BANCO CENTRAL DE RESERVA DEL PERÚ

Google Trends: Predicción del nivel de empleo agregado en Perú usando datos en tiempo real, 2005-2011

Jillie Chang* y Andrea Del Río*

* Universidad del Pacífico.

DT. N° 2013-015
Serie de Documentos de Trabajo
Working Paper series
Diciembre 2013

Google Trends: predicción del nivel de empleo agregado en Perú usando datos en tiempo real, 2005-2011*

JILLIE CHANG KCOMT Y ANDREA DEL RÍO LAZO**

Resumen

En este documento se analiza si la información proporcionada por *Google Trends* puede reflejar el comportamiento de variables macroeconómicas de Perú, como el Índice de Empleo de Lima para Empresas de 100 y más Trabajadores (IE100). Utilizando esta fuente de información se construyó un índice que representa a la población que busca trabajo, el cual fue denominado Índice de Google de Desempleo (IGD). Los resultados indican que el modelo que incluye este índice mejora la predicción del IE100. Asimismo, se encuentra que este permite realizar predicciones contemporáneas y un periodo hacia adelante; empero, no permite anticipar la senda futura para más de un periodo. La importancia de estos hallazgos radica en que *Google Trends* es una fuente de información con una frecuencia más alta (semanal) que está disponible cuatro meses antes que las fuentes oficiales. En tal sentido, resulta una herramienta útil para toma de decisiones de los hacedores de política en particular en épocas de crisis, en donde el seguimiento y predicción de las variables de la actividad económica en tiempo real es fundamental.

Clasificación JEL : C22, E24, C82, C53
Palabras Clave : Internet, motor de búsquedas, Google, Perú, empleo, análisis de series de tiempo, predicciones

* Este documento es una versión del trabajo que se presentó en el curso de investigación económica de la Universidad del Pacífico, misma que fue expuesta en el XXIX Encuentro de Economistas del BCRP en el año 2011. Las autoras agradecen a Marco Vega y Erick Lahura por sus valiosos aportes. Del mismo modo, agradecen los comentarios y sugerencias de Bruno Seminario.

** Chang: Universidad del Pacífico (e-mail: changkjv@alum.up.edu.pe). Del Río: Universidad del Pacífico (e-mail: delriolak@alum.up.edu.pe).

Google Trends: forecasting aggregate employment in Peru using real time data, 2005-2011*

JILLIE CHANG KCOMT Y ANDREA DEL RÍO LAZO**

Abstract

This paper explores whether Google Trends, Internet browsing statistics, can capture the behavior of macroeconomic variables in Peru. In particular, we analyze the employment index in companies of 100 or more employees in Lima (EI100). Using Google Trends data, we built an index that captures the behavior of the population seeking employment, which was dubbed Google unemployment index (GUI). The results show that the model that includes the GUI is better at forecasting the IE100 than the model that does not include the GUI. Furthermore, it was found that the GUI is useful for nowcasting and one-period forecasts. However, we found that the GUI is not useful for forecasting EI100 further ahead in the future. The relevance of these findings reside in the fact that Google Trends data is available on a weekly basis, around four months before data from official sources. Google Trends is a powerful tool for policy makers, particularly during times of crisis where following the economic activity in real time is crucial.

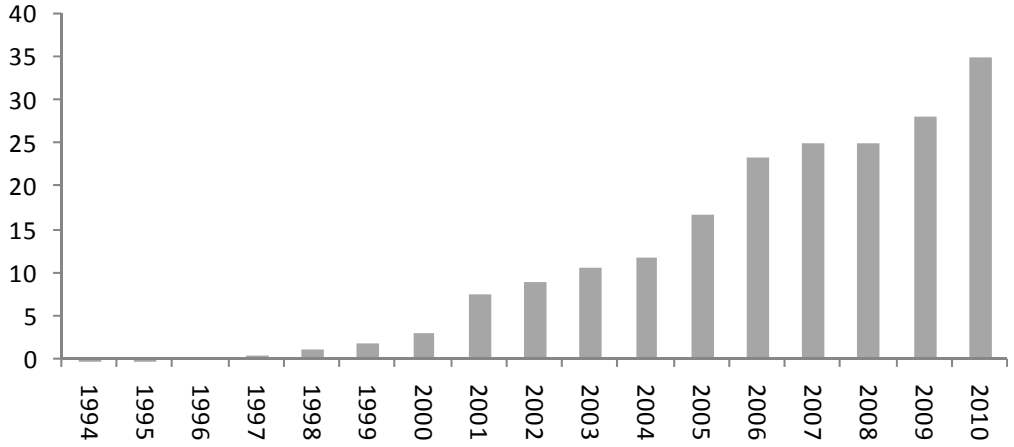
Clasificación JEL : C22, E24, C82, C53
Palabras Clave : Google, Internet, search engine, Peru, employment, time-series analysis, forecasting, nowcasting

* This document is a version of a paper that was presented for the Economic Research class at Universidad del Pacífico. This version was also part of the XXIX BCRP Meeting of Economists presentations in 2011. The authors express their thanks to Marco Vega and Erick Lahura for their comments and feedback. They also thank professor Bruno Seminario for his suggestions and input on this paper.
** Chang: Universidad del Pacífico (e-mail: changkjv@alum.up.edu.pe). Del Río: Universidad del Pacífico (e-mail: delriolak@alum.up.edu.pe).

1 INTRODUCCIÓN

En la última década, la penetración del internet en Perú, es decir, el porcentaje de usuarios con respecto a la población, ha aumentado notablemente. Como se muestra en el Gráfico N°1, esta se ha elevado aproximadamente de 3 usuarios por cada 100 habitantes en el año 2000 a 36 usuarios por cada 100 habitantes en el tercer trimestre del año 2010. Si se considera solamente Lima Metropolitana, la cantidad de usuarios es aún más elevada: en el segundo trimestre de 2011 fue de aproximadamente 52%[1]. El incremento de la penetración se puede relacionar con dos hechos fundamentales: el acceso a internet y un proceso de digitalización.

GRÁFICO 1.
Usuarios de Internet (por cada 100 personas) en Perú, 2000-2010***



Notas: **Los usuarios de Internet son personas con acceso a la red mundial.** El dato del año 2010 fue extraído del INEI, que corresponde a la población de 6 a más años que accede a internet. Fuente: Banco Mundia e INEI. Elaboración propia*

Por un lado, la rápida evolución de la tecnología ha generado que cada vez el acceso a internet tenga un menor costo y, por lo tanto, que un mayor número de personas pueda acceder a ella. Por ejemplo, en el caso peruano, el 62.7% de los usuarios alcanzó a entrar a internet a través de las cabinas públicas en el año 2010[2], en donde una hora de Internet cuesta aproximadamente un Nuevo Sol.

[1] INEI(2011). Las Tecnologías de Información y Comunicación en los Hogares, pp.17. Disponible en: http://www.inei.gob.pe/web/Biblioinei/boletinflotante.asp?file=13036.pdf, actualizado al 7 de octubre de 2011.
[2] *Ibíd.*, pp.20.

Por otro lado, la población experimenta un proceso de digitalización, mediante el cual esta se familiariza cada vez más con la tecnología. Según Tapscott (2009), en este proceso de digitalización están involucradas dos generaciones: la "generación Y" que corresponde a los nacidos entre los años 1977 y 1997 y la "Generación Z" que corresponde a los nacidos entre los años 1998 en adelante.[3] El primer grupo es conocido como la primera generación digital por su uso de tecnologías como el VHS, CD y DVD. Mientras que el segundo es conocido como la segunda generación de nativos digitales ya que sus miembros estuvieron familiarizados con grandes avances tecnológicos desde temprana edad. Entre estos avances se encuentran: los teléfonos celulares, el internet, las computadoras, entre otros.

Esta familiarización de ambas generaciones con dichos avances junto con un acceso más fácil a internet ha generado que cada vez más personas realicen un gran número de actividades virtuales, entre las cuales una de las más comunes resulta la búsqueda de información. Fosk (2011) estima que el 94% de los usuarios peruanos mayores de 15 años utiliza el internet para la búsqueda de información.[4]

En ese contexto, Google, el motor de búsqueda por internet más grande a nivel mundial[5], ha tenido un papel clave en este proceso de digitalización, ya que ha canalizado los datos disponibles en las distintas páginas web y las ha consolidado para facilitar el acceso a los usuarios y para que el servicio de búsqueda sea óptimo. Un servicio gratuito de Google que cuantifica dichas búsquedas es *Google Trends*, el cual reporta un índice semanal del volumen de búsquedas de un término o frase para un lugar determinado a lo largo del tiempo. Este servicio clasifica los datos según categorías para algunos países, como por ejemplo, electrodomésticos, alimentos y bebidas, recursos humanos, entre otros. A través de este índice, que muestra la intensidad de búsqueda para un término específico, se pueden cuantificar los intereses y/o preocupaciones de la población con acceso a internet en tiempo real. Ahora bien, ¿se puede utilizar esta información, que refleja los intereses de la población, en economía?

El presente trabajo busca analizar si la información en tiempo real de *Google Trends* durante el periodo 2005-2011 mejora la predicción del índice de empleo de Lima para empresas de 100 y más trabajadores (en adelante, IE100). La hipótesis que se maneja en este trabajo es que efectivamente dicha información es útil para mejorar las predicciones de la variable económica mencionada previamente. Ante esta hipótesis surge la pregunta: ¿existen indicios de que esta puede ser verdadera?

[3] Tapscott (2009). Grown up digital: How the net generation is changing your world, pp.16.
[4] Fosk (2011). Estado de Internet con un enfoque en el Perú, ComScore, pp.12.
Disponible en: http://www.iabperu.com/descargas/Desc_2011920162111.pdf, actualizado al 7 de octubre.
[5] A Febrero de 2010 recibía 34,000 búsquedas por segundo (lo que equivale a unas 88 billones de búsquedas mensuales) mientras que los servicios de búsqueda que le siguen en popularidad Yahoo y Bing tenían 8.4 y 4 billones de búsqueda, respectivamente.

Una señal de que la hipótesis puede ser cierta reside en el proceso de digitalización que se está observando en las actividades realizadas por la población. Una buena proporción de la PEA corresponde a la "generación Y" (en el 2009, el 45% de la PEA tenía entre 14 y 34 años)[6]. Y, como se mencionó anteriormente, esta hace un amplio uso del internet para muchas actividades de su vida diaria, entre ellas, la búsqueda de trabajo (en una publicación del 2011 del Diario Gestión, la directora general del portal de empleos bumeran.com[7] sostuvo que se han registrado de 700 mil a 800 mil visitas mensuales en todos los portales de trabajo en el Perú, de las cuales el 75% de usuarios logra ubicar empleo)[8]. Además, se observa cómo las empresas responden a los cambios que están ocurriendo en la oferta de trabajo en relación al proceso de digitalización. Es el caso del Diario El Comercio, el cual ha llevado a sus populares clasificados de empleo al internet bajo el nombre de Aptitus[9].

De esta forma, tanto por el lado de la demanda como por el de la oferta, se puede apreciar que ya está ocurriendo cierta digitalización de la búsqueda de trabajo en el mercado laboral peruano. En tal sentido, esta evidencia empírica puede ser un indicio de que la información proporcionada por *Google Trends* podría ser un indicador del mercado laboral y ayudar a mejorar las predicciones del IE100.

Para comprobar esta hipótesis, el presente trabajo de investigación se estructurará de la siguiente manera. En primer lugar, se hará una breve revisión de la literatura económica. En segundo lugar, se analizarán los datos. En tercer lugar, se describirá la metodología y, por último, se presentarán las conclusiones.

2 REVISIÓN DE LA LITERATURA

En la literatura económica se ha demostrado la efectividad del uso de los datos de *Google Trend*s para mejorar predicciones en variables microeconómicas y macroeconómicas.

Por un lado, con respecto a las variables microeconómicas, Hal R. Varian, economista principal de Google Inc., y Hyunyoung Choi realizaron uno de los primeros trabajos econométricos incluyendo al índice de *Google Trends* como un regresor y analizaron las siguientes variables: ventas minoristas, ventas de automóviles, casas y viajes. Varian y Choi (2009a) partieron de la hipótesis de que el volumen de búsqueda de las categorías que

[6] Dato calculado a partir de la Encuesta Nacional de Hogares (ENAHO) del año 2009. En el 2009 La PEA estuvo conformada por 16,202,991 personas y la población entre 14 y 34 años que correspondió a la PEA estuvo integrada por 7,250,907 personas.
[7] Portal web disponible en http://bumeran.com.pe/, actualizado al 9 de mayo de 2011.
[8] Diario Gestión, "Un millón de peruanos buscarán trabajo en portales web este año", 14 de abril de 2011, pp.7.
[9] Disponible en: http://aptitus.pe/, actualizado al 30 de abril de 2004.

corresponden a estas variables podría estar correlacionado con el nivel de actividad económica en esas industrias. Así encuentran que, al incluir los datos que brinda Google sobre el volumen de búsqueda de estas categorías, la predicción de ventas mensuales de cada industria es mejor comparada con modelos que no la incluyen. Utilizan la metodología ARMA, específicamente modelos autorregresivos estacionales.

Luego, en una investigación realizada por el Banco Central de Chile, Carrière-Swallow y Labbé (2010) construyen un índice de búsquedas relacionadas con autos para incluirlo en modelos de *nowcasting*[10] de ventas de autos en Chile. De esta forma encuentran que la inclusión de estos datos mejora el ajuste, la eficiencia y la predicción dentro y fuera de muestra de modelos autorregresivos.

Posteriormente, Song *et ál.* (2010) emplean los datos de Google para predecir la demanda por cuartos de hoteles en Charlotte, Estados Unidos. Utilizan cuatros modelos econométricos: modelo de rezagos distribuidos (ARDL), modelo autorregresivo integrado de media móvil (ARIMA), modelo de suavización espacial (ES) y modelo de parámetro variante en el tiempo (TVP) para evaluar el cambio en la capacidad predictiva al incluir información de búsquedas en Google. En los cuatro casos encuentran que este tipo de datos permite mejorar la predicción de la demanda de los turistas a un bajo costo.

Por otro lado, con respecto a las variables macroeconómicas, Della Penna y Huang (2009) y Schmidt y Vosen (2009) demuestran que es posible construir un índice de confianza del consumidor empleando datos de *Google Trends* para predecir el consumo privado en Estados Unidos. Ambos trabajos encuentran que este es más preciso que los índices existentes hechos sobre la base de dos encuestas: *The University of Michigan Consumer Sentiment Index* (ICS) y *The Conference Board Consumer Confidence Index* (CCI). En particular, Della Penna y Huang construyen un índice de confianza del consumidor sobre la base de búsquedas (SBI, por sus siglas en inglés). Encuentran que el SBI anticipa (predice) cambios en el ICS y en el CCI. Mediante dos metodologías (modelos autorregresivos y modelos de corrección de errores) demuestran que el SBI supera tanto al ICS como al CCI en predecir el crecimiento en el gasto de consumo personal y en ventas minoristas. Schmidt y Vosen, por su parte, utilizan un modelo autorregresivo simple con el consumo como variable dependiente. Comparan cómo varía el poder predictivo de este modelo simple al agregar cada indicador de confianza del consumidor. Son tres los modelos comparados, el primero con ICS, el segundo con el CCI y el tercero con el índice Google. Luego, realizan predicciones dentro y fuera de la muestra y encuentran que la capacidad predictiva aumenta más en el modelo con el índice de Google en relación a los otros dos modelos.

[10] *Nowcasting* es la predicción de corto plazo o predicción contemporánea.

Otras variables macroeconómicas son el PBI y la demanda por trabajo. Suhoy (2009) evalúa si una serie de categorías de *Google Trends* (contratación de personal, electrodomésticos, viajes, bienes raíces, alimentos y bebidas y belleza y cuidado personal) podrían haber predicho la caída en el PBI de Israel que ocurrió en 2008 en tiempo real. Encuentra con probabilidades bayesianas dentro de la muestra que estas series sí hubieran podido predecir el declive económico que experimentó Israel en una fecha muy cercana a la predicha por los datos oficiales. Es importante mencionar que este último también encontró que el índice de *Google Trends* para la categoría "contratación de personal" en Israel sirve como un leading indicator de la demanda por trabajo en ese país.

Por último, también se analiza el desempleo. Bersier (2010) y D'Amuri y Marcucci (2009) muestran que al incluir el índice de Google, las predicciones del desempleo en Estados Unidos mejoran. D'Amuri y Marcucci utilizan la metodología de modelo autorregresivo integrado de media móvil (ARIMA). Asimismo, Askitas y Zimmermann. (2009) realizaron un trabajo similar y obtuvieron resultados satisfactorios para Alemania utilizando modelos de corrección de errores con cuatro palabras clave relacionadas a la búsqueda de empleo. Sobre la base de los trabajos realizados para Alemania e Israel, Varian y Choi (2009b) encuentran que incluir al índice de *Google Trends* mejora el ajuste de modelos que pretenden predecir la ocurrencia de solicitudes de beneficios por desempleo en Estados Unidos. Por último, Oleksandr (2010) replicó el trabajo de Askitas y Zimmermann. (2009) para Ucrania; empero, encontró que la información proporcionada por Google no mejoraba la predicción del desempleo.

3 DATOS

Los datos empleados en este trabajo provienen de dos fuentes de información. Em primer lugar, el Índice de empleo para empresas de 100 y más trabajadores se obtuvo de fuentes oficiales, como el Instituto Nacional de Estadística e Informática (INEI). Y en segundo lugar, el Índice de Google de desempleo se obtuvo a partir de los datos de *Google Trends*.

3.1 ÍNDICE DE EMPLEO PARA EMPRESAS DE 100 Y MÁS TRABAJADORES

A diferencia de los trabajos señalados en la literatura económica que analizan variables del mercado laboral, no se seleccionó al desempleo como variable dependiente pues dadas las características del mercado laboral peruano y la metodología de medición, se cree que este no reflejaría la situación del mercado laboral en el Perú de manera precisa.

En primer lugar, esta variable resulta poco significativa en los países como Perú, que no cuentan con seguro de desempleo y presentan altas tasas de informalidad y de trabajo independiente (ver Cuadro N°1). Según Chacaltana (2001)[11] y Bescond *et ál* (2003)[12] , las personas no pueden permitirse estar mucho tiempo desempleadas sin tener seguro de desempleo o alguna ayuda equivalente. Por lo tanto, estas se ven obligadas a realizar trabajos por cuenta propia (trabajo independiente) o trabajos en la economía informal por la necesidad de obtener algún ingreso. En consecuencia, como la mayoría de su población activa trabaja en la economía informal, se reportan tasas de desempleo bajas.

En Bescond *et ál* (2003), se señala que el concepto clásico de desempleo es de poca utilidad en los mercados laborales dominados por el trabajo independiente. En dicho trabajo se afirma que el riesgo de desempleo es mayor en el trabajo asalariado pues este requiere un contrato entre el trabajador y el empleador que cualquiera de las dos partes puede rescindir. En contraste, un trabajador independiente que atraviesa por problemas económicos simplemente gana menos y, por lo general, no se registra como desempleado. En tal sentido, se sostiene que es normal que la mayoría de los desempleados sean asalariados que han perdido su colocación y, además, que la mayoría de los desempleados sean personas que buscan empleo asalariado.[13]

Cuadro N° 1
Indicadores del mercado laboral en Perú

indicador	Perú
Sistema de seguro de desempleo	No
Indemnizaciones por despidos	Sí
Cuentas de ahorro individual por desempleo	Sí
Sector informal con respecto al empleo urbano	54.9% (2005)
Trabajadores independientes en relación a la población ocupada	35.2% (2009)

Fuente: Velásquez, Mario (2010). "Seguros de desempleo y reformas recientes en América Latina", Macroeconomía del desarrollo, serie 99, División de Desarrollo Económico, Cepal, Chile, pp. 15 -23. OIT (2010). Panorama Laboral 2010 América Latina y el Caribe, Oficina Regional para América Latina y el Caribe, OIT, Lima, pp. 126-134. ILO(2006). Labour Overview Latin America and the Caribbean, Regional Office for Latin America and the Caribbean, ILO, Lima, pp.61-64.
Elaboración propia

[11] Chacaltana (2001). ¿Qué sabemos sobre el desempleo en el Perú?, INEI, CIDE y Programa MECOVI-PERU, pp.13-14. Disponible en: http://www.inei.gob.pe/biblioineipub/bancopub/Est/Lib0489/Libro.pdf, actualizado al 11 de diciembre de 2011.
[12] Bescond *et ál* (2003). Siete indicadores para medir el trabajo decente, Revista Internacional del Trabajo, 122 (2), p. 210. Disponible en: http://www.ila.org.pe/proyectos/observatorio/material/trabajo_decente_3_1.pdf, actualizado al 11 de diciembre de 2011.
[13] *Ibídem.*

En segundo lugar, se considera que existen problemas metodológicos en la medición del desempleo. Chacaltana (2001)[14] sostiene que el desempleo abierto estaría subestimado dado que para clasificar a una persona como desocupado se requiere una búsqueda activa de empleo. Además, Garavito (2000) sostiene que la poca variación de la tasa de desempleo abierto en Perú y su bajo nivel estaría explicada por sesgos en la elaboración de la ENAHO debido a la manera de medir el indicador y al problema aún no resuelto del desempleo oculto.[15]

Por todo lo expuesto previamente, se trató de elegir un indicador de empleo que trate de resolver los problemas señalados y que refleje la verdadera evolución del empleo en el mercado laboral. En tal sentido, se eligió el índice de empleo de Lima para empresas de 100 y más trabajadores (IE100), que mide el nivel de empleo a partir de los trabajadores sujetos al régimen laboral del sector privado para empresas de 100 y más trabajadores. La variación de dicho índice reflejaría la creación o "destrucción" de nuevos empleos formales en la economía.

Este índice se calcula a partir de la Encuesta Nacional de Variación Mensual de Empleo en Empresas de 10 y más Trabajadores (ENVME), la cua es elaborada por el Ministerio de Trabajo y Promoción del Empleo (MTPE) a nivel de establecimientos. Esta se aplica a empresas y establecimientos con trabajadores sujetos al régimen laboral del sector privado, pudiendo ser las empresas públicas o privadas. Asimismo, se realiza sobre los trabajadores asalariados, personal con contrato de servicio (con más de 25 horas semanales de trabajo), aprendices, trabajadores contratados por empresas de servicios y cooperativas de trabajo.

Sobre la base de esta encuesta, se calcula la información del empleo por área geográfica, sector económico (industria manufacturera, comercio y servicios) y tamaño de empresa. En relación al tamaño, estas pueden ser de 10-49 trabajadores (muestra), de 50-99 trabajadores (muestra) y de 100 y más trabajadores (censo). La información se muestra como índice con base octubre 1997 igual a 100.

En particular, el IE100 permite observar el comportamiento mensual del empleo asalariado en las empresas de 100 y más trabajadores en el conjunto de sectores económicos de Lima y se calcula según las siguientes fórmulas:

$$IE100_t = IE100_{t-1} * \left[1 + \frac{VE_t}{100}\right] \quad (1)$$

[14] Chacaltana, *op. cit.*, p. 9.

[15] En el Perú se define el Desempleo Abierto como una condición que presentan las personas de 14 años y más, que durante la semana de referencia (semana previa a la Encuesta), no tienen trabajo y lo buscan activamente, que estaban disponibles para trabajar de inmediato, y habían tomado medidas concretas para buscar un empleo asalariado o un empleo independiente. Esta definición se rige a la propuesta por la Organización Internacional de Trabajo (OIT, 1983).

$$VE_t = \left[\frac{L_t}{L_{t-1}} - 1 \right] * 100 \tag{2}$$

Donde, VE_t es la variación porcentual mensual del empleo para empresa de Lima de 100 y más trabajadores en el mes t y L_t es el total de trabajadores en dichas empresas en el periodo t. Los datos de este índice son calculados por el MINTRA y recopilados por el BCRP. La frecuencia de esta variable es mensual y es publicada con un retraso aproximado de 4 meses.

3.2 Índice de Google de desempleo (IGD)

El Índice de Google de desempleo se obtuvo a partir de las series de *Google Trends*, las cuales están disponibles desde el año 2004 y se pueden encontrar en: http://www.Google.com/insights/search. Este es un servicio gratuito que cuantifica las búsquedas que se realizan a través de Google para un término específico o frase, controlando por ciertos parámetros como: ubicación geográfica, periodo de tiempo y categorías (electrodomésticos, alimentos y bebidas, recursos humanos, entre otros)[16] . Sobre la base de dichos parámetros, reporta un índice con frecuencia semanal, que refleja la intensidad de búsqueda de una palabra o frase clave que se realiza a través del buscador Google.

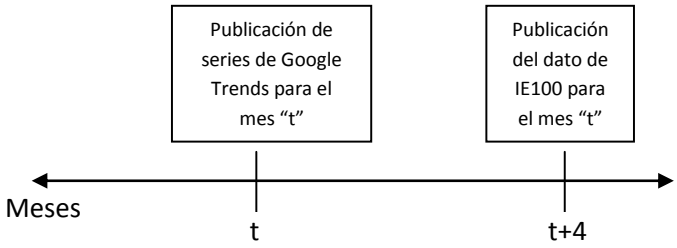
¿Cómo se construye este índice de Google? A partir del volumen de búsquedas de una palabra clave, Google lo construye sobre la base de dos procesos. El primero consiste en la normalización, en donde se divide el número de búsquedas de una palabra o frase clave entre el número total de búsquedas controlando a través de los parámetros mencionados: región geográfica, fecha y categoría. Se debe recalcar que este último parámetro no está disponible para el Perú. Como señalan Carrière-Swallow y Labbé (2010), el objetivo de este proceso es aislar los datos de los siguientes efectos: a) crecimiento total del número total de usuarios de internet y b) aumento de la popularidad de Google como buscador.

El segundo proceso corresponde al escalamiento, del cual resulta un índice de 0 a 100, en donde 100 es asignado al número de búsquedas más alto para el periodo analizado. Por ejemplo, si se toman las estadísticas de búsqueda de la palabra clave "Humala" en Perú en el periodo 2005-2011, el día en donde se realizó el mayor número de búsquedas aparecerá en la serie con el número 100 y los demás valores se asignarán en función a este último. Este proceso se realiza con el fin de que todas las series de *Google Trends* se encuentren en la misma escala y por un tema de confidencialidad del valor absoluto del volumen de búsqueda de cada palabra clave.

[16] La opción de categorías no está disponible para todos los países, entre ellos el Perú.

La información que brinda *Google Trends* tiene un valor agregado interesante pues proporciona información de los intereses de la población en tiempo real, es una fuente de información con una frecuencia más alta (semanal) en comparación con las fuentes oficiales (mensual) y es gratis. Además, una ventaja importante de *Google Trends* como fuente de información es que está disponible antes que la publicación de la información oficial. El primer lunes de cada semana de cada mes es posible descargar las series de *Google Trends* hasta el mes anterior. Por ejemplo, el 1 de octubre es posible descargar las series para todo el mes de setiembre; es decir, el retraso de la publicación es de un día. En contraste, el rezago de las fuentes oficiales es de aproximadamente 4 meses (Ver Gráfico N°2).

GRÁFICO 2. *Rezagos en la publicación de información*



Obtener la información de Google antes que la información publicada por las fuentes oficiales es relevante si se cumple que la información de Google recoge en tiempo real lo que está ocurriendo con las variables macroeconómicas, cuyas cifras estadísticas oficiales serán publicadas con varios meses de retraso. En este caso, sería una herramienta útil para el seguimiento de la actividad económica, sobre todo en épocas de crisis.

Sobre la base de los datos de *Google Trends*, se construyó el Índice de desempleo (IGD), que refleja a las personas que buscan trabajo; es decir, sería una *proxy* de la variable desempleo abierto. Se espera que este índice tenga una correlación inversa con el IE100. La construcción de este índice se realizó según el siguiente procedimiento:

a. Selección de palabras claves. Se realizó una lista con las palabras o frases que intuitivamente tenían una relación con la búsqueda de trabajo. Sobre la base de esta, se eligieron las que generaban la mayor correlación entre el índice conformado a partir de las mismas con IE100. De esta manera, algunas de las palabras seleccionadas fueron: "busco trabajo", "aptitus" y "bolsa trabajo".

b. ***Mensualización de cada serie semanal.*** Debido a que la frecuencia de la variable endógena (IE100) es mensual y los datos de *Google Trends* tienen frecuencia semanal, se convertirán las series de frecuencia más alta a frecuencia mensual a través de promedios simples.

c. ***Indexación de las series ponderando por la inversa de la desviación estándar.*** Para obtener las series indexadas, primero se calculan el promedio (ecuación 3) y la desviación estándar (ecuación 4) para cada palabra clave "n". Luego, se obtienen los ponderadores a partir de la división entre la inversa de la desviación estándar de una palabra clave y la suma de las inversas de las desviaciones estándar de las "n" palabras clave (ecuación 5). Finalmente, los índices se obtienen a partir de la suma producto de las observaciones de las palabras clave en el mes t y sus ponderadores (ecuación 6). En los Anexos, se muestra un ejemplo de la construcción de IGD a partir de dos palabas clave.

$$\bar{x}_n = \sum_{t=1}^{T} x_{tn} \tag{3}$$

$$\sigma_n = \sqrt{\frac{\sum_{t=1}^{T}(x_{tn} - \bar{x}_n)^2}{T-1}} \quad \forall \quad t = 1 \; a \; T \tag{4}$$

$$\alpha_n = \frac{\frac{1}{\sigma_n}}{\sum_{n=1}^{N} \frac{1}{\sigma_n}} \tag{5}$$

$$IGD_t = \sum_{n=1}^{N} x_{tn}\alpha_n \tag{6}$$

Se debe mencionar que se construirá dicho índice para la muestra comprendida entre enero de 2005 y agosto de 2011 por dos motivos. En primer lugar, está la disponibilidad de datos: el último dato publicado de IE100 en el momento de realizar el trabajo fue el correspondiente a agosto de 2011. En segundo lugar, debido a que la información de *Google Trends* está disponible solo a partir de enero de 2004, se consideró prudente tomar un margen de un año para empezar a utilizar las series de *Google Trends* para evitar cualquier deficiencia que podría haber existido en las series del primer año de *Google Trends* por ser un servicio nuevo.

3.3 IGD E IE100

Como se puede apreciar en los gráficos 3 y 4, la inspección visual sugiere que el comportamiento entre las dos variables es el esperado. Cuando la variación del IE100 cambia

bruscamente, la variación del índice de desempleo Google también cambia en el sentido contrario. Del mismo modo, el gráfico de dispersión confirma que ambas están negativamente correlacionadas.

GRÁFICO 3.
IE100 e IGD 2005-2011

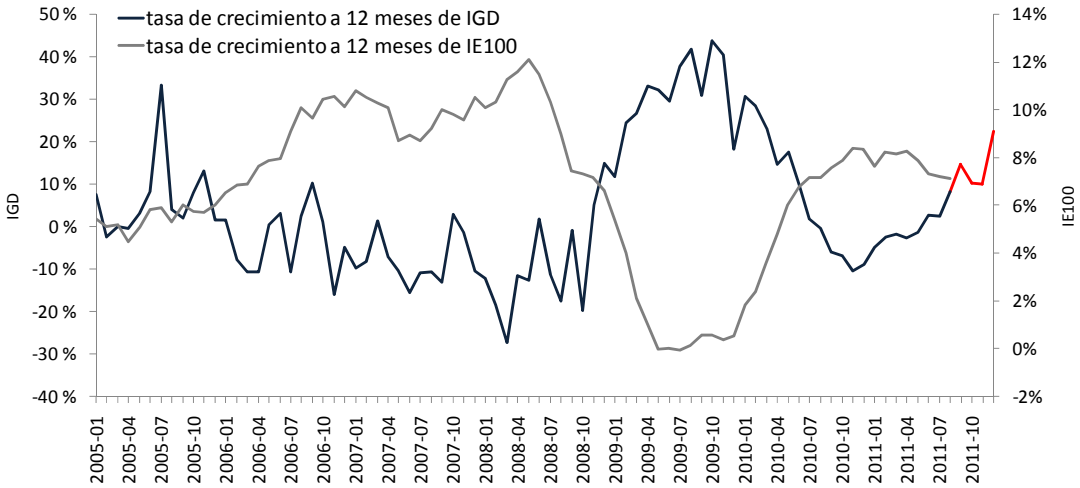
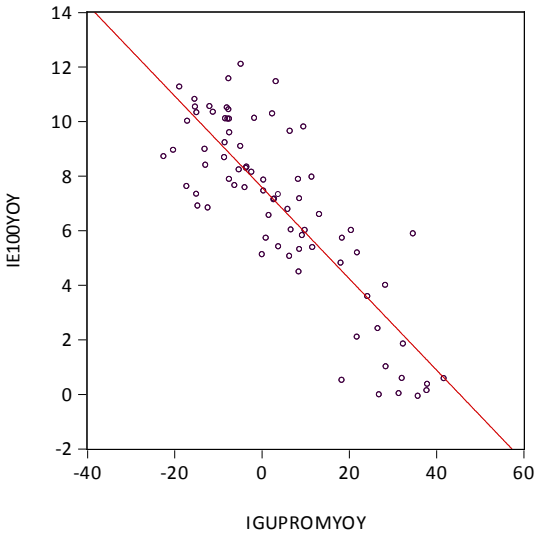


GRÁFICO 4.
Gráfico de dispersión de IE100 e IGD en Perú, 2005-2011 (tasa a 12 meses)



La evidencia de una relación entre ambas variables es un hallazgo importante. El lapso que transcurre entre un periodo de tiempo y la fecha en que las entidades encargadas hacen

la publicación de los datos macroeconómicos para ese periodo representa un limitante para las autoridades de política. Aunque el índice de desempleo de Google y el IE100 tienen una relación negativa contemporánea, ambas son publicadas con distintos rezagos. Mientras que el IE100 es publicado con un retraso aproximado de 4 meses, el índice de Google es publicado semanalmente. Por ello, la información en tiempo real proporcionada por *Google Trends* podría resultar una herramienta útil para realizar un seguimiento de los cambios que ocurren en el mercado laboral casi de manera inmediata. Como se menciona en Carrière-Swallow y Labbé (2010), este retraso tiende a ser más prolongado para los países en vías de desarrollo, por lo cual la utilización de esta herramienta gratuita resulta aún más conveniente.

De esta manera, al demostrar que las palabras que las personas buscan en Google tienen una correlación con las variables económicas, esto significaría que *Google Trends* sería una valiosa fuente de información en tiempo real disponible para todos los agentes económicos: consumidores, trabajadores, empresas y Estado. La información de Google sería publicada aproximadamente 4 meses antes en relación con las fuentes oficiales. Así, a partir del IGD, se podría observar cómo va a ser el comportamiento de IE100 luego de tres meses.

En particular, como menciona la OIT (2011) es importante analizar la variación de IE100 pues además de ser un indicador global del funcionamiento de la economía de un país, a partir de esta se pueden observar los cambios en el mercado de trabajo, que pueden estar relacionados con otros fenómenos económicos y sociales. De esta forma, este puede resultar importante para la planificación y formulación de políticas macroeconómicas y desarrollo de los recursos humanos.

4 METODOLOGÍA

Para seleccionar la metodología a seguir, el primer paso fue analizar la presencia de raíz unitaria en las series. Las pruebas implementadas sugieren que las series IE100 e IGD son procesos raíz unitaria o procesos integrados de orden 1. Sin embargo, existe incertidumbre sobre los resultados y, por lo tanto, la metodología a seguir.

La presencia de raíz unitaria sugiere evaluar la presencia de cointegración y, de ser el caso, evaluar la serie IGD sobre la base del modelo de corrección de errores. Dado que el periodo de la muestra es muy corto (casi ocho años) parece aceptable interpretar sólo estadísticamente el vector de cointegración más no necesariamente en términos económicos. Sin embargo, es posible que las pruebas de raíz unitaria tengan poco poder debido al tamaño de la muestra y por lo tanto las series sean estacionarias; en este sentido, se debería aplicar modelos ARMA y ARDL bajo el supuesto de estacionariedad de las series. Este supuesto también es válido dado

que el objetivo principal del trabajo es evaluar la predicción contemporánea y de corto plazo, por lo tanto, la estacionariedad o no estacionariedad de las series podría no ser tan relevante.

Debido a esta incertidumbre, optamos por una perspectiva ecléctica y se optó por dos enfoques. El primero es aplicar modelos ARMA y ARDL para analizar si la incorporación de los datos de Google mejora la capacidad predictiva del modelo. Para ello se aplicarán indicadores como el error cuadrático medio (RECM) y el error absoluto medio (EAM). Asimismo, se empleará la prueba Diebold y Mariano (1995) modificado por Clark y West (2007) para dar más robustez a los resultados. El segundo enfoque consiste en ver si las series cointegran y plantear un modelo de corrección de errores con el objetivo de estudiar si Google es un predictor insesgado e indicador líder. Asimismo, se evaluará si este indicador es un buen predictor para más de un periodo adelante.

4.1 ARMA y ARDL

Respecto a la metodología ARMA, el modelo planteado es de la forma:

$$y_t = \beta_0 + \sum_{i=1}^{p} \beta_i y_{t-i} + \epsilon_t + \sum_{i=1}^{q} \alpha_i \epsilon_{t-i} \tag{7}$$

donde y es la variable endógena para el periodo t ("índice de variación de empleo en Lima para empresas de 100 y más trabajadores") y es el error. La variable y_t está expresada en variaciones a 12 meses y se asume que es estacionaria. Adicionalmente, se elegirá el mejor modelo según los siguientes criterios: significancia económica y estadística de las variables, ajuste y parsimonia. Además, se verificará que los errores tengan un comportamiento de ruido blanco (no autocorrelación, no heteroscedasticidad y normalidad).

Después de la elección del mejor modelo ARMA, se replicará el mismo modelo con la inclusión de un regresor adicional. Este modelo de rezagos distribuidos (ARDL) será denominado modelo aumentado e incluirá a IGD.

$$y_t = \beta_0 + \sum_{i=1}^{p} \beta_i y_{t-i} + \epsilon_t + \sum_{i=1}^{q} \alpha_i \epsilon_{t-i} + \sum_{i=1}^{T} \gamma_i IGD_{t-i} \tag{8}$$

La muestra que se utilizará abarca el periodo enero de 2005 – agosto de 2011. Esta será dividida en dos submuestras para evaluar la capacidad predictiva de los modelos candidatos: enero de 2004 – diciembre de 2009 para la estimación y enero de 2010 – agosto de 2011 para la predicción. Para todos los modelos, con la muestra de predicción, se calculan predicciones

"un período adelante" recursivas[17] y se guardarán los errores mensuales de predicción en dos vectores.

Con dichos vectores de errores de predicción, se calcularán para cada caso la raíz del error cuadrático medio (RECM) y el error absoluto medio (EAM). A partir de estos errores, se realizará un análisis comparativo entre los estadísticos para ambos vectores. Si el RECM y el EAM son menores para el vector de errores de predicción del modelo aumentado, entonces se podrá afirmar que la hipótesis es verdadera. Adicionalmente, sobre la base de estos errores de predicción de cada modelo, se realizará una comparación de la precisión predictiva a través de la prueba modificada del test de Diebold y Mariano (1995) propuesta por Clark y West (2007), la cual permite comparar modelos anidados[18].

Antes de explicar el test de Clark y West es conveniente hacer una breve explicación de la versión original del test de Diebold y Mariano (DM). Sean e_{1i} y e_{2i} los errores de predicción de los dos modelos que se busca comparar, la precisión de cada predicción es medida por una función de pérdidas particular "g" (Algunas funciones de pérdidas comunes son: $g(e_{1i}) = (e_{1i})^2$ y $g(e_{1i}) = |e_{1i}|$. La prueba de DM está basada en el diferencial de las funciones de pérdida de cada error:

$$d_i = g(e_{1i}) - g(e_{2i}). \tag{9}$$

De esta forma, la hipótesis nula es la siguiente:

$$\text{Ho: } E(d_i) = \overline{d} = \frac{1}{H}\sum_{i=1}^{H}[g(e_{1i}) - g(e_{2i})] = 0, \tag{10}$$

donde H es el número de predicciones un periodo hacia adelante realizadas con los modelos. Si se cumple dicha hipótesis, es decir, $\overline{d} = 0$, entonces quiere decir que la precisión de ambos modelos es la misma en los pronósticos. En contraste si rechaza la hipótesis nula, el modelo uno es mejor prediciendo que el modelo dos.

Para testear la hipótesis nula, se debe calcular el estadístico DM, el cual requiere el cálculo de la varianza de $\overline{d}$[19] . Si la serie d_i es serialmente no correlacionada con la varianza muestral γ_0, la varianza estimada de $\overline{d}$ estaría dada por $\sqrt{\gamma_0/(H-1)}$. Al usar la varianza estimada de $\overline{d}$,

[17] Por ejemplo, a partir del modelo se calculan los errores de predicción para la fecha enero de 2008 y se guarda dicho error en un vector. Luego, se incorpora la fecha señalada anteriormente (enero de 2008) en el modelo y se vuelve a realizar una predicción del siguiente mes (febrero 2008) y se guarda el error de predicción en el vector señalado. Se repite este proceso hasta incorporar toda la muestra (se irá incluyendo un mes adicional a la muestra inicial hasta completar la muestra total) y, de esta forma, se guardarán todos los errores de predicción en un vector (el horizonte de predicción será un mes).

[18] Los modelos anidados son dos modelos que son idénticos excepto por el hecho que uno restringe a uno o más parámetros (el modelo nulo) y el otro no tiene esas restricciones (el modelo alternativo).

[19] Para mayor detalle ver Walter Enders (2010), *Applied Econometric Time Series*, pp. 88.

la expresión $\overline{d}/\sqrt{\gamma_0/(H-1)}$ tiene distribución t y H-1 grados de libertad. Diebold y Mariano (1995) toman a γ_i como la iésima autocovarianza de la d_t de la secuencia. Asumiendo que los primeros q valores de γ_i son diferentes de cero. La varianza de $\overline{d}$ puede ser aproximada por $var(\overline{d}) = [\gamma_0 + 2\gamma_1 + ... + 2\gamma_q](H-1)^{-1}$. La versión mejorada que proponen Harvey, Leybourne y Newbold (1998) para construir al estadístico DM es:

$$DM = \frac{\overline{d}}{\sqrt{(\gamma_0 + 2\gamma_1 + ... + 2\gamma_q)(H-1)}} \tag{11}$$

Cuando se trabaja con modelos no anidados el procedimiento adecuado es comparar el valor estimado de DM con el estadístico t con $H-1$ grados de libertad. Sin embargo,Clark y McCracken (2001) desmostraron que el test DM solo tiene una distribución t cuando los modelos que se están usando para la predicción no son anidados. Posteriormente, Clark y West (2007) desarrollaron un procedimiento simple para ajustar los errores de predicción del modelo que contiene al otro de manera que una variación simple del estadístico DM puede ser utilizada para modelos anidados.

Primero, denotando las predicciones hechas con el modelo 1 como f_{1i} y a sus errores de predicción como e_{1i}. De manera similar, las predicciones hechas con el modelos 2 son f_{2i} y los errores de predicción de ese modelo son e_{2i}. Asumiendo que es el modelo 1 el que está anidado dentro del modelo 2, la única razón para discrepancias entre f_{1i} y f_{2i} sería el error en la estimación de parámetros. Si el error de estimación es substraido de e_{2i}, los errores de predicción ajustados pueden ser usados como la base para el test DM modificado. Luego, construyendo la serie z_t con los cuadrados de estos errores como:

$$z_i = (e_{1i})^2 - \left[(e_{2i})^2 - (f_{1i} - f_{2i})^2\right] \quad , \text{ donde i=1,...H.} \tag{12}$$

Bajo la hipótesis nula que ambos modelos predicen igual de bien, z_i debería ser cero. Mientras que, bajo la hipótesis alternativa, uno predice mejor que otro. Según Enders (2009), una forma de evaluar esta hipotésis nula es regresionar a "z" contra una constante. Si el $t-$estadístico de la constante supera a 1.645 es posible rechazar la hipótesis nula de igual precisión predictiva al 5% de significancia.

4.2 Modelo de correción de errores

Si las series analizadas son no estacionarias, entonces es posible que cointegren. Para evaluar la hipótesis de cointegración, se utilizará la prueba de Johansen (1989) basadas en

los estadísticos de la traza y el valor propio máximo, asumiendo que los datos no tienen componente tendencial determinístico (como lo sugieren los gráficos)[20].

Si las series cointegran, entonces es posible representar la relación como:

$$IE100_t = \beta IGD_t + \upsilon_t \tag{13}$$

y, de acuerdo con el teorema de representación de Granger, es posible analizar la dinámica de corto plazo a través del siguiente modelo de corrección de errores para cada ecuación.

$$\Delta IE100_t = \alpha \left[IE100_{t-1} - \beta IGD_{t-1} \right] + \sum_{t=1}^{m} \phi_i \Delta IE100_{t-j} + \sum_{t=1}^{m} \theta_j \Delta IGD_{t-j} + \epsilon_t \tag{14}$$

$$\Delta IGD_t = \alpha^{'} \left[IE100_{t-1} - \beta IGD_{t-1} \right] + \sum_{t=1}^{m} \phi_i^{'} \Delta IE100_{t-j} + \sum_{t=1}^{n} \theta_j^{'} \Delta IGD_{t-j} + \epsilon_t^{'} \tag{15}$$

En este contexto, el índice de desempleo de Google, IGD, será un predictor insesgado si $\beta = 1$. Más aún, será útil como indicador líder si $\alpha \neq 0$ pues su nivel precede (causa en el sentido de Granger) a IE100. Finalmente, permitirá anticipar la senda futura de IE100 si IGD es fuertemente exógena para β, es decir, si $\alpha^{'} = 0 = \phi_1^{'} ... = \phi_m^{'}$.

5 Resultados

5.1 ARMA y ARDL

Para representar el proceso estocástico de la serie a través de sus valores pasados (componente autorregresivos) y términos estocásticos presentes y pasados (componente de medias móviles), se partió del siguiente modelo de referencia:

$$y_t \sim y_{t-1} + y_{t-2} + \epsilon_t + \epsilon_{t-12} \tag{16}$$

Donde ϵ_r es el término de error, y_t es la tasa a 12 meses de IE100. Los errores del modelo no presentan problemas de heterocedasticidad ni de autocorrelación y sí se asemejan a una distribución normal[21]. Sobre la base de estos modelos de referencia, se agregó el índice de

[20] Se debe mencionar que por un problema de unidades de medida, se colocaron a IE100 e IGD en la misma escala.
[21] Ver anexos.

Google (IGD) y, de esta forma, se obtuvo el siguiente modelo aumentado:

$$y_t \sim y_{t-1} + y_{t-2} + \epsilon_t + \epsilon_{t-12} + IGD_t \tag{17}$$

Para comprobar si la inclusión del índice de Google mejoraba la predicción de las variables, se realizaron predicciones dentro y fuera de la muestra.

5.2 PREDICCIÓN DENTRO DE LA MUESTRA

A partir de la muestra completa, se estimaron los modelos de referencia y aumentados, y se analizaron ciertos indicadores de capacidad predictiva para compararlos. Entre estos indicadores se encuentran: el R^2 ajustado, la raíz del error cuadrático medio (RECM) y el error medio absoluto (EAM).

Como se puede apreciar en el Cuadro N°2, los modelos que incluyen la información proporcionada por Google, es decir, los modelos aumentados presentan una mejor capacidad predictiva en las predicciones dinámicas. El RECM y el EAM caen 51.40% y 58.33%, respectivamente. Del mismo modo, el R^2 aumenta ligeramente[22]. Sin embargo, en el caso de las predicciones estáticas dichos indicadores casi no varian al incluir IGD (aumentan 0.03% y 0.09%, respectivamente).

5.3 PREDICCIÓN FUERA DE LA MUESTRA

Con respecto a las predicciones fuera de la muestra; para cada modelo, es decir, tanto para el modelo de referencia y para el modelo aumentado; se dividió la muestra en dos partes: la de estimación y la de predicción. Sobre la base de la muestra de estimación, comprendida entre enero de 2005 y diciembre de 2009, se realizaron pruebas recursivas mensuales para el periodo enero de 2010 y agosto de 2011 (ver Gráfico N°5).

A partir de los errores recursivos de predicción, se analizaron el RMSE, MAE y, además, se aplicó la prueba Diebold y Mariano. A continuación se muestran los resultados.

Como se aprecia en las predicciones fuera de la muestra, los indicadores de capacidad predictiva mejoran al incorporar la información proporcionada por Google. En el modelo

[22] Se debe mencionar que los valores del R^2 son bastane altos probablemente por un problema de raíz unitaria, que para fines de este trabajo no es muy relevante dado que el objetivo principal es la predicción contemporánea y de corto plazo.

Cuadro N° 2

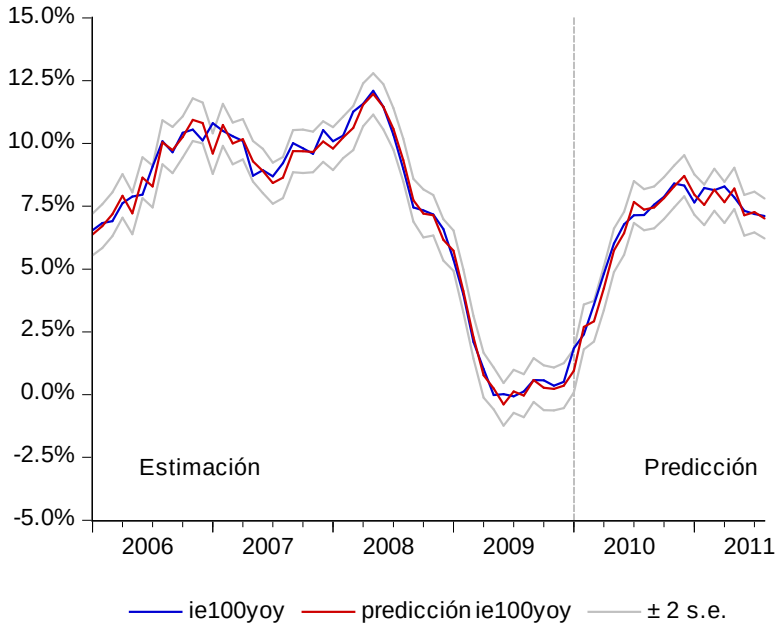
Parámetros estimados y coeficientes de capacidad predictiva para predicciones dentro de la muestra

Variable endógena: IE100 (y_t)

Variables exógenas	Modelo de referencia	Modelo aumentando
y_{t-1}	1.44	1.38
	(0.11)	(0.10)
y_{t-2}	-0.48	-0.51
	(0.11)	(0.10)
$MA(12)$	-0.92	-0.91
	(0.03)	(0.03)
IGD_t		-0.02
		(0.01)
Observaciones	68	68
$R^2_{ajustado}$	0.986	0.987
$RECM_{dinmico}$	1.07	0.52
$EAM_{dinmico}$	0.96	0.40
$RECM_{esttico}$	0.29	0.30
$EAM_{esttico}$	0.23	0.25

Notas: *Nivel de significancia al 5%. Todas las variables se encuentran en tasas a 12 meses. El modelo cuenta con intercepto.*

GRÁFICO 5.
Predicción de IE100



de referencia, se observa que en la predicción dinámica, el RECM cae 79.16% y el EAM disminuye 503.45%. Además, en el caso de la predicción estática, estos caen 16.95% y 20%, respectivamente.

Cuadro N° 3
Indicadores de capacidad predictiva para predicciones fuera de la muestra

Variables exógenas	Modelo de referencia	Modelo aumentando
$Observaciones$	20	20
$R^2_{ajustado}$	0.986	0.989
$RECM_{dinmico}$	7.58	1.58
$EAM_{dinmico}$	7.00	1.16
$RECM_{esttico}$	0.59	0.49
$EAM_{esttico}$	0.50	0.40

En el caso de la prueba de Diebold y Mariano para modelos anidados, con un nivel de confianza de 95%, hay suficiente evidencia estadística para no rechazar la hipótesis nula, que indica que ambos son iguales en términos de predicción. Dado que el signo es positivo, se puede afirmar que los errores del modelo 1 son mayores a los del modelo 2; es decir la capacidad predictiva del modelo aumentado es mejor que la del modelo de referencia.

Cuadro N° 4
Prueba Diebold y Mariano

Diebold y Mariano para modelos anidados	
$Observaciones$	20
$EstadsticoDM$	2.51
$Pvalue$	0.02

Ho: modelos 1 y 2 son similares en términos de predicción.
Ha: el modelo 2 predice mejor que el modelo 1

En síntesis, los resultados muestran que la predicción de IE100 mejora al incorporar los datos de *Google Trends* sobre todo en predicciones fuera de la muestra.

5.4 MODELO DE CORRECCIÓN DE ERRORES

Dado que las series IE100 e IGD son $I(1)$, es posible analizar si ambas están o no cointegradas. Es decir, si existe o no una combinación lineal de ambas variables que sea $I(0)$. Para ello, como se muestra en el Cuadro N°5, se realizó una prueba de raíz unitaria a los residuos de la relación de largo plazo: $\epsilon_t = IE100_t - \beta IGD_t$. Adicionalmente, para tener una mayor evidencia de que estas cointegran, se realizó las pruebas de traza y valor propio máximo de Johansen (Ver Cuadro N°6).

Cuadro N° 5
ADF a residuos de Largo plazo

ADF sobre los residuos de la relación de largo plazo	
Conconstante	0.00
Estadsticot(*Ho* : *interceptonulo*)	0.07
Pvalue	0.00

Notas: *Nivel de significancia al 0.05 usando los valores críticos de McKinnon.*
Los valores mostrados son las probabilidades asociadas.

Cuadro N° 6
Metodología de Johansen

Estadístico Traza	
Noexistevectordecointegracin	0.00
*Hasta*1*vectordecointegracin*	0.12
Estadístico Valor propio máximo	
Noexistevectordecointegracin	0.00
*Hasta*1*vectordecointegracin*	0.12
¿Cointegran?	Sí

Notas: *Nivel de significancia al 0.05 usando los valores críticos de McKinnon.*

Los resultados de la prueba de cointegración basada en la estacionariedad de los residuos (usando la prueba ADF) y las pruebas de Johansen confirman que existe una relación de

cointegración. Esta relación está determinada por:

$$IE100_t = 109.33 - 1.10IGD_t. \tag{18}$$

Como se aprecia, la pendiente de la función de IE100 es negativa, cercana a uno y estadísticamente significativa. Para analizar si IGD es un predictor insesgado de la tendencia, adicionalmente se realizó un modelo restricto, en donde se fijó los coeficientes asociados a un valor igual a 1. Como se señala en el Cuadro N°7, con un nivel de significancia de 5%, hay suficiente evidencia estadística para no rechazar la hipótesis nula; es decir, IGD es un predictor insesgado.

Cuadro N° 7
Modelo VEC con restricción 1

Restricción: $B(1,1) = 1, B(1,2) = 1$	Probabilidad
Chi − square	1.14
Probabilidad	0.29

Al analizar los coeficientes del error de cointegración (velocidad de ajuste) en cada ecuación de corto plazo, se puede apreciar que IGD no es débilmente exógena.

Cuadro N° 8
Exogeneidad débil (metodología de Johansen)

	Ecuación de IE100	Ecuación de IGD
Velocidaddeajuste	-0.20	-0.37
Estadsticot	-3.10	-2.35
conclusin	IE100 no es débilmente exógena	

Sin embargo, al imponer la restricción de exogeneidad débil no se rechaza la hipótesis nula. En este contexto, luego de realizar las pruebas mencionadas y al realizar la prueba de causalidad en el sentido de Granger, se encontró que IGD permite realizar predicciones contemporáneas en "t" y "t+1"; no obstante, no permite conocer la senda futura de esta variable macroeconómica para más de un periodo hacia adelante.

Cuadro N° 9
Modelo VEC con restricción 1

Restricción: $B(1,1) = 1, B(1,2) = 1, A(2,1) = 0$	Probabilidad
Chi − square	4.47
Probabilidad	0.11

Cuadro N° 10
Causalidad a lo Granger

Hipótesis nula	Probabilidad
IGD no causa a IE100	0.42
IE100 no causa a IGD	0.80

6 CONCLUSIONES

Se examinó si la información proporcionada por *Google Trends* era relevante para predecir el índice de empleo para empresas de 100 y más trabajadores (IE100). Para ello, se elaboró un índice llamado IGD a partir de la información de *Google Trends*, que reflejaría a la población que busca trabajo en el mercado laboral. De esta forma, se plantearon modelos ARMA y ARDL: un modelo de referencia, que no incluía a IGD y otro modelo aumentado que sí lo incluía. Se encontró que en predicciones dentro de la muestra, IGD mejora las predicción dinámica, sin embargo, los indicadores de la predicción estática aumentan ligeramente. Además, se halló que en predicciones fuera de la muestra, la capacidad predictiva del modelo mejora al incorporar la información de Google en predicciones estáticas y dinámicas.

Por otro lado, dado que las variables eran integradas de orden uno, también se aplicó un modelo de corrección de errores. Sin embargo, es importante señalar que la muestra solo estaba conformada por 68 observaciones y eso era una limitación para la aplicación de este modelo.

Los resultados del estudio indican que IGD permite realizar predicciones contemporáneas y un periodo hacia adelante; empero, no permite anticipar la senda futura de esta variable macroeconómica para más de un periodo hacia adelante. Sobre la base de dichos resultados, se puede concluir que IGD es un buen predictor en el corto plazo de IE100.

Demostrar que efectivamente la información proporcionada por *Google Trends* ayuda a predecir variables económicas, como el empleo, resulta importante pues *Google Trends* es una fuente de información con una frecuencia más alta (semanal), está disponible antes que la publicación de información de fuentes oficiales, brinda información sobre las necesidades de las personas y, por último, es gratis. Esta resultaría una herramienta importante sobre todo en épocas de crisis, en donde el seguimiento en tiempo real y predicción de variables económicas es relevante para la formulación de políticas oportunas.

7 ANEXOS

GRÁFICO 6. *Ejemplo de indexación para el IGD*

fecha	promedio mensual de "busco trabajo"	promedio mensual de "bolsa de trabajo"
2011-04 $\}x_{i,t}$	16.5	49.5
2011-05	17.2	50.8
Promedio (A)	16.85	50.15
Desviación Estándar (B)	0.49	0.92
1/Desviación Estándar (C)	2.02	1.09
Suma de la inversa de las desviaciones estándar (2.02+1.09) (D)	3.11	
ponderadores (C/D) α_i	0.65	0.35

IGD para el mes 04 (indexado) = $\sum \alpha_i x_{i,t=4}$=16.5*0.65+49.5*0.35

GRÁFICO 7. *Inspección gráfica y pruebas de raíz unitaria para la serie IE100*

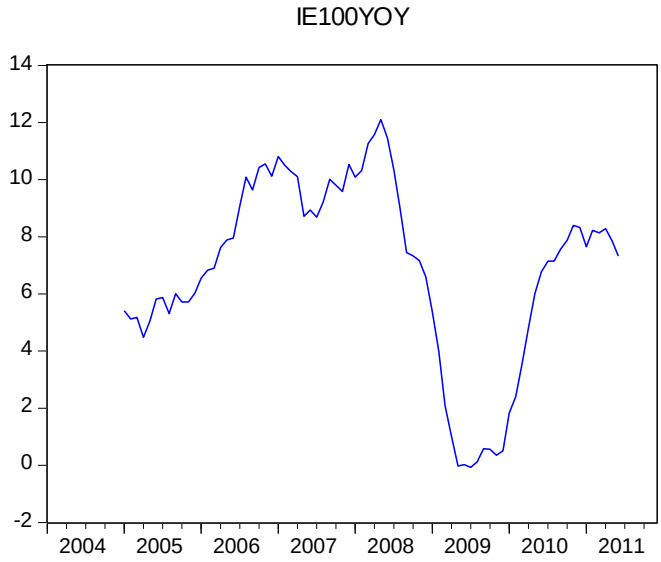
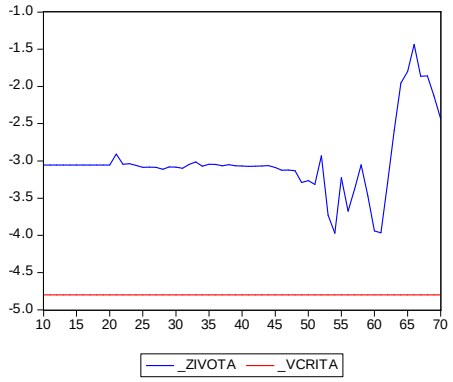


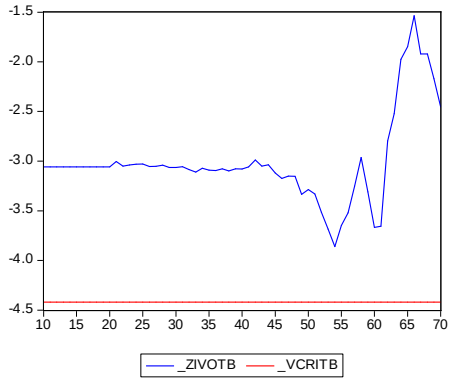
GRÁFICO 8. *Pruebas de raíz unitaria para la serie IE100*

Prueba	*ADF*	*PP*
Con constante y tendencia	0.50	0.74
Con constante	0.26	0.46
Sin constante ni tendencia	0.49	0.48

Notas:
*-*Nivel de significancia al* 0.05
- Los valores mostrados son las probabilidades asociadas

GRÁFICO 9. *Prueba Zivot y Andrews para IE100*





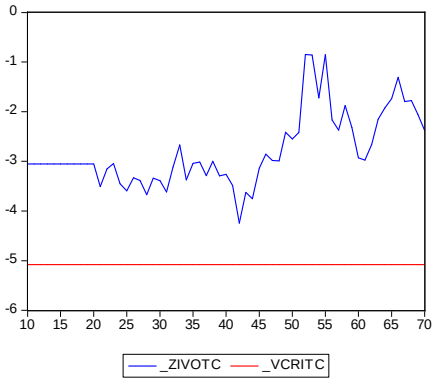


GRÁFICO 10. *pruebas de raíz unitaria para la diferencia de IE100*

Prueba	*ADF*	*PP*
Con constante y tendencia	0.00	0.00
Con constante	0.00	0.00
Sin constante ni tendencia	0.00	0.00

Notas:
*-*Nivel de significancia al* 0.05
- Los valores mostrados son las probabilidades asociadas

GRÁFICO 11. *Inspección gráfica IGD*

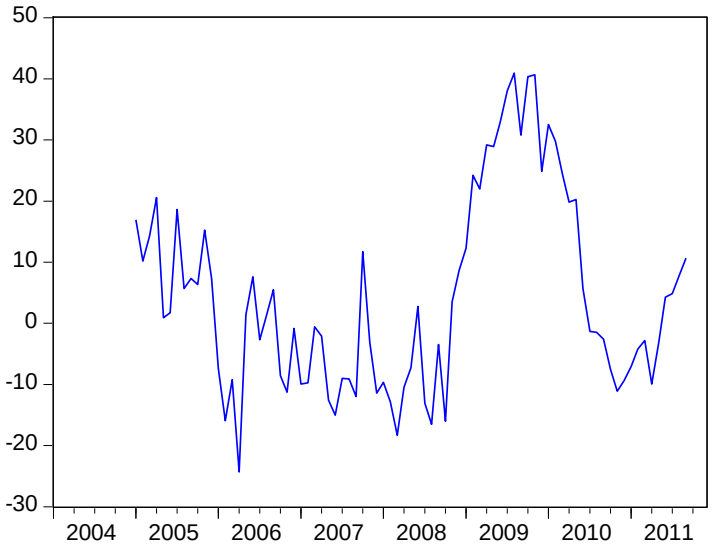
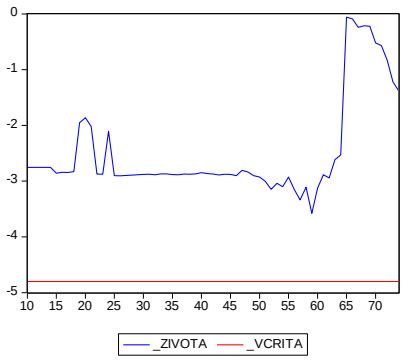


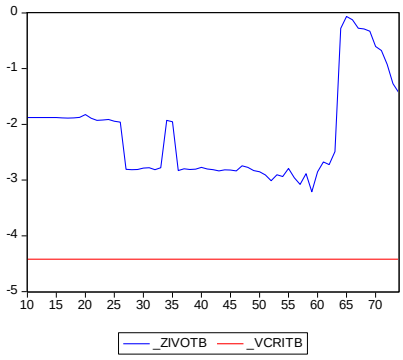
GRÁFICO 12. *Pruebas de raíz unitaria para la serie IGD*

Prueba	*ADF*	*PP*
Con constante y tendencia	0.17	0.21
Con constante	0.08	0.11
Sin constante ni tendencia	0.01	0.02

Notas:
*-*Nivel de significancia al* 0.05
- Los valores mostrados son las probabilidades asociadas

GRÁFICO 13. *Prueba Zivot y Andrews para IGD*





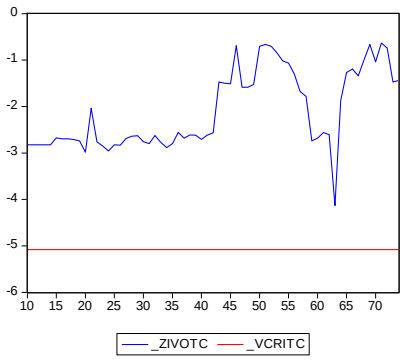


GRÁFICO 14. *Pruebas de raíz unitaria para la serie Digd*

Prueba	*ADF*	*PP*
Con constante y tendencia	0.00	0.00
Con constante	0.00	0.00
Sin constante ni tendencia	0.00	0.00

Notas:
*-*Nivel de significancia al* 0.05
- Los valores mostrados son las probabilidades asociadas

GRÁFICO 15. *Correlograma del modelo de referencia*

Date: 12/15/11 Time: 00:19
Sample: 2006M01 2011M08
Included observations: 68
Q-statistic probabilities adjusted for 1 ARMA term(s)

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
.*| . |	.*| . |	1	-0.120	-0.120	1.0306	
. |*. |	. |*. |	2	0.158	0.146	2.8340	0.092
. |*. |	. |*. |	3	0.097	0.135	3.5201	0.172
.*| . |	.*| . |	4	-0.177	-0.184	5.8461	0.119
. |*. |	. |*. |	5	0.183	0.118	8.3833	0.079
. |*. |	. |** |	6	0.192	0.299	11.221	0.047
. | . |	. | . |	7	0.003	0.035	11.222	0.082
. |*. |	. | . |	8	0.167	0.019	13.445	0.062
.*| . |	.*| . |	9	-0.112	-0.090	14.460	0.071
.*| . |	.*| . |	10	-0.088	-0.101	15.096	0.088
. | . |	.*| . |	11	-0.033	-0.122	15.185	0.125
.*| . |	.*| . |	12	-0.159	-0.192	17.340	0.098
. | . |	.*| . |	13	-0.052	-0.169	17.575	0.129
. | . |	.*| . |	14	-0.066	-0.112	17.954	0.159
.*| . |	. | . |	15	-0.108	-0.050	19.007	0.165
.*| . |	.*| . |	16	-0.094	-0.087	19.812	0.179
.*| . |	. | . |	17	-0.081	0.007	20.429	0.202
.*| . |	. | . |	18	-0.158	-0.023	22.799	0.156
.*| . |	. | . |	19	-0.098	-0.042	23.737	0.164
.*| . |	.*| . |	20	-0.125	-0.066	25.279	0.152
.*| . |	.*| . |	21	-0.131	-0.107	27.024	0.135
. |*. |	. |*. |	22	0.106	0.140	28.189	0.135
.*| . |	.*| . |	23	-0.141	-0.077	30.285	0.112
. | . |	. | . |	24	0.017	-0.056	30.317	0.141
. | . |	. | . |	25	-0.049	-0.063	30.586	0.166
.*| . |	. | . |	26	-0.104	-0.028	31.809	0.164
. | . |	. | . |	27	0.058	-0.017	32.197	0.187
. | . |	. | . |	28	0.015	-0.056	32.224	0.224

GRÁFICO 16. *Prueba Breusch-Godfrey*

Breusch-Godfrey Serial Correlation LM Test:

F-statistic	1.971367	Prob. F(2,62)	0.1479
Obs*R-squared	4.034666	Prob. Chi-Square(2)	0.1330

GRÁFICO 17. *Prueba de heteroscedasticidad de White para modelo de referencia*

Heteroskedasticity Test: White

F-statistic	1.337778	Prob. F(14,53)	0.2176
Obs*R-squared	17.75525	Prob. Chi-Square(14)	0.2182
Scaled explained SS	13.08225	Prob. Chi-Square(14)	0.5201

REFERENCIAS BIBLIOGRÁFICAS

Askitas, N. y K. F. Zimmermann (2009). Google Econometrics and UnemploymentForecasting, Applied Economics Quarterly, 55(2), 107-120, Alemania.

Bersier, F. (2010). Towards Better Policy and Practice Using Real-Time Data, Oxford Internet Institute, Inlaterra.

Bescond, D.; A. Châtaignier y Farhad Mehran (2003). Siete indicadores para medir el trabajo decente. Comparación internacional". Revista Internacional del Trabajo, vol. 122, N° 2.

Castle, J; N.W.P. Fawcett y D. F. Hendry (2009). Nowcasting is not just Contemporaneous Forecasting, Oxford University.

Carrière-Swallow, Y. y F. Labbé (2010). Nowcasting with Google Trends in an emerging market, Documento de Trabajo, 2010-588, Banco Central de Chile.

Chacaltana, Juan (2001). ¿Qué sabemos sobre el desempleo en el Perú?, INEI, CIDE y Programa MECOVI-PERU.

Clark, T. E., y K. D. West (2007), "Approximately normal tests for equal predictive accuracy in nested models", Journal of Econometrics, 138, 1, pp. 291-311.

D'Amuri, F. y J. Marcucci (2009).Google it! Forecasting the US unemployment rate with a Google job search index, working paper 2009-32, Institute for social and Economic Research, Inglaterra.

Della Penna, N. y H. Huang (2009). Constructing Consumer Sentiment Index for U.S. Using Google Searches, Working Papers, 2009-26, Universidad de Alberta, Canadá.

Diebold F.X. y Mariano R.S. (1995). Comparing predictive accuracy. Journal of Business and Economic Statistics 13, 253-263.

Enders, Walter (2009). Applied Econometric Times Series.

Enders, W. (2010), Applied Econometric Time Series.

Fosk, A. (2011). Estado de Internet con un enfoque en el Perú, ComScore.

Garavito, Cecilia (2000). Empleo y desempleo: un análisis de la elaboración de estadísticas, PUCP, Documento de Trabajo N°180.

Granger C.W.J. y Newbold P. (1986).Forecasting economic time series. New York, Academic Press.

Harvey, L. y Newbold (1997). Testing the equality of prediction mean squared errors.International Journal of Forecasting 13, 281-291.

ILO (2006). 2006 Labour Overview. Latin America and the Caribbean, Regional Office for Latin America and the Caribbean, ILO, Lima, pp.61-64

OIT (2010). Panorama Laboral 2010 América Latina y el Caribe, Oficina Regional para América Latina y el Caribe, OIT, Lima, pp. 126-134.

Oleksandr, B.(2010). Can Google's search engine be used to forecast unemployment in Ukraine?, Kyiv School of Economics, Ucrania.

Schmidt, T. y S. Vosen (2009). Forecasting Private Consumption: Survey-based Indicators vs. Google Trends, Ruhr Economic Papers, 2009-155, Alemania.

Suhoy, T. (2009). Query Indices and a 2008 Downturn: Israeli Data,Discussion Paper , 2009-06, Banco de Israel.

Song, H, P. Bing y D. Y-L Ng (2010).Forecasting demand for hotel rooms with search engine query volume data, School of Hotel and Tourism Management, Hong Kong y College of Charleston, U.S.

Tapscott, D. (2009). Grown up digital: How the net generation is changing your world.

Choi, H. y H. Varian (2009a). Predicting the Present with Google Trends, Technical Report, Economics Research Group, Google, EstadosUnidos.

Choi, H. y H. Varian (2009b). Predicting Initial Claims for Unemployment Benefits, Technical Report, Economics Research Group, Google, EstadosUnidos.

Velásquez, Mario (2010). Seguros de desempleo y reformas recientes en América Ltina, Macroeconomía del desarrollo, serie 99, División de Desarrollo Económico, Cepal, Chile, pp. 15 -23.

West, K.D. (1996). Asymptotic Inference about Predictive Ability. Econometrica 64(5): 1067-84.