,51*	0,30	0,65*
EPA_{t-2}				-0,56*
EPA_{t-3}				0,43
EPA_{t-4}			0,54*	0,24
N	32	32	32	32
R^2 ajustado	0,74	0,79	0,83	0,85
AIC	11,87	11,67	11,38	11,28
MAE	10,42	9,38	7,22	6,89

Nota: * p<0,05

Fuente: Elaboración propia

Para el **modelo 3.a** se puede observar que las especificaciones 3 y 4 quedan descartadas al obtener términos autorregresivos que no son significativos. Por lo tanto, entre las especificaciones 1 y 2 se puede observar que:

- La especificación 2 dispone de un R^2 ajustado de 0,79 , mientras que la especificación 1 dispone de un R^2 ajustado de 0,74. Por lo tanto, la especificación 2 tiene mejor valor de R^2 ajustado.

- Con respecto al valor del AIC, la especificación 2 obtiene un valor de 11,67, mientras que el valor de AIC para la especificación 1 es de 11,87. Por consiguiente, la especificación 2 obtiene un mejor valor de AIC.

- En lo referente al MAE, la especificación 1 obtiene un valor de 10,42, mientras que la especificación 2 obtiene un valor de 9,38. Así pues, la especificación 2 obtiene un mejor valor de MAE.

Por todo lo expuesto se puede concluir que la mejor especificación es la **número 2**.

Tabla 11.Estimaciones del modelo 3.b

				Modelo 3.b
	Coeficientes			
	Especificación 1	Especificación 2	Especificación 3	Especificación 4
Variables				
Constante	342,17*	298,47*	197,15*	435,1*
GT3	6,84*	7,59*	8,10*	5,37*
EPA_{t-1}		0,42*	0,40	0,96*
EPA_{t-2}				-0,92*
EPA_{t-3}				0,85*
EPA_{t-4}			0,47*	-0,10
N	32	32	32	32
R^2 ajustado	0,79	0,84	0,85	0,90
AIC	11,64	11,40	11,24	10,84
MAE	9,61	8,29	7,57	5,38

Nota: * p<0,05

Fuente: Elaboración propia

Para el **modelo 3.b** se puede apreciar que las especificaciones 3 y 4 quedan descartadas al obtener términos autorregresivos que no son significativos. Por lo tanto, entre las especificaciones 1 y 2 se puede observar que:

- La especificación 1 dispone de un R^2 ajustado de 0,79 , mientras que la especificación 2 dispone de un R^2 ajustado de 0,84. Por lo tanto, la especificación 2 tiene mejor valor de R^2 ajustado.

- Con respecto al valor del AIC, la especificación 1 obtiene un valor de 11,64, mientras que el valor de AIC para la especificación 2 es de 11,40. Por consiguiente, la especificación 2 obtiene un mejor valor de AIC.

- En lo referente al MAE, la especificación 1 obtiene un valor de 9,61, mientras que la especificación 2 obtiene un valor de 8,29. Así pues, la especificación 2 obtiene un mejor valor de MAE.

Por todo ello se puede concluir que la mejor especificación es la **número 2**.

4.2.4 Modelo 4

Tabla 12.Estimaciones del modelo 4.a

				Modelo 4.a
	Coeficientes			
	Especificación 1	Especificación 2	Especificación 3	Especificación 4
Variables				
Constante	114,88	32658,31	3755,23	6498,73
GT1	61,78	25,00*	23,81*	0,92
EPA_{t-1}		0,99*	1,05*	1,90*
EPA_{t-2}				-1,81
EPA_{t-3}				1,64
EPA_{t-4}			-0,08	-0,74
N	32	32	32	32
R^2 ajustado	0,92	0,98	0,98	0,99
AIC	14,75	13,31	13,40	12,81
MAE	11,32	5,92	5,54	4,49

Nota: * p<0,05

Fuente: Elaboración propia

Para el **modelo 4.a** se puede observar que solamente la especificación 2 obtiene parámetros significativos, por lo que la especificación elegida es la **número 2**.

Tabla 13 Estimaciones del modelo 4.b

				Modelo 4.b
	Coeficientes			
	Especificación 1	Especificación 2	Especificación 3	Especificación 4
Variables				
Constante	73,91	-3256,76	32928,32	6323,20
GT3	52,41*	-1,42	-0,65	-0,65
EPA_{t-1}		1,01*	1,05*	1,91*
EPA_{t-2}				-1,81
EPA_{t-3}				1,63*
EPA_{t-4}			-0,06	-0,73*
N	32	32	32	32
R^2 ajustado	0,83	0,97	0,97	0,99
AIC	15,48	13,61	13,69	12,80
MAE	24,13	7,10	6,64	4,60

Nota: * p<0,05

Fuente: Elaboración propia

Con respecto al **modelo 4.b** se puede observar en la Tabla 13 que las especificaciones 3 y 4 presentan términos autorregresivos no significativos, por lo que quedan descartadas. Por lo tanto, entre las especificaciones 1 y 2 se puede observar que:

- La especificación 1 dispone de un R^2 ajustado de 0,83, mientras que la especificación 2 dispone de un R^2 ajustado de 0,97. Por lo tanto, la especificación 2 tiene mejor valor de R^2 ajustado.

- Con respecto al valor del AIC, la especificación 1 obtiene un valor de 15,48, mientras que el valor de AIC para la especificación 2 es de 13.61. Por consiguiente, la especificación 2 obtiene un mejor valor de AIC.

- En lo referente al MAE, la especificación 1 obtiene un valor de 24,13, mientras que la especificación 2 obtiene un valor de 7,10. Así pues, la especificación 2 obtiene un mejor valor de MAE.

Por todo lo expuesto se debe seleccionar la especificación **número 2**. Sin embargo, como se puede apreciar, el valor de GT3 no resulta significativo; el modelo 4.b (especificación 2) se trata de un modelo válido *a priori*, pero la variable GT3 no está relacionada con el nivel de paro estimado por la EPA.

4.3. Validación del modelo 1.a

En este apartado se lleva a cabo la validación econométrica del **modelo 1.a**. Así pues, los cálculos pertinentes para el resto de los modelos se pueden encontrar en el Anexo 2. Los cálculos han sido realizados con el programa Eviews.

4.3.1 Hipótesis utilizadas en la especificación

Se han considerado las siguientes hipótesis básicas para poder aplicar el método de estimación de Mínimos Cuadrados Ordinarios (MCO).

1. El valor esperado de la perturbación es cero.

$$E(\varepsilon_t) = 0$$

2. Homocedasticidad: La varianza de la perturbación es constante

$$V\,AR(\varepsilon_t) = E[\varepsilon_t - \varepsilon^*]^2 = \text{constante}$$

3. Las perturbaciones no están relacionadas entre sí, es decir, no hay autocorrelación

$$COV(\varepsilon_t, \varepsilon_j) = 0$$

4. Las variables explicativas son fijas, es decir, no les afecta la incertidumbre o la aleatoriedad

$$COV(X_t, \varepsilon_t) = 0$$

5. Ausencia de multicolinealidad perfecta, es decir, no hay una relación perfecta entre las variables explicativas

4.3.2 Contraste de errores de especificación

Para averiguar si existen errores de especificación, ya sea por utilizar una forma funcional incorrecta o por omitir alguna variable explicativa relevante, se aplica el test RESET de Ramsey. Dicho test consiste en estimar un modelo transformado en el que se introduce como variable explicativa la variable estimada del modelo original.

Se parte, por lo tanto, del modelo original (especificación 2):

$$EPA_t = \beta_0 + \beta_1 GT1_t + \beta_2 EPA_{t-1} + \varepsilon_t$$

Figura 2. Estimación del modelo 1.a (especificación 2)

Dependent Variable: EPA
Method: Least Squares
Date: 04/12/13 Time: 13:10
Sample(adjusted): 2005:2 2012:4
Included observations: 31 after adjusting endpoints
Convergence achieved after 71 iterations

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	28560.86	376069.0	0.075946	0.9400
GT1	33.70995	6.789467	4.965036	0.0000
AR(1)	0.997690	0.032787	30.42987	0.0000

R-squared	0.990235	Mean dependent var	3509.574
Adjusted R-squared	0.989538	S.D. dependent var	1537.422
S.E. of regression	157.2555	Akaike info criterion	13.04539
Sum squared resid	692420.5	Schwarz criterion	13.18416
Log likelihood	-199.2035	F-statistic	1419.724
Durbin-Watson stat	1.360199	Prob(F-statistic)	0.000000

Inverted AR Roots	1.00

Fuente: Elaboración propia

La estimación por MCO (Mínimos Cuadrados Ordinarios) proporciona la estimación mostrada en la Figura.1. Para aplicar el test de RESET de Ramsey se define la variable FITTED como el valor estimado de EPA_t (por claridad, sólo se muestran dos cifras significativas de cada parámetro estimado):

$$FITTED = 28560{,}86 + 33{,}71 \cdot GT1_t + 0{,}99 \cdot EPA_{t-1} + \varepsilon_t$$

El modelo transformado quedaría de la siguiente manera, cuya estimación se presenta en la Figura 2.

$$EPA_t = \beta_0 + \beta_1 \cdot GT1_t + \beta_2 \cdot EPAt_{-1} + \beta_3 \cdot FITTED^2 + \varepsilon_t$$

El último paso del test RESET consiste en efectuar un contraste de significatividad individual sobre β_3. Para ello se utiliza la probabilidad exacta asociada al estadístico $t^*(\alpha^*(t^*_{\beta 3}))$, ofrecida por EViews en la columna Prob. Las hipótesis de esta prueba son las siguientes:

- $H_0 : \beta_i = 0 \rightarrow \beta_i$ no es significativo
- $H_1 : \beta_i \neq 0 \rightarrow \beta_i$ sí es significativo

La regla de decisión de esta prueba es:

- Si $\alpha^*(t^*_{\beta 3}) > \alpha = 0{,}05 \rightarrow$ se acepta H_0
- Si $\alpha^*(t^*_{\beta 3}) < \alpha = 0{,}05 \rightarrow$ se rechaza H_0

Como el valor estimado por Eviews para β_3 es 0,778 > α = 0,05, se acepta H_0, es decir, β_3 no es significativo y, por tanto, se concluye que no existen errores de especificación.

Figura 3. Test RESET de Ramsey: Estimación del modelo transformado

Test Equation:
Dependent Variable: EPA
Method: Least Squares
Date: 04/16/13 Time: 10:56
Sample: 2005:2 2012:4
Included observations: 31
Convergence achieved after 28 iterations

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-4973.617	25744.41	-0.193192	0.8483
GT1	35.37688	10.64558	3.323151	0.0026
FITTED^2	-6.59E-06	2.32E-05	-0.284440	0.7782
AR(1)	1.009020	0.032589	30.96238	0.0000

R-squared	0.990300	Mean dependent var	3509.574
Adjusted R-squared	0.989222	S.D. dependent var	1537.422
S.E. of regression	159.6098	Akaike info criterion	13.10326
Sum squared resid	687832.8	Schwarz criterion	13.28829
Log likelihood	-199.1005	F-statistic	918.8273
Durbin-Watson stat	1.270722	Prob(F-statistic)	0.000000

Fuente: Elaboración propia

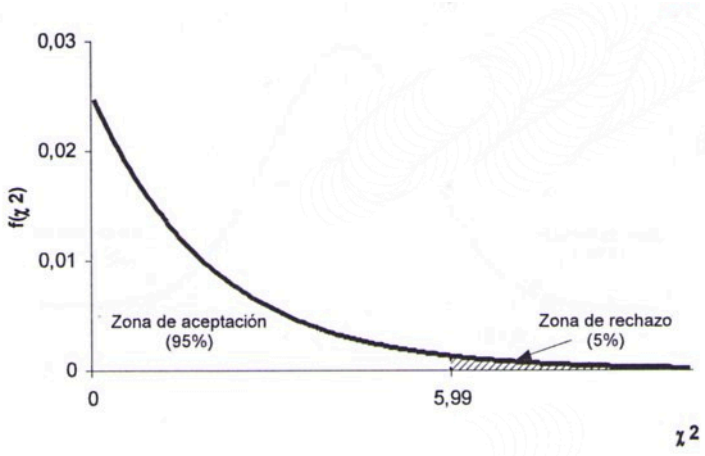
4.3.3 Análisis de la normalidad de las perturbaciones

Para analizar la normalidad de las perturbaciones se lleva a cabo la prueba de Jarque-Bera. Las hipótesis de esta prueba son:

- $H_0 : \upsilon_t \sim N(0,\sigma^2)$ → Las perturbaciones se distribuyen como una normal
- $H_1 : \upsilon_t \nsim N(0,\sigma^2)$ → No hay normalidad en las perturbaciones

Para llevar a cabo la prueba, se calcula el estadístico Jarque-Bera, que se distribuye como una χ^2 con 2 grados de libertad: $JB \sim \chi^2_2 = 5{,}99$, como muestra la Figura 4.

Figura 4 Distribución del estadístico Jarque-Bera



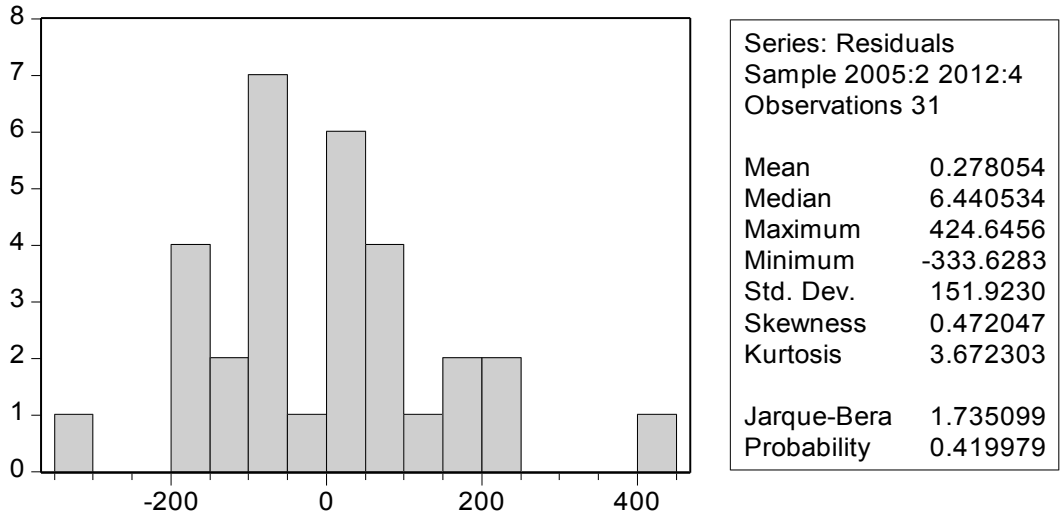
Fuente: Elaboración propia

De esta forma, la regla de decisión es la siguiente:

- Si JB* < 5,99 → se acepta H_0, las perturbaciones se distribuyen como una normal
- Si JB* > 5,99 → se rechaza H_0, las perturbaciones no son normales

EViews ofrece el cálculo del estadístico Jarque-Bera. Según muestra la Figura 4, este estadístico presenta un valor de 1,73 <5,99, por lo que se acepta H_0 y se concluye que las perturbaciones se distribuyen conforme a la distribución normal. Este resultado también indica que el resto de contrastes que se van a efectuar son fiables, por lo que sí se puede realizar inferencia en el modelo.

Figura 5. Resultado de la prueba de normalidad de los residuos



Fuente: Elaboración propia

4.3.4 Contrastes de significatividad

En este epígrafe se realizan los contrastes de significatividad de los parámetros del modelo, primero de forma individual para cada parámetro y después, de forma conjunta.

Contrastes de significatividad individual

Se van a efectuar contrastes de significatividad sobre las pendientes parciales de las variables explicativas (β_1 y β_2). El objetivo de estos contrastes es determinar si la variación de cada variable explicativa incide de forma significativa en la variable endógena. Para ello, se empleará la prueba *t* y la probabilidad exacta asociada al estadístico *t**. Ambas pruebas comparten las mismas hipótesis, que son las que se enuncian a continuación:

- $H_0 : \beta_i = 0 \rightarrow \beta_i$ no es significativo
- $H_1 : \beta_i \neq 0 \rightarrow \beta_i$ sí es significativo

Contrastes de significatividad de β_1

La prueba *t* para 1,699 y -1,699.del cálculo del estadístico $t^*_{\beta 2}$, que se calcula de la siguiente forma:

$$t * \beta 1 = \frac{\beta 1}{ee(\beta 2)} = \frac{33{,}71}{6{,}79} = 4{,}96$$

Los valores de β_1, su desviación típica así como el propio estadístico calculado, los proporciona EViews en la pantalla de resultados recogida en la Figura 2. En este caso, el estadístico $t^*_{\beta 1}$, se distribuye como una distribución t con 29 (=32-3 parámetros) grados de libertad. Por lo tanto, los valores críticos para cada una de las dos colas son 1,699 y -1,699.

De esta forma, la regla de decisión es la siguiente:

- Si $t^*_{\beta 1} \in [1{,}699; -1{,}699]$ → Se acepta H_0, β_2 no es significativo
- Si $t^*_{\beta 1} \notin [1{,}699; -1{,}699]$ → Se rechaza H_0, β_2 sí es significativo

Como $t^*_{\beta 1} = 4{,}96 \notin [1{,}699; -1{,}699]$, se rechaza H_0 y se concluye que β_1 sí que es significativo. Es decir, un cambio en el valor de la variable GT1 afectará de forma estadísticamente significativa a la variable Y_t.

Otra manera de calcular la significatividad de β_1 es mediante la probabilidad exacta asociada al estadístico t*, la cual se denota como $\alpha^*(t^*_{\beta 1})$ y viene ofrecida por EViews en la columna Prob. (Véase la Figura 2)

La regla de decisión es la siguiente:

- Si $\alpha^*(t^*_{\beta 1}) > \alpha = 0{,}05$ → se acepta H_0
- Si $\alpha^*(t^*_{\beta 1}) < \alpha = 0{,}05$ → se rechaza H_0

Como el valor proporcionado por EViews para $\alpha^*(t^*_{\beta 1})$ es $0{,}00 < \alpha = 0{,}05$, se rechaza H_0, es decir, β_1 sí que es significativo.

Contrastes de significatividad de β_2

La prueba *t* para 1,699 y -1,699.del cálculo del estadístico $t^*_{\beta 2}$, que se calcula de la siguiente forma:

$$t * \beta 2 = \frac{\beta 2}{ee(\beta 2)} = \frac{0{,}99}{0{,}03} = 30{,}42$$

Los valores de β_2, su desviación típica así como el propio estadístico calculado, los proporciona EViews en la pantalla de resultados recogida en la figura 2. En este caso, el estadístico $t^*_{\beta 2}$, se distribuye como una distribución t con 29 (=32-3 parámetros) grados de libertad. Por lo tanto, los valores críticos para cada una de las dos colas son 1,699 y -1,699.

De esta forma, la regla de decisión es la siguiente:

- Si $t^*_{\beta 2} \in [1{,}699;-1{,}699]$ → Se acepta H_0, β_2 no es significativo
- Si $t^*_{\beta 2} \notin [1{,}699;-1{,}699]$ → Se rechaza H_0, β_2 sí es significativo

Como $t^*_{\beta 2} = 30{,}42 \notin [1{,}699;-1{,}699]$, se rechaza H_0 y se concluye que β_2 sí que es significativo.

De acuerdo con la probabilidad exacta asociada al estadístico t^*, la regla de decisión es la siguiente:

- Si $\alpha^*(t^*_{\beta 2}) > \alpha = 0{,}05$ → se acepta H_0
- Si $\alpha^*(t^*_{\beta 2}) < \alpha = 0{,}05$ → se rechaza H_0

Como el valor proporcionado por EViews para $\alpha^*(t^*_{\beta 2})$ es $0{,}00 < \alpha = 0{,}05$, se rechaza H_0, es decir, β_2 sí que es significativo.

Contrastes de significatividad conjunta

Se van a efectuar contrastes de significatividad global sobre las pendientes parciales de las variables explicativas (β_1 y β_2). El objetivo de estos contrastes es determinar si existe regresión de las variables explicativas sobre la variable endógena. Para ello, se empleará la prueba *F* y la probabilidad exacta asociada al estadístico *F**. Ambas pruebas comparten las mismas hipótesis, que son las que se enuncian a continuación:

- $H_0 : \beta_1 = \beta_2 = 0$ → no existe regresión
- H_1 : Algún $\beta_I \neq 0$ → sí existe regresión

La prueba F parte del cálculo del estadístico *F*,* que se distribuye como una *F* con 2 grados de libertad en el numerador y 29 grados de libertad en el denominador. (Los datos numéricos están redondeados a la segunda unidad por claridad)

$$F* = \frac{\frac{R2}{k-1}}{\frac{(1-R2)}{n-k}} = \frac{\frac{0,99}{3-1}}{\frac{(1-0,99)}{32-3}} = 1419,72$$

El valor de R^2, así como el propio estadístico calculado, es proporcionado por EViews en la pantalla de resultados recogida en la figura 2. En este caso, el estadístico *F** se distribuye como una distribución F con 2(= 3 parámetros -1) grados de libertad en el numerador y 29 (32 observaciones – 3 parámetros) grados de libertad en el denominador. El valor crítico para $\alpha = 0,05$ según las tablas es de 3,33.

- Si $F^* < F_{(2,29)} = 3,33$ → se acepta H_0, no existe regresión
- Si $F^* > F_{(2,29)} = 3,3$ → se rechaza H_0, sí existe regresión.

Como $F^* = 1419,72 > F_{(2,29)} = 3,3$, se rechaza H_0 y se concluye que algún $\beta_i \neq 0$ y, por tanto, sí que existe regresión.

Otra manera de verificar si existe regresión es mediante la probabilidad exacta al estadístico F*, la cual viene denotada como $\alpha^*(F^*)$ y es ofrecida por EViews como Prob(F-Statistic). (Véase la Figura 2)

La regla de decisión de esta prueba es la siguiente:

- Si $\alpha^*(F^*) > \alpha = 0,05$ → se acepta H_0
- Si $\alpha^*(F^*) < \alpha = 0,05$ → se rechaza H_0

Como el valor estimado por EViews para la probabilidad asociada al estadístico F* es $0,00 < \alpha = 0,05$, se rechaza H_0, es decir, sí que existe regresión.

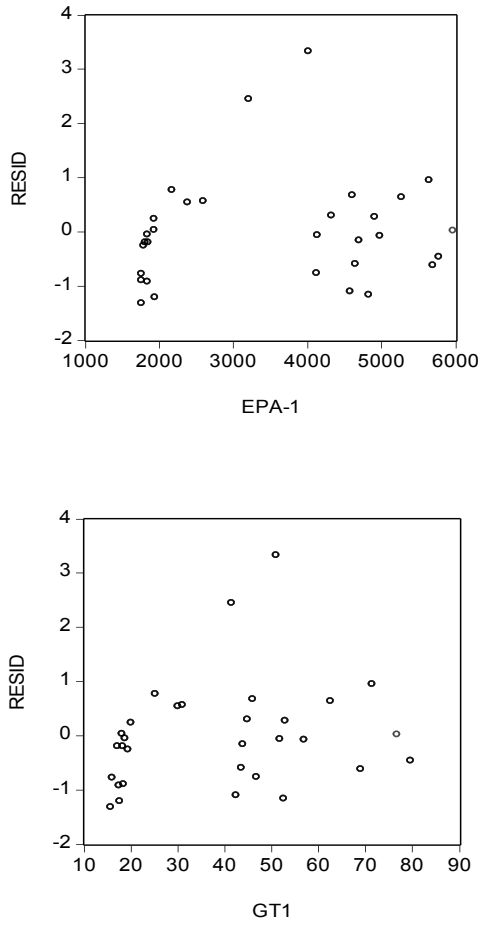
4.3.5 Análisis de la heterocedasticidad

En este apartado se comprobará la existencia de heterocedasticidad mediante dos métodos. En primer lugar se llevará a cabo el método gráfico y posteriormente se realizará el método analítico.

Método gráfico

El método gráfico de detección de heterocedasticidad consiste en analizar gráficamente en un diagrama de dispersión que compara el valor absoluto de los residuos con cada una de las variables explicativas. Estos diagramas de dispersión pueden verse en la Figura 6, de los que no se aprecia una clara relación funcional. En cualquier caso, será el método analítico el que confirme o rechace esta posible relación.

Figura 6. Método gráfico de detección de la heterocedasticidad



Fuente: Elaboración propia

Método analítico

El método analítico de detección de heterocedasticidad consiste en aplicar la prueba de White. Ésta se apoya en una regresión auxiliar en la que la variable endógena es el cuadrado de los residuos de la regresión original. Como variables explicativas se incluyen las variables exógenas del modelo original además del producto de las variables explicativas del modelo original.

En este caso, la regresión auxiliar sería la siguiente:

$$e = \alpha_1 + \alpha_1 GT1 + \alpha_2 GT1^2 + \alpha_3 GT1^2 \cdot Y_{t-1} + \alpha_4 Y_{t-1}^2 + \alpha_5 GT1 \cdot Y_{t-1}^2$$

Las hipótesis de esta prueba son:

- $H_0 : e^2_i \neq f(\chi_i) \rightarrow e^2_i = \alpha_i \rightarrow$ Homocedasticidad
- $H_1 : e^2_i = f(\chi_i) \rightarrow e^2_i \neq \alpha_i \rightarrow$ Heterocedasticidad

Figura 7. Prueba de White

White Heteroskedasticity Test:

F-statistic	1.624893	Probability	0.214958
Obs*R-squared	3.223809	Probability	0.199507

Test Equation:
Dependent Variable: RESID^2
Method: Least Squares
Date: 04/12/13 Time: 10:38
Sample: 2005:2 2012:4
Included observations: 31

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	17736.88	30404.12	0.583371	0.5643
GT1	-465.7784	1612.303	-0.288890	0.7748
GT1^2	11.86721	18.28102	0.649155	0.5215

R-squared	0.103994	Mean dependent var	22336.14
Adjusted R-squared	0.039993	S.D. dependent var	37141.22
S.E. of regression	36390.94	Akaike info criterion	23.93379
Sum squared resid	3.71E+10	Schwarz criterion	24.07257
Log likelihood	-367.9738	F-statistic	1.624893
Durbin-Watson stat	1.849769	Prob(F-statistic)	0.214958

Fuente: Elaboración propia

Para llevar a cabo la prueba, se calcula el estadístico de White como $n \cdot R^2$ del modelo auxiliar. Este estadístico se distribuye asintóticamente como una χ^2 con k-1 grados de libertad, donde k es el número de regresores de la regresión auxiliar (4 = 5-1 grados de libertad en este caso).

El valor crítico para $\alpha = 0{,}05$ es $\chi^2_4 = 9{,}49$. De esta forma, la regla de decisión es la siguiente.

- Si $n \cdot R^2 < \chi^2_4 = 9{,}49$ → Se acepta H_0, por tanto, existe homocedasticidad
- Si $n \cdot R^2 > \chi^2_4 = 9{,}49$ → Se rechaza H_0, por tanto, existe heterocedasticidad

EViews ofrece el cálculo del estadístico de White. Según muestra la Figura 7, este estadístico presenta un valor de $3{,}22 < \chi^2_4 = 9{,}49$, por lo que se acepta H_0 y se concluye que existe homocedasticidad.

4.3.6 Análisis de la autocorrelación

Para determinar la existencia de autocorrelación se empleará el contraste de Box-Ljung. Este contraste tiene como objetivo determinar la autocorrelación de cualquier orden. Dado que los datos que se dispone son trimestrales, se incluirán 8 términos *p* en el contraste.

Para llevar a cabo este contraste se utilizará la siguiente regresión auxiliar, donde p_p es el coeficiente de correlación de orden *p*:

$$e_t = p_1 \cdot e_{t-1} + p_2 \cdot e_{t-2} + p_3 \cdot e_{t-3} + p_4 \cdot e_{t-4} + p_5 \cdot e_{t-5} + p_6 \cdot e_{t-6} + p_7 \cdot e_{t-7} + p_8 \cdot e_{t-8} + \varepsilon_t$$

Las hipótesis de este contraste son las siguientes:

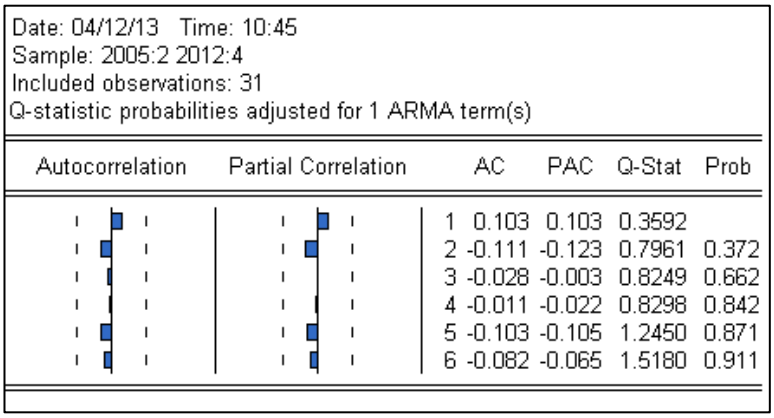
- H_0 *:* $p_1 = p_2 = \ldots = p_p = 0$ → no existe autocorrelación de orden p
- H_1 *:* Algún $p_p \neq 0$ → existe autocorrelación de orden p

El contraste requiere del cálculo del estadístico *Q* para cada orden de autocorrelación que se quiera contrastar. Este estadístico se distribuye como una χ^2 con *p* grados de libertad. De esta forma, la regla de decisión es la siguiente para cada orden de autocorrelación.

- Si $Q_p < \chi^2_p$ → se acepta H_0, por tanto, no existe autocorrelación de orden p
- Si $Q_p > \chi^2_p$ → se rechaza H_0, por tanto, existe autocorrelación de orden p

EViews ofrece el cálculo del estadístico Q para cada orden de autocorrelación, según muestra la Figura 8. Analizando la tabal y utilizando $\alpha = 0{,}05$ para calcular los valores críticos, se puede concluir que:

Figura 8. Detección de la autocorrelación



Fuente: Elaboración propia

- Como $Q_1 = 0{,}36 < \chi^2_1 = 3{,}84$ → No existe autocorrelación de orden 1
- Como $Q_2 = 0{,}79 < \chi^2_2 = 3{,}84$ → No existe autocorrelación de orden 2
- Como $Q_3 = 0{,}82 < \chi^2_3 = 5{,}99$ → No existe autocorrelación de orden 3
- Como $Q_4 = 0{,}83 < \chi^2_4 = 7{,}81$ → No existe autocorrelación de orden 4
- Como $Q_5 = 1{,}25 < \chi^2_5 = 9{,}49$ → No existe autocorrelación de orden 5
- Como $Q_6 = 1{,}51 < \chi^2_6 = 11{,}07$ → No existe autocorrelación de orden 6
- Como $Q_7 = 1{,}97 < \chi^2_7 = 12{,}59$ → No existe autocorrelación de orden 7
- Como $Q_8 = 2{,}54 < \chi^2_8 = 15{,}51$ → No existe autocorrelación de orden 8

Por consiguiente y, de acuerdo con todo lo establecido en el apartado 4.3, se puede validar el uso del modelo 1.a y corroborar su poder predictivo. Como se especificó en al introducción, la validación de los modelos restantes se encuentra en el Anexo 2.

4.4. Discusión

En este epígrafe se realiza una discusión de los modelos previamente formulados y validados. Asimismo, se determina la validez de las hipótesis ligadas a cada uno de los modelos.

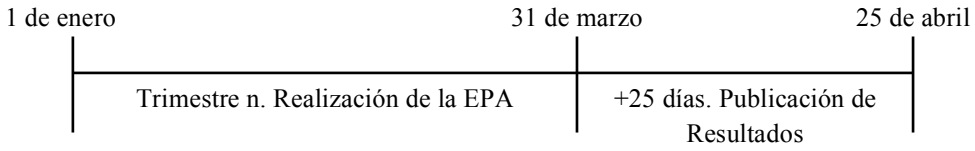
4.4.1 Hipótesis 1

H_1: "La popularidad del término de búsqueda *Paro* ayuda a predecir el número total de desempleados en España"

La hipótesis 1 **se puede aceptar** puesto que el valor β de GT1 es distinto de cero, existiendo una relación entre el número de personas desempleada en España y la cantidad de búsquedas relacionadas con el Paro.

A modo de conceptualizar las implicaciones económicas del presente modelo se presenta la Figura 9.

Figura 9. Esquema temporal de la publicación de los resultados de la EPA



Fuente: Elaboración propia

Como se puede apreciar, los resultados de la Encuesta de Población Activa tardan en ser publicados una media de 25 días. Con el modelo 1.a se podría predecir el número total de desempleados en España (con un error medio del 3,82%) el último día del trimestre *n*, adelantándose casi 1 mes a la publicación oficial de los resultados.

Ello implica un mayor margen de maniobra por parte del Gobierno y una mejor perspectiva de la situación económica del país. Además, se podrían llevar a cabo predicciones semanales con el volumen de búsquedas semanal.

4.4.2 Hipótesis 2

H_2: "La popularidad del término de búsqueda *Trabajo* ayuda a predecir el número total de desempleados en España."

El modelo 1.b, modelo que predice el total de desempleados en España a través del término de búsqueda *Trabajo*, no ha podido ser validado puesto que las perturbaciones no se distribuyen de manera normal (Véase Anexo 2). Ello implica que la **hipótesis 2 no puede ser aceptada**.

A diferencia de lo que sugerían diversos autores que han realizado estudios similares y han utilizado el término de búsqueda *Trabajo* (D'Amuri y Marcucci, 2009; Askitas y Zimmermann, 2009), en España dicho término no evoluciona de manera similar al total de desempleados que existen.

Los motivos por los cuales la Hipótesis 2 no puede ser aceptada podrían estar relacionados con la manera de buscar trabajo de los ciudadanos. Así pues, es más común buscar en Google el nombre de un portal de trabajo que el término de búsqueda *Trabajo* + *Ciudad/Puesto*… etc.

4.4.3 Hipótesis 3

H_3: "La popularidad término de búsqueda *Paro* mejora la predicción del número total de desempleados en España basados en los datos del desempleo disponibles."

Los modelos 2.a y 2.b han sido validados y, por lo tanto, su utilización es correcta. Sin embargo, como se esperaba, el modelo 2.a (modelo que incluye los datos de Google Trends para el término de búsqueda *Paro*) ha proporcionado mejores resultados. Dicho esto**, se puede aceptar la hipótesis 3**, puesto que los datos procedentes de Google Trends mejoran la predicción del número de desempleados en España.

Los resultados se encuentran recogidos en el epígrafe 4.2.2. Se puede comprobar que el modelo 2.a obtiene mejores resultados para el R^2 ajustado, para el AIC y para el MAE.

Por lo tanto, para realizar un *nowcasting* de la situación actual del paro sería conveniente formular un modelo en el que se recogieran los datos del SEPE y de GT, ya que su uso conjunto mejora la predicción.

4.4.4 Hipótesis 4

H_4: La popularidad del término de búsqueda *Cobrar Paro* está relacionada con aquellas personas que han trabajado anteriormente y llevan desempleadas poco tiempo.

Ambos modelos han logrado pasar las pruebas de validación y, por lo tanto, su poder predictivo ha sido corroborado. Sin embargo, como se esperaba, el modelo 3.b (modelo que incluye los datos de Google Trends para el término de búsqueda *Cobrar Paro*) ha proporcionado mejores resultados. Dicho esto, se puede **aceptar la hipótesis 4**, puesto que existe una relación entre la cantidad de gente que busca en Google el término *Cobrar Paro* y la cantidad de gente que ha sido despedida y lleva en paro menos de 3 meses

Los resultados se encuentran recogidos en el apartado 4.2.3. Se puede comprobar que el modelo 2.b obtiene mejores resultados para el R^2 ajustado, para el AIC y para el MAE.

Desde un punto de vista económico, la monitorización de este estrato del desempleo resulta de gran utilidad, puesto que un aumento en las búsquedas relativas a prestaciones económicas se corresponden con un aumento en el gasto público destinado al desempleo.

4.4.5 Hipótesis 5

H_5: La popularidad del término de búsqueda *Cobrar Paro* **no** está relacionada con aquellas personas que han trabajado anteriormente y llevan desempleadas mucho tiempo.

Tal y como se esperaba, el modelo 4.b no ha podido ser validado puesto que las perturbaciones no se distribuyen de manera normal. Ello implica que las búsquedas del término *Cobrar Paro* no están relacionadas con la cantidad de gente que ha sido despedida y lleva desempleada más de 3 meses. Por tanto, se puede **aceptar la hipótesis 5**.

La explicación reside en que las búsquedas relativas a información sobre prestaciones por desempleo se suelen efectuar en las primeras semanas de haber sido despedido, puesto que existe un límite de 15 días laborables para exigir dichas prestaciones.

Así pues, el hecho de que el modelo 4.b no sea válido ratifica en mayor manera el uso del anterior modelo (3.b) para monitorizar dicho estrato del desempleo.

CAPÍTULO 5

CONCLUSIONES

CAPÍTULO 5. CONCLUSIONES

El presente Trabajo Fin de Carrera ha demostrado la pertinencia del uso de Google Trends para predecir el nivel de desempleo en España. Así pues, se han confirmado los resultados de investigaciones previas con respecto al uso de Google Trends para predecir el nivel de desempleo.

La naturaleza innovadora del presente estudio se basa en predecir la estructura del desempleo, así como en validar el uso de Google Trends para países en los que el número de hogares con acceso a Internet no es muy elevado. Los resultados proporcionados demuestran que la información facilitada por Google Trends puede mejorar la predicción no solo del nivel de desempleo, sino también de sus componentes. A lo largo del presente trabajo se hace referencia a dos grandes bloques principalmente. En primer lugar, al nivel agregado de desempleo que conforma el nivel de Paro total en España y, en segundo lugar, al número de parados basados en el tiempo que llevan desempleados (menos de 3 meses o más). Para ello, es necesario descubrir los términos de búsquedas correctos e incluirlos en los respectivos modelos. De este modo, se concluyó que el término *Paro* era el idóneo para predecir el nivel global del desempleo y el término *Cobrar Paro* para predecir la cantidad de desempleados que demandarán prestaciones económicas.

Desde un punto de vista prospectivo, dichos modelos previamente presentados representan una oportunidad para monitorizar el ámbito macroeconómico español. Así pues, disponer de datos actualizados semanalmente permite un margen de maniobra más amplio y propicia el uso de políticas económicas más eficaces. A modo de conceptualización, el impacto de una reforma laboral podría apreciarse desde las primeras semanas aplicando dichos modelos, siendo innecesaria la espera de 3 meses para consultar los datos de la Encuesta de Población Activa.

Con respecto a las limitaciones que este trabajo ha tenido que hacer frente, se pueden resumir en dos grandes bloques. En primer lugar, la periodicidad del indicador principal del desempleo (la EPA) es trimestral, lo cual dificulta el uso de la información obtenida a través de Google Trends. Dicha dificultad consiste en tener que realizar medias aritméticas para ajustar los datos semanales de Google Trends con los datos trimestrales de la EPA. Por otra parte, de acuerdo con los datos del Eurostat en 2012, España se

encuentra por debajo de la media europea en el Índice de Hogares con Acceso a Internet. A pesar de que en los últimos años dicha brecha ha ido convergiendo, los datos procedentes de los primeros años de este estudio constataban esta gran diferencia, incrementando la dificultad de elaborar un modelo válido para toda la serie temporal. Sin embargo, a pesar de dichas limitaciones, el uso de Google Trends para predecir el nivel y la estructura del desempleo en España ha sido validado, demostrando así su robustez.

Como trabajo futuro se podría proponer la realización de modelos econométricos que distinguieran entre las diferentes Comunidades Autónomas. Actualmente, los datos ofrecidos para cada región de España no resultan del todo significativos. También se podría plantear la búsqueda de nuevos *keywords* que ayudaran a predecir nuevos segmentos del Paro. Finalmente, dicha metodología podría ser aplicada a otro tipo de variables económicas, como el nivel de PIB, el nivel de inflación, etc.

BIBLIOGRAFÍA

BIBLIOGRAFÍA

- AASTVEIT, K. A.; GERDRUP, K. R.; JORE, A. S.; THORSRUD, L. A. (2011) Nowcasting GDP in real-time. A density combination approach. *Norges Bank, trabajo en curso* nº 2011/11.

- AASTVEIT, K.; TROVIK, T. (2012) Nowcasting norwegian GDP: the role of asset prices in a small open economy. *Empirical Economics*, nº 42(1) pp. 95-119.

- ANDREOU, E.; GHYSELS, E.; KOURTELLOS, A. (2008) Should macroeconomic forecasters look at daily financial data? *Manuscrito*, Universidad de Chipre.

- ANGELINI, E.; BANBURA, M.; RÜNSLER, G. (2010) Estimating and forecasting the euro área monthly national accounts from a dynamic factor model. *OECD Journal of Business Cycle Measurement and Analysis*, nº 2010(1) pp. 7.

- ANGELINI, E.; CAMBA-MÉNDEZ, G.; GIANNONE, D.; REICHLIN, L.; RÜNSTLER, G. (2011) Short-Term forecasts of euro área GDP growth. *Econometrics Journal*, nº 14(1) pp. 25-44.

- ASKITAS, N.; ZIMMERANN, K.F. (2009) Google Econometrics and Unemployment Forecasting. *Applied Economics Quarterly*, nº 55(2) pp. 107-120.

- BANBURA, M; GIANNONE, D; MODUGNO, M; REICHLIN, L (2012). *Trabajo en curso ECARES* nº 2012-026.

- BANBURA, M.; MODUGNO, M. (2010) Maximum likehood estimation of large factor modelling datasets with arbitrary pattern of missing data. *Banco Central Europeo, trabajo en curso* nº 1189.

- BANBURA, M.; RÜNSTLER, G. (2011) A look into the factor model black box: Publitacion lags and the role of hard and soft data in forecasting GDP. *International Journal of Forecasting*, nº 27(2) pp. 333-346.

- BARHOUMI, K.; DARN, O. K.; FERRARA, L. (2010) Are disaggregate data useful for factor analysis in forecasting French GDP? *Journal of Forecasting*, nº 29 (1-2) pp. 132-144.

- CAMACHO, M.; PEREZ-QUIROS, G. (2010) Introducing the EURO-STING: Short term INdicator of Euro Area Growth. *Journal of Applied Econometrics*, 25(4) pp. 663-694.

- CHADWICK, M. G.; SENGUL, G. (2012) Nowcasting Unemployment Rate in Turkey: Let's Ask Google. *Banco Central de Turquía, trabajo en curso* nº12/18.

- CHOI, H.; VARIAN, H. (2009) Predicting the present with Google Trends. *Economic Record*, nº88 pp. 2-9.

- CORROCHER, N.; ORDANINI, A. (2002) Measuring the digital divide: a framework for the analysis of cross-country differences. *Journal of Information Technology*, nº 17(1) pp. 9-19.

- D'AMURI, F. (2009) Predicting unemployment in short samples with internet job search data. *MPRA, trabajo en curso* nº 18403.

- D'AMURI, F.; MARCUCCI, J. (2009) Google it! Forecasting the US Unemployment Rate with A Google Jobs Search Index. *FEEM, trabajo en curso* nº 31.2010.

- DA, Z.; ENGELBERG, J.; GAO, P. (2010) The sum of all FEARS: Investors sentiment and Asset Price. *Universidad de Notre Dame y Universidad de Carolina del Norte, Trabajo en curso.*

- DE ANTONIO, D.; FERNANDEZ, E. (2010) Nowcasting Spanish GDP Growth in real time: "one and a half months earlier". *Banco de España*, *documento de trabajo* nº1037.

- DE WINTER, J. (2011) Forecasting GDP growth in times of crisis: private sector forecasts vs. Statistical models. *Banco Central de Holanda, trabajo en curso* nº 320.

- DZIELINSKI, M. (2012) Measuring economic uncertainty and its impacto n the stock market. *Finance Research Letters*, nº 9 (3) pp. 167-175.

- ETTREDGE, M.; GERDES, J.; KARUGA, G. (2005) Using web-based search data to predict macroeconomic statistics. *Communications of the ACM*, nº 48(11) pp. 87-92.

- EUROSTAT (2012): Internet use in households and by individuals in 2012 (http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-SF-12-050/EN/KS-SF-12-050-EN.PDF) (Consulta: 14 de abril de 2013)

- GIANNONE, D.; REICHLIN, L.; SMALL, D. (2008) Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics*, nº 55(4) pp. 665-676.

- GILL, T.; PERERA, D.; SUNNER, D. (2012) Electronic Indicators of Economic Activity. *Reserve Bank of Australia Bulletin.* 2012 (Junio) pp. 1-11.

- GINSBERG, J.; MOHEBBI, M. H.; PATEL, R. S. ; BRAMMER, L.; SMOLINSKI, M. S.; BRILLIANT, L. (2009) Detecting influenza epidemics using search engine querey data. *Nature*, nº457 pp. 1012-1014.

- GOEL, S; HOFMAN, J. M.; LAHAIE, S.; PENNOCK, D. M.; WATTS, D. J. (2010) Predicting consumer behaviour with web search. *Proceedings of the National Academy of Sciencis*, nº 7 (41) pp. 17846-17490.

- GOOGLE (2012): Información sobre Google Trends (https://support.google.com/trends/topic/13973?hl=es&ref_topic=13761) (Consulta: 22 de abril de 2013)

- GUZMAN, G. (2011) Internet search behaviour as an economic forecasting tool. *The Journal of Economic and Social Measurament*, nº 36 (1-3) pp. 337-386.

- INSTITUTO NACIONAL DE ESTADÍSTICA (2009): Informe sobre la Encuesta de Población Activa (http://www.ine.es/docutrab/epa05_disenc/epa05_disenc.pdf) (Consulta: 18 de abril de 2013)

- INSTITUTO NACIONAL DE ESTADÍSTICA (2012): Encuesta de Población Activa (http://www.ine.es/jaxiBD/tabla.do?per=03&type=db&divi=EPA) (Varias consultas)

- LAHIRI, K.; MONOKROUSSOS, G. (2011) Nowcasting US GDP: The role of ISM Business Surveys. *SUNY trabajo en curso* nº 11-01.

- LIEBERMANN, J. (2012) Real-time forecasting in a data-rich environment. *MPRA*, *trabajo en curso* nº 39452

- MCLAREN, N.; SHANBOGHE, R. (2011) Using Internet Search Data as Economic Indicators. *Bank of England Quarterly Bulletin* 2011 Q2.

- ORDEN MINISTERIAL 11/3 1985. (Consulta: 14 de abril de 2013)

- PREIS, T.; REITH, D.; EUGENE, S. H. (2010) Complex dynamics of our economic life with different scales: insights from search engine query data. *Phil. Trans. R. Soc. A*, pp. 5707-5719.

- SERVICIO PÚBLICO DE EMPLEO ESTATAL (2012): Datos sobre el Paro Registrado (http://www.sepe.es/contenido/estadisticas/datos_avance/paro/) (Varias consultas)

- SUHOY, T. (2009) Query indices and a 2008 downturn. *Banco de Israel, trabajo en curso* nº 2009.06

ANEXO 1

ARTÍCULO: ON THE USE OF GOOGLE TRENDS TO NOWCAST THE UNEMPLOYMENT LEVEL AND STRUCTURE IN SPAIN

ON THE USE OF GOOGLE TRENDS DATA TO NOWCAST THE UNEMPLOYMENT LEVEL AND STRUCTURE IN SPAIN

JORGE REDONDO

Dept. Economía y Ciencias Sociales
Universitat Politecnica de Valencia
Camí de Vera, s/n
46022 Valencia

e-mail: jorreca @ade.upv.es
Telefono: +34 963877007

JOSEP DOMENECH

Dept. Economía y Ciencias Sociales
Universitat Politecnica de Valencia
Camí de Vera, s/n
46022 Valencia

e-mail: jdomenech@upvnet.upv.es
Telefono: +34 963877007

Abstract

Google Trends (GT) is increasingly being explored to nowcast a wide range of social and economic series, including macro variables such as the unemployment level. Previous research has successfully applied GT data to improve predictions on the unemployment level in countries with high digital literacy rates, such as USA and Germany.

This research work aims to apply GT-based prediction models to the unemployment level of Spain, which is a country that is lagging behind in the digital era and having particularly high unemployment rates. We pursue not only to validate the use of GT data for forecasting the aggregate unemployment level in Spain, but also to find correlations between different parts of the unemployment structure and GT series for different keywords.

Our results show that GT series can contribute to nowcast the unemployment rate in a labor market with particularly high unemployment level such as the prevalent in Spain. Our results also show that, by using different keywords, it is possible to anticipate separately different segments of the unemployment structure.

Key Words: Google Trends, Nowcasting, Unemployment, Spain.

Thematic Area: Quantitative Methods for Economics and Business.

Resumen

El uso de Google Trends (GT) como herramienta de *nowcast* se encuentra en fase de expansión. Ya se aplica con éxito a una gran variedad de series sociales y económicas, incluyendo variables macroeconómicas como el nivel de desempleo. Trabajos previos han empleado los datos de GT para mejorar notablemente la predicción del nivel de desempleo en países con elevada cultura digital, tales como EE.UU. o Alemania.

Este trabajo pretende aplicar modelos de predicción basados en GT al nivel de desempleo en España, país que parte con cierto retraso en la era digital y con tasas de desempleo particularmente altas. Aquí pretendemos no sólo validar el uso de datos de GT para predecir el nivel de desempleo en España, sino también encontrar correlaciones entre diferentes partes de la estructura del desempleo y distintas series de GT con diferentes palabras clave.

Los resultados experimentales muestran que las series de GT pueden contribuir a realizar un *nowcast* del nivel de desempleo en un mercado laboral con tasas de desempleo anormalmente altas, como las existentes en España. Estos resultados también muestran que, mediante el uso de distintas palabras clave, es posible anticipar la evolución de distintos segmentos de desempleados por separado.

Palabras clave: Google Trends, Nowcasting, Desempleo, España.

Área temática: Métodos Cuantitativos para la Economía y la Empresa.

1. INTRODUCTION

The Internet has become a second reality for human kind, in which we can see our thoughts and concerns reflected. Every day, the online activity generates tons of data, leaving a digital footprint on its way. Although such amounts of data may be overwhelming, there are tools that allow researchers to analyse and comprehend our behaviour from an economic and social point of view. Economic indicators built on online data have a number of advantages over the traditional ways. Most of them are related to the fact that data for these indicators are inexpensive to collect and, therefore, the indicators produced are timely. This allows real time analysis and intervention at both business and policy levels (Varian, 2010).

In this context, there is a growing literature dealing with the systematic extraction of real-economy indicators from the internet. One of the popular approaches to do so is to employ Google Trends (GT), which is an online service that provides the evolution of the popularity of queries in Google search engine. This tool has given researchers the opportunity to generate real-time indicators for a wide variety of purposes, such as estimating the evolution of flu activity (Ginsberg et al. 2009), forecasting retail sales (Choi and Varian 2009) or predicting transaction volumes in the stock market (Preis et al. 2010). About the unemployment level, previous research (D'Amuri and Marcucci 2009; Askitas and Zimmermann 2009) demonstrated the pertinence of using Google Trends to improve the prediction of the level of unemployment in countries which have a high level of digital literacy such as the USA and Germany (Corrocher and Ordanini 2002).

This paper aims to apply GT-based prediction models to the unemployment level of Spain, which is a country that is lagging behind in the digital era and having particularly high unemployment rates. We pursue not only to validate the use of GT data for forecasting the aggregate unemployment level in Spain, but also to find correlations between different parts of the unemployment structure and GT series for different keywords.

This paper is organized as follows. Section 2 reviews recent research on Google Trends and its application to nowcasting the real world. Section 3 discusses the research method and findings from the data analysis. Finally, Section 4 presents some concluding remarks.

2. BACKGROUND

Google Trends is an online tool provided by Google Inc. that generates a time series index of the volume of the queries that users enter into Google in a certain geographic area. Since 2009, GT has been the topic of a larger number of articles with a wide variety of purposes. For example, in epidemiology, Polgreen et al. (2008) and Ginsberg et al. (2009) presented GT as a tool to improve the prediction of influenza-like diseases. In the field of economics, Choi and Varian (2009) used GT to predict different economic metrics: initial claims of unemployment, automobile demand and vacation destinations. Similarly, McLaren and Sanbhongue (2011) analysed the UK housing and labour markets, whereas Gill et al. (2012) focused on a variety of aspects of the Australian economic activity, both using online-query data.

Using Google Trends to help predict the level of unemployment has been a recurrent topic in the literature, especially applied to countries with a high digital literacy. In this sense, the relation of GT to the level of unemployment was found to be significant in Germany (Askitas and Zimmermann 2009), the United States (D'Amuri and Marcucci 2009), and the United Kingdom (McLaren and Shanbhogue 2011). Other countries for which this relation has been explored include Israel (Suhoy 2009), Italy (D'Amuri 2009) and Turkey (Chadwick et al. 2012). In all of them Google Trends was successfully applied to improve the forecasting of the unemployment level.

A wide variety of terms and categories has been used to find the GT data that better predict the unemployment level. On the one hand, Choi and Varian (2009) considered all the terms under categories "jobs" and "welfare" for their research about initial claims in US. Similarly, D'Amuri (2009) used the category "job offers", in which all queries regarding job offers are included. On the other hand, D'Amuri and Marcucci (2009) employed the keyword "jobs" due to the fact that represented the highest incidence among different job-search-related keywords. Likewise, McLaren and Sanbhongue (2011) used the terms "jobs", "unemployment" and "jobseeker allowance".

A different approach is considered by Askitas and Zimmermann (2009), since they considered four groups of terms to predict the aggregate level of unemployment in Germany. The main novelty relies on the fact that each group of terms is presumably related to some inflows or outflows in the labour market. However, the prediction power of these groups of terms is only checked against the aggregate unemployment level, and not against each different segment of the unemployment. Our research pursues to confirm that different query terms are related to different parts of the structure of the unemployment, thus allowing a more efficient design of the nowcast models.

3. METHODS

3.1. DATA

This section describes the data used to carry out this research. We introduce the Labour Force Survey (LFS) time series and explain the use of Google Trends, as well as the keywords employed.

The unemployment level was measured by means of the results of the Labour Force Survey (LFS), which is carried out in Spain by the National Institute of Statistics (see Figure 1). It is conducted quarterly and its results are published about 25 days after the end of each quarter. To relate the evolution of web searches to the unemployment structure, the LFS series were broken down into two: LFS_L3M represents the number of persons who are unemployed for less than three months; while LFS_M3M represents those who are unemployed for more than three months. Figures 2(a) and 2(b) shows the evolution of these variables and Table 1 presents their main statistics.

Table 1. Descriptive statistics on the LFS variables

	Mean	Std. Dev.	Min	Max
LFS	3384.85	1487.57	1760.00	5778.10
LFS_M3M	2685.79	1395.97	1037.50	5000.30
LFS_L3M	650.06	167.58	391.30	906.01

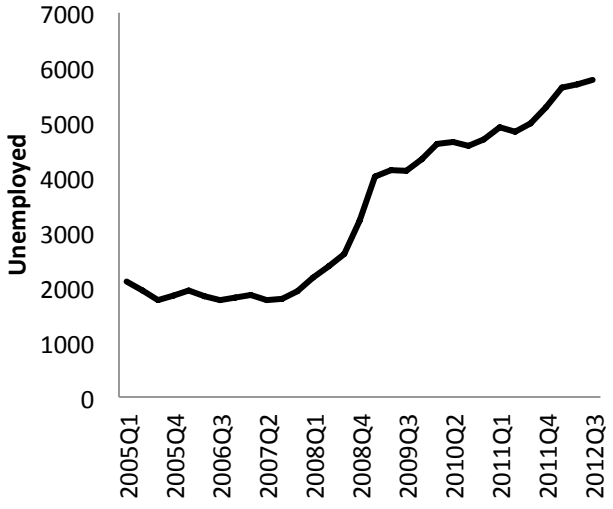
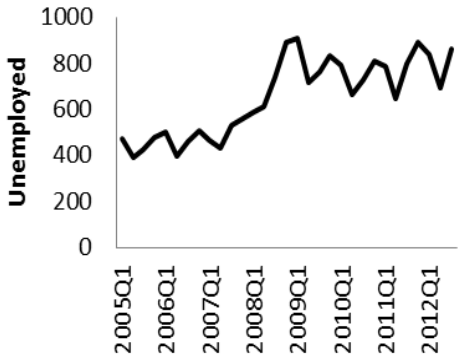


Figure 1. Evolution of the unemployment level in Spain (LFS), in thousands.

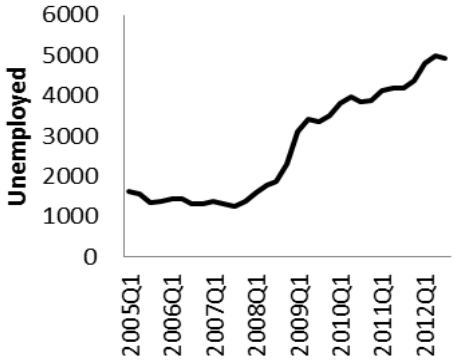
Google Trends provides series gathering the volume of searches for a given query in a certain geographic area. Data are presented as a normalized index ranging from 0 to 100 rather than in absolute number of searches. Unlike LFS series, the frequency of GT series is weekly. Hence,

the average of the weekly data during the quarter was computed to allow comparisons of both series, in line with prior research work, e.g., D'Amuri (2009).

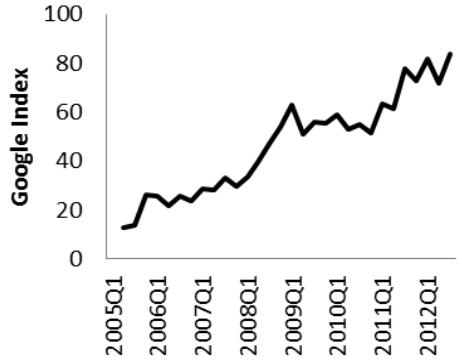
GT series for two different terms were selected as predictors for the LFS data (see Table 2). GT1 is expected to be connected with the evolution of the overall amount of unemployed (LFS), while GT2 is presumably related to only those people recently unemployed who are looking for information about how to receive unemployment benefits. The evolution of both search terms is shown in Figures 2(b) and 2(c). It can be observed that GT1 follows a pattern similar to LFS and LFS_M3M, while GT2 seems closer to LFS_L3M.



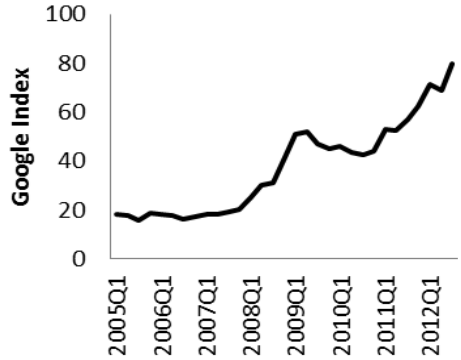
(a) LFS_L3M: Persons who are less than three months unemployed (in thousand)



(b) LFS_M3M: Persons who are more than three months unemployed (in thousand)



(c) GT1: Google Trends series for the term "unemployment"



(d) GT2: Google Trends series for the term "receive unemployment benefits"

Figure 2. Time series plots of the LFS components and of their corresponding queries in Google Trends

Table 2. Terms used to build Google Trends based series

Variable	Term	Translation
GT1	Paro	Unemployment
GT2	Cobrar Paro	Receive unemployment benefits

3.2. GT AND UNEMPLOYMENT LEVEL IN SPAIN

This section checks the use of Google Trends data as predictor of the overall unemployment level in Spain. Following the methods used by Choi and Varian (2011), a simple autoregressive model is compared to a similar model augmented by including GT data:

$$(1)\quad LFS_t = \beta_0 + \beta_1 LFS_{t-1} + e_t$$

$$(2)\quad LFS_t = \beta_0 + \beta_1 LFS_{t-1} + \beta_2 GT1_t + e_t$$

This way, Model (1) works as a baseline model, while Model (2) measures the extent to which GT can improve the prediction of the level of unemployment level.

Table 3 reports the estimation results of Equations (1) and (2). As one can observe, the statistically significant coefficient for GT1 in Model (2) points out that this variable contributes to improving the prediction of the unemployment level. The power of GT1 in out-of-sample forecasting is evidenced by the fact that the mean absolute error (MAE) drops from 5.03 percent in (1) to 3.76 in Model (2). That is, the inclusion of this variable represents an improvement of 25.2 percent.

3.3. INSIGHTS ON THE STRUCTURE OF THE UNEMPLOYMENT

This section pursues to show how GT data can provide some insights on the structure of the unemployment. To this end, a separate analysis on each unemployment segment is carried out. By dividing this series into two parts, it is possible to identify the most appropriate query term for predicting each unemployment segment. The rationale behind this division is that people who have been unemployed for short time are more likely to have been concerned with the procedure for claiming unemployment benefits. Thus, GT2 is expected to be better predictor for the recent unemployed segment. However, people who have been unemployed for longer time rarely look for this kind of information, making GT1 predict better this segment of the unemployment.

The following equations show the different combinations of GT terms and unemployment segments considered in the analysis:

$$(3)\quad LFS\_L3M_t = \beta_0 + \beta_1 LFS\_L3M_{t-1} + \beta_2 GT1_t + e_t$$

$$(4)\quad LFS\_L3M_t = \beta_0 + \beta_1 LFS\_L3M_{t-1} + \beta_2 GT2_t + e_t$$

$$(5)\quad LFS\_M3M_t = \beta_0 + \beta_1 LFS\_M3M_{t-1} + \beta_2 GT1_t + e_t$$

$$(6)\quad LFS\_M3M_t = \beta_0 + \beta_1 LFS\_M3M_{t-1} + \beta_2 GT2_t + e_t$$

The estimates for these models are shown in Table 3. The results for Models (3) and (4) exhibit that, as expected, GT2 predicts the number of people who are unemployed for less than 3 months better than GT1. When comparing out-of-sample forecasts, the mean absolute error of Model (4) is 1.54 percentage points lower than Model (3). This means that, in relative terms, the MAE is reduced by 16.2% if GT2 is used instead of GT1.

However, the opposite is found when predicting the evolution of the number of unemployed people for longer time. Models (5) and (6) pursue to predict this variable (LFS_M3M) using GT1 and GT2, respectively. Results show that the mean average error when including GT1 is 27% lower than the model including GT2. Even more, the estimates for Model (6) exhibits that the coefficient for GT2 is non-significant.

Table 3. Estimation results for the models

	(1)	(2)	(3)	(4)	(5)	(6)
Variables	*LFS*	*LFS*	*LFS_L3M*	*LFS_L3M*	*LFS_M3M*	*LFS_M3M*
GT1		35.966*	6.981*		28.128*	
GT2				7.132*		1.382
AR(1)	1.024*	0.994*	0.528*	0.513*	0.996*	0.998*
Constant	-1787.836*	10991.970*	389.168*	322.801*	14341.090*	54197.080
N	30	30	30	30	30	30
R^2 adj.	0.980	0.990	0.787	0.832	0.983	0.976
AIC	13.619	13.001	11.695	11.402	13.294	13.695
MAE	5.032	3.759	9.569	8.022	4.989	6.838

* p<0.05

4. CONCLUSIONS

This paper has shown the pertinence of Google Trends data for nowcasting the unemployment level in Spain, which lags behind in the digital era and has particularly high unemployment rates. This way, the results of previous research concerning the relation of GT with the labour market have been confirmed in a more extreme case of unemployment.

Our work also extends prior research by considering the structure of the unemployment. The results provided demonstrate that GT data can improve not only the prediction of the unemployment level, but also the forecasts of its components. To do so, the proper query terms for each segment have to be discovered and included in each predictive model. The keyword we used for short-term unemployment forecasting was quite intuitive, as long as the procedure for claiming unemployment benefits is mainly demanded by the person who lost their job recently.

REFERENCES

ASKITAS, N.; ZIMMERMANN K. F. (2009). Google econometrics and unemployment forecasting. *Applied Economics Quarterly*. 55(2), 107-120.

CHADWICK, M. G.; SENGUL G. (2012). Nowcasting Unemployment Rate in Turkey: Let's Ask Google. *Central Bank of Turkey. Working paper* No.12/18.

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

CHOI, H.; VARIAN, H. (2009b). Predicting Initial Claims for Unemployment Insurance Using Google Trends Technical report, Google. Available from: http://research.google.com/archive/papers/initialclaimsUS.pdf.

CHOI, H.; VARIAN H. (2012). Predicting the Present with Google Trends. *Economic Record*. 88, 2 - 9.

CORROCHER, N.; ORDANINI A. (2002). Measuring the digital divide: a framework for the analysis of cross-country differences. *Journal of Information Technology*. 17(1), 9 - 19.

D'AMURI, F. (2009). Predicting unemployment in short samples with internet job search query data," *MPRA Paper* No: 18403.

D'AMURI, F.; J. MARCUCCI (2009). "Google It!' Forecasting the US Unemployment Rate with A Google Job Search Index," *FEEM Working Paper* No. 31.2010.

GILL, T., PERERA D.; SUNNER D. (2012). Electronic Indicators of Economic Activity. *Reserve Bank of Australia Bulletin*. 2012(June Quarter), 1-11.

GINSBERG, J., MOHEBBI M. H., PATEL R. S., BRAMMER L., SMOLINSKI M. S.; BRILLIANT L. (2009). Detecting influenza epidemics using search engine query data. *Nature*. 457(7232), 1012 - 1014.

MCLAREN, N.; SHANBHOGUE R. (2011). Using Internet Search Data as Economic Indicators. *Bank of England Quarterly Bulletin*. 2011 Q2.

POLGREEN, P. M., CHEN Y., PENNOCK D. M.; NELSON F. D. (2008). Using Internet Searches for Influenza Surveillance. *Clinical Infectious Diseases*. 47(11), 1443 – 1448.

PREIS, T., REITH D.; STANLEY E. H. (2010). Complex dynamics of our economic life on different scales: insights from search engine query data. *Philosophical Transactions Of The Royal Society A-Mathematical Physical And Engineering Sciences*. 5707-5719.

SUHOY, T. (2009). Query Indices and a 2008 Downturn: Israeli Data, *Technical report, Bank of Israel.* Available from: http://www.bankisrael.gov.il/deptdata/mehkar/papers/dp0906e.pdf.

VARIAN, H. R. (2010). Computer Mediated Transactions. *American Economic Review*. 100(2), 1 - 10.

ANEXO 2

VALIDACIONES ECONOMÉTRICAS

Validación Modelo 1.b

- **Contraste de errores de especificación**

Figura 10 Modelo 1.b Test RESET

Ramsey RESET Test:

F-statistic	4.647097	Probability	0.040180
Log likelihood ratio	4.923096	Probability	0.026500

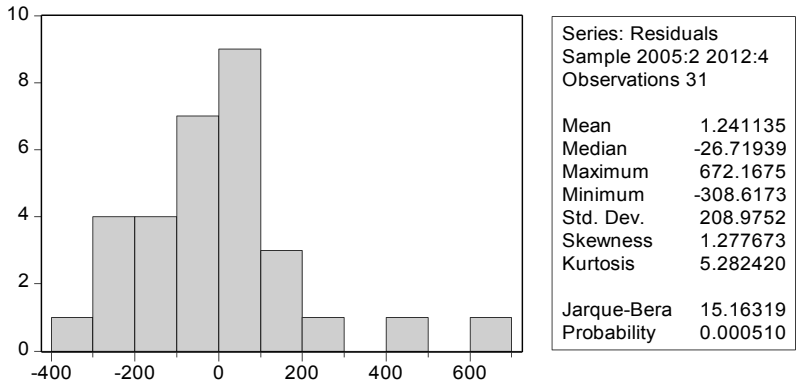
Test Equation:
Dependent Variable: EPA
Method: Least Squares
Date: 06/02/13 Time: 18:03
Sample: 2005:2 2012:4
Included observations: 31
Convergence achieved after 31 iterations

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	15276.35	105805.2	0.144382	0.8863
GT2	-0.402352	4.254231	-0.094577	0.9253
FITTED^2	5.30E-05	2.83E-05	1.868415	0.0726
AR(1)	0.994288	0.046595	21.33900	0.0000
R-squared	0.984237	Mean dependent var		3509.574
Adjusted R-squared	0.982485	S.D. dependent var		1537.422
S.E. of regression	203.4681	Akaike info criterion		13.58881
Sum squared resid	1117780.	Schwarz criterion		13.77384
Log likelihood	-206.6265	F-statistic		561.9442
Durbin-Watson stat	1.459854	Prob(F-statistic)		0.000000
Inverted AR Roots	.99			

Como el valor estimado por Eviews para FITTED^2 es 0,07 > α = 0,05, se acepta H_0, es decir, FITTED^2 no es significativo y, por tanto, se concluye que no existen errores de especificación.

- **Análisis de la normalidad de las perturbaciones**

Figura 11 Modelo 1.b Jarque-Bera



EViews ofrece el cálculo del estadístico Jarque-Bera. Este estadístico presenta un valor de 15,16 > 5,99, por lo que se rechaza H_0 y se concluye que las perturbaciones no se distribuyen conforme a la distribución normal. Este resultado también indica que el modelo no resulta válido.

Validación Modelo 2.A

- **Contraste de errores de especificación**

Figura 12 Modelo 2.a Test Reset

Ramsey RESET Test:

F-statistic	0.034846	Probability	0.853265
Log likelihood ratio	0.039799	Probability	0.841874

Test Equation:
Dependent Variable: EPA
Method: Least Squares
Date: 06/02/13 Time: 18:25
Sample: 2005:1 2012:4

Included observations: 32

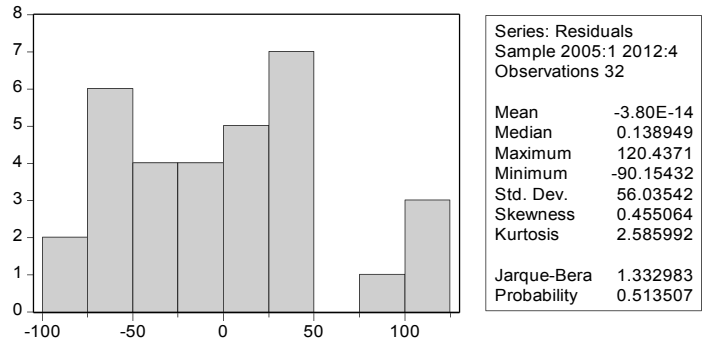
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-830.5378	128.4002	-6.468351	0.0000
GT1	11.47455	2.272559	5.049174	0.0000
SEPE	0.001208	6.62E-05	18.24807	0.0000
FITTED^2	1.49E-06	7.96E-06	0.186670	0.8533

R-squared	0.998665	Mean dependent var	3465.494
Adjusted R-squared	0.998522	S.D. dependent var	1532.840
S.E. of regression	58.92429	Akaike info criterion	11.10685
Sum squared resid	97218.02	Schwarz criterion	11.29007
Log likelihood	-173.7096	F-statistic	6983.373
Durbin-Watson stat	1.537256	Prob(F-statistic)	0.000000

Como el valor estimado por Eviews para FITTED^2 es 0,85 > α = 0,05, se acepta H_0, es decir, FITTED^2 no es significativo y, por tanto, se concluye que no existen errores de especificación. (SEPE = PR)

- **Análisis de la normalidad de las perturbaciones**

Figura 13 Modelo 2.a Jarque-Bera



EViews ofrece el cálculo del estadístico Jarque-Bera. Este estadístico presenta un valor de 1,33 <5,99, por lo que se acepta H_0 y se concluye que las perturbaciones se distribuyen conforme a la distribución normal. Este resultado también indica que el resto de contrastes que se van a efectuar son fiables, por lo que sí se puede realizar inferencia en el modelo.

- **Contrastes de significatividad**

En este epígrafe se realizan los contrastes de significatividad de los parámetros del modelo, primero de forma individual para cada parámetro y después, de forma conjunta.

Figura 14 Modelo 2.a Estimaciones

Dependent Variable: EPA
Method: Least Squares
Date: 06/02/13 Time: 18:25
Sample: 2005:1 2012:4
Included observations: 32

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-852.8755	45.77175	-18.63323	0.0000
GT1	11.73067	1.781234	6.585696	0.0000
SEPE	0.001219	3.33E-05	36.60080	0.0000
R-squared	0.998664	Mean dependent var		3465.494
Adjusted R-squared	0.998571	S.D. dependent var		1532.840
S.E. of regression	57.93546	Akaike info criterion		11.04560
Sum squared resid	97339.00	Schwarz criterion		11.18301
Log likelihood	-173.7295	F-statistic		10835.67
Durbin-Watson stat	1.521599	Prob(F-statistic)		0.000000

Contrastes de significatividad de β_1

Como el valor proporcionado por EViews para $\alpha^*(t^*_{\beta 1})$ es 0,00 < α = 0,05, se rechaza H_0, es decir, β_1 sí que es significativo.

Contrastes de significatividad de β_2

Como el valor proporcionado por EViews para $\alpha^*(t^*_{\beta 2})$ es 0,00 < α = 0,05, se rechaza H_0, es decir, β_2 sí que es significativo.

Contrastes de significatividad conjunta

Como el valor estimado por EViews para la probabilidad asociada al estadístico F* es 0,00 < α = 0,05, se rechaza H_0, es decir, sí que existe regresión.

- **Análisis de la heterocedasticidad**

Figura 15 Modelo 2.a Estadístico White

White Heteroskedasticity Test:

F-statistic	0.830640	Probability	0.517477
Obs*R-squared	3.506364	Probability	0.476911

Test Equation:
Dependent Variable: RESID^2
Method: Least Squares
Date: 06/02/13 Time: 18:27
Sample: 2005:1 2012:4

Included observations: 32

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-16955.73	23391.06	-0.724881	0.4748
GT1	-644.9619	655.1131	-0.984505	0.3336
GT1^2	4.321528	6.182305	0.699016	0.4905
SEPE	0.019049	0.021934	0.868473	0.3928
SEPE^2	-2.12E-09	3.06E-09	-0.691102	0.4954
R-squared	0.109574	Mean dependent var		3041.844
Adjusted R-squared	-0.022341	S.D. dependent var		3892.078
S.E. of regression	3935.314	Akaike info criterion		19.53597
Sum squared resid	4.18E+08	Schwarz criterion		19.76499
Log likelihood	-307.5755	F-statistic		0.830640
Durbin-Watson stat	2.157102	Prob(F-statistic)		0.517477

EViews ofrece el cálculo del estadístico de White. Este estadístico presenta un valor de $3{,}50 < \chi^2_4 = 9{,}49$, por lo que se acepta H_0 y se concluye que existe homocedasticidad.

- **Análisis de la autocorrelación**

Figura 16 Modelo 2.a Análisis de la autocorrelación

Date: 06/02/13 Time: 18:28
Sample: 2005:1 2012:4
Included observations: 32

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.050	0.050	0.0877	0.767
		2	0.289	0.287	3.1068	0.212
		3	-0.021	-0.050	3.1233	0.373
		4	0.186	0.116	4.4637	0.347
		5	-0.179	-0.193	5.7607	0.330
		6	0.001	-0.067	5.7607	0.451
		7	-0.117	-0.012	6.3614	0.498
		8	-0.102	-0.124	6.8361	0.554

Se puede apreciar que no existe autocorrelación, puesto que no se sobrepasan las líneas límite de la columna "Partial Correlation" (Este procedimiento será llevado a cabo a lo largo de todo el Anexo 2). Así pues y, de acuerdo con todo lo establecido en el apartado 4.3, se puede validar el uso del modelo 2.a y corroborar su poder predictivo.

Validación Modelo 2.b

- **Contraste de errores de especificación**

Figura 17 Modelo 2.b Test RESET

Ramsey RESET Test:

F-statistic	0.860461	Probability	0.361537
Log likelihood ratio	0.938308	Probability	0.332713

Test Equation:
Dependent Variable: EPA
Method: Least Squares
Date: 06/02/13 Time: 18:53
Sample: 2005:2 2012:4
Included observations: 31

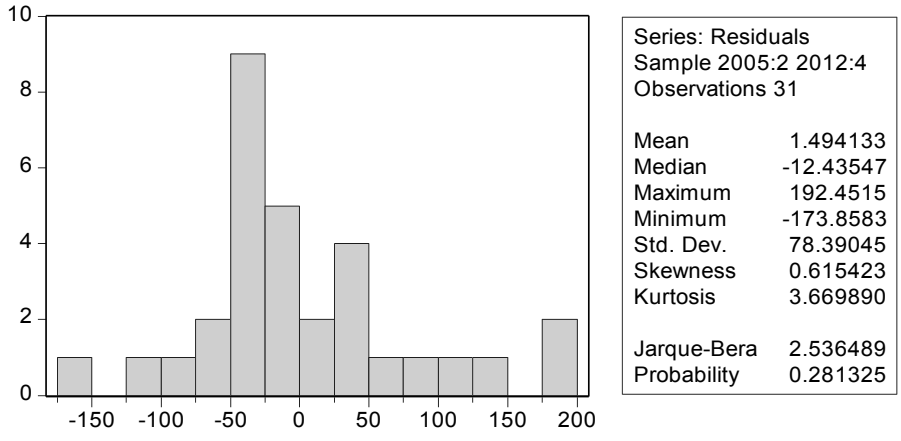
Convergence achieved after 21 iterations

Variable	Coefficient	Std. Error	t-Statistic	Prob.
SEPE	0.001394	0.000197	7.075138	0.0000
FITTED^2	-1.56E-05	1.73E-05	-0.898930	0.3764
AR(1)	0.977546	0.024871	39.30528	0.0000
R-squared	0.997477	Mean dependent var		3509.574
Adjusted R-squared	0.997297	S.D. dependent var		1537.422
S.E. of regression	79.93807	Akaike info criterion		11.69215
Sum squared resid	178922.7	Schwarz criterion		11.83092
Log likelihood	-178.2283	Durbin-Watson stat		1.790857
Inverted AR Roots	.98			

Como el valor estimado por Eviews para FITTED^2 es 0,38 > α = 0,05, se acepta H_0, es decir, FITTED^2 no es significativo y, por tanto, se concluye que no existen errores de especificación.

- **Análisis de la normalidad de las perturbaciones**

Figura 18 Modelo 2.b Jarque-Bera



EViews ofrece el cálculo del estadístico Jarque-Bera. Este estadístico presenta un valor de 2,53 <5,99, por lo que se acepta H_0 y se concluye que las perturbaciones se distribuyen conforme a la distribución normal. Este resultado también indica que el resto de contrastes que se van a efectuar son fiables, por lo que sí se puede realizar inferencia en el modelo.

- **Contrastes de significatividad**

En este epígrafe se realizan los contrastes de significatividad de los parámetros del modelo, primero de forma individual para cada parámetro y después, de forma conjunta.

Figura 19 Modelo 2.b Estimaciones

Dependent Variable: EPA
Method: Least Squares
Date: 06/02/13 Time: 18:53
Sample(adjusted): 2005:2 2012:4
Included observations: 31 after adjusting endpoints
Convergence achieved after 11 iterations

Variable	Coefficient	Std. Error	t-Statistic	Prob.
SEPE	0.001243	7.15E-05	17.38006	0.0000
AR(1)	0.969452	0.033126	29.26520	0.0000
R-squared	0.997399	Mean dependent var		3509.574
Adjusted R-squared	0.997310	S.D. dependent var		1537.422
S.E. of regression	79.74552	Akaike info criterion		11.65790
Sum squared resid	184421.1	Schwarz criterion		11.75041
Log likelihood	-178.6974	Durbin-Watson stat		1.911378
Inverted AR Roots	.97			

Contrastes de significatividad de β_1

Como el valor proporcionado por EViews para $\alpha^*(t^*_{\beta 1})$ es 0,00 < α = 0,05, se rechaza H_0, es decir, β_1 sí que es significativo.

Contrastes de significatividad de β_2

Como el valor proporcionado por EViews para $\alpha^*(t^*_{\beta 2})$ es 0,00 < α = 0,05, se rechaza H_0, es decir, β_2 sí que es significativo.

Contrastes de significatividad conjunta

Como el valor estimado por EViews para la probabilidad asociada al estadístico F* es 0,00 < α = 0,05, se rechaza H_0, es decir, sí que existe regresión.

- **Análisis de la heterocedasticidad**

Figura 20 Modelo 2.b Estadístico White

White Heteroskedasticity Test:

F-statistic	0.547933	Probability	0.584217
Obs*R-squared	1.167583	Probability	0.557780

Test Equation:
Dependent Variable: RESID^2
Method: Least Squares
Date: 06/02/13 Time: 18:55
Sample: 2005:2 2012:4

Included observations: 31

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-13236.71	27550.32	-0.480456	0.6346
SEPE	0.011348	0.018656	0.608305	0.5479
SEPE^2	-1.51E-09	2.86E-09	-0.527025	0.6023

R-squared	0.037664	Mean dependent var	5949.067
Adjusted R-squared	-0.031074	S.D. dependent var	9968.222
S.E. of regression	10121.91	Akaike info criterion	21.37456
Sum squared resid	2.87E+09	Schwarz criterion	21.51333
Log likelihood	-328.3057	F-statistic	0.547933
Durbin-Watson stat	2.423665	Prob(F-statistic)	0.584217

EViews ofrece el cálculo del estadístico de White. Este estadístico presenta un valor de $1{,}16 < \chi^2_4 = 9{,}49$, por lo que se acepta H_0 y se concluye que existe homocedasticidad.

- **Análisis de la autocorrelación**

Figura 21 Modelo 2.b Análisis de la autocorrelación

Date: 06/02/13 Time: 19:03
Sample: 2005:2 2012:4
Included observations: 31
Q-statistic probabilities adjusted for 1 ARMA term(s)

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	-0.034	-0.034	0.0387	
		2	-0.002	-0.003	0.0388	0.844
		3	-0.221	-0.221	1.8230	0.402
		4	0.331	0.333	5.9726	0.113
		5	-0.199	-0.240	7.5269	0.111
		6	0.115	0.129	8.0661	0.153
		7	-0.157	-0.067	9.1229	0.167
		8	0.140	-0.031	9.9987	0.189

Se puede apreciar que no existe autocorrelación. Por consiguiente y, de acuerdo con todo lo establecido, se puede validar el uso del modelo 2.b y corroborar su poder predictivo.

Validación Modelo 3.a

- **Contraste de errores de especificación**

Figura 22 Modelo 3.a Test RESET

Ramsey RESET Test:

F-statistic	0.331020	Probability	0.569823
Log likelihood ratio	0.377749	Probability	0.538810

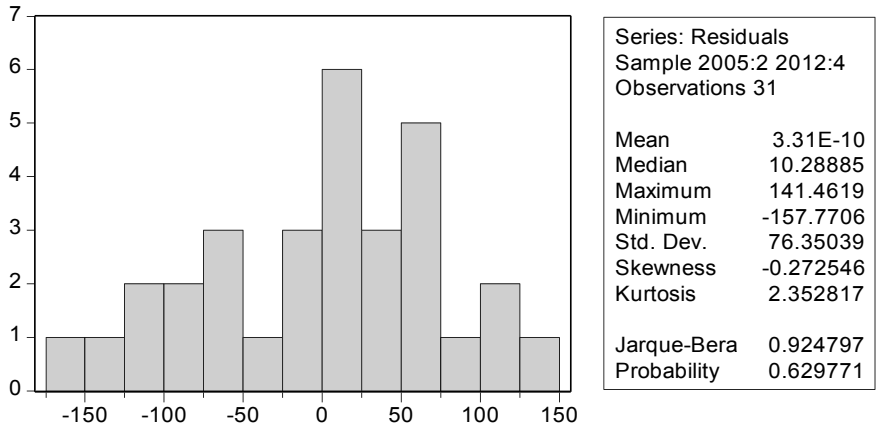
Test Equation:
Dependent Variable: EPA_L3M
Method: Least Squares
Date: 06/02/13 Time: 18:59
Sample: 2005:2 2012:4
Included observations: 31
Convergence achieved after 31 iterations

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	369.8702	76.50836	4.834376	0.0000
GT1	5.436506	5.304538	1.024878	0.3145
FITTED^2	0.000178	0.000538	0.330363	0.7437
AR(1)	0.454379	0.344238	1.319954	0.1979
R-squared	0.804457	Mean dependent var		664.8677
Adjusted R-squared	0.782730	S.D. dependent var		171.6104
S.E. of regression	79.99152	Akaike info criterion		11.72163
Sum squared resid	172763.4	Schwarz criterion		11.90666
Log likelihood	-177.6853	F-statistic		37.02560
Durbin-Watson stat	1.680873	Prob(F-statistic)		0.000000
Inverted AR Roots	.45			

Como el valor estimado por Eviews para FITTED^2 es 0,74 > α = 0,05, se acepta H_0, es decir, FITTED^2 no es significativo y, por tanto, se concluye que no existen errores de especificación.

- **Análisis de la normalidad de las perturbaciones**

Figura 23 Modelo 3.a Jaque-Bera



EViews ofrece el cálculo del estadístico Jarque-Bera. Este estadístico presenta un valor de 0,92 <5,99, por lo que se acepta H_0 y se concluye que las perturbaciones se distribuyen conforme a la distribución normal. Este resultado también indica que el resto de contrastes que se van a efectuar son fiables, por lo que sí se puede realizar inferencia en el modelo.

- **Contrastes de significatividad**

En este epígrafe se realizan los contrastes de significatividad de los parámetros del modelo, primero de forma individual para cada parámetro y después, de forma conjunta.

Figura 24 Modelo 3.a Estimaciones

Dependent Variable: EPA_L3M
Method: Least Squares
Date: 06/02/13 Time: 18:56
Sample(adjusted): 2005:2 2012:4
Included observations: 31 after adjusting endpoints
Convergence achieved after 6 iterations

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	382.1206	63.09320	6.056447	0.0000
GT1	7.234961	1.360275	5.318748	0.0000
AR(1)	0.506451	0.164447	3.079716	0.0046
R-squared	0.802059	Mean dependent var		664.8677
Adjusted R-squared	0.787921	S.D. dependent var		171.6104
S.E. of regression	79.03016	Akaike info criterion		11.66930
Sum squared resid	174881.5	Schwarz criterion		11.80807
Log likelihood	-177.8742	F-statistic		56.72825
Durbin-Watson stat	1.661749	Prob(F-statistic)		0.000000
Inverted AR Roots	.51			

Contrastes de significatividad de β_1

Como el valor proporcionado por EViews para $\alpha^*(t^*_{\beta 1})$ es 0,00 < α = 0,05, se rechaza H_0, es decir, β_1 sí que es significativo.

Contrastes de significatividad de β_2

Como el valor proporcionado por EViews para $\alpha^*(t^*_{\beta 2})$ es 0,00 < α = 0,05, se rechaza H_0, es decir, β_2 sí que es significativo.

Contrastes de significatividad conjunta

Como el valor estimado por EViews para la probabilidad asociada al estadístico F* es 0,00 < α = 0,05, se rechaza H_0, es decir, sí que existe regresión.

- **Análisis de la heterocedasticidad**

Figura 25 Modelo 3.a Estadístico White

White Heteroskedasticity Test:

F-statistic	2.191599	Probability	0.130538
Obs*R-squared	4.195977	Probability	0.122703

Test Equation:
Dependent Variable: RESID^2
Method: Least Squares
Date: 06/02/13 Time: 18:59
Sample: 2005:2 2012:4

Included observations: 31

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-4600.633	5363.667	-0.857740	0.3983
GT1	487.2000	284.4304	1.712897	0.0978
GT1^2	-4.606146	3.225000	-1.428262	0.1643
R-squared	0.135354	Mean dependent var		5641.337
Adjusted R-squared	0.073594	S.D. dependent var		6669.938
S.E. of regression	6419.816	Akaike info criterion		20.46393
Sum squared resid	1.15E+09	Schwarz criterion		20.60271
Log likelihood	-314.1909	F-statistic		2.191599
Durbin-Watson stat	2.282678	Prob(F-statistic)		0.130538

EViews ofrece el cálculo del estadístico de White. Este estadístico presenta un valor de $4{,}19 < \chi^2_4 = 9{,}49$, por lo que se acepta H_0 y se concluye que existe homocedasticidad.

- **Análisis de la autocorrelación**

Figura 26 Modelo 3.a Análisis de la autocorrelación

Date: 06/02/13 Time: 19:00
Sample: 2005:2 2012:4
Included observations: 31
Q-statistic probabilities adjusted for 1 ARMA term(s)

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	-0.039	-0.039	0.0506	
		2	-0.010	-0.012	0.0542	0.816
		3	-0.075	-0.076	0.2608	0.878
		4	0.253	0.249	2.6894	0.442
		5	-0.197	-0.197	4.2129	0.378
		6	0.017	0.020	4.2242	0.518
		7	-0.160	-0.148	5.3188	0.504
		8	0.008	-0.082	5.3217	0.621

Se puede apreciar que no existe autocorrelación. Por consiguiente y, de acuerdo con todo lo establecido, se puede validar el uso del modelo 3.a y corroborar su poder predictivo.

Validación Modelo 3.a

- **Contraste de errores de especificación**

Figura 27 Modelo 3.b Test RESET

Ramsey RESET Test:

F-statistic	0.462698	Probability	0.502157
Log likelihood ratio	0.526745	Probability	0.467979

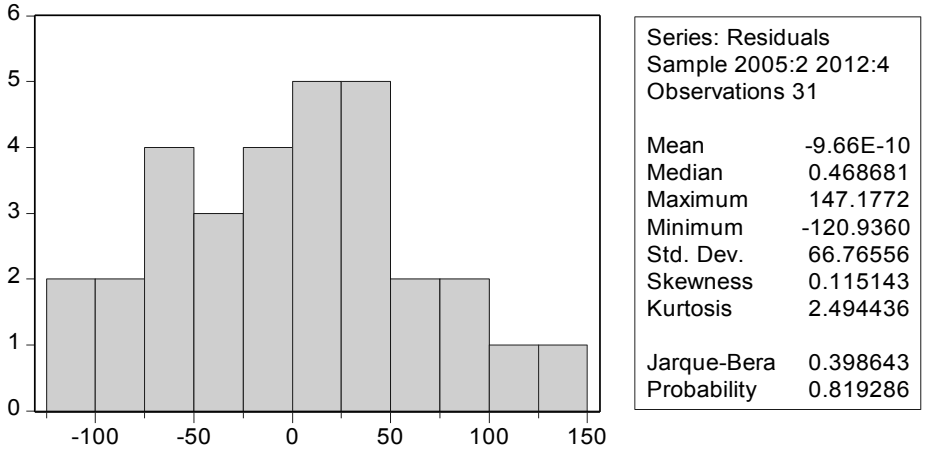
Test Equation:
Dependent Variable: EPA_L3M
Method: Least Squares
Date: 06/02/13 Time: 19:03
Sample: 2005:2 2012:4
Included observations: 31
Convergence achieved after 32 iterations

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	306.8731	57.10359	5.373972	0.0000
GT3	5.292908	5.801891	0.912273	0.3697
FITTED^2	0.000221	0.000546	0.404867	0.6888
AR(1)	0.388567	0.297069	1.308005	0.2019
R-squared	0.851188	Mean dependent var		664.8677
Adjusted R-squared	0.834653	S.D. dependent var		171.6104
S.E. of regression	69.78169	Akaike info criterion		11.44853
Sum squared resid	131476.1	Schwarz criterion		11.63357
Log likelihood	-173.4523	F-statistic		51.47897
Durbin-Watson stat	1.638078	Prob(F-statistic)		0.000000
Inverted AR Roots		.39		

Como el valor estimado por Eviews para FITTED^2 es 0,68 > α = 0,05, se acepta H_0, es decir, FITTED^2 no es significativo y, por tanto, se concluye que no existen errores de especificación.

- **Análisis de la normalidad de las perturbaciones**

Figura 28 Modelo 3.b Jarque-Bera



EViews ofrece el cálculo del estadístico Jarque-Bera. Este estadístico presenta un valor de 0,39 <5,99, por lo que se acepta H_0 y se concluye que las perturbaciones se distribuyen conforme a la distribución normal. Este resultado también indica que el

resto de contrastes que se van a efectuar son fiables, por lo que sí se puede realizar inferencia en el modelo.

- **Contrastes de significatividad**

En este epígrafe se realizan los contrastes de significatividad de los parámetros del modelo, primero de forma individual para cada parámetro y después, de forma conjunta.

Figura 29 Modelo 3.b Estimaciones

Dependent Variable: EPA_L3M
Method: Least Squares
Date: 06/02/13 Time: 19:02
Sample(adjusted): 2005:2 2012:4
Included observations: 31 after adjusting endpoints
Convergence achieved after 4 iterations

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	298.4686	54.45650	5.480861	0.0000
GT3	7.589095	0.998942	7.597131	0.0000
AR(1)	0.419532	0.157876	2.657355	0.0129
R-squared	0.848638	Mean dependent var		664.8677
Adjusted R-squared	0.837826	S.D. dependent var		171.6104
S.E. of regression	69.10892	Akaike info criterion		11.40101
Sum squared resid	133729.2	Schwarz criterion		11.53978
Log likelihood	-173.7157	F-statistic		78.49334
Durbin-Watson stat	1.575561	Prob(F-statistic)		0.000000
Inverted AR Roots	.42			

Contrastes de significatividad de β_1

Como el valor proporcionado por EViews para $\alpha^*(t^*_{\beta 1})$ es 0,00 < α = 0,05, se rechaza H_0, es decir, β_1 sí que es significativo.

Contrastes de significatividad de β_2

Como el valor proporcionado por EViews para $\alpha^*(t^*_{\beta 2})$ es 0,01 < α = 0,05, se rechaza H_0, es decir, β_2 sí que es significativo.

Contrastes de significatividad conjunta

Como el valor estimado por EViews para la probabilidad asociada al estadístico F* es 0,00 < α = 0,05, se rechaza H_0, es decir, sí que existe regresión.

- **Análisis de la heterocedasticidad**

Figura 30 Modelo 3.b Estadístico White

White Heteroskedasticity Test:

F-statistic	1.291384	Probability	0.290762
Obs*R-squared	2.618004	Probability	0.270089

Test Equation:
Dependent Variable: RESID^2
Method: Least Squares
Date: 06/02/13 Time: 19:03
Sample: 2005:2 2012:4

Included observations: 31

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-2510.614	5136.227	-0.488805	0.6288
GT3	255.4633	233.5668	1.093748	0.2834
GT3^2	-1.981558	2.355083	-0.841396	0.4073
R-squared	0.084452	Mean dependent var		4313.845
Adjusted R-squared	0.019055	S.D. dependent var		5360.723
S.E. of regression	5309.401	Akaike info criterion		20.08411
Sum squared resid	7.89E+08	Schwarz criterion		20.22288
Log likelihood	-308.3037	F-statistic		1.291384
Durbin-Watson stat	2.249036	Prob(F-statistic)		0.290762

EViews ofrece el cálculo del estadístico de White. Este estadístico presenta un valor de $2{,}61 < \chi^2_4 = 9{,}49$, por lo que se acepta H_0 y se concluye que existe homocedasticidad.

- **Análisis de la autocorrelación**

Figura 31 Análisis de la autocorrelación

Date: 06/02/13 Time: 19:03
Sample: 2005:2 2012:4
Included observations: 31
Q-statistic probabilities adjusted for 1 ARMA term(s)

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	-0.034	-0.034	0.0387	
		2	-0.002	-0.003	0.0388	0.844
		3	-0.221	-0.221	1.8230	0.402
		4	0.331	0.333	5.9726	0.113
		5	-0.199	-0.240	7.5269	0.111
		6	0.115	0.129	8.0661	0.153
		7	-0.157	-0.067	9.1229	0.167
		8	0.140	-0.031	9.9987	0.189

Se puede apreciar que no existe autocorrelación. Por consiguiente y, de acuerdo con todo lo establecido, se puede validar el uso del modelo 3.b y corroborar su poder predictivo.

Validación Modelo 4.a

- **Contraste de errores de especificación**

Figura 32 Modelo 4.a Test RESET

Ramsey RESET Test:

F-statistic	0.415633	Probability	0.524564
Log likelihood ratio	0.473573	Probability	0.491348

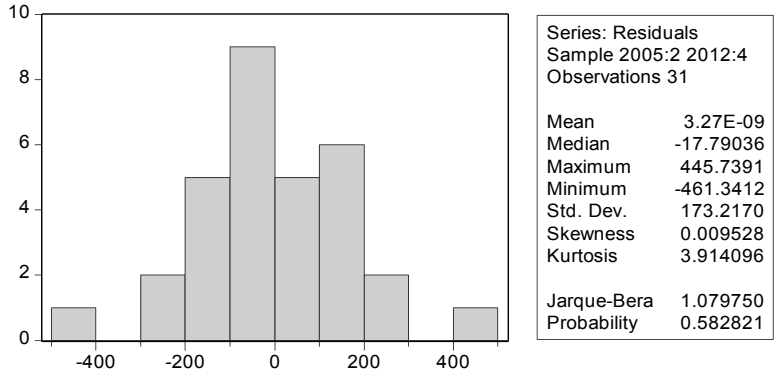
Test Equation:
Dependent Variable: EPA_M3M
Method: Least Squares
Date: 06/02/13 Time: 19:05
Sample: 2005:2 2012:4
Included observations: 31
Convergence achieved after 36 iterations

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-6099.191	37422.27	-0.162983	0.8717
GT1	27.77101	9.131524	3.041224	0.0052
FITTED^2	-1.83E-05	2.74E-05	-0.667590	0.5101
AR(1)	1.007828	0.038326	26.29589	0.0000
R-squared	0.982396	Mean dependent var		2536.032
Adjusted R-squared	0.980440	S.D. dependent var		1295.606
S.E. of regression	181.1974	Akaike info criterion		13.35697
Sum squared resid	886477.8	Schwarz criterion		13.54200
Log likelihood	-203.0330	F-statistic		502.2597
Durbin-Watson stat	1.468056	Prob(F-statistic)		0.000000
Inverted AR Roots		1.01		
	Estimated AR process is nonstationary			

Como el valor estimado por Eviews para FITTED^2 es 0,51 > α = 0,05, se acepta H_0, es decir, FITTED^2 no es significativo y, por tanto, se concluye que no existen errores de especificación.

- **Análisis de la normalidad de las perturbaciones**

Figura 33 Modelo 4.a Jarque-Bera



EViews ofrece el cálculo del estadístico Jarque-Bera. Este estadístico presenta un valor de 1,07 <5,99, por lo que se acepta H_0 y se concluye que las perturbaciones se distribuyen conforme a la distribución normal. Este resultado también indica que el resto de contrastes que se van a efectuar son fiables, por lo que sí se puede realizar inferencia en el modelo.

- **Contrastes de significatividad**

En este epígrafe se realizan los contrastes de significatividad de los parámetros del modelo, primero de forma individual para cada parámetro y después, de forma conjunta.

Figura 34 Modelo 4.a Estimaciones

Dependent Variable: EPA_M3M
Method: Least Squares
Date: 06/02/13 Time: 19:04
Sample(adjusted): 2005:2 2012:4
Included observations: 31 after adjusting endpoints
Convergence achieved after 20 iterations

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	32658.31	708761.7	0.046078	0.9636
GT1	25.00778	7.711552	3.242899	0.0031
AR(1)	0.998221	0.040416	24.69855	0.0000
R-squared	0.982125	Mean dependent var		2536.032
Adjusted R-squared	0.980849	S.D. dependent var		1295.606
S.E. of regression	179.2966	Akaike info criterion		13.30773
Sum squared resid	900124.1	Schwarz criterion		13.44650
Log likelihood	-203.2698	F-statistic		769.2359
Durbin-Watson stat	1.657690	Prob(F-statistic)		0.000000
Inverted AR Roots		1.00		

Contrastes de significatividad de β_1

Como el valor proporcionado por EViews para $\alpha^*(t^*_{\beta 1})$ es 0,00 < α = 0,05, se rechaza H_0, es decir, β_1 sí que es significativo.

Contrastes de significatividad de β_2

Como el valor proporcionado por EViews para $\alpha^*(t^*_{\beta 2})$ es 0,00 < α = 0,05, se rechaza H_0, es decir, β_2 sí que es significativo.

Contrastes de significatividad conjunta

Como el valor estimado por EViews para la probabilidad asociada al estadístico F* es 0,00 < α = 0,05, se rechaza H_0, es decir, sí que existe regresión.

- **Análisis de la heterocedasticidad**

Figura 35 Modelo 4.a Estadístico White

White Heteroskedasticity Test:

F-statistic	4.265605	Probability	0.024150
Obs*R-squared	7.239496	Probability	0.026789

Test Equation:
Dependent Variable: RESID^2
Method: Least Squares
Date: 06/02/13 Time: 19:05
Sample: 2005:2 2012:4

Included observations: 31

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	32231.18	38148.78	0.844881	0.4053
GT1	-1612.127	2022.995	-0.796901	0.4322
GT1^2	31.16121	22.93763	1.358519	0.1851
R-squared	0.233532	Mean dependent var		29036.26
Adjusted R-squared	0.178784	S.D. dependent var		50386.34
S.E. of regression	45660.58	Akaike info criterion		24.38762
Sum squared resid	5.84E+10	Schwarz criterion		24.52640
Log likelihood	-375.0082	F-statistic		4.265605
Durbin-Watson stat	1.922590	Prob(F-statistic)		0.024150

EViews ofrece el cálculo del estadístico de White. Este estadístico presenta un valor de $7{,}23 < \chi^2_4 = 9{,}49$, por lo que se acepta H_0 y se concluye que existe homocedasticidad.

- **Análisis de la autocorrelación**

Figura 36 Modelo 4.a Análisis de la autocorrelación

Date: 06/02/13 Time: 19:05
Sample: 2005:2 2012:4
Included observations: 31
Q-statistic probabilities adjusted for 1 ARMA term(s)

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.140	0.140	0.6712	
		2	-0.103	-0.125	1.0431	0.307
		3	-0.097	-0.066	1.3872	0.500
		4	0.080	0.096	1.6291	0.653
		5	-0.110	-0.161	2.1014	0.717
		6	-0.039	0.016	2.1647	0.826
		7	-0.112	-0.126	2.6951	0.846
		8	-0.105	-0.114	3.1837	0.868

Se puede apreciar que no existe autocorrelación. Por consiguiente y, de acuerdo con todo lo establecido, se puede validar el uso del modelo 4.a y corroborar su poder predictivo.

Validación Modelo 4.b

- **Contraste de errores de especificación**

Figura 37 Modelo 4.b Test RESET

Ramsey RESET Test:

F-statistic	0.436266	Probability	0.514527
Log likelihood ratio	0.496894	Probability	0.480868

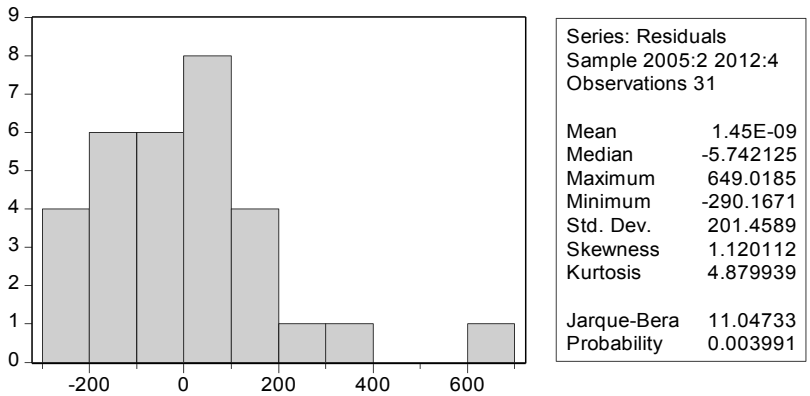
Test Equation:
Dependent Variable: EPA_M3M
Method: Least Squares
Date: 06/02/13 Time: 19:43
Sample: 2005:2 2012:4
Included observations: 31
Convergence achieved after 42 iterations

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	662.7041	216.8540	3.055992	0.0050
GT3	16.50905	5.650382	2.921759	0.0070
FITTED^2	0.000132	1.85E-05	7.149787	0.0000
AR(1)	0.556135	0.165981	3.350593	0.0024
R-squared	0.976206	Mean dependent var		2536.032
Adjusted R-squared	0.973562	S.D. dependent var		1295.606
S.E. of regression	210.6612	Akaike info criterion		13.65829
Sum squared resid	1198209.	Schwarz criterion		13.84332
Log likelihood	-207.7035	F-statistic		369.2480
Durbin-Watson stat	1.750441	Prob(F-statistic)		0.000000
Inverted AR Roots		.56		

Como el valor estimado por Eviews para FITTED^2 es 0,00 < α = 0,05, se rechaza H_0, es decir, FITTED^2 es significativo y, por tanto, se concluye que existen errores de especificación y que el modelo no está completo.

- **Análisis de la normalidad de las perturbaciones**

Figura 38 Modelo 4.b Jarque-Bera



EViews ofrece el cálculo del estadístico Jarque-Bera. Este estadístico presenta un valor de 11,04 > 5,99, por lo que se rechaza H_0 y se concluye que las perturbaciones no se distribuyen conforme a la distribución normal. Este resultado también indica que el modelo no cumple con los requisitos de los modelos de regresión para realizar predicciones.