Google Correlate y Google Trends como herramientas para realizar un nowcast de las ventas minoristas

Camusso, María Florencia*; Jorge, Ramiro Emmanuel*

* Centro de Estudios y Servicios de la Bolsa de Comercio de Santa Fe (e-mail: ces@bcsf.com.ar) y Facultad de ciencias Económicas (FCE) de la Universidad Nacional del Litoral (UNL)[1].

Resumen

El trabajo internaliza información proveniente de las herramientas Google Trends y Google Correlate, con el objetivo de realizar un *nowcast* de las ventas de supermercados de la Provincia de Santa Fe; indicador que se publica con algunos meses de rezago.

En primer lugar se identifican un conjunto de variables *proxies* con alto poder predictivo y luego se plantea un método de agregación para incorporar los patrones de búsqueda a la serie *target*.

Las estimaciones obtenidas con el modelo, son contrastadas con datos reales de la serie *target* (ex post) y con los *forecasts* que arroja el X13-ARIMA-SEATS. 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 proposes a nowcasting model for Santa Fe's supermarkets retail sales, an indicator that is released within two months of delay, internalizing information from Google Trends and Google Correlate. The procedure identifies an array of proxy variables with high predictive capabilities and then uses the data in order to estimate the target series considering searching patterns.

Estimations computed by the model are compared to X13-ARIMA-SEATS's forecasts. Obtained output suggests that results are not only consistent but also more opportune that official statistical releases.

JEL classification codes: [E27], [E32]

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

[1] Los autores agradecen especialmente los comentarios y aportes del Dr. Juan Mario Jorrat.

1. Introducción

Desde el año 2007 el Centro de Estudios y Servicios de la Bolsa de Comercio de Santa Fe (CES-BCSF) lleva adelante un programa que estudia ciclos económicos a nivel sub-nacional. Su principal producto es el ICASFe[2], 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 objeto 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. En esta línea, hace algunos años se elaboró un *paper* con los avances que se implementaron utilizando los *forecasts* que arroja el X-13ARIMA-SEATS en el proceso de filtrado de las series componentes (Cohan, D'Jorge, & Lazzaroni, 2016).

En esta oportunidad, se decide aprovechar la disponibilidad de información proveniente de los motores de búsqueda de Internet para internalizar el comportamiento de las ventas minoristas, uno de los cuatro elementos fundamentales del ciclo económico (Achuthan & Banerji, 2004) y una de las series cuya publicación presenta mayor rezago en Argentina.

Google, el motor de búsqueda más difundido en la actualidad, proporciona dos herramientas al respecto: Google Trends y Google Correlate. La primera permite visualizar el flujo de búsquedas de palabras claves[3] a lo largo del tiempo en un espacio geográfico; mientras que Google Correlate proporciona un listado con las palabras cuyas búsquedas presentan mayor correlación con una serie de datos específica.

Un breve relevamiento de antecedentes señala que algunos autores (Artola, Galán, Askitas, Zimmermann, Schmidt y Vosen, entre otros), ya vienen utilizando la herramienta Google Trends para desarrollar indicadores económicos y algunas estimaciones de coyuntura. En todos estos casos, la principal dificultad fue establecer un criterio objetivo para seleccionar las palabras claves con mejores resultados. Por tal motivo, se decide complementar la selección subjetiva de los términos disparadores con la aplicación de Google Correlate para identificar aquellos potencialmente vinculados a la serie de referencia con mayor rigurosidad y precisión.

En cuanto a la estructura del *paper*, el primer punto que aborda el documento expone el marco de referencia y algunas generalidades conceptuales. Luego se detalla el proceso seguido por los autores para identificar y seleccionar las palabras claves con alto potencial predictivo. Dicha información es utilizada posteriormente para modelar la variable *target*. Finalmente se realizan estimaciones de coyuntura y se las compara con las proyecciones del X13. El último apartado presenta una síntesis de resultados junto a las principales conclusiones.

[2] ICASFe: Índice Compuesto Coincidente de Actividad Económica de Santa Fe.

[3] 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.

2. Marco de referencia y generalidades conceptuales

2.1. Google Trends

Google Trends es una herramienta diseñada para analizar las tendencias de las palabras clave que se ingresan en el motor de búsqueda, durante un periodo determinado y para una zona geográfica especifica.

Una característica relevante es que los datos del sistema se exponen en términos relativos. En este sentido, Google identifica el momento de mayor popularidad para el término en cuestión, asignándole un valor de 100. Para el resto de los registros, se establecen valores según la proporción de consultas realizadas con respecto a dicha referencia.

Por lo tanto, la información disponible no representa específicamente la cantidad de registros, sino un valor índice del número de búsquedas de una palabra clave (delimitado por región geográfica, fecha y categoría seleccionada). Luego se procede a un re escalamiento, siendo 100 el valor que representa el máximo de búsquedas en el periodo analizado (Choi & Varian, 2009).

El mecanismo de construcción del índice presentado por Google Trends, consta de la selección de una muestra imparcial de los datos de búsquedas. Adicionalmente, para evitar sesgos, se excluyen los registros referidos a un grupo reducido de usuarios, como también aquellos duplicados y no se tienen en cuenta los caracteres especiales.

De forma complementaria, esta herramienta brinda información respecto a “consultas relacionadas”. Estas consisten en una lista de términos que habitualmente son consultados conjuntamente con la palabra en cuestión. Por ejemplo, para el caso de “turismo”, una búsqueda relacionada es “hotel”, lo que implica que una considerable proporción de usuarios que consultó “turismo”, hizo lo propio con “hotel”.

2.2. Google Correlate

Otra de las herramientas que ofrece la plataforma es Google Correlate. La principal ventaja de este instrumento es que proporciona un método automatizado para la selección de consultas relacionadas. En este sentido, la elección de las palabras no requiere de conocimientos previos sobre el fenómeno bajo análisis.

Puntualmente, dado un patrón de interés temporal o espacial, se determina qué consultas imitan mejor los datos. Estas consultas de búsqueda pueden servir para construir una estimación del verdadero valor del fenómeno (*proxy*). Cada base de datos contiene decenas de millones de consultas, las cuales provienen de los registros anónimos de búsquedas en la web de Google, desde enero de 2004 hasta marzo de 2017.

El objetivo de Google Correlate es resaltar aquellas consultas en la base de datos cuyo patrón espacial o temporal está más altamente correlacionado (mediante contraste de coeficientes de correlación) con un patrón objetivo. Para esto, 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 (Mohebbi, y otros, 2011), utilizando un proceso inverso. En este caso, se parte de una serie de datos, y se obtiene una lista de palabras claves que presentan consultas similares a dicha serie; contrariamente, Google Trends parte de un término específico y arroja como resultado la tendencia de las búsquedas de esta palabra.

La unidad de medida de las salidas obtenidas mediante el uso de Google Correlate, se expresan en desviaciones estándar por encima y por debajo de la media. Es importante aclarar que estos datos están estandarizados, por lo que presentan una media de 0 y un desvió estándar de 1.

Una limitación importante de estas aplicaciones es que la existencia de una fuerte correlación con una serie durante un largo periodo de tiempo, no valida su continuidad a futuro, dado que es posible que se den cambios (estructurales) en el comportamiento de búsqueda de los usuarios.

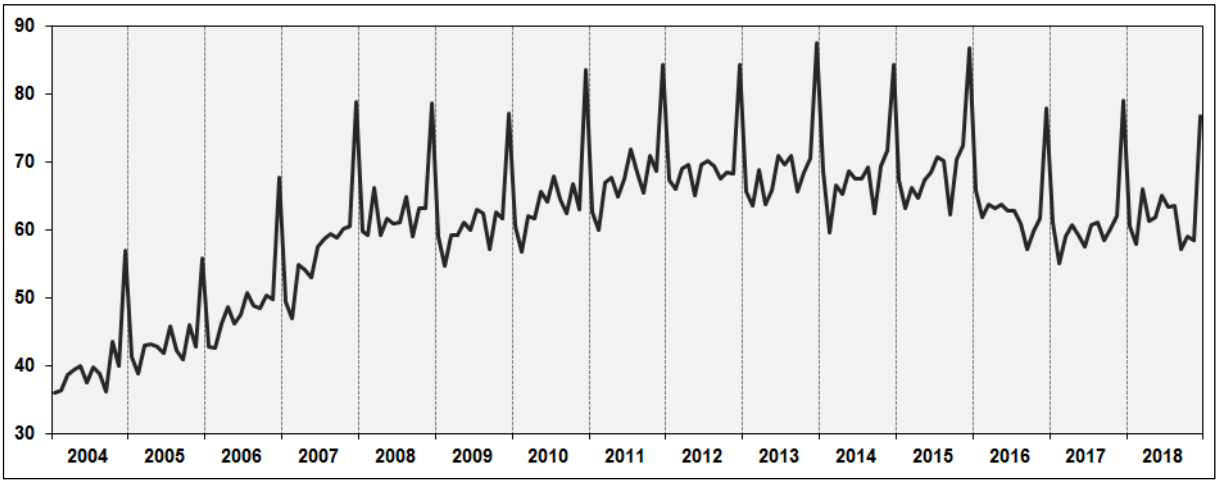
2.3. Variable *target*

Como ya se ha mencionado en la introducción, uno de los elementos más importantes para estudiar el comportamiento del ciclo económico es el flujo de las ventas minoristas. En la provincia de Santa Fe, dicha variable se internaliza por medio de una *proxy:* la serie publicada por el Instituto Nacional de Estadísticas y Censos (INDEC) que refiere concretamente a las ventas nominales de supermercados de la provincia de Santa Fe. El indicador refleja las ventas correspondientes a 68 bocas de expendio de grandes superficies localizadas en el territorio sub-nacional.

Para expresar la información en términos reales la serie de datos se deflacta utilizando un índice de precios minorista[4]. En el Gráfico 1 se exponen los datos de la serie publicada por INDEC desde enero de 2004 a diciembre de 2018 (límite propuesto como período temporal de análisis en el *paper),* expresados en valores constantes de 1993.

[4] Dada la falta de continuidad y consistencia que han tenido algunos indicadores en Argentina, se utiliza un empalme: Índice de Precios Implícitos (IPI) de la encuesta de supermercados (hasta mayo de 2007), luego el IPC-SFE hasta diciembre de 2011, el IPC Congreso entre enero de 2012 y agosto de 2013, y a partir de septiembre de 2013 se deflacta por el componente de alimentos y bebidas no alcohólicas del Índice de Precios al Consumidor de la Ciudad de Buenos Aires (IPCBA).

Gráfico 1: Ventas de supermercados en la provincia de Santa Fe. Millones de pesos de 1993. Período: 2004.01 a 2018.12.

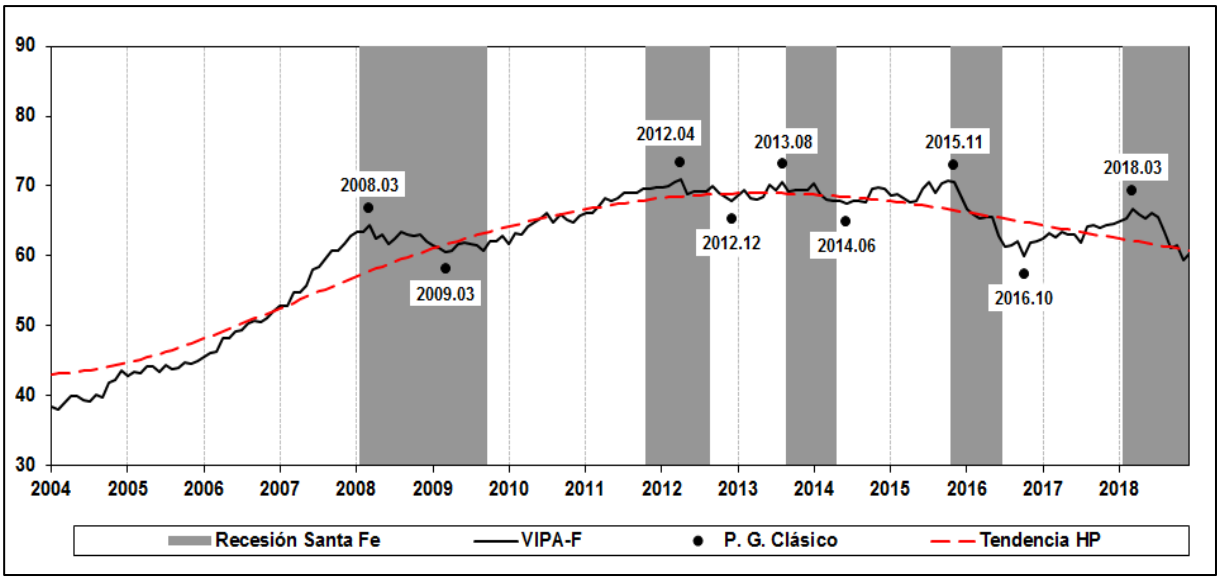


Fuente: Elaboración propia en base a datos de Google Inc. y CES-BCSF

Como se observa en el Gráfico 1, la serie posee un marcado patrón estacional con mayor actividad regular en torno a los meses de junio, julio y fundamentalmente diciembre; coincidiendo con los períodos vacacionales y las festividades de fin de año. Asimismo, como ya hemos mencionado, el indicador capta con claridad el movimiento cíclico de la economía, con una leve tendencia ascendente de largo plazo.

El Gráfico 2 muestra la serie filtrada, y para mejorar la lectura, incluye además los puntos de giro clásicos[5], y las recesiones y expansiones de la actividad económica de la provincia de Santa Fe. Por último, se muestra la tendencia de largo plazo, obtenida a través del filtro de Hodrick-Prescott (Tendencia HP).

Gráfico 2: Serie filtrada, filtro HP, puntos de giro clásicos y fases del ciclo de la economía santafesina (recesiones y expansiones). Millones de pesos de 1993. Período: 2004.01 a 2018.12.



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

[5] Representan mínimos y máximos relativos de la serie, en un entorno reducido.

Cabe aclarar que, como insumo del indicador coincidente, las ventas se filtran previamente por estacionalidad e irregularidad, utilizando específicamente el componente cíclico para internalizar el nivel de actividad[6].

Sin embargo, en adelante, se decide trabajar con la serie bruta sin filtrar, con el objetivo de identificar variables que capturen un patrón estacional análogo; el criterio adoptado se fundamenta en que el flujo de la serie debe contrastarse con el flujo de búsquedas que incorpora Google Correlate sobre la base de series temporales que no fueron sujetas a filtrados estacionales.

3. Variables *proxy*: identificación y selección

3.1. Utilización de la herramienta Google Correlate

La idea de fondo es ingresar a Google Correlate los datos de la serie "Ventas de supermercados de la provincia de Santa Fe a precios constantes de 1993 (VIPA)", obteniendo como resultado una lista de palabras cuyo movimiento presente un alto poder predictivo de la variable *target*.

Para ello, la serie *target* es introducida a Google Correlate en niveles, siguiendo requerimientos de formato establecidos por la aplicación[7]. Como primer paso, se decide utilizar un paquete que inicia en enero de 2004, a pesar de que la serie objetivo cuenta con datos mensuales desde 1994. Esto, por cuanto la base de datos disponible en Correlate contiene información desde enero de 2004 a marzo de 2017.

Una vez realizado este proceso, el programa arroja una lista de los 100 términos con mejor ajuste respecto de la variable objetivo: la serie de ventas de supermercados. La aplicación efectúa los cálculos de correlación luego de estandarizar el *input*[8]. Del total de palabras expuestas, se opta por tomar en cuenta cinco indicadores con alto nivel de correlación (ver Cuadro 1) y, al mismo tiempo, razonabilidad desde el punto de vista económico. A saber: "Carrefour", "El Entrerriano", "Falabella", "microcomponente" y "nuevo".

[6] Para realizar el filtrado por estacionalidad y valores irregulares se utiliza el software X13-ARIMA-SEATS.

[7] 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).

[8] La estandarización consiste en estimar $Z = (x - \mu)/\sigma$, donde x es el valor de la variable en cuestión, μ es la media poblacional de x y σ su desvío estándar. El resultado es un conjunto de series con media = 0 y desvió estándar = 1.

Cuadro 1: Nivel de correlación entre las palabras claves sugeridas por Google Correlate y la serie VIPA. Datos mensuales correspondientes al período: 2004.01 a 2017.03.

	R
Carrefour	0.853
ElEntrerriano	0.869
Falabella	0.869
Microcomponente	0.861
Nuevo	0.815

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

Aunque se seleccionaron palabras conceptualmente razonables (como variable explicativa), lo interesante de esta herramienta es que permite 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: “cumbias viejas”, “tablas de fiambres”, “frases de felicitaciones”, entre otros.

Además de evaluar los resultados correspondientes al *input* de la serie a precios contantes, se verificaron las salidas del sistema utilizando la serie de VIPA expresada en precios corrientes, y ajustada por estacionalidad y valores extremos. Reforzando la decisión adoptada a priori, ninguno de estos ejercicios arrojó palabras claves con significado económico.

Otro punto a mencionar respecto al proceso, es 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 en casos que la estructura económica del espacio sub-nacional sea radicalmente distinta a la nacional. Algo que no ocurre particularmente con las ventas minoristas, sujetas fundamentalmente al contexto macroeconómico general. Más aún, la actividad económica de la provincia de Santa Fe guarda una sincronía significativa con el flujo de actividad nacional en términos cíclicos (Centro de Estudios y Servicios de la Bolsa de Comercio de Santa Fe, 2019).

3.1.1. Las salidas de Google Correlate

Como resultado de esta primera etapa del proceso, la herramienta arroja un listado de 100 palabras cuyos coeficientes de correlación superan el 0.80. Adicionalmente, se puede descargar una serie correspondiente al historial de búsqueda de cada una de estas palabras, 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 el desenvolvimiento 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

Siguiendo con el procedimiento antes descripto, las palabras seleccionadas en Google Correlate se ingresaron a Google Trends. De esta forma se obtuvo el patrón de búsqueda actualizado de dichas variables.

Un segundo aporte que permite la herramienta como parte de la salida, es que identifica consultas relacionadas al término original que también poseen aptitudes predictivas. En este caso, se obtuvieron las siguientes: “Coto”, “Easy”, “Fravega”, “Garbarino”, “MercadoLibre”, “Musimundo” y “Walmart”. Estas sugerencias son incorporadas como base de análisis en los puntos sub-siguientes.

Por lo tanto, en esta instancia, se cuenta con doce *keywords* pre-seleccionadas[9] con un comportamiento similar al de las ventas de supermercados, e información disponible de forma oportuna.

Es importante aclarar que, si bien esta herramienta permite obtener resultados a nivel provincial, se decidió utilizar datos nacionales para armonizar 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 (en nuestro caso desde enero de 2004, acotado geográficamente al ámbito nacional). Es decir que, el valor 100 indica el momento en el cual el término alcanza su máxima popularidad.

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

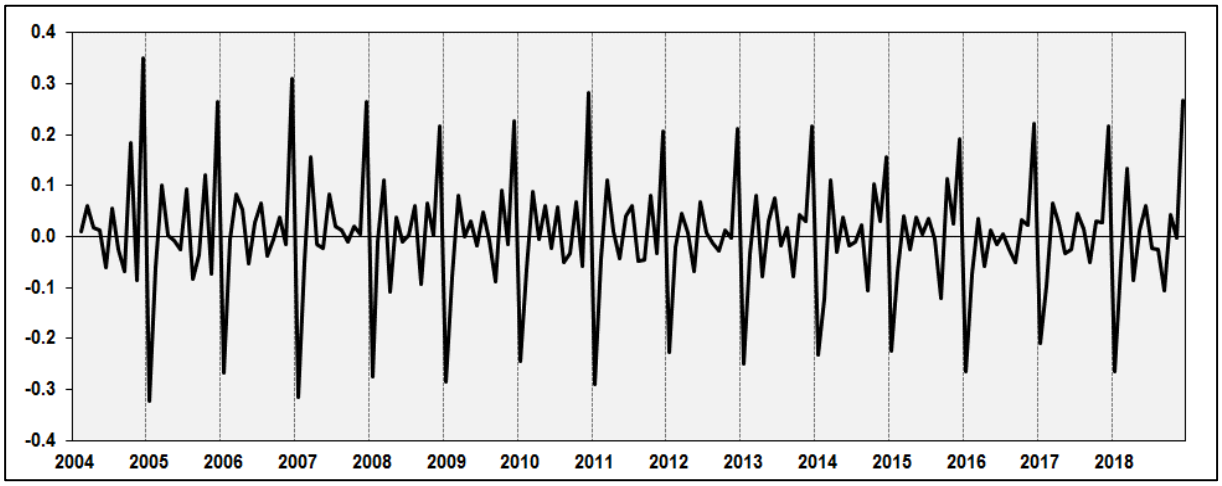
Una vez identificadas las variables *proxies*, el siguiente paso fue generar un marco procedimental que permitiera estimar los movimientos coyunturales de la variable *target* en función de la información disponible. En este marco, una nueva revisión de antecedentes permitió reconocer distintas metodologías. Por mencionar algunas de las alternativas, Blanco (2014) utiliza un modelo *Auto Regressive Moving Average (ARMA)*, aplicado a tasas de cambio mensuales, las cuales no fueron filtradas por estacionalidad. En otra línea, un trabajo realizado por la *Ruhr-Universität Bochum*, utiliza un modelo multivariable con internalización de Media Móvil (MM) con 3 *lags*. Por su parte, en el caso de Askitas y Zimmermann (2009), y Morán (2016), realizan el modelado a través del método de regresión lineal.

4.1. Análisis econométrico de las series

[9] “Carrefour”, “El Entrerriano”, “Falabella”, “microcomponente”, “nuevo”, “Coto”, “Easy”, “Fravega", “Garbarino”, “MercadoLibre”, “Musimundo” y “Walmart”.

Inicialmente se analizan las características de la serie *target* para corroborar sus propiedades en relación a requisitos estadísticos fundamentales. En este sentido, se evalúa si la serie es o no estacionaria[10]. Como se puede inferir del Gráfico 1, al igual que muchas series económicas, VIPA en niveles denota la presencia de una tendencia positiva, por lo que su media no es constante. En contraposición, la tasa de cambio mensual logarítmica (TCML) de la serie, representada en el Gráfico 3, tiene media y varianza aproximadamente constantes a lo largo del tiempo.

Gráfico 3: Tasa de cambio mensual logarítmica de las ventas de supermercados en la provincia de Santa Fe. Período: 2004.01 a 2018.12.



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

Para corroborar la estacionariedad o no de la serie, se realiza el test de Dickey-Fuller aumentado (Augmented Dickey-Fuller, ADF). Los resultados para VIPA en nivel y para su TCML se muestran en el Cuadro 2. El contraste fue realizado con doce retardos, utilizando el criterio de Akaike (AIC), considerando constante y tendencia.

Cuadro 2: Resultados del test de Augmented Dickey-Fuller de la variable VIPA en niveles y de su tasa de cambio mensual logarítmica. Datos mensuales: 2004.01 a 2018.12.

	Valor P
Contraste ADF VIPA niveles	0.964
Contraste ADF TCML(VIPA)	0.028

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

Para las variables en niveles, no puede rechazarse la hipótesis de raíz unitaria al 95.0% de confianza, mientras que si aplicamos primeras diferencias, sí. Por lo tanto, a partir de aquí, se utiliza la TCML, tanto de la variable *target* como de las *proxies*[11]. Los correspondientes ADF de cada variable explicativa arrojan resultados favorables, siendo todas estacionarias[12].

[10] Una serie es estacionaria cuando es estable a lo largo del tiempo, es decir, cuando su media y varianza son constantes en el tiempo.
[11] Ver gráficos en Anexo I.
[12] Ver resultados en Anexo II.

4.2. Determinación de un modelo de agregación

4.2.1. Regresión lineal múltiple: una primera aproximación

Una vez definido trabajar con la TCML de las variables, se realiza una regresión lineal múltiple para sondear el aporte explicativo que presenta cada *proxy*. En cuanto al rango de datos, debido a que el objetivo del trabajo es realizar una estimación para el valor actual de las ventas, se considera un paquete acotado a los últimos cinco años. En este sentido, el modelo de regresión lineal múltiple para aproximar la relación entre la variable dependiente Y – en nuestro caso VGT[13] –, y las variables independientes X_i – las TCML de las *proxies* de VIPA – puede ser expresado como:

$$\hat{y}_t = B_0 + \sum_{i=1}^{n} B_i x_{it} + \epsilon_t$$

$$n = 1, \dots, 12$$

$\hat{y}_t$: Valor estimado de las TCML de la variable dependiente, en el momento t.

x_{it} : Valor de las TCML de la variable independiente i, en el momento t.

B_0 : Intersección o término constante de la regresión.

B_i : Mide la influencia de cambios en x_i sobre $\hat{y}$.

ϵ_t : Término que representa todos aquellos factores diferentes a las variables explicativas, que ayudan a explicar $\hat{y}$, en el momento t.

En el Cuadro 3 se observan los resultados de la regresión de VGT en relación a las TCML de las variables *target,* aplicando el modelo de regresión lineal múltiple con un nivel de significancia del 95.0% (rango de datos: 2014.01-2018.12).

Cuadro 3: Estadísticos de la regresión lineal múltiple del modelo con doce variables independientes. Período 2014.01 a 2018.12.

	Coeficiente	Desv. Típica	Valor P
const	-0.004	0.007	0.584
TCML(Carrefour)	0.082	0.075	0.280
TCML(Coto)	**0.361**	**0.099**	**0.001**
TCML(Easy)	-0.157	0.125	0.217
TCML(ElEntrerriano)	0.001	0.062	0.984
TCML(Falabella)	-0.072	0.127	0.573
TCML(Fravega)	0.067	0.172	0.697
TCML(Garbarino)	-0.108	0.151	0.478
TCML(MercadoLibre)	-0.212	0.176	0.234
TCML(Microcomponente)	-0.042	0.029	0.159
TCML(Musimundo)	**0.255**	**0.092**	**0.008**
TCML(Nuevo)	**0.185**	**0.087**	**0.037**
TCML(Walmart)	0.004	0.045	0.938

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

[13] VGT: Ventas Google Trends.

Las variables significativas al modelo se resaltaron en negrita en el cuadro precedente, siendo estas: “Coto”, “Musimundo” y “Nuevo”. A continuación se exponen los resultados de una nueva regresión de TCML(VIPA) pero en relación a las tres variables previamente seleccionadas (Cuadro 4).

Cuadro 4: Estadísticos de la regresión lineal múltiple del modelo con tres variables independientes. Período 2014.01 a 2018.12.

	Coeficiente	Desv. Típica	Valor P
const	-0.002	0.007	0.818
TCML(Coto)	0.299	0.078	0.000
TCML(Musimundo)	0.144	0.044	0.002
TCML(Nuevo)	0.245	0.070	0.001

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

Esta regresión obtuvo un R^2 ajustado de 0.717 y se realizaron los correspondientes test para constatar la correcta especificación del modelo.

Los principales estadísticos, resumidos en el Cuadro 5, evalúan problemas de autocorrelación de los residuos (mediante la prueba LM de Breusch-Godfrey), y de ausencia de normalidad en la distribución de los mismos (prueba de normalidad Jarque-Bera). Adicionalmente se realiza una prueba ARCH que evalúa la no existencia de heterocedasticidad en doce rezagos y un RESET que contrasta la linealidad del modelo. Por su parte el estadístico VIF (*Variance Inflation Factor*) evalúa la colinealidad de las variables explicativas.

Cuadro 5: Resultados de contrastes sobre la correcta especificación del modelo bajo los supuestos de la regresión lineal múltiple para el modelo con tres variables.

	Valor P		VIF
RESET (cuadrado)	0.206	TCML-Coto	2.326
RESET (cubo)	0.462		
RESET (cuadrado-cubo)	0.341	TCML-Musimundo	1.497
Contraste de autocorr. de residuos (12 rezagos)	0.000		
Contraste de ARCH (12 rezagos)	0.341	TCML-Nuevo	1.842
Contraste Jarque-Bera	0.080		

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

Los resultados indican que un modelo lineal es adecuado, que los residuos presentan una distribución aproximadamente normal y que no hay evidencia de heteroscedasticidad ni colinealidad entre las variables explicativas. Sin embargo, existe evidencia de autocorrelación de los residuos, por lo tanto, siguiendo recomendaciones econométricas (ver Gujarati & Porter, 2009 y Wooldridge, 2005), y en función al correlograma de los residuos[14], se decide incorporar variables autorregresivas (AR) para lograr una mejor especificación del modelo[15].

[14] Ver en Anexo III.

[15] Una forma alternativa para salvar los problemas de autocorrelación es el uso de variables dicotómicas, sin embargo, este método implica incorporar once variables *dummies*, que agrega

4.2.2. Regresión lineal múltiple: incorporando AR(p)

Al modelo de regresión especificado con tres variables se le incorpora un AR(1). El nuevo R^2 ajustado del modelo se incrementa a 0.758. Sus estadísticos y contrastes se resumen en los cuadros 6 y 7.

Cuadro 6: Estadísticos de la regresión lineal múltiple del modelo con tres variables independientes y VIPA con un rezago. Período 2014.01 a 2018.12.

	Coeficiente	Desv. Típica	Valor P
const	-0.003	0.007	0.701
TCML(Coto)	0.259	0.073	0.001
TCML(Musimundo)	0.131	0.041	0.003
TCML(Nuevo)	0.286	0.066	0.000
TCML(VIPA t-1)	-0.218	0.068	0.002

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

Cuadro 7: Resultados de contrastes sobre la correcta especificación del modelo bajo los supuestos de la regresión lineal múltiple con tres variables y VIPA con un rezago.

	Valor P		VIF
RESET (cuadrado)	0.580	TCML-Coto	2.395
RESET (cubo)	0.786	TCML-Musimundo	1.513
RESET (cuadrado-cubo)	0.849	TCML-Nuevo	1.911
Contraste de autocorr. de residuos (12 rezagos)	0.000	TCML-VIPA (t-1)	1.073
Contraste de ARCH (12 rezagos)	0.610		
Contraste Jarque-Bera	0.031		

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

El *output* obtenido señala que el nuevo modelo no presenta problemas adicionales de especificación, aunque el contraste de autocorrelación denota la persistencia de este efecto aun en el modelo con AR(1). Es por esto que se decide utilizar un AR(2).

Cuadro 8: Estadísticos de la regresión lineal múltiple del modelo con tres variables independientes y VIPA con uno y dos rezagos. Período 2014.01 a 2018.12.

	Coeficiente	Desv. Típica	Valor P
const	-0.004	0.006	0.432
TCML(Coto)	0.236	0.058	0.000
TCML(Musimundo)	0.119	0.033	0.001
TCML(Nuevo)	0.314	0.053	0.000
TCML(VIPA t-1)	-0.320	0.057	0.000
TCML(VIPA t-2)	-0.313	0.055	0.000

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

La incorporación conjunta de $VIPA_{t-1}$ y $VIPA_{t-2}$, presenta coeficientes negativos con un alto nivel de significancia estadística. El valor de R^2 ajustado se incrementa en este caso a 0.846. Finalmente, como puede observarse en el Cuadro 9, el resultado del contraste de

complejidad al modelo. Se optó nuevamente por el criterio de parsimonia y se descartó esta alternativa.

autocorrelación para el modelo con AR(2) respalda el rechazo de la hipótesis nula, e indica que esa limitación queda salvada.

Cuadro 9: Resultados de contrastes sobre la correcta especificación del modelo bajo los supuestos de la regresión lineal múltiple con tres variables y VIPA con uno y dos rezagos.

	Valor P		VIF
RESET (cuadrado)	0.376	TCML-Coto	2.406
RESET (cubo)	0.460	TCML-Musimundo	1.519
RESET (cuadrado-cubo)	0.629	TCML-Nuevo	1.928
Contraste de autocorr. De residuos (12 rezagos)	0.206	TCML-VIPA (t-1)	1.189
Contraste de ARCH (12 rezagos)	0.598	TCML-VIPA (t-2)	1.110
Contraste Jarque-Bera	0.719		

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

Para reforzar la elección del modelo con AR(2), se contrastó el mismo con las variantes de AR(1) y sin AR a través de criterios estadísticos ampliamente aceptados: Criterio de Akaike (AIC), Criterio de Schwarz (BIC) y Durbin-Watson (D-W). Los resultados de los contrastes se exponen a continuación.

Cuadro 10: Criterios estadísticos para comparación de los modelos con tres variables sin corrección de autocorrelación, con AR(1) y con AR(2).

	Akaike	Schwarz	Durbin-Watson
Modelo de 3 variables	-168.447	-160.070	2.593
Modelo de 3 variables con AR(1)	-176.865	-166.393	2.686
Modelo de 3 variables con AR(2)	**-203.333**	**-190.767**	**2.552**

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

El Cuadro 10 evidencia que el modelo de tres variables incorporando uno y dos rezagos de la variable dependiente arroja mejores valores estadísticos.

Concluyendo este apartado, el modelo de estimación de la TCML de las ventas incorporando variables autorregresivas hasta el orden 2 se denomina $VGT_t^{AR(2)}$ y queda planteado de la siguiente manera:

$$VGT_t^{AR(2)} = -0.004 + 0.236\,TCML(Coto)_t + 0.119\,TCML(Musimundo)_t + 0.314\,TCML(Nuevo)_t - 0.320\,TCML(VIPA)_{t-1} - 0.313\,TCML(VIPA)_{t-2}$$

Sin embargo, como puede observarse, en este modelo los coeficientes de las variables rezagadas tienen un peso relativo elevado. Es decir, que las estimaciones se apoyan en gran medida en los datos históricos de la variable *target*. En este sentido, aunque el modelo de tres variables señale ciertas limitaciones, sus estimaciones presentan mayor autonomía respecto al pasado, pues se focalizan con mayor firmeza sobre los datos de variables independientes a la serie *target*. Este es un factor relevante, al menos desde el punto de vista teórico, dada su mayor capacidad para predecir puntos de giro; en contraste con la limitación que suelen

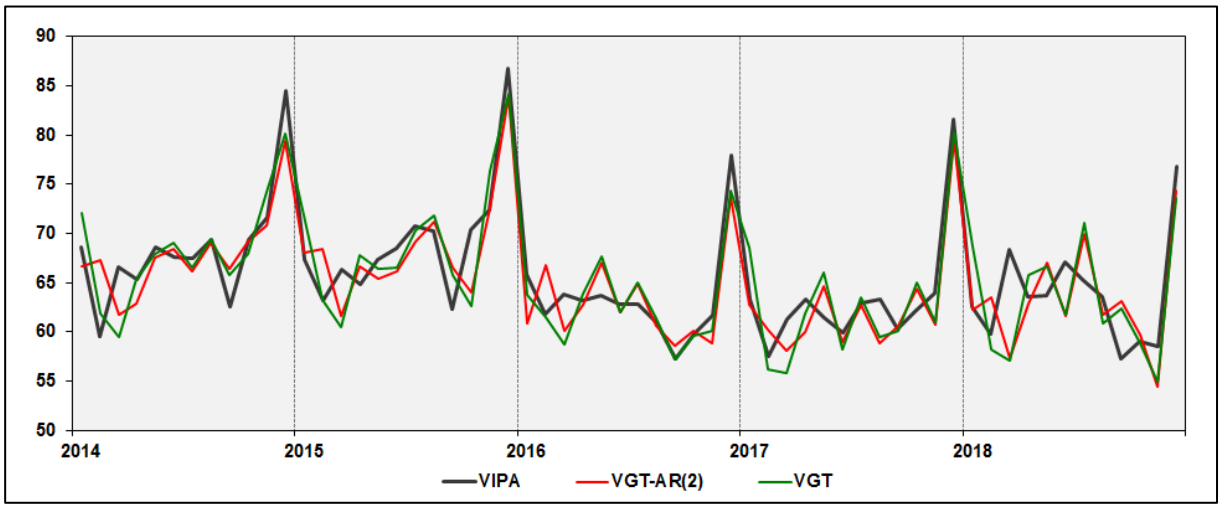
presentar los *forecasts* tradicionales. La tasa de cambio mensual logarítmica estimada sin corrección de autocorrelación (sin AR) queda expresada como sigue:

$$VGT_t = -0.002 + 0.299\, TCML(Coto)_t + 0.144\, TCML(Musimundo)_t + 0.245\, TCML(Nuevo)_t$$

4.3. Estimación mediante aplicación del modelo

En línea con las consideraciones y el procedimiento llevado adelante hasta esta instancia, se utilizan los modelos de regresión lineal múltiple sin AR y con AR(2), para realizar una estimación de VIPA desde 2014.01 hasta 2018.12. Los valores obtenidos representan una reconstrucción estimada de la tasa de cambio mensual logarítmica de las ventas de supermercados de la provincia de Santa Fe, por medio de variables *proxies*. Para poder expresar las series en valores representativos de las ventas, es decir, en unidades monetarias, se aplicó la TCML de $VGT_t^{AR(2)}$ y VGT_t obtenida a través de ambas estimaciones, al valor consolidado de las ventas en el período anterior.

Gráfico 3: Ventas de supermercados en la provincia de Santa Fe en millones de pesos de 1993 y estimación de las ventas mediante regresión lineal múltiple. Período: 2014.01 a 2018.12.



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

Como se observa en el Gráfico 3, mediante el procedimiento de mínimos cuadrados ordinarios, aplicando el modelo de tres variables sin AR y con AR(2), se obtiene una sólida estimación del valor de las ventas minoristas a precios constantes.

En primer lugar, se capta la estacionalidad, presentando ambas series picos coincidentes en el mes de diciembre de cada año. Por otra parte, el componente tendencia-ciclo logra replicarse a través de las estimaciones realizadas; quedando de esta manera captados los principales componentes de la serie de tiempo.

5. Estimaciones para 2019 y errores de predicción

Una vez planteados los modelos que mejor replican la serie de referencia, se procede a utilizar los mismos para estimar los meses de enero a junio de 2019. Dicho *nowcast* se realiza ponderando las variaciones mensuales de las *proxies* por los coeficientes obtenidos en la regresión[16].

La estimación efectuada mes a mes (cada dato se calcula con información disponible al mes anterior) representa una tasa de cambio, la cual se aplica al último dato disponible de la serie bruta de ventas. Luego, se filtra por estacionalidad e irregularidad.

Los valores reales de VIPA-F (ex-post) se comparan con las estimaciones de cada modelo y con el *forecast* tradicional obtenido con el X13-ARIMA-SEATS[17]. El Cuadro 11 expone los resultados y los errores de predicción para los primeros cinco meses de 2019.

Cuadro 11: Series filtradas por estacionalidad y valores irregulares mes a mes mediante X13-ARIMA-SEATS. Millones de pesos contantes de 1993. Periodo: 2019.01 a 2019.06.

Meses	VIPA-F (valores reales ex-post)	Forecast X13-ARIMA-SEATS	Nowcast (1)	Nowcast (2)	Error de predicción (%)		
					X13-ARIMA-SEATS	Nowcast (1)	Nowcast (2)
2019.01	60.057	61.143	61.042	60.713	1.8%	1.6%	1.1%
2019.02	59.366	60.165	59.851	59.340	1.3%	0.8%	0.0%
2019.03	59.266	59.671	59.692	59.083	0.7%	0.7%	-0.3%
2019.04	57.784	59.669	58.273	58.143	3.3%	0.8%	0.6%
2019.05	56.648	57.483	56.331	56.580	1.5%	-0.6%	-0.1%
2019.06	#N/A	56.246	57.366	57.522	#N/A	#N/A	#N/A

VIPA-F (valores reales ex-post): Serie filtrada de las ventas de supermercado en la provincia de Santa Fe.
Forecast X13-ARIMA-SEATS: datos estimados para VIPA-F mediante el X13-ARIMA-SEATS.
Nowcast (1): Nowcast de las ventas incorporando estimaciones del modelo RLM para tres variables mes a mes.
Nowcast (2): Nowcast de las ventas incorporando estimaciones del modelo RLM para tres variables + autorregresivos mes a mes.

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

El sentido de las tasas de cambio de la variable real y el de las estimaciones coinciden en todos los casos salvo en junio. Esto puede ir en línea con la idea teórica de que los *nowcasts* se comportan de manera más adecuada ante posibles puntos de giro (aún no se encuentra disponible el dato real a la fecha de cierre de este *paper*).

En cuanto a la precisión de las estimaciones, los errores de predicción de la proyección efectuada por el X13-ARIMA-SEATS se ubican en un rango de 0.7 a 3.3%.

Por su parte, los resultados obtenidos con los *nowcast* se muestran aún más certeros. El modelo de tres variables, que internaliza datos completamente independientes (como son las búsquedas en Google) permitió estimar las ventas para los meses de enero a mayo con errores que van de -0.6 a 1.6%. La incorporación de autorregresivos, logró obtener errores

[16] Los coeficientes podrían recalcularse cada 12 meses.
[17] Para mayor detalle consultar Cohan, D'Jorge & Lazzaroni (2016).

aún menores, ubicados entre -0.3 y 1.1%, debiendo considerarse que los AR incorporan información pasada de la variable dependiente, a la hora de predecir el valor actual de la misma.

Por todo esto, se puede afirmar que con este mecanismo, se logra un alto grado de precisión en las estimaciones, permitiendo además ganar hasta dos meses de oportunidad.

6. Síntesis de resultados y comentarios finales

Los resultados de este *paper* nos permiten afirmar que el uso conjunto de Google Trends y Google Correlate ha sido satisfactorio para identificar variables *proxies* y realizar un *nowcast* de las ventas de supermercados en la provincia de Santa Fe.

El procedimiento desarrollado toma mayor relevancia en cuanto puede replicarse fácilmente a otros indicadores que también presentan rezagos en sus publicaciones. Comparativamente con 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 de adelanto.

Aunque existen otras alternativas, la investigación realizada permitió identificar dos modelos de estimación (*nowcast*) con resultados de especificación muy sólidos que superaron las expectativas. Las proyecciones para 2019 fueron aún más precisas que las sugeridas por el *forecast* del X13-ARIMA-SEATS.

7. Bibliografía

Achuthan, L., & Banerji, A. (2004). *Beating the business cycle.* Nueva York: Currency Doubleday.

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?* ASOCIACION ARGENTINA DE ECONOMIA POLITICA.

Centro de Estudios y Servicios de la Bolsa de Comercio de Santa Fe. (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.

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.

Jimenez, L. (2017). *Can Google Trends data help to improve the nowcasting and short-term forecasting of the arrivals of tourists to the Dominican Republic?* International Center for Research and Study on Turismo (CIRET).

Jorrat, J. (2005). Construcción de índices compuestos mensuales coincidente y líder de Argentina. En M. (. Marchionni, *Progresos en econometría* (págs. 43-100). Ciudad Autonoma de Buenos Aires: Asociación Argentina de Economía Política.

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

Morán, J. (2016). *Google Trends: una nueva herramienta para la predicción económica.* Victoria, Buenos Aires: Universidad de San Andrés.

Schmidt, T., & Vosen, S. (2009). *Forecasting Private Consumption: Survey-based Indicators vs. Google Trends.* Alemania: Ruhr-Universität Bochum (RUB), Department of Economics Universitätsstr.

Anexo I

Los siguientes gráficos muestran la evolución de las búsquedas en Google de las doce variables *proxy* que surgieron del proceso de selección. Se las compara con la serie de referencia, VIPA, expresadas en tasas de cambio mensual logarítmicas desde enero de 2004 hasta diciembre de 2018.

Gráfico A: evolución de las búsquedas de “Carrefour”

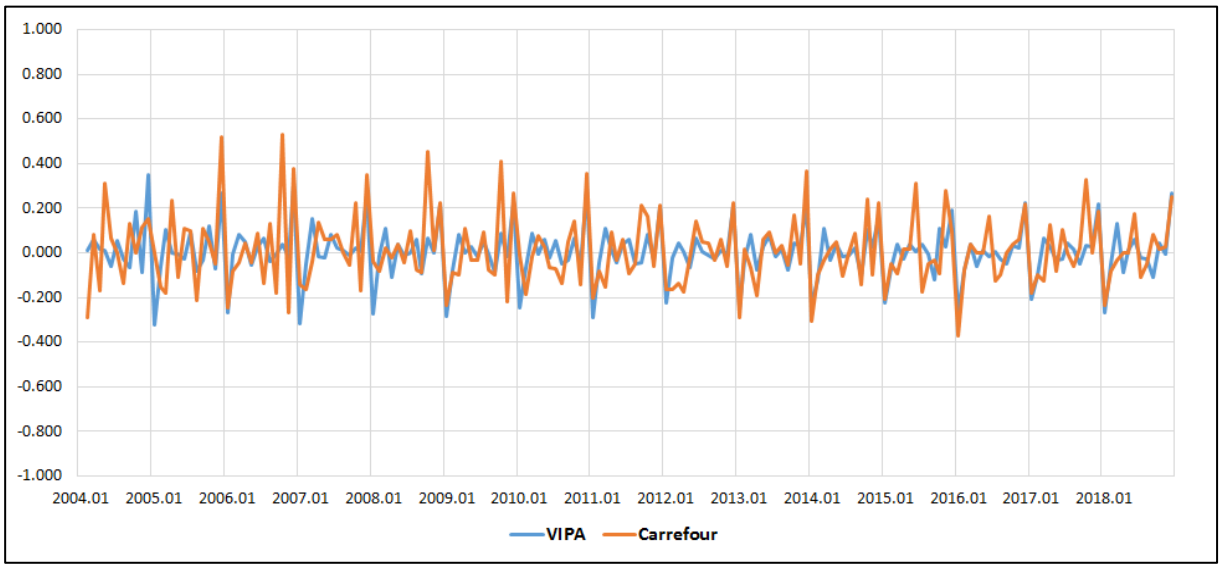


Gráfico B: evolución de las búsquedas de “Coto”

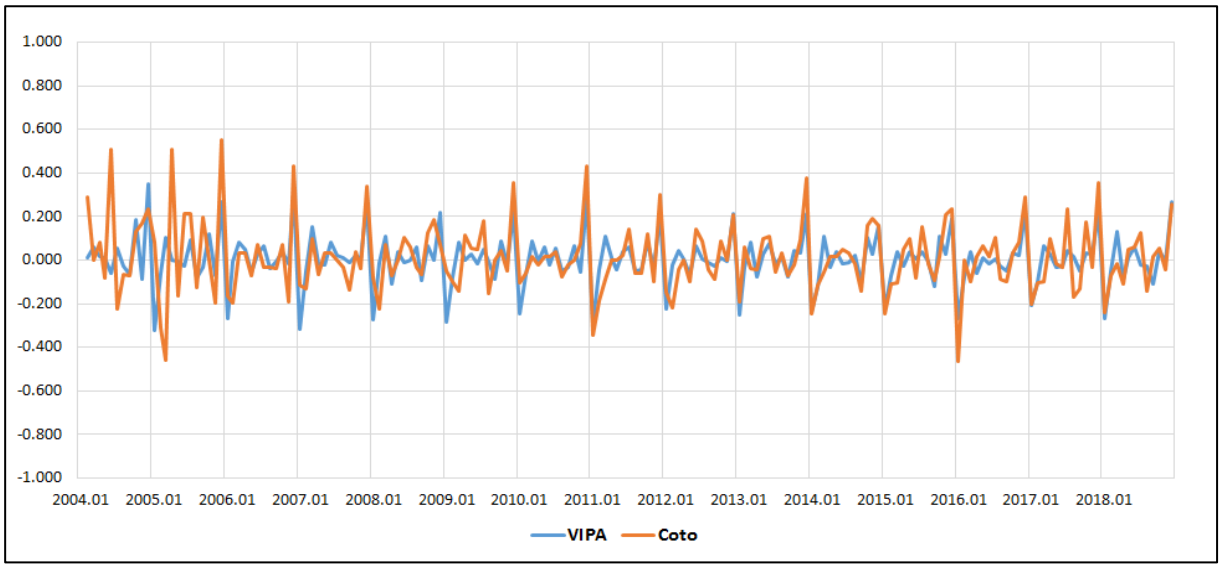


Gráfico C: evolución de las búsquedas de “El Entrerriano”

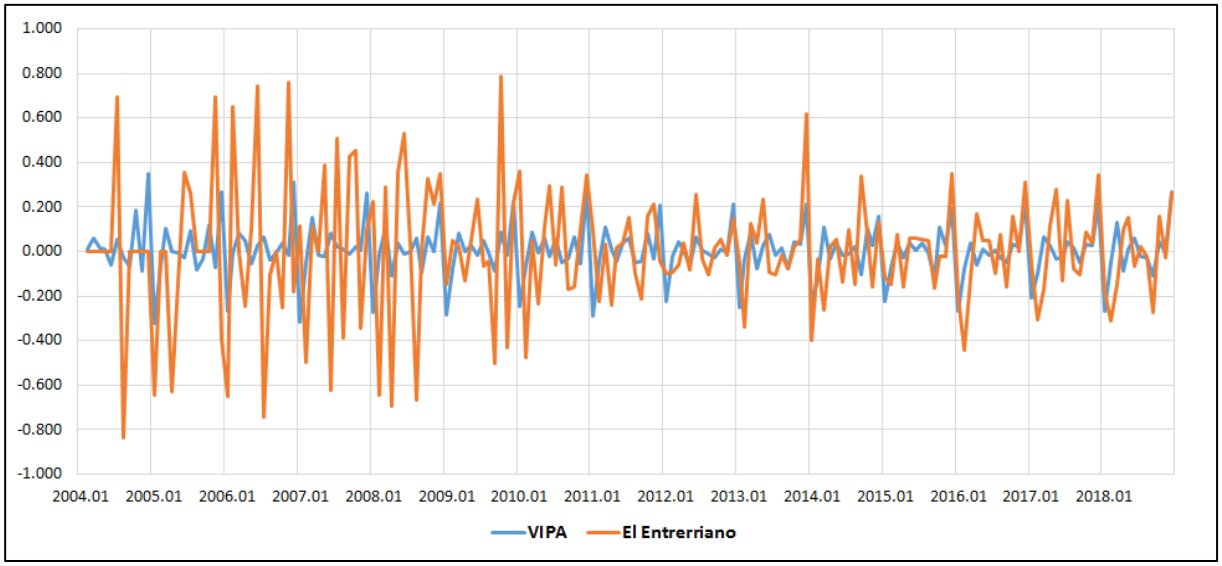


Gráfico D: evolución de las búsquedas de “Easy”

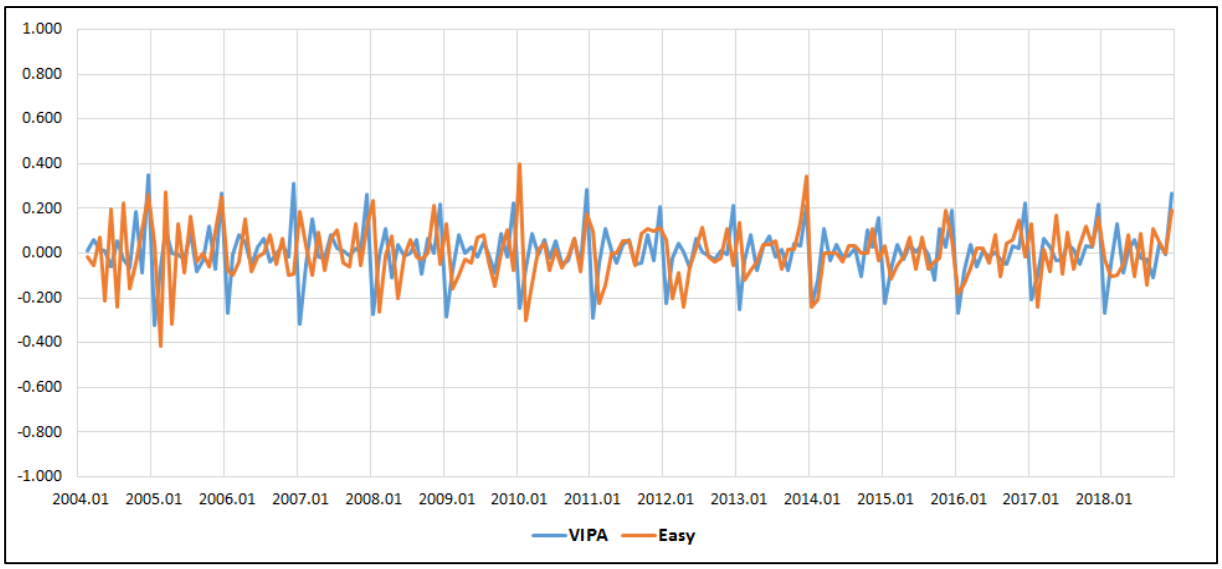


Gráfico E: evolución de las búsquedas de “Falabella”

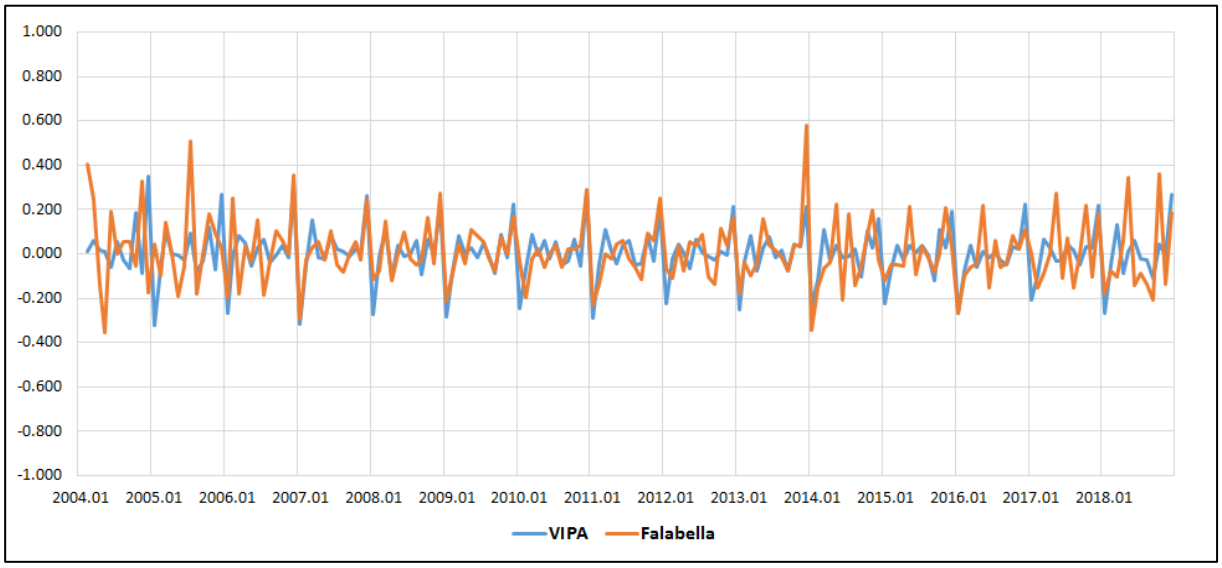


Gráfico F: evolución de las búsquedas de “Fravega”

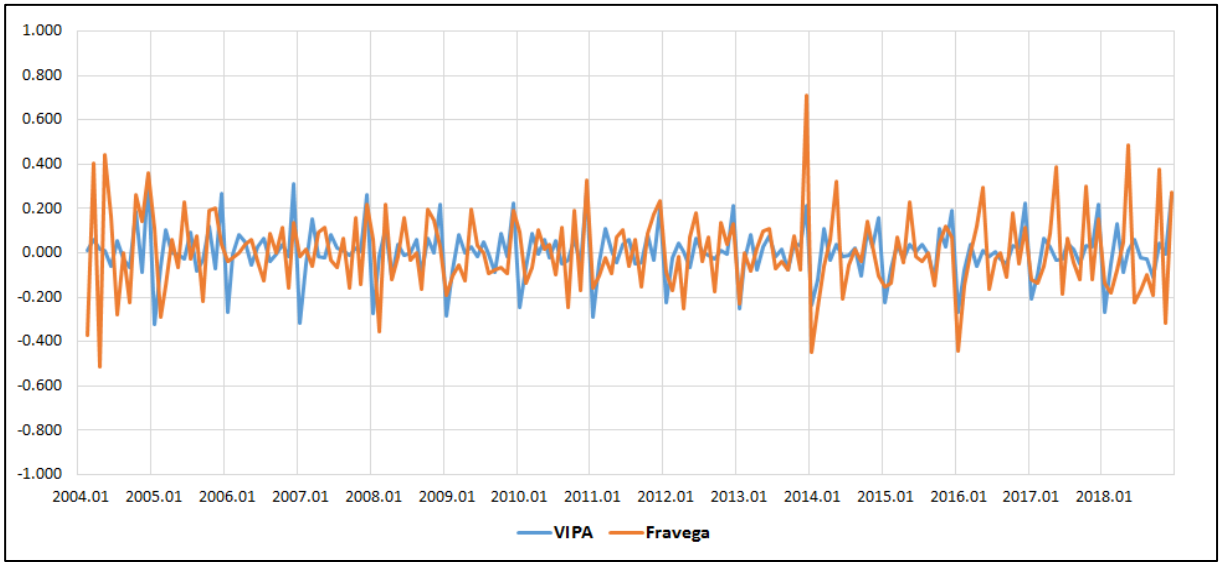


Gráfico G: evolución de las búsquedas de “Garbarino”

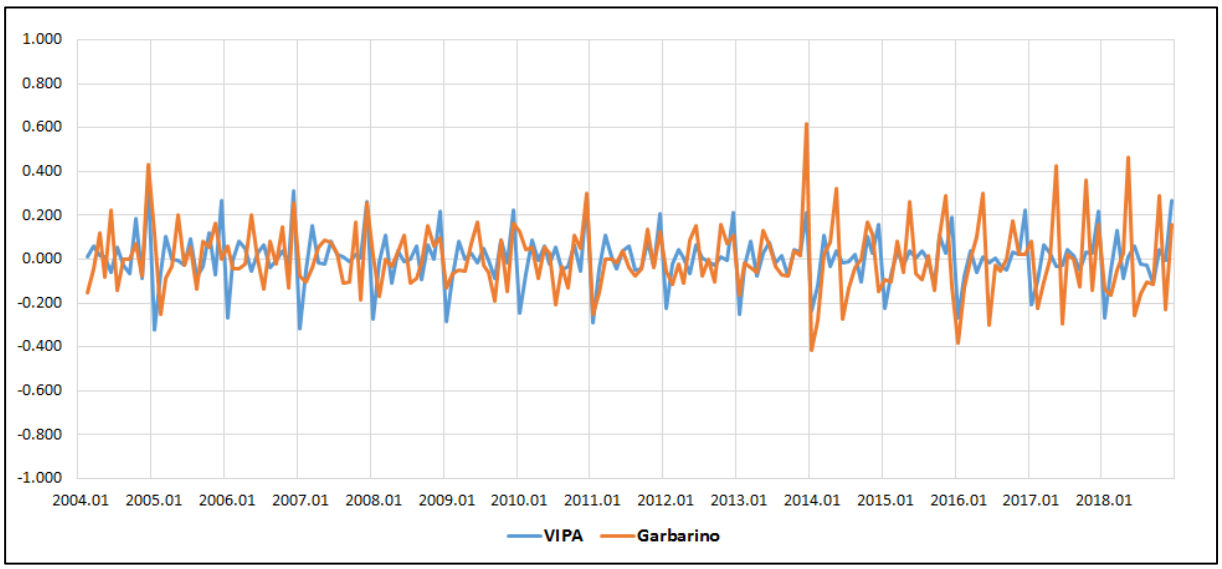


Gráfico H: evolución de las búsquedas de “MercadoLibre”

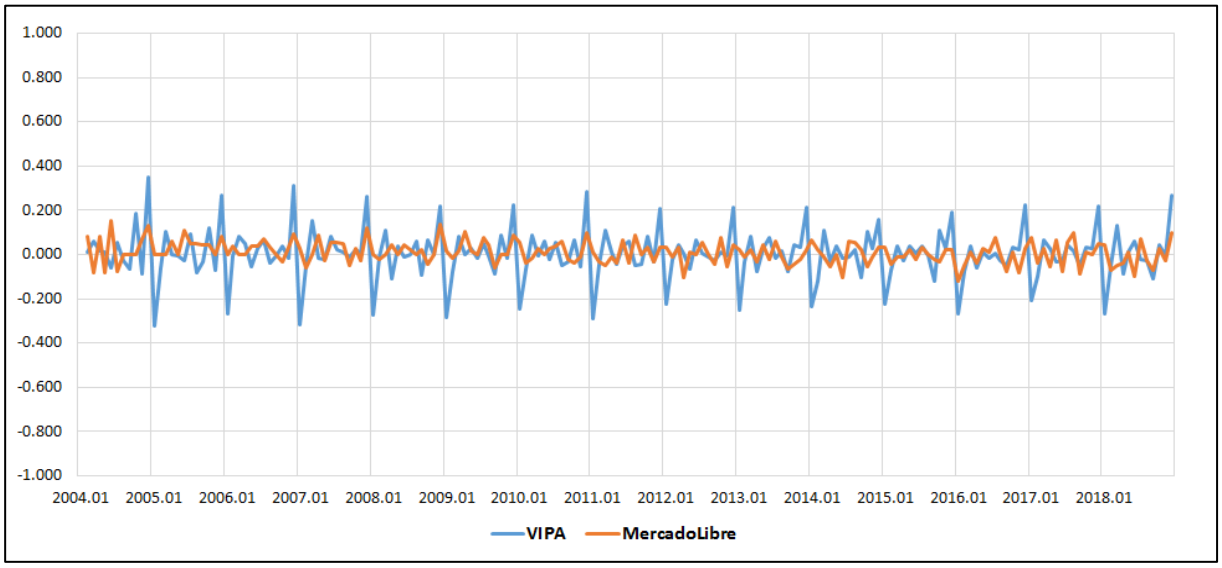


Gráfico I: evolución de las búsquedas de “Microcomponente”

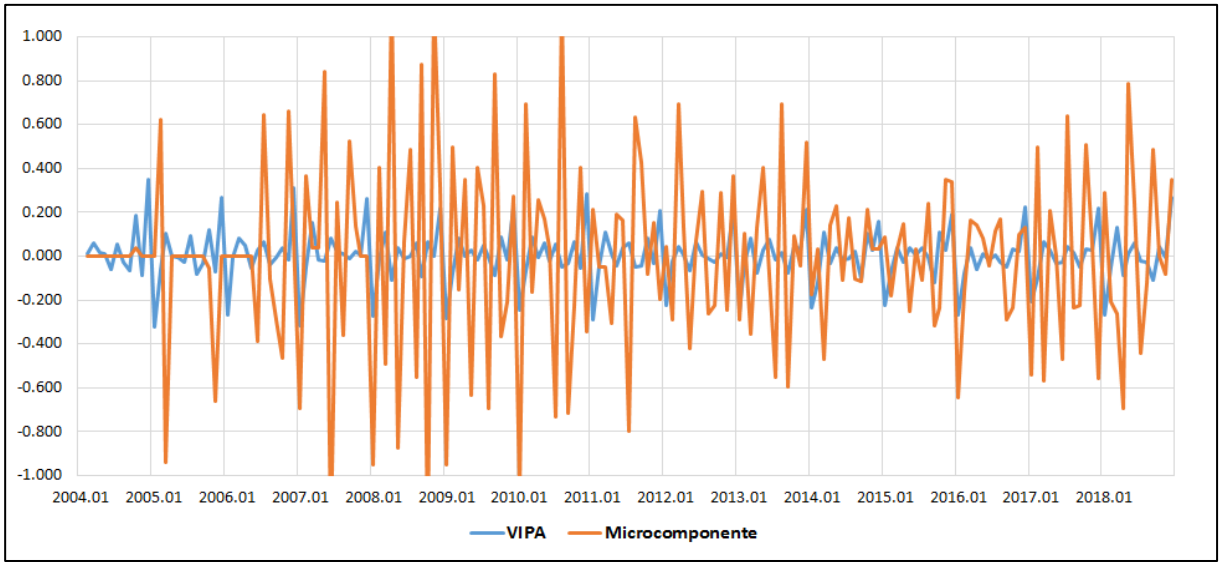


Gráfico J: evolución de las búsquedas de “Musimundo”

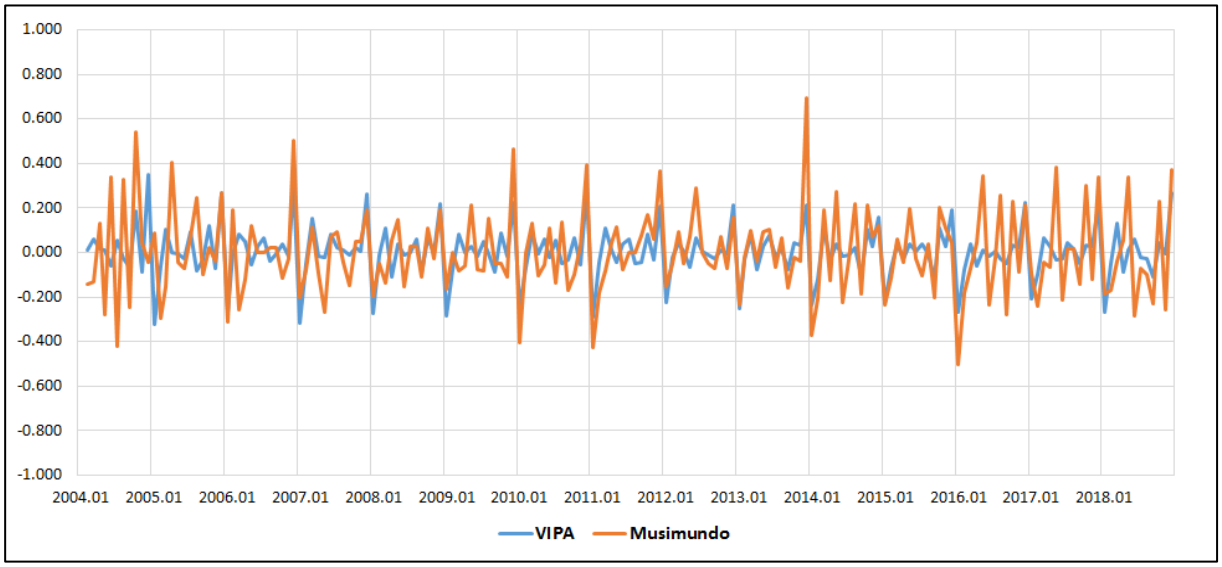


Gráfico K: evolución de las búsquedas de “Nuevo”

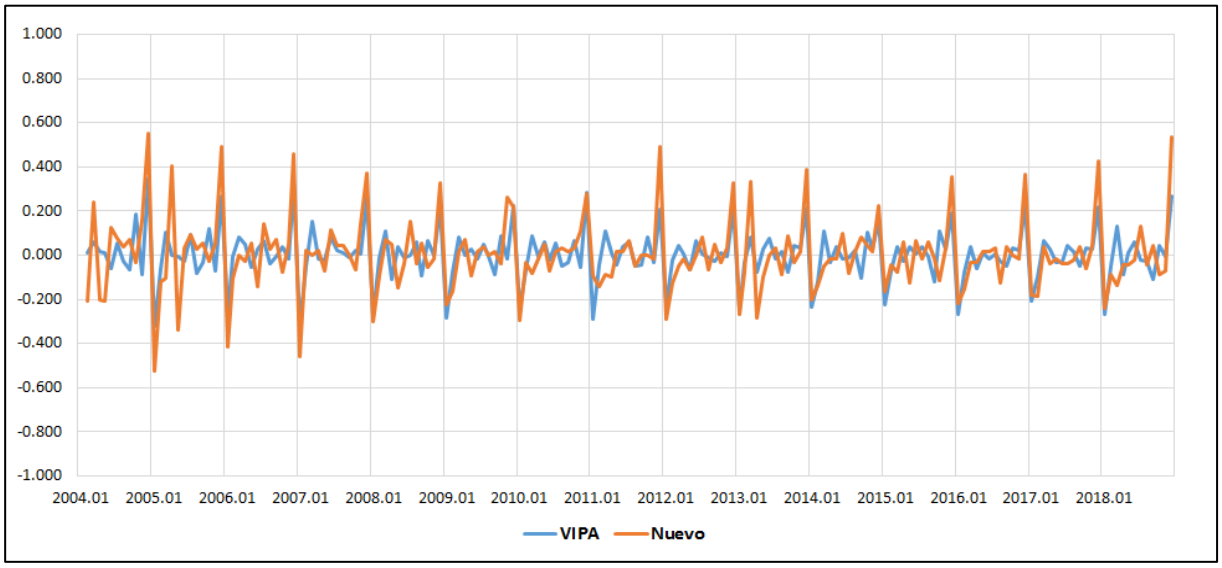
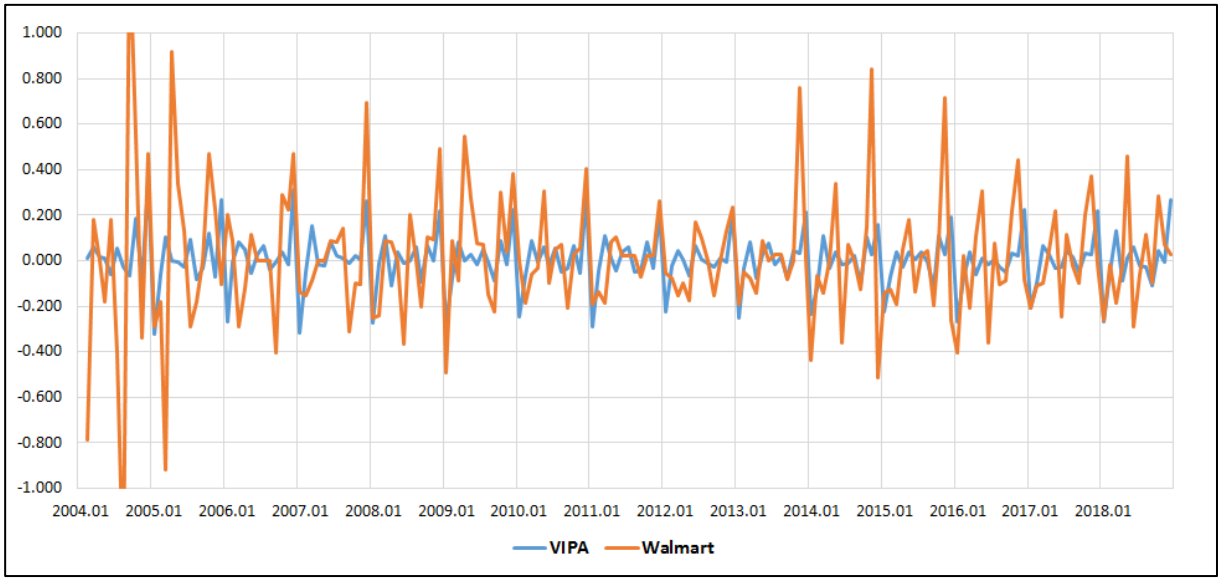


Gráfico L: evolución de las búsquedas de “Walmart”



Anexo II

Se constata que las series expresadas en variaciones logarítmicas son I(0), pudiendo asumir que son estacionarias.

Cuadro A: Resultados del test de Augmented Dickey-Fuller de las variables explicativas expresadas en tasas de cambio mensual logarítmicas. Datos mensuales: 2004.01 a 2018.12.

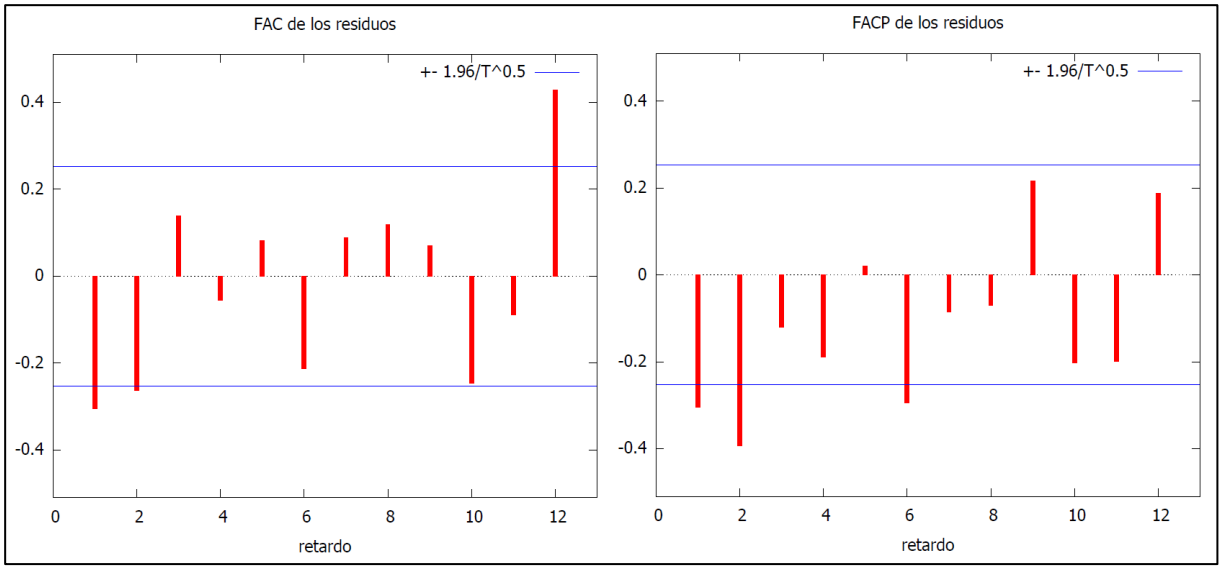
Contraste ADF TCML	Valor P
Carrefour	0.058
Coto	0.000
Easy	0.005
El Entrerriano	0.000
Falabella	0.003
Fravega	0.000
Garbarino	0.001
MercadoLibre	0.013
Microcomponente	0.000
Musimundo	0.011
Nuevo	0.000
Walmart	0.018

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

Anexo III

Se evidencia existencia de autocorrelación respecto a los residuos en uno y dos rezagos, como así también en doce.

Gráfico M: Correlograma simple y parcial de los residuos del modelo con tres variables explicativas. Datos mensuales: 2014.01 a 2018.12.



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