

# Herramientas de Google para la predicción de variables económicas. Una aplicación al Índice Compuesto Coincidente de Actividad Económica de la Provincia de Santa Fe (ICASFe)

Jorge, Ramiro Emmanuel\*

\* Centro de Estudios y Servicios de la Bolsa de Comercio de Santa Fe (e-mail: [ces@bcsf.com.ar](mailto:ces@bcsf.com.ar)) y Facultad de ciencias Económicas (FCE) de la Universidad Nacional del Litoral (UNL).

## Resumen

El *paper* internaliza información proveniente de las herramientas Google Trends y Google Correlate con el objetivo de predecir de manera oportuna el valor del Índice Compuesto Coincidente de Actividad Económica de la Provincia de Santa Fe (ICASFe), indicador que se publica con dos meses de rezago.

Para esto, se identifican aquellos términos cuyos patrones de búsqueda tienen mayor correlación con el ICASFe y luego se plantea un método de agregación para incorporarlos a la serie *target*.

Las estimaciones obtenidas con el modelo son contrastadas con datos reales de la serie *target* (*ex post*). Los resultados indican que las herramientas y el procedimiento adoptado permiten realizar una estimación consistente y ganar oportunidad respecto a las publicaciones oficiales.

## Abstract

*The paper internalizes information from Google Trends and Google Correlate in order to predict movements over the Coincident Composite Index of Economic Activity for the Province of Santa Fe (ICASFe). The main pursued goal is to improve the gauge's opportunity since its monthly output is being published within two months of lag.*

*In a first instance, terms and search patterns highly correlated to ICASFe's performance are identified. Best matches are selected and, furthermore, incorporated by a proposed aggregation method into the target series.*

*Finally, estimations obtained by the model are contrasted against real data (ex post). To this regard, results indicate that adopted procedures allow a consistent prediction of the target series, gaining opportunity in terms of present publications' dates.*

**JEL classification codes:** [E27], [E32]

**Keywords:** Cycles, nowcast, big data, Google tools

## 1. Introducción

El Centro de Estudios y Servicios de la Bolsa de Comercio de Santa Fe (CES-BCSF) lleva adelante, desde el año 2007, un programa de estudio de ciclos económicos a nivel sub-nacional. Su principal producto es el ICASFe<sup>1</sup>, un índice coincidente de actividad económica de periodicidad mensual que permite datar las fases de contracción y expansión económica de la provincia de Santa Fe con un rezago de dos a tres meses (Cohan & D’Jorge, 2015).

Con el objetivo de realizar mediciones más oportunas, desde hace tiempo se evalúan vías alternativas para estimar la coyuntura de los componentes del índice que presentan mayor demora en su publicación. Algunos ejemplos de esta labor son los *papers* elaborados por dicha institución en años anteriores. En los documentos mencionados se describe la evolución de las series componentes del índice, como así también los criterios de selección de las mismas (Cohan et al., 2007), se detallan los avances implementados en la utilización de *forecasts* obtenidos del *software* X-13ARIMA-SEATS en el proceso de filtrado de las series componentes (Cohan, D’Jorge, & Lazzaroni, 2016); y en trabajos más recientes, se emprende un estudio que evalúa la utilidad de Google Trends y Google Correlate en la estimación oportuna del valor de las ventas minoristas, una de las series utilizadas en la elaboración del ICASFe con mayor rezago en su publicación (Camusso & Jorge, 2019).

En línea con el último antecedente mencionado, el objetivo de esta investigación<sup>2</sup> es comprobar la factibilidad de realizar una predicción del ICASFe de manera anticipada, es decir, antes de disponer de la información necesaria para su actualización mensual, utilizando para ello las herramientas Google Trends y Google Correlate. La primera permite visualizar el flujo de búsquedas de palabras claves<sup>3</sup> a lo largo del tiempo en un espacio geográfico determinado; mientras que Google Correlate proporciona un listado con las palabras cuyas búsquedas presentan mayor correlación con una serie de datos específica.

La herramienta Google Trends ha sido utilizada por algunos autores para desarrollar indicadores económicos y estimaciones de coyuntura. Se destacan los trabajos de Askitas y Zimmermann (2009), Choi y Varian (2009), Carrière-Swallow y Labbé (2011), Artola y Galán (2012), Blanco (2014), Bortoli y Combes (2015), Park et al. (2016), Jun et al. (2017), Dergiades et al. (2018) y Naccarato et al. (2018). En todos estos casos, una de las principales dificultades a la que se enfrentan los autores es establecer un criterio objetivo para seleccionar las palabras claves con mayor potencial predictivo. Por tal motivo, se decide complementar la selección subjetiva de los términos con la aplicación de Google

---

<sup>1</sup> ICASFe: Índice Compuesto Coincidente de Actividad Económica de Santa Fe.

<sup>2</sup> Esta investigación tiene como punto de partida el trabajo final de tesina de grado del autor. Se agradece la contribución fundamental de la Dra. Jimena Vicentín Masaro para la elaboración de dicha tesina y los aportes del Mg. Pedro Pablo Cohan a este *paper*.

<sup>3</sup> Se entiende por “palabras clave” aquellas ingresadas en el motor de búsquedas de Google como también los resultados obtenidos aplicando filtros por categoría sugeridos por dicha base.

Correlate para identificar aquellas palabras potencialmente vinculadas a la serie de referencia de manera más rigurosa, objetiva y precisa (Camusso & Jorge, 2019).

En cuanto a la estructura del *paper*, a continuación de este primer apartado introductorio, se expone el marco de referencia y algunas generalidades conceptuales. Luego se detalla el proceso de identificación y selección de palabras claves con alto potencial predictivo. Posteriormente se describe el método mediante el cual dicha información es internalizada en un modelo de agregación. Una vez definido el modelo, se procede a realizar las estimaciones de la variable *target* tanto dentro como fuera de la muestra, comparando los resultados con el verdadero valor del ICASFe con el objetivo de evaluar la precisión del método. El último apartado presenta una síntesis de resultados junto a las principales conclusiones.

## **2. Marco de referencia y generalidades conceptuales**

### **2.1. Google Trends**

Es una herramienta de libre acceso y gratuita que permite conocer la evolución de las búsquedas de un término o palabra clave (*keyword*) a lo largo de un período de tiempo determinado. La firma comienza a publicar los datos en el año 2004, y su publicación continúa hasta la actualidad.

Esta herramienta brinda la posibilidad de obtener información de las búsquedas con diferentes periodicidades, i.e. anual, mensual, semanal, y hasta por hora. El mecanismo resulta sencillo e intuitivo; el usuario ingresa una palabra clave en el buscador, pudiendo aplicar diferentes filtros de búsqueda: por área geográfica (cuya dimensión mínima es por provincia), período de tiempo, categoría (arte, ciencias, compras, deportes, etc.) y tipo de búsqueda (en la web, en noticias, imágenes).

La información obtenida respecto al nivel de búsqueda de la palabra clave ingresada no se muestra en términos absolutos, sino en forma de índice, con escala de 0 a 100. El valor 100 representa el momento con mayor frecuencia de búsquedas del término en cuestión durante el período temporal seleccionado. Paralelamente, la herramienta expone un cuadro de “consultas relacionadas” que lista los términos que también fueron consultados por usuarios que ingresaron esa palabra clave.

### **2.2. Google Correlate<sup>4</sup>**

Google Correlate utiliza un método automatizado para la selección de consultas relacionadas a una serie de referencia. Lo hace a través de un algoritmo que, a partir de la

---

<sup>4</sup> Esta herramienta ya no se encuentra disponible. La firma decidió dar de baja la plataforma a partir del 15 de diciembre de 2019.

aplicación de coeficientes de correlación, devuelve un conjunto de palabras cuyas búsquedas poseen mayor correlación con la serie en cuestión. Este proceso de identificación tiene en cuenta las dimensiones temporales y espaciales.

La herramienta emplea un algoritmo de aproximación sobre millones de consultas en un árbol de búsqueda en línea con el objetivo de arribar a resultados similares al enfoque empleado por Google Trends, pero utilizando un proceso inverso (Mohebbi et al., 2011). Puntualmente, dado un patrón de interés temporal o espacial, se determina qué consultas imitan mejor los datos. Estas búsquedas poseen potencial para construir una estimación del valor del fenómeno (*proxy*).

Respecto a la unidad de medida de las salidas de Google Correlate, las mismas se expresan en términos estandarizados, por lo que los datos presentan media de 0 y desvío estándar de 1. Es decir, las salidas exponen la evolución de un indicador que se expresa en desviaciones estándar por encima y por debajo de la media de búsquedas. Al igual que Google Trends, los datos se encuentran disponibles desde enero de 2004, pero la actualización fue discontinuada en marzo de 2017.

### 2.3. Variable *target*

La variable de interés predictivo en el presente trabajo es el ICASFe. Un índice compuesto coincidente de actividad económica de periodicidad mensual que permite identificar las fases de contracción y expansión económica de la provincia de Santa Fe con un rezago de dos a tres meses.

Cuenta con datos desde enero de 1994 (con base 1994=100) hasta el presente y sintetiza la situación económica de la provincia de Santa Fe. El índice constituye una herramienta de información de gran relevancia para la toma de decisiones en ámbitos públicos y privados. De acuerdo a sus características, resulta una fuente de consulta para un heterogéneo grupo de usuarios (empresarios, docentes, alumnos, funcionarios del sector público, medios de comunicación y público en general).

En particular, insume información proveniente de catorce series temporales de interés económico de alcance provincial. La inclusión de cada componente se fundamenta en el cumplimiento de los siguientes requisitos: (1) brindan información referida al espacio geográfico, (2) son representativas de variables con significancia económica, (3) tienen una periodicidad mensual, y (4) poseen una disponibilidad y una fecha de inicio común (D'Jorge et al., 2018). Al mismo tiempo, la selección de las mismas responde a un proceso metodológico basado en los criterios utilizados por el *Conference Board's Business Cycle Indicators Program* para la elaboración del Índice Compuesto de Actividad Coincidente de EEUU (Cohan et al., 2007).

La metodología utilizada para construir el ICASFe fue transferida al CES mediante un convenio firmado entre la Bolsa de Comercio de Santa Fe y el Dr. Juan Mario Jorrat, director del “Programa de Ciclos Económicos Argentinos” de la Universidad Nacional de Tucumán (UNT), en diciembre de 2006 (D’Jorge et al., 2018).

Siguiendo los lineamientos metodológicos y habiendo realizado revisiones periódicas, actualmente se mantienen vigentes catorce series componentes del ICASFe, las mismas se exponen en la Cuadro 1.

**Cuadro 1:** Series componentes del Índice Compuesto Coincidente de Actividad Económica de Santa Fe.

Bloque	Variable
Empleo	- Número de puestos de trabajo registrados en la provincia
	- Índice de demanda laboral
Producción Industrial	- Consumo de energía eléctrica industrial
	- Consumo de gas industrial
	- Consumo de hidrocarburos líquidos
	- Faena de ganado bovino y porcino
	- Producción industrial de lácteos
	- Molienda de oleaginosas
Ventas Minoristas	- Ventas de maquinaria agrícola
	- Ventas reales de supermercados
	- Consumo de cemento Pórtland
Ingreso Disponible	- Patentamiento de vehículos nuevos
	- Recaudación tributaria de la provincia y coparticipación
	- Masa de remuneraciones reales percibida por los asalariados

Fuente: elaboración propia en base a datos del CES-BCSF

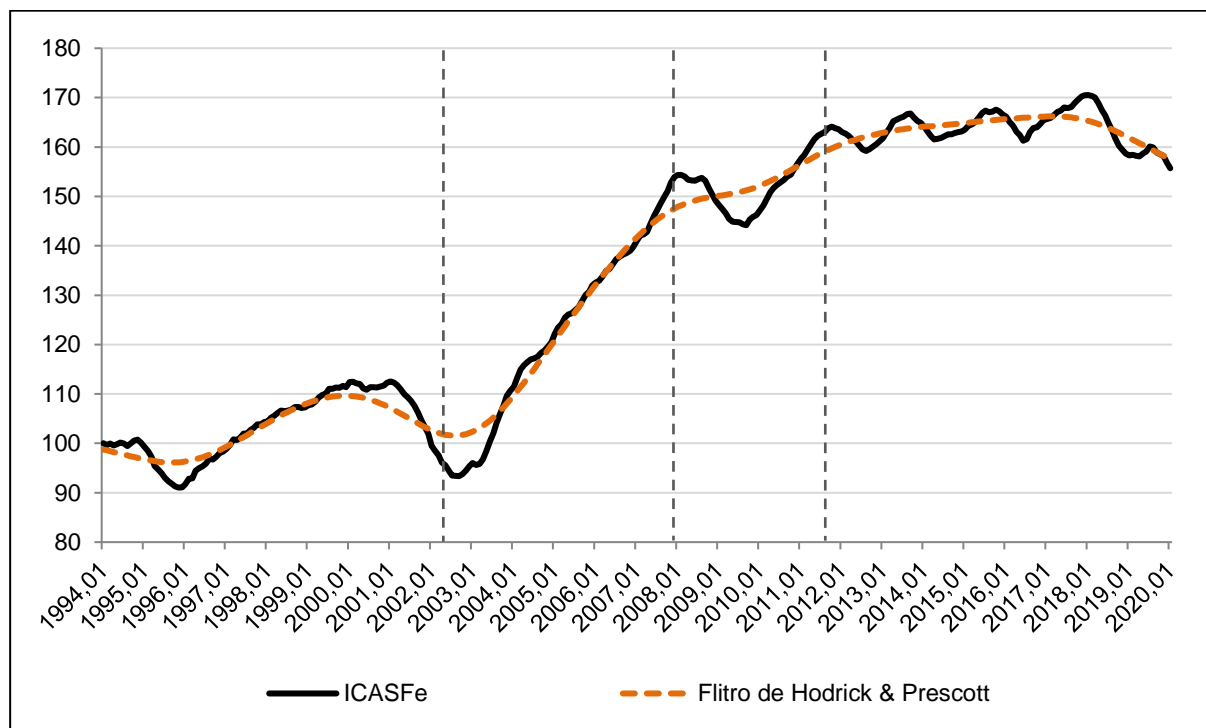
Una vez que las catorce series son ajustadas por estacionalidad y valores irregulares mediante el *software* X-13ARIMA-SEATS, se procede a estandarizar las tasas de cambio mensuales logarítmicas de cada una, siendo la tasa de cambio del índice, el promedio de las variaciones mensuales estandarizadas de las series que lo componen (D’Jorge et al., 2018). Aplicando la tasa de cambio del mes  $t$ , al valor del índice en el mes  $t - 1$ , se obtiene el valor del ICASFe en el mes  $t$ .

En el Gráfico 1 se puede ver la evolución del ICASFe desde enero de 1994 hasta enero de 2020. Este indicador muestra una tendencia creciente (que puede visualizarse a través de la tendencia que resulta de aplicar el filtro de Hodrick & Prescott) y un patrón cíclico recurrente, con mayor frecuencia a partir de 2008. El índice no presenta estacionalidad ni valores irregulares, ya que es el resultado de la agregación de un conjunto de series filtradas<sup>5</sup>. Dicha característica es precisamente deseable en la construcción de este tipo de indicadores,

<sup>5</sup> Para más detalles acerca de la metodología de selección y filtrado de las series componentes del ICASFe, visitar: <https://www.bcsf.com.ar/ces/icasfe.php>.

dato que el propósito es estimar el componente tendencia-ciclo de la serie de actividad económica, y no sus oscilaciones recurrentes de corto plazo.

**Gráfico 1:** Serie del Índice Compuesto Coincidente de Actividad Económica de Santa Fe y tendencia de Hodrick & Prescott. Índice base 1994=100. Período 2004.01 a 2020.01.



Fuente: elaboración propia en base a datos del CES-BCSF

En términos desagregados, el período 1994-2003 presenta oscilaciones en el nivel de actividad, sin denotar una tendencia marcada. Entre 2003 y 2008 se observa el mayor crecimiento de la serie mientras que, a partir de 2008, el nivel de actividad económica de la provincia comienza a crecer a una tasa menor, y hacia 2012 pasa a registrar un estancamiento relativo con un movimiento de largo plazo a la baja.

### 3. Variables proxy: identificación y selección

#### 3.1. Utilización de la herramienta Google Correlate

La serie de datos del ICASFe se ingresa a Google Correlate, y se obtiene una lista de las 100 consultas cuyos patrones de búsquedas tienen mayor correlación con la misma. Esta herramienta permite detectar términos de búsqueda con un comportamiento similar al de la serie ingresada, algunos de los cuales de manera intuitiva podrían no ser tenidos en cuenta. La serie *target* es introducida a Google Correlate siguiendo requerimientos de formato establecidos por la aplicación<sup>6</sup>. Se decide utilizar un paquete que inicia en enero de 2004 – a

<sup>6</sup> Se exige el uso de un archivo tipo “.csv”, donde la primera columna refiere a las fechas, siguiendo un formato de cuatro dígitos para el año calendario, dos para el mes y dos para el día; separados cada uno por guiones (aaaa-mm-dd).

pesar de que el ICASFe cuenta con datos mensuales desde 1994 – debido a que la base de datos disponible en Correlate contiene información a partir de dicho mes y hasta marzo de 2017.

De la lista obtenida, se decide considerar las 5 consultas con mayor coeficiente de correlación respecto a la serie del ICASFe. Cada término y su nivel de correlación se exponen en el Cuadro 2.

**Cuadro 2:** Palabras cuyo patrón de búsquedas posee mayor coeficiente de correlación respecto a la serie del Índice Compuesto Coincidente de Actividad Económica de Santa Fe.

Palabras Correlacionadas a ICASFe	Coeficiente de Correlación con ICASFe
Coopeplus	0,9608
Vallan	0,9495
Cualquier	0,9484
Pasar	0,9478
Bancopatagonia	0,9440

Fuente: elaboración propia en base a CES-BCSF y Google Inc.

Dentro de los términos con mayor correlación existen dos palabras que poseen significado económico, estas son “coopeplus” y “bancopatagonia” (Banco Patagonia). El primer caso corresponde a una tarjeta de crédito emitida por Nueva Card S.A., mientras que Banco Patagonia es un banco privado que surge de la fusión entre Banco Mercantil Argentino y Banco Caja de Ahorro en el año 1999 y que tiene operatoria en la provincia de Santa Fe. Tanto la tarjeta Coopeplus como el Banco Patagonia tienen alcance nacional y por su significado económico, es de esperar que posean cierto grado de correspondencia con el nivel de actividad del país.

Por su parte, los términos “vallan”, “cualquier” y “pasar”, no poseen un significado económico evidente. Esto pone en manifiesto una característica interesante de esta herramienta, a saber, su capacidad para identificar relaciones de comportamiento entre variables que pueden no responder a una hipótesis previa. En esta línea, algunos de los términos descartados que tuvieron buen ajuste fueron: “nuevos videos”, “barra invertida”, “en la mañana”, entre otros.

Es importante destacar que la aplicación toma como referencia geográfica búsquedas efectuadas a nivel país, es decir que, a pesar de que la variable *target* sea de alcance provincial, sus movimientos se contrastan con *proxies* de alcance nacional. Esto no necesariamente implica una limitación desde el punto de vista estadístico, pero sí reduce la posibilidad de identificar relaciones de largo plazo ante la eventualidad de que la estructura económica provincial sea radicalmente distinta a la nacional en algún período de tiempo. Aun así, y como estudios previos lo corroboran, se puede afirmar que la actividad económica

de la provincia de Santa Fe guarda una sincronía significativa con el flujo de actividad nacional en términos cíclicos (CES-BCSF, 2019).

### 3.1.1. Las salidas de Google Correlate

La lista de términos obtenida está compuesta por 100 palabras cuyos coeficientes de correlación con la serie del ICASFe superan, en todos los casos, el valor 0,75. Respecto a cada una de las palabras, se puede descargar una serie correspondiente al historial de búsqueda, que contiene datos mensuales estandarizados para el período 2004.01-2017.03. Dado que el paquete de datos finaliza en marzo de 2017, Google Correlate no permite conocer la evolución de las variables con posterioridad a dicha fecha. Para salvar esta limitación se utilizó de manera complementaria la herramienta Google Trends, cuya base de datos se encuentra disponible con rezagos menores a las dos semanas respecto de la fecha de búsqueda.

### 3.2. Utilización de la herramienta Google Trends

Una vez obtenidas las cinco palabras seleccionadas de Google Correlate, las mismas se ingresan a Google Trends. De esta forma se obtiene el patrón de búsqueda actualizado de dichas variables.

Un segundo aporte de esta herramienta es que, como parte de las salidas, pone a disposición de los usuarios una lista con 25 consultas relacionadas al término original, las cuales también poseen potenciales aptitudes predictivas. Para cada término obtenido mediante Google Correlate, se decidió seleccionar sólo las dos primeras palabras de la lista de consultas relacionadas de Google Trends. De esta forma, se cuenta con un total de 15 *keywords*, todas ellas representadas en el Cuadro 3.

**Cuadro 3:** Palabras clave de Google Correlate y consultas relacionadas obtenidas mediante Google Trends.

Salidas de Google Correlate	Consultas Relacionadas de Google Trends
Coopeplus	Coopeplus tarjeta Coopeplus bahía blanca
Vallan	Colectivos Valla
Cualquier	Cualquier coincidencia con la realidad es pura coincidencia Cualquier coincidencia con la realidad
Pasar	Como pasar credito Como pasar
Bancopatagonia	Bancopatagonia e bank Banco

Fuente: elaboración propia en base a datos de Google Inc.



Es importante aclarar que, si bien esta herramienta permite obtener resultados a nivel provincial, se decidió utilizar datos nacionales para armonizar con el criterio definido al trabajar con Google Correlate.

### 3.2.1. Las salidas de Google Trends

Al igual que con Google Correlate, las salidas del Trends arrojan una serie de datos para cada variable. Sin embargo, esta herramienta presenta la información a través de un índice con base=100 en el mes con máximo nivel de búsqueda dentro del periodo temporal determinado. Se toma como referencia temporal el período que va desde enero de 2004 hasta enero de 2019 – acotado geográficamente al ámbito nacional – para obtener los datos correspondientes y realizar estimaciones dentro de la muestra.

## 4. Internalización de las series *proxies* para realizar un *nowcast* de la serie *target*: proceso de transformación y agregación

Una vez seleccionadas las variables *proxies*, fue necesario generar un marco procedimental que permitiera estimar los movimientos coyunturales de la variable *target* en función de la información disponible. Considerando esto, una nueva revisión de antecedentes permitió reconocer distintas metodologías. En lo que respecta al trabajo de Askitas y Zimmermann (2009), los autores optan por utilizar diversos modelos de corrección de errores, fundamentando la selección del mejor de estos a través del Criterio de Bayes (BIC<sup>7</sup>). Por su parte, el resto de los autores recurre a diferentes variantes de modelos autorregresivos. Choi y Varian (2009), Blanco (2014) y Bortoli y Combes (2015) aplican modelos autorregresivos simples, recurriendo a *dummies* para valores irregulares y criterios de información para la selección del modelo más adecuado; Carrière-Swallow y Labbé (2011) internalizan la información a través de un modelo autorregresivo de medias móviles (ARMA<sup>8</sup>), mientras que Artola y Galán (2012), Park et al. (2016) y Naccarato et al. (2018) aplican modelos autorregresivos integrados de medias móviles (ARIMA<sup>9</sup>). Naccarato et al. (2018) compara los resultados del modelo ARIMA con los de un modelo de vectores autorregresivos (VAR<sup>10</sup>), metodología a la que también recurre Dergiades et al. (2018), complementando el análisis mediante *tests* de causalidad de Granger.

### 4.1. Determinación de un modelo de agregación

#### 4.1.1. Modelo Autorregresivo con Rezagos Distribuidos

---

<sup>7</sup> Por sus siglas en inglés: *Bayesian Information Criterion*.

<sup>8</sup> Por sus siglas en inglés: *Autoregressive Moving Average*.

<sup>9</sup> Por sus siglas en inglés: *Autoregressive Integrated Moving Average*.

<sup>10</sup> Por sus siglas en inglés: *Vector Autoregressive*.

Bajo un criterio de parsimonia, se decide evaluar el poder explicativo de las variables *proxy* obtenidas. Para esto, y teniendo en cuenta los antecedentes antes mencionados, se puso en práctica un modelo autorregresivo con rezagos distribuidos (ARDL<sup>11</sup>) contemplando diferentes variantes que alternan en relación a la cantidad de variables explicativas y rezagos de la variable explicada a incluir. Los datos comprenden el espacio temporal que va desde enero de 2004 hasta enero de 2019, con el fin de obtener un modelo para estimar los valores del ICASFe en ese período y contrastarlos con el verdadero valor del índice. De esta manera, se pretende evaluar la capacidad predictiva dentro de la muestra, para luego, realizar predicciones fuera de esta.

Se opta por un modelo ARDL, debido a que estos son utilizados para modelar la relación entre variables de series de tiempo en una sola ecuación y son útiles para la predicción y la separación de relaciones de corto y largo plazo entre las variables de interés. Así, suponiendo que  $y_t$  es la variable de interés predictivo y  $x_t = (x_{1t}, x_{2t}, \dots, x_{nt})$  son variables de series de tiempo exógenas, entonces un modelo ARDL( $p, q$ ) puede representarse de la siguiente manera:

$$y_t = \alpha + \varphi t + \sum_{i=1}^p \gamma_i y_{t-i} + \sum_{j=0}^q \beta_j' x_{t-j} + u_t . \quad (1)$$

Con  $p \geq 1$ , y  $q \geq 0$ , donde  $x_*$  es un vector  $n \times 1$ . La parte de la ecuación (1) referida a la autorregresión ( $y_{t-i}$ ) utiliza información del pasado para predecir el valor presente de  $y$ . A su vez, los términos contenidos en el vector  $x_*$  incorporan información proveniente de variables exógenas, tanto en su valor presente como en valores pasados.

Una de las ventajas de estos modelos es que no tienen una exigencia sobre el orden de integración de las series, pudiéndose aplicar tanto en series  $I(0)$  como  $I(1)$ , o una combinación de ambas. Además, permiten tratar el problema de relaciones espurias por medio de la incorporación de rezagos.

#### 4.1.2. Selección del modelo

Se analizan cuatro modelos alternativos: uno con 15 variables explicativas (5 de GC más 10 de GT), uno con 5 variables (obtenidas de GC), y dos modelos con 10 variables (uno con 5 de GC y 5 de GT, el otro con las 10 de GT). Para todos los casos se tienen en cuenta las variables con 1 y con 2 rezagos. De esta forma, se totalizan ocho alternativas. La selección del modelo más adecuado se realiza en función del modelo que presenta mejores criterios de información de Akaike ( $AIC^{12}$ ) y de Bayes (BIC).

<sup>11</sup> Por sus siglas en inglés: *Autoregressive Distributed Lag*.

<sup>12</sup> Por sus siglas en inglés: *Akaike Information Criterion*.

En el Cuadro 4 se presentan los estadísticos de las ocho alternativas consideradas. Las medidas de bondad de ajuste evidencian que la opción más adecuada es la representada mediante la inclusión de las 5 variables obtenidas de Google Correlate, incorporando dos rezagos para las mismas y dos para el ICASFe, es decir, el modelo ARDL(2,2).

**Cuadro 4:** Principales estadísticos de ajustes de las diferentes variantes de modelos autorregresivos con rezagos distribuidos.

Nº variables	Origen	ARDL	R <sup>2</sup>	R <sup>2</sup> ajustado	Error estándar residual	AIC	BIC
5	GC	(1,1)	0,998	0,998	0,704	410,596	452,531
5	GC	(2,2)	<b>0,999</b>	<b>0,999</b>	<b>0,433</b>	<b>234,001</b>	<b>295,193</b>
10	GT	(1,1)	0,998	0,998	0,731	433,932	508,124
10	GT	(2,2)	<b>0,999</b>	<b>0,999</b>	0,440	253,170	362,662
10	GC (5) GT (5)	(1,1)	0,998	0,998	0,714	425,224	499,416
10	GC (5) GT (5)	(2,2)	<b>0,999</b>	<b>0,999</b>	0,440	252,785	362,277
15	GC (5) GT (10)	(1,1)	0,998	0,998	0,726	439,530	545,980
15	GC (5) GT (10)	(2,2)	<b>0,999</b>	<b>0,999</b>	0,434	258,337	416,134

Nota: en negrita se muestran los valores mínimos o máximos, según corresponda.

Fuente: elaboración propia en base a datos de CES-BCSF y Google Inc.

Una vez seleccionado el modelo ARDL(2,2) en base a los valores obtenidos de AIC y BIC, se procede a aplicar los correspondientes contrastes de normalidad de los residuos (Shapiro-Wilk y Jarque-Bera), correlación serial (Breusch-Godfrey) y autocorrelación de los residuos (Durbin-Watson y Ljung-Box). Los resultados obtenidos en cada una de las pruebas se encuentran en la Cuadro 5.

**Cuadro 5:** Resultados de contrastes de normalidad, correlación serial y autocorrelación de los residuos del modelo autorregresivo con rezagos distribuidos seleccionado.

Contraste	Hipótesis nula	Valor p	Conclusión sobre H <sub>0</sub>
Shapiro-Wilk	H <sub>0</sub> : Normalidad de los residuos	0,063	No se rechaza al 99%
Jarque-Bera	H <sub>0</sub> : Normalidad de los residuos	0,537	No se rechaza al 99%
Breusch-Godfrey	H <sub>0</sub> : No correlación serial	0,764	No se rechaza al 99%
Durbin-Watson	H <sub>0</sub> : No autocorrelación de residuos	0,449	No se rechaza al 99%
Ljung-Box	H <sub>0</sub> : No autocorrelación de residuos	0,812	No se rechaza al 99%

Fuente: elaboración propia en base a datos de CES-BCSF y Google Inc.

El modelo seleccionado genera residuos que no evidencian problemas de no normalidad, ya que tanto en la prueba Jarque-Bera como en el caso de Shapiro-Wilk no se rechaza la hipótesis nula de normalidad. En relación a la autocorrelación, los resultados de las pruebas

de Durbin-Watson, Ljung-Box y Breusch-Godfrey indican que no hay problemas de autocorrelación de los residuos<sup>13</sup>.

En la Cuadro 6 se muestran los coeficientes y desvíos estándar de cada variable del modelo seleccionado; adicionalmente, se indica el nivel al que cada variable es significativa. De las variables incorporadas en el modelo sólo el ICASFe resulta significativa al 99% de confianza, con ambos rezagos. Si bien el mejor modelo multivariado es el que incorpora como variables exógenas las de GC, ninguna de ellas resulta significativa incluso al nivel del 10% de significancia. Sin embargo, con motivo de comparar la capacidad predictiva de este con el verdadero valor de la variable *target*, se decide mantenerlas como variables.

**Cuadro 6:** Coeficientes estimados y desvíos estándar de los coeficientes del modelo autorregresivo con rezagos distribuidos seleccionado.

Variable	Coeficiente	Desvío Estándar
Intercepto	2,013 *	0,697
Coopeplus <sub>t</sub>	0,001	0,002
Coopeplus <sub>t-1</sub>	-0,002	0,002
Coopeplus <sub>t-2</sub>	0,002	0,002
Vallan <sub>t</sub>	0,003	0,005
Vallan <sub>t-1</sub>	0,002	0,006
Vallan <sub>t-2</sub>	-0,004	0,006
Cualquier <sub>t</sub>	0,008	0,008
Cualquier <sub>t-1</sub>	-0,000	0,008
Cualquier <sub>t-2</sub>	0,006	0,008
Pasar <sub>t</sub>	0,003	0,002
Pasar <sub>t-1</sub>	-0,003	0,002
Pasar <sub>t-2</sub>	-0,001	0,002
Bancopatagonia <sub>t</sub>	0,002	0,004
Bancopatagonia <sub>t-1</sub>	-0,006	0,004
Bancopatagonia <sub>t-2</sub>	0,003	0,004
ICASFe <sub>t-1</sub>	1,783 **	0,046
ICASFe <sub>t-2</sub>	-0,802 **	0,046

Nota: significativo al nivel \*5% y \*\*1%

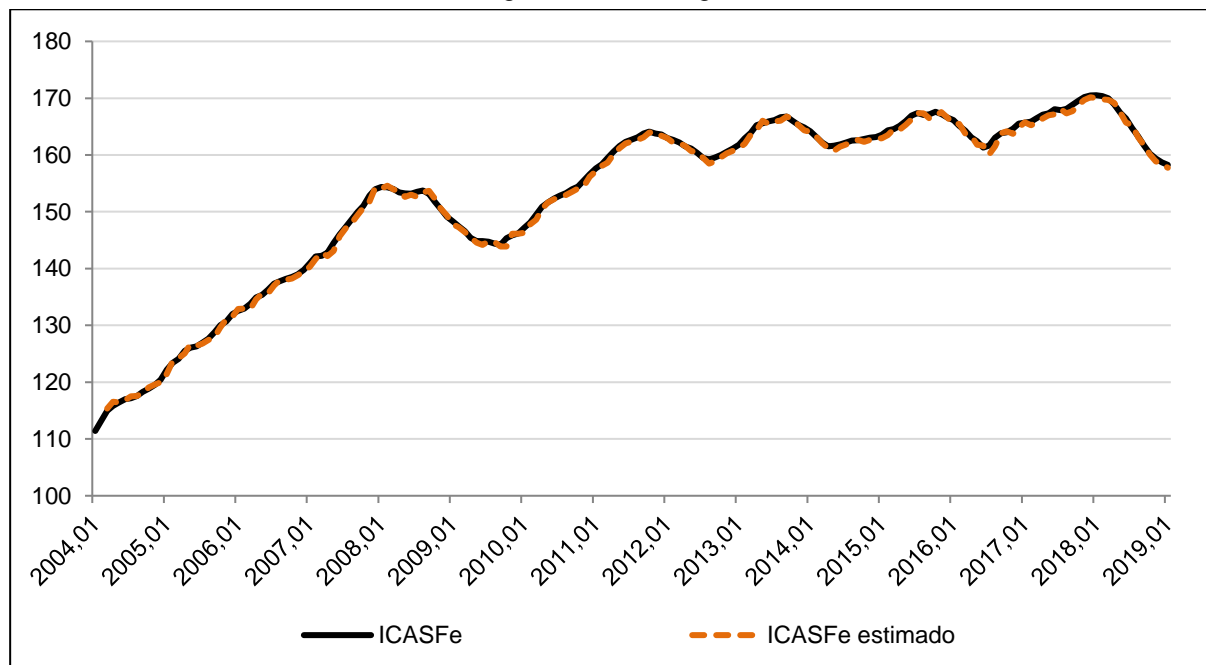
Fuente: elaboración propia en base a datos de CES-BCSF y Google Inc.

Una vez obtenidos los parámetros, se procede a realizar la estimación dentro de la muestra, ponderando los valores mensuales de búsqueda de cada palabra clave desde enero de 1994 a enero de 2019, y ponderando cada uno por el coeficiente correspondiente.

<sup>13</sup> Para mayor detalle, los residuos del modelo multivariado, su correlograma y su distribución se encuentran expuestos de manera gráfica en el Anexo del documento.

El Gráfico 2 muestra el valor del ICASFe y la predicción obtenida mediante el modelo ARDL(2, 2). Se puede observar que la evolución de la predicción asume valores muy similares a los reales (determinados por el índice coincidente).

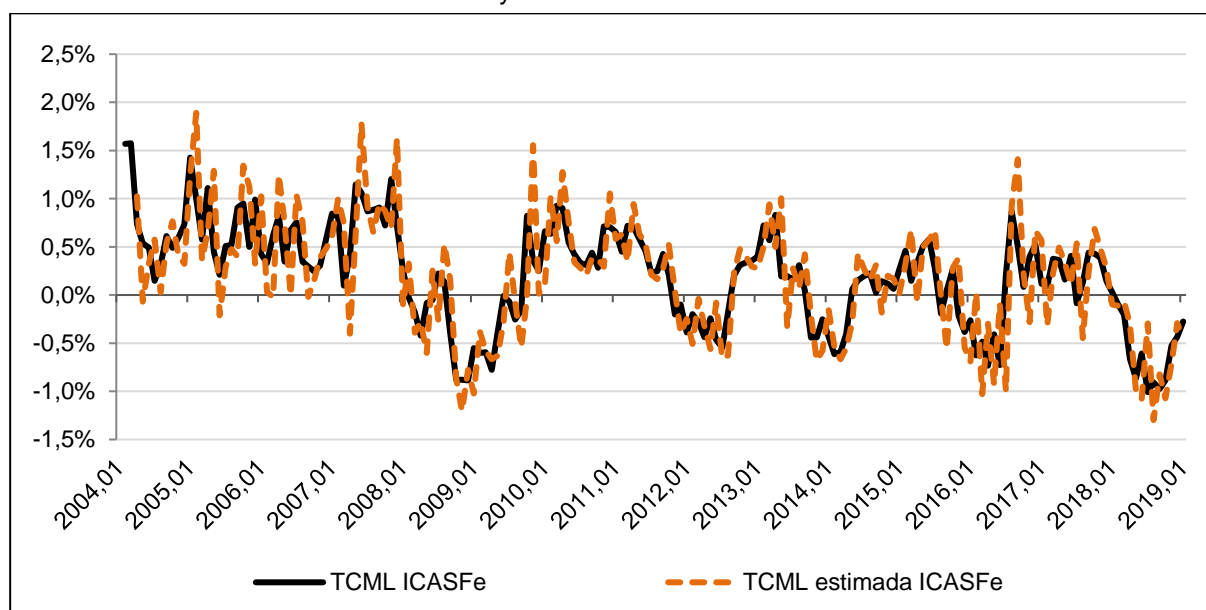
**Gráfico 2:** Índice Compuesto Coincidente de Actividad Económica de Santa Fe y su valor estimado mediante el modelo autorregresivo con rezagos distribuidos seleccionado.



Fuente: elaboración propia en base a datos de CES-BCSF y Google Inc.

Para lograr una mejor visualización, en el Gráfico 3 se muestra la tasa de cambio mensual logarítmica del ICASFe y el valor estimado de la misma mediante el modelo multivariado seleccionado.

**Gráfico 3:** Evolución de la tasa de cambio mensual logarítmica del Índice Compuesto Coincidente de Actividad Económica de Santa Fe y su valor estimado mediante el modelo seleccionado.



Fuente: elaboración propia en base a datos de CES-BCSF y Google Inc.

Como se puede observar en esta nueva gráfica, si bien la tasa de cambio de ambas series se asemeja, existen meses en los que se aprecian diferencias.

Por todo lo mencionado anteriormente se puede afirmar que, en líneas generales, la internalización de la evolución de las *proxies* por medio del modelo ARDL(2,2) permite obtener una buena estimación del verdadero valor del ICASFe. Esto se cumple para las observaciones dentro de la muestra (período desde enero de 2004 a enero de 2019). Resta corroborar si, efectivamente, este comportamiento se mantiene para datos no considerados como insumos a la hora de realizar la estimación. Con tal fin, se procede a realizar la estimación del valor del ICASFe entre los meses de febrero de 2019 a enero de 2020, completando el período de doce meses posterior a los considerados como muestra y utilizados para obtener los parámetros del Cuadro 6.

## **5. Predicción a corto plazo del ICASFe a través del modelo seleccionado**

Con el objetivo de analizar la capacidad predictiva del modelo seleccionado, se realizan estimaciones para los meses de febrero de 2019 a enero de 2020.

La principal particularidad del proceso predictivo es la actualización mensual de los parámetros. Esto implica que, para poder estimar el valor del ICASFe del mes de febrero de 2019, se utilizan los parámetros expuestos en el Cuadro 6, mientras que la estimación correspondiente al mes de marzo de 2019, se realiza internalizando el verdadero valor del índice del mes de febrero y reestimando dichos parámetros.

La predicción de cada mes se obtiene a través de coeficientes reestimados, incorporando el último dato disponible. De este modo, se repite el procedimiento hasta llegar a enero de 2020, cuyos parámetros de estimación cuentan con información consolidada del ICASFe de diciembre de 2019.

Una vez obtenidos los valores estimados del ICASFe, se obtiene la tasa de cambio de los mismos y se las compara con las tasas de cambio del verdadero valor del índice. El Cuadro 7, muestra los verdaderos valores del ICASFe, sus predicciones mediante el modelo ARDL(2, 2), las tasas de cambio de ambos y el error de estimación de dichas tasas.

Dado que el interés predictivo se centra en estimar el verdadero valor del ICASFe, e indirectamente, predecir de manera adecuada su tasa de cambio, los datos en el Cuadro 7 permiten un interesante análisis.

**Cuadro 7:** Valor del Índice Compuesto Coincidente de Actividad Económica de Santa Fe, su valor estimado mediante el modelo seleccionado, tasas de cambio y error de estimación de la variación. Período 2019,02-2020,01.

Fecha	ICASFe	ICASFe estimado ARDL(2, 2)	Variación del ICASFe	Variación del ICASFe estimado	Error de estimación de la variación
2019,02	158,65	158,33	0,12%	0,05%	0,07%
2019,03	158,32	158,77	-0,21%	0,27%	-0,48%
2019,04	158,09	158,10	-0,14%	-0,42%	0,28%
2019,05	158,59	158,03	0,32%	-0,04%	0,36%
2019,06	159,05	159,10	0,29%	0,68%	-0,39%
2019,07	160,11	159,52	0,66%	0,26%	0,40%
2019,08	159,78	160,96	-0,21%	0,90%	-1,11%
2019,09	158,78	159,52	-0,63%	-0,89%	0,27%
2019,10	158,20	158,09	-0,37%	-0,90%	0,53%
2019,11	157,55	157,82	-0,41%	-0,17%	-0,24%
2019,12	156,63	157,08	-0,59%	-0,47%	-0,12%
2020,01	155,70	155,29	-0,59%	-1,14%	0,55%

Fuente: elaboración propia en base a datos de CES-BCSF y Google Inc.

El modelo ARDL(2, 2) arroja valores predichos del ICASFe cuyas tasas de cambio coinciden en signo con la tasa de cambio del verdadero valor del índice, en nueve de los doce meses analizados. Esto implica que en el 75% de los meses bajo estudio, el modelo acierta el sentido de la variación del índice. El error de estimación se encuentra en un rango que va de -1,11 a 0,07 puntos porcentuales, con un valor promedio de 0,40 puntos porcentuales, considerando incluso aquellos meses en los que el modelo estima variaciones con signo matemático diferente a las del índice consolidado.

Se puede afirmar entonces, que con este mecanismo se logra un grado considerable de precisión en las estimaciones del ICASFe, con la ventaja adicional de que permite ganar hasta dos meses de oportunidad.

## 6. Síntesis de resultados y comentarios finales

Los resultados de este *paper* permiten afirmar que el uso conjunto de Google Trends y Google Correlate ha sido satisfactorio para identificar variables *proxies* y realizar estimaciones del ICASFe.

El procedimiento desarrollado toma mayor relevancia en cuanto puede replicarse fácilmente a otros indicadores que también presentan rezagos en sus publicaciones. Comparativamente con la mayoría de los antecedentes relevados, el mayor aporte del trabajo refiere a la aplicación de Google Correlate como una herramienta objetiva de selección de palabras claves y series altamente correlacionadas (sin requerir de conocimientos previos sobre el fenómeno bajo análisis).

Complementariamente, Google Trends posibilita obtener información sobre los patrones de búsqueda de forma oportuna (presentando datos consolidados con dos semanas de rezago aproximadamente). En el caso particular de la variable *target* analizada en este trabajo, implica incrementar la oportunidad de la información en dos meses.

Aunque existen otras alternativas, la investigación realizada permitió identificar un modelo que estima de manera precisa la tasa de cambio del ICASFe, tanto en su signo matemático como en su magnitud, con la ventaja adicional de una mayor oportunidad en la disponibilidad del dato.



## 7. Bibliografía

- Artola, C., & Galán, E. (2012). *Tracking the future on the web: construction of leading indicators using Internet searches*. Madrid, España: Banco de España.
- Askitas, N., & Zimmermann, K. (2009). *Google Econometrics and Unemployment Forecasting*. Bonn: Forschungsinstitut zur Zukunft der Arbeit Institute for the Study of Labor.
- Blanco, E. (2014). *Herramientas de Big Data: ¿Podemos Aprovechar Google Trends para Pronosticar Algunas Variables Macro Relevantes?* Asociación Argentina de Economía Política (AAEP).
- Bortoli, C., & Combes, S. (2015). *Contribution from Google Trends for forecasting the short-term economic outlook in France: limited avenues*.
- Camusso, M. F., & Jorge, R. E. (2019). *Google Correlate y Google Trends como herramientas para realizar un nowcast de las ventas minoristas*. Santa Fe: CES-BCSF.
- Carrière-Swallow, Y., & Labbé, F. (2011). *Nowcasting with Google Trends in an Emerging Market*. Santiago, Chile: Journal of Forecasting.
- CES-BCSF. (2019). *Análisis: cíclico económico argentino y de la provincia de Santa Fe. 2002-2018*. Santa Fe.
- Choi, H., & Varian, H. (2009). *Predicting the Present with Google Trends*. Google Inc.
- Choi, S., Jun, S.-P., & Yoo, H. S. (2018). *Ten years of research change using Google Trends: From the perspective of big data utilizations and applications. Technological forecasting and social change*, 130, 69-87.
- Cohan, P. P., & D'Jorge, M. L. (2015). *Índice compuesto coincidente de actividad económica para la provincia de Santa Fe (Argentina): indicador mensual de alcance sub-nacional*. Santa Fe: Centro de Estudios y Servicios de la Bolsa de Comercio de Santa Fe.
- Cohan, P. P., D'Jorge, M. L., Henderson, S. J., & Sagua, C. E. (2007). *Proceso de contrucción del Índice Compuesto Coincidente Mensual de Actividad Económica de la Provincia de Santa Fe (ICASFe)*. Santa Fe: Centro de Estudios y Servicios de la Bolsa de Comercio de Santa Fe.
- Cohan, P., D'Jorge, M. L., & Lazzaroni, M. (2016). *Forcasts del X-13 ARIMA-SEATS aplicados al Índice de Actividad Económica Coincidente de la provincia de Santa Fe*. Santa Fe: Centro de Estudio de la Bolsa de Comercio de Santa Fe.
- D'Jorge, M. L., Cohan, P. P., Lazzaroni, M., Cherri, A., Camusso, F., & Zanini, L. (2018). *14 Sub-indicadores Considerados por el Índice Compuesto Coincidente de Actividad Económica de la Provincia de Santa Fe (ICASFe)*. Santa Fe, Argentina: Centro de

Estudios y Servicios - Bolsa de Comercio de Santa Fe.

Dergiades, T., Mavragani, E., & Pan, B. (2018). Google Trends and tourists' arrivals: Emerging biases and proposed corrections. *Tourism Management*, 66, 108-120.

Mohebbi, M., Vanderkam, D., Kodysh, J., Schonberger, R., Choi, H., & Kumar, S. (2011). *Google Correlate Whitepaper*.

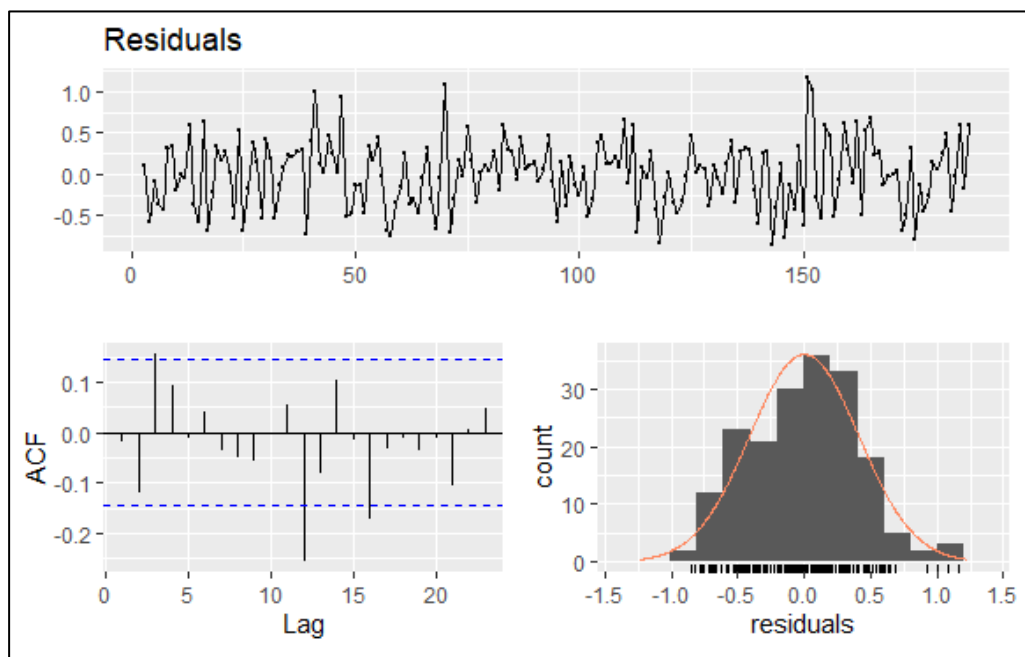
Naccarato, A., Falorsi, S., Loriga, S., & Pierini, A. (2018). Combining official and Google Trends data to forecast the Italian youth unemployment rate. *Technological Forecasting and Social Change*, 130, 114-122.

## Anexo

### A. Análisis de los residuos

La figura A1 muestra los residuos del ARDL(2,2), su correlograma y distribución. Como se puede observar, no se evidencia correlación significativa antes del orden 12, y la distribución es aproximadamente simétrica y acampanada, dando sustento a la hipótesis de normalidad y a los resultados obtenidos a través de las pruebas correspondientes.

**Figura A1:** Valores, correlograma y distribución de los residuos del modelo con rezagos distribuidos.



Fuente: elaboración propia en base a datos del CES-BCSF y Google Inc.