CENTRO DE INVESTIGACIÓN Y DOCENCIA ECONÓMICAS, A.C.

EL PRESAGIO DE GOOGLE:
PREDICIENDO EL DESEMPLEO EN EL CASO DE MÉXICO

TESINA

QUE PARA OBTENER EL TÍTULO DE

LICENCIADO EN ECONOMÍA

PRESENTA

SANTIAGO DE BUEN URRUTIA

DIRECTOR DE TESINA: DR. RODOLFO CERMEÑO

CIUDAD DE MÉXICO SEPTIEMBRE, 2016

"I'm a great believer in luck, and I find the harder I work the more I have of it."
- Thomas Jefferson

Agradecimientos

Quiero agradecer:

A mi mamá porque siempre ha estado presente, y siempre con la mejor disposición. Por el apoyo diario e incondicional. Gracias por heredarme mi mejor característica: tu sentido del humor. Me has enseñado a ser positivo ante todo y a disfrutar de cada momento. He aprendido de ti que el enojo y el estrés es una decisión que rara vez vale la pena.

A mi papá, por ser un extraordinario ejemplo a seguir. Por su esfuerzo constante para siempre darme lo mejor. Por enseñarme el valor del trabajo, la contundencia de la lógica, y la nobleza de la honestidad. Por mostrarme que con dedicación y trabajo se pueden lograr todas mis metas.

A mis hermanos, Daniel y María. A Daniel por todas las pláticas y experiencias compartidas. Por ir dos pasos adelante y siempre ofrecerme el mejor consejo. A María porque su motivación puede ser contagiosa. Me ha mostrado que de pronto hay que dejar de lado la racionalidad e intentar cosas nuevas.

A Alejandro Villagómez, quien ha sido un extraordinario mentor y amigo. Por guiarme a lo largo de la carrera hacia un camino exitoso y recordarme que las calificaciones sí importan. Por siempre darme el mejor consejo y apoyarme como lo ha hecho. A Rodolfo Cermeño, por su voluntad para dirigir este proyecto y su vocación como profesor. A Daniel Ventosa, porque sin su ayuda este trabajo simplemente no existiría. A Fausto Hernández, por todos los consejos durante el seminario de titulación.

A Roberto Gómez, quien ha sido mucho más que sólo un mejor amigo. Por las risas y los momentos inigualables, porque sería imposible pedir más en un amigo. A Carlos Aguilar y Gonzalo Ares de Parga, porque nunca pensé que compartir un salón de clases diario con las mismas personas por cuatro años sería tan divertido. Por todas las discusiones y bromas, por el privilegio de haber compartido ese tiempo con ustedes. A Roberto Romero, porque su intensidad siempre me hace reír. A Pac, por ser el mejor integrante de un gran equipo y un excelente amigo. Finalmente, gracias a Simon, Andrés y Wash por hacer de cada invierno y de cada verano un periodo de grandes reencuentros con grandes amigos.

Sobre todo, gracias a María Fernanda Porras por ser una inmejorable compañía a lo largo de estos inolvidables cuatro años. Por enriquecer cada experiencia y agrandar cada sonrisa, o por regalarme una cuando más la necesitaba. Por ser una motivación constante cuando no encontraba razón para seguir trabajando. Porque haber compartido estos cuatro años ha sido una de las mejores cosas que me ha dejado el CIDE. Por todas las risas y momentos increíbles, gracias.

Índice general

Índice de figuras

Índice de cuadros

Capítulo 1

Introducción

El rezago en la publicación de información macroeconómica es un obstáculo importante para el desarrollo de política pública y la toma de decisiones. Tener mediciones más oportunas de variables clave permitiría reaccionar ante nuevas tendencias mientras estas se desarrollan. Una de las variables esenciales para medir el funcionamiento económico de un país es la tasa de desempleo.

En México, la tasa de desempleo se mide de forma mensual por el Instituto Nacional de Estadística y Geografía (INEGI). La información tiene un rezago de publicación de un mes, lo cual presenta una oportunidad para estimar la tasa de desempleo en tiempo real.

Por otra parte, Google publica estadísticas de las búsquedas que realizan sus usuarios en tiempo real a través de una plataforma gratuita llamada *Google Trends*. Una gran cantidad de información se genera diariamente en el internet. Con el paso del tiempo la accesibilidad a información generada en tiempo real aumenta, y puede resultar un gran insumo para mejorar el poder de las estimaciones actuales. Este trabajo se enfoca en utilizar una palabra clave de búsqueda en Google para hacer pronósticos contemporáneos y futuros de la tasa de desempleo en México.

La idea de utilizar datos de búsquedas en línea es aprovechar la revelación de preferencias y características de las personas. Por ejemplo, entre más usuarios busquen palabras relacionadas

con la compra y venta de casas, ¿podrá estar esto relacionado con un posible incremento en la actividad del mercado inmobiliario? De la misma forma, se espera que si más personas buscan palabras relacionadas con la búsqueda de empleo, la tasa de desempleo este aumentando al mismo tiempo.

Las estimaciones con datos obtenidos del internet han ganado popularidad pero siguen siendo pocas y relativamente recientes; es una rama de la literatura aún incipiente.

Algunos de los avances más interesantes que se han hecho son en el campo de la epidemiología. Estudios importantes de Polgreen, Chen, Pennock, y Nelson (2008) y Ginsberg y cols. (2009) mostraron que las búsquedas en Google se pueden usar para predecir la incidencia de influenza. Las personas enfermas utilizan el motor de búsqueda para evaluar posibles síntomas y así, los investigadores pueden estimar en tiempo real qué tanto se ha esparcido una cierta enfermedad. Considerando una rama de la literatura más relevante, Choi y Varian (2011) muestran que se puede usar la misma base de datos para estimar en tiempo real algunos indicadores económicos importantes como ventas de automóviles, turismo, confianza del consumidor y, notablemente, solicitudes del seguro de desempleo. Todas sus estimaciones son para Estados Unidos.

El objetivo de este trabajo es mostrar que los datos de Google se pueden usar para predecir los niveles de desempleo medidos de forma tradicional. Además, se hace énfasis en el poder predictivo del modelo para detectar cambios abruptos en la serie. Así, se podría ver cuando exista un cambio de tendencia en los indicadores de desempleo en México. Lo anterior es útil no solamente porque el desempleo es una medición de bienestar importante sino también porque está directamente ligado a la producción, y refleja el panorama del estado de la economía en general.

Los datos de Google se presentan como un índice en donde 100 representa el momento en el que el término fue más buscado en el periodo de tiempo determinado. Los datos están disponibles desde diciembre del 2013 hasta la actualidad, y se actualizan en tiempo real. La información se recaba a partir de una palabra clave que se elige cuidadosamente.

Los resultados indican que el índice de Google logra mejorar el poder predictivo cuando se agrega a modelos tradicionales. Esta es una herramienta eficaz para revelar el comportamiento de las personas y además es una medición innovadora. La metodología que se emplea para verificar la relevancia del índice es simple; consiste en estimar modelos con términos autoregresivos y de media móvil para hacer comparaciones entre inferencias que incluyan el índice de búsquedas y modelos básicos.

El estudio es importante porque es el primer trabajo que analiza los datos de *Google Trends* para México. Cada país es distinto por las condiciones del mercado laboral y el uso de motores de búsqueda en internet. Para el caso de México, el uso de internet ha crecido de manera importante en los últimos años. Al segundo trimestre del 2015, la penetración era del 57 % de la población (AMIPCI). De las actividades que se realizan en internet, buscar información es la tercera más popular, solamente detrás de consultar correos y navegar por redes sociales. Adicionalmente, Google acapara el 93 % del mercado de motores de búsqueda en México, por lo que resulta pertinente enfocar el análisis hacia esos datos. Existen múltiples estudios que se centran en el caso de países desarrollados, pero hay muy pocos que se enfoquen en países en vías de desarrollo, especialmente latinoamericanos.

El trabajo se estructura de la siguiente manera. En el segundo capítulo se discute la literatura relevante. Posteriormente, en el capítulo tres se hace una descripción de los datos. En el capítulo cuatro se formula la metodología, y en el cuarto capítulo se presentan los resultados empíricos. En el capítulo seis se discuten los resultados y las limitaciones, mientras que el séptimo capítulo concluye.

Capítulo 2

Revisión de Literatura

Recientemente, ha habido un incremento importante en la cantidad de información cuantitativa generada a través del internet y que ahora es de acceso público, por lo que se ha convertido en una buena fuente de datos. Una de las más relevantes es *Google Trends*. La literatura económica ha incorporado a las búsquedas en tiempo real en modelos predictivos y se ha probado la efectividad de datos en tiempo real para mejorar predicciones tanto en variables micro como macroeconómicas. En esta sección primero se presentan los diversos artículos que han usado datos de Google. Posteriormente, se analizan aquellos que usan los datos para predecir el desempleo en países desarrollados. Finalmente, se hace énfasis en trabajos que se han hecho para países en vías de desarrollo y con menor penetración de internet.

Una de las ramas más interesantes de la literatura se ha enfocado en estudios de epidemiología. Estos son relevantes porque la gente enferma frecuentemente busca sus síntomas en internet para saber qué padecen. Polgreen y cols. (2008) y Ginsberg y cols. (2009) crean índices compuestos de síntomas para predecir la incidencia de enfermedades contagiosas; específicamente, hacen un análisis enfocado a la influenza. Los autores encuentran que las búsquedas por internet revelan la propagación de una enfermedad de manera precisa para el caso de Estados Unidos. Hallan que los datos de Google permiten identificar las zonas geográficas más afectadas y la evolución de la epidemia. Es importante poder analizar la incidencia de una enfermedad en

tiempo real ya que, al conocer la evolución de la misma, se pueden efectuar políticas públicas de salubridad que contengan la epidemia.

En el campo económico, los primeros en usar los datos de Google fueron Choi y Varian (2009a). Su trabajo resalta la relevancia que pueden tener los datos de búsqueda. Utilizando modelos muy sencillos, encuentran que una variable compuesta por un índice de búsquedas puede mejorar el poder predictivo de muchas variables: peticiones de seguro de desempleo en Estados Unidos, demanda de automóviles, destinos vacacionales y confianza del consumidor. En este trabajo también se resalta la importancia de elegir las variables adecuadamente: solamente puede predecir variables para las cuales la gente se tome el tiempo de buscar antes en internet, con la finalidad de obtener información para poder tomar alguna decisión. Destinos de vacaciones y automóviles son claros ejemplos de productos para los cuales los consumidores querrían conocer más información antes de efectuar alguna compra.

Los datos de búsqueda de Google se han usado en varias otras aplicaciones. Guzman (2011) usa la información de *Google Trends* para construir índices que reflejen el comportamiento de la inflación en Estados Unidos, logrando mejorar las estimaciones de manera sustancial. Paralelamente, Huang y Penna (2009) hacen un análisis detallado de una predicción de confianza del consumidor para Estados Unidos.

Respecto al tema del desempleo, una de las aplicaciones más importantes que se le ha dado a los datos de búsquedas ha sido en el campo de la economía laboral. Askitas y Zimmermann (2009) predicen el nivel de desempleo en Alemania. Estos autores resaltan la importancia del poder predictivo en momentos de crisis para evaluar los cambios de tendencia en las series de desempleo. Si se puede hacer una predicción precisa acerca de cambios tan importantes, la política pública se puede ajustar rápidamente para responder a presiones que antes ni se sabía que existían. D'Amuri y Marcucci (2010) usan el mismo tipo de datos para predecir el desempleo en Estados Unidos. Los autores encuentran que su índice de búsquedas de Google tiene más precisión en estimaciones fuera de muestra, para horizontes de tiempo de uno, dos o tres meses. Adicionalmente, los autores comparan modelos que incluyen su indicador con muchos otros al

hacer un análisis para determinar los mejores predictores entre más de 500 especificaciones. D'Amuri y Marcucci encuentran que las mejores predicciones consistentemente contienen el índice de búsquedas. El artículo de Smith (2016) evalúa la utilidad de *Google Trends* para el Reino Unido. El autor utiliza una metodología de estimaciones con muestreo de datos mixtos para explotar la estructura semanal de la información del motor de búsqueda y encuentra que la incorporación del índice mejora la habilidad predictiva del modelo. Sin embargo, la reducción en los errores de pronóstico se acentúa durante periodos de crisis y disminuye a partir del 2012.

La gran mayoría de los estudios se concentran en analizar búsquedas en países desarrollados y de altos ingresos. Específicamente, se usan países con altas tasas de penetración en el uso de internet. Suhoy (2009) analiza el desempleo para Israel, un país que no tiene el mismo nivel de desarrollo que los otros previamente estudiados. Sin embargo, sigue siendo una nación con una alta penetración de uso del internet. Al igual que los autores anteriores, encuentra que los mejores modelos incluyen los datos de Google.

En Latinoamérica se han hecho dos estudios relevantes. El primero fue una investigación realizada por investigadores del Banco Central de Chile: Carrière-Swallow y Labbé (2013) construyen un índice para predecir las compras de automóviles en Chile. La inclusión del índice mejora el poder predictivo y la eficiencia de las estimaciones. Es importante mencionar que, en el 2010, los usuarios de internet como porcentaje de la población en Chile estaban alrededor del 45 %, según datos del Banco Mundial. Aunque Chile es un país con una tasa de penetración más alta que en México, la diferencia no es tan grande como con otros países que han sido analizados en la literatura relevante.

El otro estudio importante para el caso de Latinoamérica fue realizado por Chang y Río (2013) para el Banco Central de Reserva del Perú. El estudio cuestiona si la información proporcionada por *Google Trends* puede predecir el índice de empleo para empresas de 100 y más trabajadores en Perú; al utilizar un índice de empleo para empresas grandes en vez de la tasa de desempleo nacional buscan evitar el problema del mercado laboral informal. El supuesto relevante es que los usuarios de internet son más propensos a también ser trabajadores formales.

Las autoras encuentran que el índice de Google mejora sustancialmente las predicciones tanto dentro como fuera de la muestra. El estudio indica que el índice se puede utilizar para hacer predicciones contemporáneas y un periodo hacia adelante, pero no es eficiente para estimar el resto de la senda futura de la variable. Así, solamente es útil para el muy corto plazo. Sin embargo, eso puede ser crucial para la implementación de política pública, ya que el desempleo es un indicador importante de actividad económica. Esto es especialmente relevante en tiempos de crisis, cuando el seguimiento en tiempo real es difícil y la predicción de actividad económica tan importante.

En México no se han utilizado los datos de Google para hacer predicciones. Es por eso que este artículo contribuye a la literatura, al analizar si la dinámica de búsqueda favorece el poder predictivo de los modelos. Adicionalmente, este análisis contiene un número significativo de observaciones adicionales durante periodos menos volátiles; lo cual también supone una diferencia con respecto a la literatura existente. Antes de detallar la metodología es necesario conocer la estructura de la base de datos.

Capítulo 3

Datos

En este trabajo se utilizan dos series históricas: la tasa de desempleo para México y la búsqueda de la palabra *empleo* en Google. La figura 3.1 muestra ambas series desestacionalizadas.

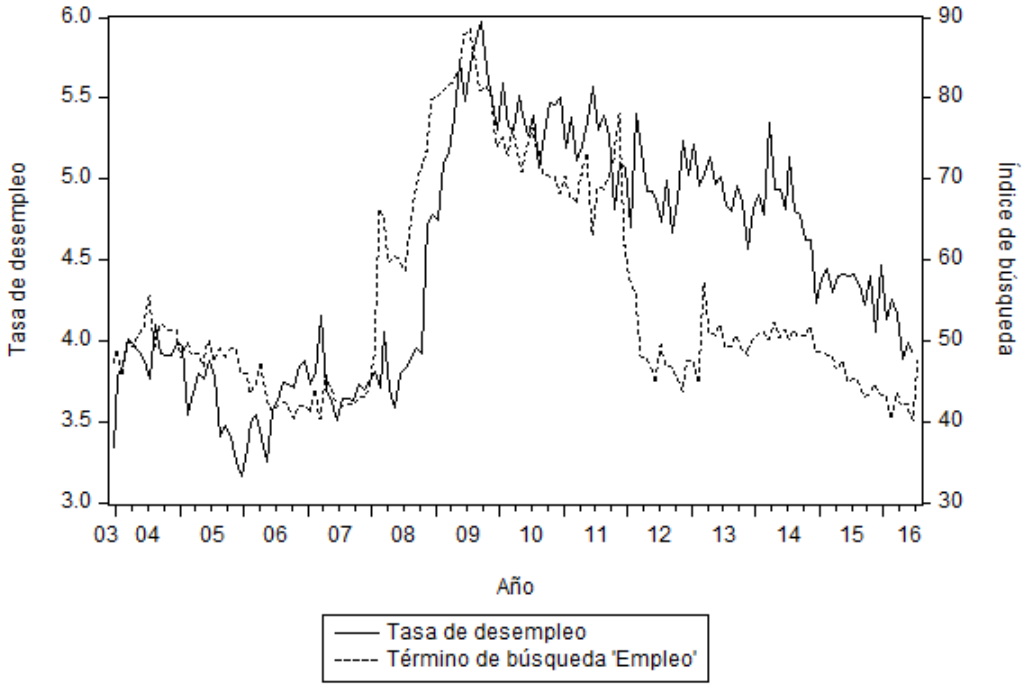

Figura 3.1: Tasa de desempleo y búsqueda por la palabra *Empleo*

3.1. Google Trends

Los datos que se obtienen de *Google Trends* abarcan desde diciembre del 2003 hasta julio del 2016. La información se genera en tiempo real y está disponible a diversos grados de agregación. Para términos de este trabajo, se utiliza una serie mensual ya que es fácilmente comparable con la tasa de desempleo. La serie es un índice que refleja la intensidad de búsqueda de una palabra clave sobre el periodo de tiempo analizado.

Es importante destacar un par de detalles acerca de la generación del índice. En primera instancia, se normalizan los datos tal que cada punto en la serie represente las búsquedas en ese mes entre las búsquedas totales a lo largo del tiempo. Posteriormente se modifica la escala del índice para que tome valores entre cero y cien. El rango representa la intensidad de búsqueda para cada mes, donde valores mayores indican más búsquedas.

El mes en el que el índice toma un valor de 100 es cuando hubo una mayor intensidad de búsqueda. Para comparar, un mes en el que el índice tome un valor de 25 quiere decir que el término se buscó una cuarta parte de las veces que se buscó en el mes más popular —es decir, en el mes que tiene un índice con valor de 100. Finalmente, se introduce un ruido al índice por motivos de confidencialidad. El ruido cambia diariamente, por lo que descargar la serie en días distintos y tomar promedios de cada observación tiende a aclarar el proceso generador de datos.

Existen diversos enfoques para seleccionar los términos de búsqueda de *Google Trends*. Indiscutiblemente, la selección de palabras es un tema fundamental en cualquier análisis de este tipo. El objetivo es encontrar una palabra o una serie de palabras que los usuarios buscarían al encontrarse en desempleo o al pensar que se acerca un periodo de desempleo. Naturalmente, la mayoría de los estudios relacionados utilizan palabras que se identifiquen con dos vertientes: beneficios de desempleo y búsqueda de empleo. Sin embargo, la selección de palabras clave depende del idioma del pais en cuestión y sus características institucionales propias, por lo que no se puede utilizar los mismos términos de búsqueda para distintas regiones.

El primer enfoque consiste en elegir un término representativo del desempleo. Una opción es ‘Seguro de Desempleo’ o ‘Portal del Empleo’; no obstante, utilizar el nombre de algún programa

gubernamental tiene dos posibles desventajas. En primera instancia, puede que en cualquier momento las búsquedas correspondan a cambios o anuncios relacionados con la política pública y no con el desempleo en sí. Además, el nombre del programa no está fijo y es incierto si se podrá utilizar para estimaciones futuras. El segundo método utiliza una mezcla de términos, típicamente ponderados por su importancia. Se construye un índice que refleja el panorama general del desempleo. Finalmente, *Google Trends* publica 'categorías' de series, entre las que destaca una categoría denominada 'categoría - empleo'.

Para el caso de México, se decidió no seleccionar una palabra relacionada con beneficios en el desempleo porque los términos están ligados a programas gubernamentales que han sido sujeto de discusión de política pública. Es por ello que el enfoque utilizado consta en elegir una palabra que refleje la búsqueda de empleo. Para ello, existen dos palabras clave: *empleo* y *trabajo*. Mientras que la palabra *empleo* solamente se utiliza en un contexto laboral, el término *trabajo* se puede referir a la acción de realizar algún esfuerzo y, por lo tanto, puede ser utilizado en otros ámbitos. Bajo la premisa de elegir algo parsimonioso, invariante en el tiempo y representativo, se utiliza la serie del término de búsqueda *empleo*. Otros estudios usan *desempleo* como término principal, la razón por la cual no se elige es porque se espera que ante una situación de desempleo, los usuarios intenten reincorporarse al mercado laboral buscando empleo.

3.2. Desempleo

Las cifras oficiales de desempleo en México son publicadas por el INEGI. La serie se obtiene de la Encuesta Nacional de Ocupación y Empleo (ENOE) que se elabora desde el primer trimestre del 2005. Sin embargo, con el objetivo de aprovechar la serie completa de *Google Trends*, que comienza en diciembre del 2003, se tomaron los datos de desempleo para México de la Organización para la Cooperación y Desarrollo Económico (OCDE). La OCDE publica, dentro de la base de datos de indicadores económicos, una serie histórica armonizada de la tasa de desempleo en México.

El desempleo en México se mide como la población desocupada perteneciente a la Población Económicamente Activa (PEA). Esto quiere decir que son individuos que han buscado trabajo activamente en un periodo de tiempo reciente y que no han logrado encontrar una fuente de empleo. Específicamente, se utiliza la tasa de desempleo: el número de personas desocupadas entre la PEA. Una modificación a la Constitución Política de los Estados Unidos Mexicanos elevó la edad legal mínima para trabajar de los 14 a los 15 años, la tasa de desempleo se refiere al universo de la población de 15 años de edad en adelante. La tasa de desempleo en México para cada mes se publica a finales del siguiente mes, lo cual implica un rezago de publicación efectivo de un mes.

Existe una consideración adicional para el mercado laboral mexicano: la informalidad. El mercado informal en México es una fuente importante de empleo. Las estadísticas de ocupación y empleo que recaba el INEGI se basan en censos poblacionales, por lo cual el sector informal está incluído en la tasa de desempleo. El mercado informal podría implicar un problema si un buscador en línea solo se utilizara para encontrar trabajos formales. Chang y Río (2013) utilizan únicamente una tasa de desempleo en grandes empresas para el caso de Perú, donde el mercado laboral también tiene un alto grado de informalidad.

Capítulo 4

Metodología

El objetivo principal es utilizar la serie *Empleo* para estimar la tasa de desempleo de manera contemporánea. Sin embargo, también se investiga si las búsquedas se pueden usar para elaborar pronósticos a futuro.

Para evitar confusión en la notación, de ahora en adelante se hará referencia al término *Empleo* de *Google Trends* únicamente como la serie *Google*. Asimismo, la serie de tasa de desempleo se va a escribir como *Desempleo*.

4.1. Propiedades de las series

Es bien conocido que la tasa de desempleo típicamente exhibe un comportamiento estacional. Un análisis visual indica que tanto la tasa de desempleo como la serie *Google* son procesos altamente estacionales. Se utiliza el ajuste estacional X-12 ARIMA desarrollado por la Oficina del Censo de Estados Unidos para remover el componente estacional de ambas series.

Como un primer paso es necesario verificar la presencia de raíces unitarias en las series de interés. Aunque no es el enfoque del artículo, es prudente mencionar que hay dos hipótesis principales con respecto a las series de desempleo. Por un lado está la teoría de la tasa natural del desempleo propuesta por Friedman (1968), la cual sugiere que choques a la serie solamente tienen efectos transitorios y eventualmente el proceso se revierte a la tasa natural. La teoría

contrastante fue propuesta por Blanchard y Summers (1986), quienes promueven que choques transitorios pueden tener efectos permanentes, lo cual se ha conocido como la 'hipótesis de histéresis' - Yilanci (2008).

La literatura empírica al respecto es mixta, como se puede ver en Khraief, Shahbaz, Heshmati, y Azam (2015), mientras algunos autores indican que el proceso no es estacionario, otros sugieren que al incorporar no-linealidad y quiebres estructurales, la serie exhibe un comportamiento estacionario. En el periodo de tiempo analizado para este artículo (2004-2016) hubo una gran crisis mundial, lo cual elevó sustancialmente la tasa de desempleo. Es por ello que es necesario que la prueba de raíz unitaria tome en consideración la posibilidad de quiebres estructurales.

La estacionariedad de la serie se verifica utilizando la metodología desarrollada por Kapetanios, Shin, y Snell (2003) (KSS). La prueba KSS añade a la prueba estándar de Dickey-Fuller Aumentada. La hipótesis nula consiste en una raíz unitaria lineal y se prueba contra la hipótesis alternativa de estacionariedad no-lineal. Si se considera la siguiente regresión no lineal:

$$\Delta y_t = \gamma y_{t-1}[1 - exp(-\theta y_{t-d}^2)] + \epsilon_t$$

El parámetro θ es cero bajo la hipótesis nula y mayor que cero bajo la alternativa. Los autores hacen una aproximación de Taylor para poder identificar los parámetros. Además de permitir la no-linealidad, una de las ventajas de la prueba KSS es que permite incorporar la posibilidad de múltiples quiebres estructurales.

Los resultados de la prueba son altamente robustos, tanto para la serie desestacionalizada de la tasa de desempleo como para el índice de Google, se rechaza la hipótesis nula de la presencia de raíz unitaria. Se hacen distintas especificaciones que incluyen cambiar el número de rezagos, el porcentaje de la muestra que tiene que existir entre cualquier par de quiebres estructurales y el número máximo de quiebres. A través de todas las especificaciones, se verifica que las series son estacionarias con un nivel de significancia de 1 %.

Una vez que se determina que las series son estacionarias, es necesario identificar los quie-

bres estructurales. Se sigue la metodología elaborada por Bai y Perron (1998) para la identificación de múltiples quiebres. El objetivo es determinar la existencia de m quiebres, lo cual crea en la muestra $m + 1$ segmentos. En cada segmento, los coeficientes son constantes. El modelo se puede escribir como una regresión simple de la siguiente manera:

$$y_i = x_i^T \beta + u_i \qquad (i = i_{j-1} + 1, \ldots, i_j, \quad j = 1, \ldots, m + 1)$$

El subíndice j representa el segmento mientras que i_j identifica los quiebres. La hipótesis nula de ausencia de quiebres estructurales se evalúa contra la hipótesis alternativa de la existencia de quiebres. El método de Bai y Perron puede seleccionar el número de quiebres estructurales de manera secuencial o a través de una maximización global. La prueba únicamente actúa sobre el intercepto y no en la tendencia porque la teoría económica no sugiere que haya una razón para la existencia de una tendencia en una serie de desempleo. Todas las pruebas se hacen sobre las series desestacionalizadas.

Al efectuar las pruebas, se encuentra evidencia robusta a favor de tres quiebres estructurales en la serie de desempleo, tanto las pruebas secuenciales como las globales indican el mismo número de quiebres. No es el caso para *Google*, donde las pruebas secuenciales indican que no hay quiebres estructurales. Sin embargo, la evolución del índice de búsquedas cumple al pie de la letra con uno de los escenarios detallados en el artículo de Bai y Perron (2003). Los autores explican que en el caso de las series en las que se encuentren dos quiebres estructurales y que el valor del coeficiente regrese al valor original después del segundo quiebre, las pruebas secuenciales no se comportan de manera adecuada. En este caso, es difícil rechazar la hipótesis nula de cero contra un quiebre, pero también es trivial rechazar la hipótesis nula de cero quiebres contra la alternativa de más de un quiebre. Es por ello que los autores recomiendan identificar si es que existe algún quiebre con las pruebas globales y una vez determinado que sí existen, usar la prueba de $L + 1$ quiebres contra L quiebres globales. Al utilizar el método detallado anteriormente, se encuentra que la serie *Google* en efecto cuenta con dos quiebres estructurales.

Los resultados se muestran en la figura 4.1 para la serie de desempleo y en la figura 4.2 para

la serie de *Google Trends*.

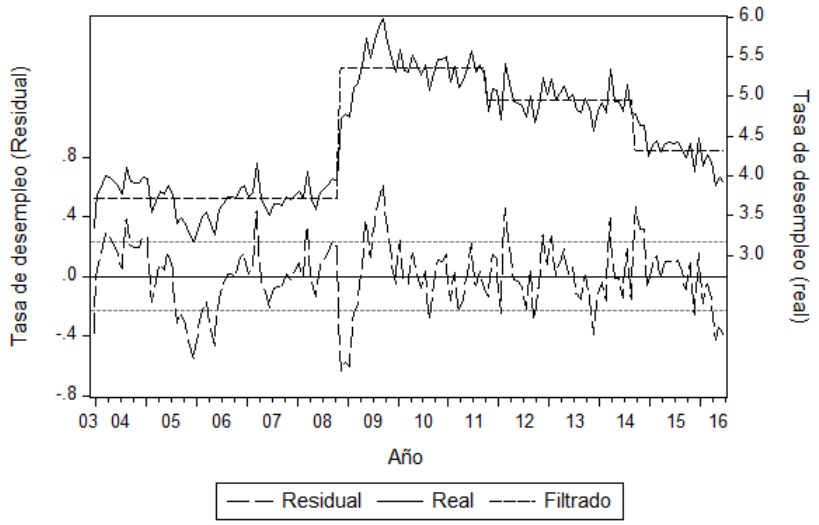

Figura 4.1: Prueba de Bai Perron para la serie *Desempleo*

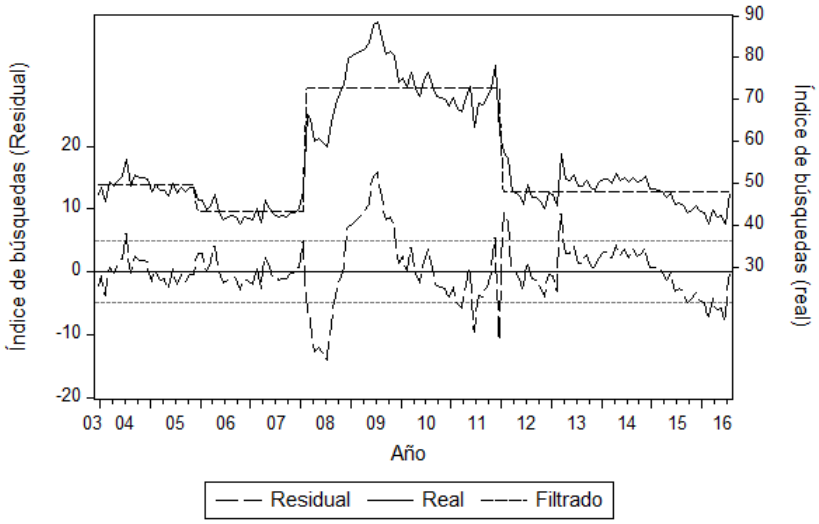

Figura 4.2: Prueba de Bai Perron para la serie *Google*

Para cada serie, se muestran los valores realizados, los quiebres que se encuentran y los residuales. Los residuales de cada serie se convierten en las nuevas series filtradas por quiebres y por estacionalidad. Como se puede observar en ambas gráficas, la crisis financiera representó un gran choque a las estadísticas de desempleo. La prueba en la serie de tasa de desempleo determina que el quiebre ocurrió en noviembre del 2008. Para el caso de *Google*, el quiebre se encuentra en febrero del 2008. La diferencia de nueve meses entre ambos quiebres sugiere que el índice de búsquedas podría ser útil para identificar crisis económicas de manera oportuna.

4.2. Análisis conjunto

Una vez determinado el comportamiento estacionario de la serie y confirmada la presencia de quiebres estructurales, es necesario ahondar en la relación que mantienen la serie de *Google Trends* y la tasa de desempleo. Un primer paso consiste en estudiar el correlograma cruzado de ambas series. El correlograma cruzado indica la fortaleza y dirección de la relación que sostienen los rezagos de las dos series. Es importante verificar si los rezagos de la serie *Google* están correlacionados con los valores actuales de la tasa de desempleo.

Se utilizan pruebas de causalidad de Granger, detalladas en Granger (1969), para verificar si hay causalidad predictiva entre las dos series y, si la hay, en qué dirección va. Es decir, es importante analizar si la serie *Google* causa a la serie de tasa de desempleo. Se dice que una serie X causa en el sentido de Granger a una serie Y si las preddiciones de la serie Y basadas en sus propios valores anteriores y en los valores anteriores de la serie X son mejores que las predicciones de la serie Y basadas en sus valores anteriores únicamente.

Formalmente, la prueba consiste en estimar dos ecuaciones:

$$y_t = a_0 + \sum_{i=1}^{m} a_i y_{t-i} + \sum_{j=1}^{n} b_j x_{t-j} + \epsilon_t$$

$$x_t = c_0 + \sum_{i=1}^{p} c_i x_{t-i} + \sum_{j=1}^{q} d_j y_{t-j} + u_t$$

Se evalúa la hipótesis nula $H_0 : b_1 = b_2 = \ldots = b_p = 0$ contra la hipótesis alternativa de que al menos uno de los parámetros sea distinto de cero. Si se rechaza la hipótesis nula, se dice que X causa en el sentido de Granger a Y. De la misma forma, se hace la prueba sobre los parámetros d para la segunda ecuación.

Aunque la prueba de Granger es ampliamente utilizada, tiene ciertas limitaciones. Gujarati (1995) afirma que una prueba de causalidad es sensible al número de rezagos y a la especificación del modelo porque puede existir otra variable que no se toma en cuenta que ocasione un sesgo de especificación. Con el fin de analizar la robustez de la causalidad, se incorpora la metodología de Toda y Yamamoto (1995).

4.3. Modelo

Todas las estimaciones se basan en una comparación de las siguientes dos ecuaciones autorregresivas lineales:

$$Desempleo_{\ t} = \alpha + \sum_{i=1}^{p} \beta_i Desempleo_{t-i} + \epsilon_t \tag{4.1}$$

$$Desempleo_{\ t} = \alpha + \sum_{i=1}^{p} \beta_i Desempleo_{t-i} + \theta\ Google_t + \sum_{j=1}^{q} \gamma_j Google_{t-j} + \epsilon_t \tag{4.2}$$

La ecuación (4.1) es un modelo autorregresivo puro; es decir, solamente usa la información pasada de la tasa de desempleo para estimar el valor actual. El objetivo del modelo es que funja como una estimación de referencia. La ecuación (4.2) se obtiene al agregar los términos de Goo-

gle. Si la segunda ecuación logra pronosticar mejor que la primera, es recomendable incorporar el término de *Google Trends*. El rezago de publicación en la tasa de desempleo permite usar el valor contemporáneo de la serie *Google* en la segunda ecuación.

Tanto la serie de desempleo como la serie *Google* son consideradas en dos variantes. La primera variante simplemente sigue a la literatura, utilizando las dos series desestacionalizadas. La segunda variante controla tanto por la estacionalidad como por los quiebres estructurales. Por lo tanto, la segunda especificación busca hacer estimaciones con series filtradas.

Ambos métodos tienen ventajas y desventajas. Al filtrar las series con respecto a los quiebres, se pierde la ganancia predictiva en momentos volátiles. Se espera que los pronósticos sean mejores para periodos de baja volatilidad ya que se conserva una mayor consistencia en los parámetros. La desventaja evidente es que se reduce la capacidad de predecir quiebres en la tasa de desempleo, tanto las series filtradas por quiebres y estacionalidad como las series desestacionalizadas se evalúan para comparar su efectividad.

Es necesario elegir p, el número de rezagos que tiene el modelo de referencia. En primer lugar, se hace un análisis de la función de autocorrelación y la función de autocorrelación parcial. La función de autocorrelación decae progresivamente; el primer rezago de la función de autocorrelación parcial es altamente significativo, mientras que el segundo rezago es ligeramente significativo, todos los otros rezagos son insignificantes. Aunado al análisis de las funciones de autocorrelación, se elige el mejor modelo con base en el criterio de información de Akaike y de Schwartz. Se verifica la significancia de los parámetros y el poder predictivo dentro de la muestra. Sobre todo, se le da mayor importancia al criterio de Schwartz porque tiende a escoger modelos más parsimoniosos.

Las especificaciones que mejor cumplen con los criterios anteriores son un modelo AR(1) para la serie filtrada y un AR(2) para la serie desestacionalizada. Incorporar más rezagos resulta en parámetros no significativos y modelos sobreajustados. Si el objetivo es elaborar estimaciones fuera de muestra, es preferible tener un modelo sencillo que no incurra en el riesgo de sobreespecificación.

4.3.1. Prediciendo el presente

El rezago en la publicación de cifras oficiales de desempleo en México permite utilizar la naturaleza de tiempo real de la serie de *Google Trends* para elaborar pronósticos de forma contemporánea. El modelo de referencia es modificado con la incorporación del índice de búsqueda. Para pronosticar la tasa de desempleo en tiempo t se utiliza el valor de *Google* en tiempo t, porque la información ya está disponible. Adicionalmente se incluyen los rezagos relevantes de *Google*.

La metodología consiste en hacer pronósticos estáticos fuera de muestra; es decir, se toma un momento en el pasado como si se estuviera en esa fecha. Se estiman los modelos como si no hubiera información después de ese momento y se evalúa la capacidad predictiva del modelo al evaluar la diferencia entre el pronóstico y el valor realizado. Los pronósticos únicamente se hacen un mes hacia adelante; en cada momento se estima cuál es el siguiente valor de la tasa de desempleo, y para el siguiente periodo se toma el valor realizado, mas no el estimado, para hacer la siguiente predicción. Los pronósticos estáticos tienen márgenes de error bajos, pero solamente funcionan para predicciones de corto plazo. Específicamente, son adecuados para hacer predicciones contemporáneas porque solamente se busca estimar el valor de la tasa de desempleo en el siguiente mes.

La diferencia entre el valor pronosticado y el valor real de la serie para un mes dado se conoce como el error de predicción. Para evaluar el poder predictivo del modelo, se utiliza la raíz del error cuadrático medio (RMSE, por sus siglas en inglés), el error absoluto medio (MAE) y el error porcentual absoluto medio (MAPE):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

donde y_i se refiere al valor realizado mientras que $\hat{y}_i$ es el valor pronosticado. Los errores de predicción del modelo de referencia se comparan con los del modelo con la serie de *Google Trends* para determinar si existe una mejora en el poder predictivo. La prueba elaborada por Diebold y Mariano (1995) se utiliza para verificar si hay una diferencia estadística en la precisión del pronóstico.

4.3.2. Prediciendo una crisis

La información en tiempo real cobra una relevancia especial cuando es capaz de indicar cambios abruptos en los niveles de las series. Además de pronosticar la tasa de desempleo un periodo para adelante en momentos de baja volatilidad, se evalúa si la serie de *Google Trends* mejora la capacidad predictiva del modelo en tiempos de crisis o en cambios de tendencia. Esto es especialmente útil, ya que pronosticar el desempleo se vuelve particularmente complicado durante un periodo de recesión.

Al igual que en la sección anterior, en ésta se utiliza el método de regresión recursiva. La innovación del método recursivo recae en que, para cada periodo se hace una predicción para el siguiente mes y se reestima todo el modelo para hacer el siguiente pronóstico. Estimar el modelo en periodos subsecuentes permite que los parámetros se adapten a los cambios estructurales y mejore la capacidad predictiva. Además, el método recursivo es el enfoque que se tomaría para hacer predicciones en la práctica. No tendría sentido usar las series filtradas por quiebres, así que solamente se hacen las estimaciones con ambas series ajustadas por estacionalidad.

4.3.3. Prediciendo el futuro

En general, la literatura acerca del uso de *Google Trends* para pronosticar series económicas únicamente lo ha hecho para pronósticos de un periodo hacia adelante. Una excepción se puede encontrar en Tuhkuri (2015), quien usa la serie de Google para el desempleo en Estados Unidos

pero concluye que la mejora en los pronósticos no es estadísticamente significativa.

Intuitivamente, hay dos tipos de trabajadores que vale la pena analizar en este enfoque: el primero es el trabajador desempleado, quien está buscando activamente trabajo para reicorporarse a la fuerza laboral; el otro tipo de trabajador es quien tiene trabajo pero por algún motivo espera quedarse desempleado dentro de un par de meses, ya sea por decisión propia o por terminación de contrato. El segundo tipo de trabajador puede empezar a buscar otros trabajos antes de caer en el desempleo, lo cual sugiere que el índice de búsquedas de Google podría funcionar como un indicador adelantado.

A diferencia de las secciones anteriores, en este caso los pronósticos tienen un carácter dinámico; es decir, para el periodo t, se pronostica el valor en el periodo $t + 1$. Posteriormente, se utiliza el valor estimado en $t + 1$ para pronosticar $t + 2$. Uno de los retos de usar esta metodología es que hay que pronosticar la serie *Google* hacia adelante para poder usar esos valores en la regresión de la tasa de desempleo.

Hacer pronósticos varios meses a futuro no quiere decir que se pierda la relevancia del tiempo de publicación de la serie *Google*. En efecto, en cualquier momento del tiempo se tiene una observación más de la misma. Si se va a realizar una predicción de seis meses hacia adelante en la serie *Desempleo*, el primer valor de *Google* ya se conoce, y solamente se tienen que hacer pronósticos para los siguientes cinco meses.

El modelo para reconstuir la serie *Google* varios periodos hacia adelante se elige con base en los mismos criterios que se usaron para la tasa de desempleo. Se hace un énfasis especial en el criterio de Schwartz ya que castiga más la incorporación de nuevos regresores que el criterio de Akaike, y lo que se busca siempre es un modelo parsimonioso. Para el modelo filtrado por quiebres, el mejor modelo es un AR(2). El modelo AR(1) resulta ser el más apropiado para la serie desestacionalizada. Se utilizan los valores pronosticados de *Google* para constuir el modelo ARMAX que describe el comportamiento de la serie *Desempleo*.

La utilidad de poder pronosticar varios meses hacia adelante es evidente. También es fácil ver que los pronósticos dinámicos van a tener errores mayores a los estáticos. No obstante, es

relevante tener una idea de la dirección en la que va a evolucionar la tasa de desempleo varios meses hacia adelante.

De forma análoga a la sección anterior, se utiliza la prueba de Mariano-Diebold para evaluar estadísticamente la mejora predictiva. Además, se construyen intervalos de confianza para las predicciones. Es importante considerar que hay dos fuentes de errores que tienen que ser consideradas para la construcción de los intervalos de confianza. Se tienen que incorporar los errores de los pronósticos de *Google* a los errores de las predicciones del modelo ARMAX.

Capítulo 5

Resultados

En esta sección se discuten los resultados de la relación entre las dos series y los modelos utilizados para pronosticar.

5.1. Diagnóstico de series

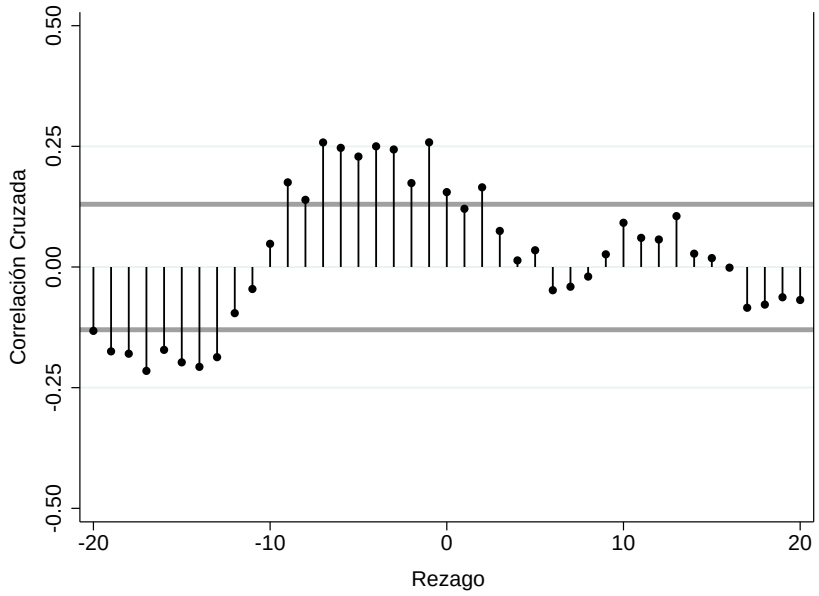

Figura 5.1: Correlación cruzada entre *Google* y *Desempleo*

El correlograma cruzado de las series filtradas por quiebres y estacionalidad se presenta en la figura 5.1. El gráfico muestra la correlación entre los rezagos de *Google* y la variable de interés, *Desempleo*. Es importante notar que los valores pasados de *Google* son estadísticamente significativos, mientras que los valores futuros de la misma serie prácticamente no lo son. Lo anterior tiene implicaciones relevantes. Por un lado, la serie de *Google Trends* tiene información relevante para estimar la tasa de desempleo. Por el otro, las correlaciones son positivas y significativas hasta el noveno rezago, lo cual quiere decir que es posible que los usuarios de internet estén buscando empleo hasta nueve meses antes de caer en desocupación.

Aunque el correlograma cruzado ofrece evidencia visual a favor de una relación interesante, resulta importante investigar si existe causalidad predictiva. El cuadro 5.1 presenta los resultados de las pruebas de causalidad de Granger y de Toda y Yamamoto. Se presentan los resultados para las series filtradas, para las que únicamente están desestacionalizadas y también para las series originales.

Especificación		Causalidad de Granger		Toda & Yamamoto	
¿Controla por quiebres?	¿Controla por estacionalidad?	*Desempleo* a *Google*	*Google* a *Desempleo*	*Desempleo* a *Google*	*Google* a *Desempleo*
Sí	Sí	$1\,\%$	$> 10\,\%$	$1\,\%$	$10\,\%$
No	Sí	$1\,\%$	$5\,\%$	$1\,\%$	$5\,\%$
No	No	$1\,\%$	$5\,\%$	$1\,\%$	$1\,\%$

Cuadro 5.1: Pruebas de causalidad

El porcentaje dentro de cada recuadro es el *p-value* de la prueba, es decir, el nivel de confianza con el que se puede rechazar la hipótesis nula. La hipótesis nula "H_0: la serie *Google* no causa en el sentido de Granger a la serie *Desempleo*" siempre se rechaza a un mayor nivel que la hipótesis nula que establece una causalidad hacia el lado contrario. En efecto, se prueba que la serie de *Google Trends* ofrece información valiosa para pronosticar los valores del desempleo a través de todas las especificaciones.

5.2. Prediciendo el Presente

Los resultados de la regresión dentro de muestra se presentan en el cuadro 5.2.

Variable Dependiente	SA Referencia	SA Google	BR Referencia	BR Google
Y_{t-1}	0.6456***	0.6632***	0.5162***	0.5243***
	(0.0754)	(0.0841)	(0.0702)	(0.0678)
Y_{t-2}	0.3236***	0.2920***		
	(0.0755)	(0.0867)		
$Google_t$		-0.0002		-0.0030
		(0.0057)		(.0053)
$Google_{t-1}$		0.0131***		0.0154***
		(0.0045)		(0.0042)
Constante	4.1918***	3.6336***		
	(0.4343)	(0.4878)		
R^2	0.916	0.920	0.255	0.312
AIC	-0.256	-0.306	-0.403	-0.481
BIC	-0.176	-0.186	-0.363	-0.401
n	151	150	151	150

SA: Series desestacionalizadas

BR: Series filtradas por quiebres y estacionalidad

* p<0.1, ** p<0.05, *** p<0.001

$(\cdot)$: Error estándar

Cuadro 5.2: Resultados de regresión dentro de muestra

Las primeras dos columnas se refieren a los modelos con las variables desestacionalizadas, a diferencia de las últimas dos columnas que además de controlar por estacionalidad, también filtran por los quiebres estructurales en las series. Se presentan resultados de los modelos de referencia y de los modelos con la incorporación del término de búsqueda.

Los resultados de las estimaciones son inusuales. Se puede ver que en todos los modelos, los términos autorregresivos son altamente significativos; sin embargo, el valor contemporáneo de *Google* resulta no serlo. Un valor rezagado del índice de búsquedas también es altamente significativo. Lo anterior tiene grandes implicaciones. La literatura relevante para otros países encuentra que el valor contemporáneo de la serie de búsquedas sí es significativo; no es el caso

para México. La ventaja en el tiempo de publicación de la serie de *Google Trends* se pierde ya que el valor que realmente importa está rezagado un periodo.

El parámetro rezagado de *Google* es altamente significativo. Aunque el valor contemporáneo no lo es, incorporar los términos del índice de búsqueda provoca una mejoría en el ajuste del modelo. Se puede observar que, para ambas especificaciones, todas las mediciones de ajuste tienen mejores resultados para los modelos con el término de búsquedas incorporado.

De cualquier forma, se evalúan los modelos a la luz de estimaciones fuera de muestra. Todas las predicciones se hacen en un entorno estático con una metodología recursiva. Para determinar con más detalle la utilidad del índice de *Google Trends* se estiman tres clases de modelos que incorporan los términos de búsquedas para poder hacer comparaciones con los modelos de referencia.

$$Desempleo_t = \alpha + AR(\rho) + \theta_0 Google_t + \theta_1 Google_{t-1} + \epsilon_t \tag{5.1}$$

$$Desempleo_t = \alpha + AR(\rho) + \theta_1 Google_{t-1} + \epsilon_t \tag{5.2}$$

$$Desempleo_t = \alpha + AR(\rho) + \theta_0 Google_t + \epsilon_t \tag{5.3}$$

Si en efecto el valor contemporáneo no es de ninguna utilidad, el modelo (5.2) debería ser el que mejor prediga los valores de la tasa de desempleo. La especificación (5.3) no debería de ofrecer ninguna ganancia con respecto al modelo de referencia ya que el parámetro asociado a la variable de interés no es estadísticamente significativo.

Es importante recordar que tanto el modelo desestacionalizado como el modelo filtrado se reconstruyen para que los errores de predicción se juzguen con respecto a la serie original. De esa forma, los distintos criterios de evaluación son comparables a través de distintas especificaciones. Los modelos se estiman con información hasta junio del 2014 y se usan los últimos dos años para evaluar la capacidad predictiva del modelo. El cuadro 5.3 presenta los criterios de evaluación para todos los modelos.

Modelo	RMSE	MAE	MAPE	Δ
Desestacionalizado				
Referencia	0.173	0.134	3.107	
5.1	0.160	0.121	2.801	-7.9 %
5.2	0.159	0.121	2.792	-8.1 %
5.3	0.160	0.124	2.851	-7.3 %
Filtrado				
Referencia	0.209	0.167	3.814	
5.1	0.200	0.160	3.651	-4.2 %
5.2	0.199	0.160	3.636	-4.8 %
5.3	0.205	0.164	3.747	-2.0 %

Cuadro 5.3: Resultados de criterios de evaluación de pronósticos

El cuadro 5.3 reporta los resultados de la raíz del error cuadrático medio, el error absoluto medio y el error porcentual absoluto medio. Además, la última columna indica la mejora porcentual de cada modelo con respecto al modelo de referencia en términos de la raíz del error cuadrático medio. La primera sección hace referencia a los modelos desestacionalizados mientras que la segunda lo hace a los modelos filtrados por quiebres y estacionalidad.

Lo primero que hay que resaltar es que todos los modelos exhiben menores errores de predicción que los de referencia. La mejora, a través de los tres criterios y de los tres modelos desestacionalizados, fluctúa entre el $7\,\%$ y el $10\,\%$. Agregar términos de búsquedas en Google sí mejora el poder predictivo del modelo de desempleo. De hecho, el modelo (5.2) es el que mejores pronósticos genera, aunque la diferencia con los otros modelos que incorporan el índice de *Google Trends* no es grande. El resultado anterior indica que, en efecto, el valor contemporáneo de *Google* es inútil. La información más relevante para pronosticar se puede obtener desde un periodo anterior. Un resultado sorprendente es que la mejora porcentual del modelo (5.3) con respecto al modelo de referencia es comparable en magnitud con los efectos encontrados en la literatura, donde el parámetro asociado al valor contemporáneo en la serie de búsquedas es altamente significativo.

La disminución en los errores de predicción se tiene que evaluar estadísticamente. Es por ello que los pronósticos se diagnostican con una prueba Mariano-Diebold. La prueba indica que para la tasa de desempleo desestacionalizada, el modelo con el índice de búsquedas predice estadísticamente mejor que el modelo puramente autorregresivo. La hipótesis nula que establece igualdad en la precisión de los pronósticos se puede rechazar a un nivel de significancia de $1\,\%$. Aunque ya se ha determinado que al filtrar los datos por quiebres se obtienen pronósticos con mayores errores, de cualquier forma se evalúa si la incorporación del índice de *Google Trends* mejora estadísticamente las predicciones. La prueba Mariano-Diebold indica que la especificación con *Google* mejora los pronósticos con respecto al modelo de referencia. Sin embargo, la hipótesis nula solamente se puede rechazar con un nivel de significancia de $10\,\%$.

El segundo resultado importante es que las series desestacionalizadas pronostican mejor que aquellas en las que se controla tanto por estacionalidad como por quiebres estructurales. Todos los criterios de evaluación indican que las series filtradas por quiebres se ajustan peor a los valores realizados. El resultado anterior se debe a que las búsquedas en internet pueden indicar con anticipación la presencia de un quiebre estructural en la tasa de desempleo.

5.3. Prediciendo una crisis

En la sección anterior se determinó que incorporar el índice de búsquedas mejora la capacidad predictiva del modelo; lo cual es más evidente para las series en las que únicamente se controla por estacionalidad. Una posible explicación para la inferioridad de los modelos que filtran por quiebres es que la serie de búsquedas en internet contiene información valiosa para poder identificar quiebres estructurales.

En esta sección se evalúa si *Google* mejora la capacidad predictiva durante la gran crisis mundial del 2008-2009. Con este fin, se utilizan estimaciones recursivas en el periodo que se considera como crisis, que en la serie de desempleo abarca desde mayo del 2008 hasta noviembre del 2011. Al igual que en la sección anterior, el mejor modelo en términos predictivos está dado

por la ecuacion (5.2). Utilizar los datos de *Google Trends* permite reducir el RMSE en $14.82\,\%$. Se hace una prueba Mariano-Diebold para evaluar la significancia de la mejora en la precisión de pronóstico, la cual indica que la mejora es estadísticamente significativa al $1\,\%$.

Adicionalmente, se estudian los puntos de cambio en tendencia buscando evaluar si la incorporación del índice de búsquedas en internet mejora la capacidad predictiva del modelo. La figura 5.2 muestra los periodos en los que la serie de desempleo tuvo un cambio en tendencia. Dos de ellos marcan el comienzo y el final de la gran crisis financiera.

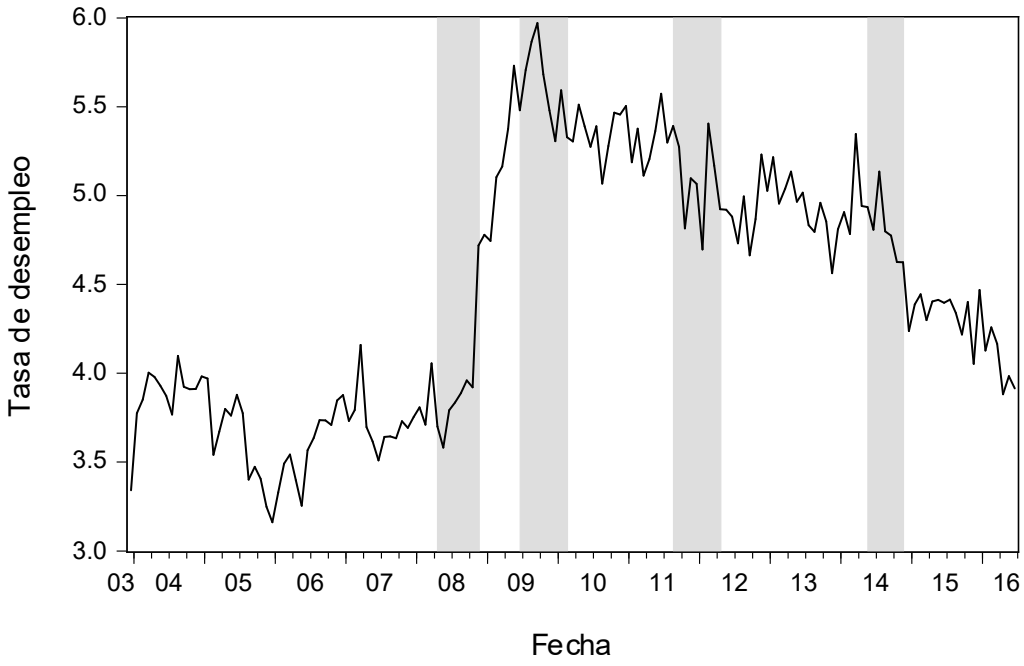

Figura 5.2: Tasa de desempleo. Cambios en tendencia en gris.

Fecha inicial	Fecha final	RMSE Referencia	RMSE Google	Δ
2008m05	2008m11	0.3227	0.2963	-8.2 %
2009m08	2009m12	0.2565	0.1879	-26.7 %
2011m07	2011m11	0.2743	0.2276	-17.0 %
2014m08	2014m11	0.1676	0.1116	-33.4 %

Cuadro 5.4: Error de predicción durante periodos de cambio de tendencia

El cuadro 5.4 presenta los resultados de las estimaciones durante los periodos de cambio de tendencia. En los cuatro periodos, las mejoras en el poder predictivo del modelo son sustanciales. En términos estadísticos, la prueba Mariano-Diebold indica que, para el primer periodo (2008m05 a 2008m11), no se puede afirmar que la diferencia en la precisión de los pronósticos sea estadísticamente significativa. En cambio, para los otros tres puntos en los que cambia la tendencia, la hipótesis nula de la prueba se puede rechazar con un nivel de significancia de $1\,\%$.

La incorporación del índice de búsquedas no solamente mejora el ajuste del modelo durante periodos regulares, sino que también funciona como una herramienta para poder identificar cambios importantes en tendencia o una crisis económica mientras está sucediendo.

5.4. Prediciendo el Futuro

Una aplicación de los datos de *Google Trends* que ha sido poco explorada consiste en hacer pronósticos más allá de un periodo hacia adelante. El correlograma cruzado entre la tasa de desempleo y el índice de búsquedas demuestra que hay una correlación significativa entre los rezagos de *Google* y el nivel de desempleo. En efecto, la información de búsquedas en un mes dado podría utilizarse para pronosticar la tasa de desempleo varios meses hacia adelante; el correlograma cruzado indica una relación significativa y positiva hasta el noveno rezago.

Las estimaciones se hacen para distintos horizontes de predicción; de dos a seis meses en adelante. Los resultados indican que utilizar un rezago de *Google* mejora la raíz del error cuadrático medio entre 2.5 y $4.5\,\%$; sin embargo, la mejora no es estadísticamente significativa al examinar la prueba de Mariano-Diebold. No obstante, es relevante apuntar una observación más:

para estimaciones a cuatro o más meses en adelante, las series filtradas por quiebres y estacionalidad exhiben menores errores de predicción que las series desestacionalizadas. En suma, según los resultados obtenidos, para horizontes de pronóstico cortos, es mejor conservar la estructura de quiebres de las variables; para pronosticar varios meses hacia adelante, es preferible filtrar por quiebres estructurales para tener una serie limpia.

Capítulo 6

Discusión

Los resultados al incorporar el índice de búsquedas de *Google Trends* son prometedores. Al comparar los resultados con la literatura relevante, queda claro que las mejoras alcanzadas al incorporar el índice de busquedas son comparables a las que se han encontrado en otros países. Sin embargo, existe una gran diferencia en las estimaciones para México contra las del resto de la literatura; el parámetro asociado al valor contemporáneo de *Google* no es significativo. Para México, más que funcionar como un indicador contemporáneo con un menor rezago de publicación, el índice de Google toma la función de un indicador adelantado. Por la estructura de tiempo de publicación de los datos, la mejora es igual a lo que se ha encontrado en otros países; se puede mejorar el error de pronóstico un mes antes de que salga la siguiente estadística oficial de tasa de desempleo.

En este artículo se presentan resultados para tres distintos tipos de estimaciones. Para pronosticar un periodo hacia adelante en épocas normales, incorporar el índice mejora la predicción en un $8\,\%$. En tiempos de crisis y cambios de tendencia, la mejora es mucho más grande, y permite identificar de manera oportuna un cambio en la tendencia en la serie de desempleo. Finalmente, la mejora es más modesta para la estimación varios meses hacia adelante. Queda claro que *Google Trends* es más útil para pronósticos a corto plazo. En general, los resultados son comparables en magnitud con la mayoría de los artículos de la literatura relacionada.

La relación existente entre las búsquedas de empleo en Google y la tasa de desempleo verdadera es indiscutible. Es evidente que los usuarios buscan trabajo en internet, y que lo hacen incluso antes de caer en desempleo. Si la relación existe con el desempleo, es probable que se pueda usar *Google Trends* para hacer pronósticos en otras variables de interés. Incluso, las irregularidades del mercado laboral mexicano no inhiben la utilidad de las búsquedas. Esta es una noticia positiva, ya que no era evidente que el método funcionara igual para un país con un mercado laboral dominado por una alta tasa de informalidad y menor acceso a internet.

Existen varias consideraciones para evaluar a más profundidad la utilidad de *Google Trends*. En principio, Google reporta un índice mensual, aunque es posible obtener datos con frecuencia semanal. Una mayor desagregación permitiría verificar si ciertas semanas son más útiles para pronosticar el desempleo, y se ganaría incluso más tiempo por la diferencia entre el tiempo real y el rezago de publicación de datos oficiales. La metodología más utilizada para incorporar datos semanales de *Google Trends* y compararlos con una tasa de desempleo mensual consiste en usar el método MIDAS (por sus siglas en inglés, Mixed Data Sampling). Un ejemplo se puede encontrar en Smith (2016), quien concluye que una desagregación semanal disminuye los errores de predicción y permite tener información todavía más actualizada.

La literatura que se ha enfocado en países desarrollados típicamente usa encuestas de expertos como un regresor adicional. La utilidad de *Google Trends* solamente se confirma si es capaz de mejorar modelos que ya cuentan con una predicción de la tasa de desempleo a futuro en la forma de encuestas. En México no existe ninguna encuesta de acceso público con la cual comparar el modelo. Sin embargo, se podrían considerar regresores adicionales, como indicadores adelantados del nivel del producto, para construir el mejor modelo posible y ver si la incorporación de un índice de búsquedas todavía es significativa. No obstante, ello queda fuera del enfoque del artículo actual.

Finalmente, se podría incorporar un análisis más detallado en cuanto a la forma funcional de los errores de predicción. La metodología analizada solamente considera los criterios estándar en la literatura. Sin embargo, se puede construir una función de pérdida asimétrica para evaluar

con más detalle los errores de pronóstico. En efecto, puede que se le quiera dar una mayor penalización a una subestimación del desempleo que a una sobreestimación. Utilizar una función de pérdida es un avance que no se ha explorado en la literatura relevante.

Capítulo 7

Conclusión

Este artículo explora la utilidad de un índice de búsquedas en internet para pronosticar la tasa de desempleo en México. Los resultados indican que la incorporación del término de búsquedas mejora las predicciones a corto plazo. Un análisis de ambas variables revela que están correlacionadas de manera importante. Los resultados son prometedores, ya que prueban la utilidad de datos generados en internet para México.

Las mejoras en predicción son evidentes en análisis de corto plazo. Aunque se encuentra una diferencia en la significancia de los parámetros con respecto a la literatura, el beneficio obtenido al usar el índice de búsquedas es comparable. El uso de un indicador de búsquedas es especialmente relevante durante periodos de crisis o de cambio en tendencia en la serie. En momentos en los que el mercado laboral está sujeto a cambios fuertes y repentinos, tener una medida en tiempo real de la actividad de búsqueda de trabajo de los individuos en la economía resulta ser muy provechoso.

En general, el artículo aporta a la literatura al analizar el índice de búsquedas en el contexto de México. No parece haber ningún otro artículo que explore la relevancia de *Google Trends* para el país; además de ser útil para México, los resultados sugieren que puede también ser benéfico para países con características similares. Específicamente, países con tasas de penetración de internet más bajas que los países desarrollados y con mercados laborales con una fuerte presencia

de informalidad; características comunes en países latinoamericanos.

Cada vez se genera más información en el internet, y su utilidad ya es evidente. Mientras crece el acceso a diversas fuentes de información, también crece la precisión en los pronósticos de distintos indicadores económicos. Este resultado es increíblemente relevante, ya que nuevas fuentes de información podrían guiar la toma de decisiones y la formulación de política pública en un futuro cercano.

Referencias

Askitas, N., y Zimmermann, K. F. (2009). "Google econometrics and unemployment forecasting." *Applied Economics Quarterly*, *55*(2), 107–120.

Bai, J., y Perron, P. (1998). "Estimating and testing linear models with multiple structural changes." *Econometrica*, *66*(1), 47–78.

Bai, J., y Perron, P. (2003). "Computation and analysis of multiple structural change models." *Journal of Applied Econometrics*, *18*(1), 1–22.

Bangwayo-Skeete, P. F., y Skeete, R. W. (2015). "Can google data improve the forecasting performance of tourist arrivals? mixed-data sampling approach." *Tourism Management*, *46*, 454–464.

Blanchard, O. J., y Summers, L. H. (1986). "Hysteresis and the european unemployment problem." *NBER Macroeconomics Annual 1986*, *1*, 15–90.

Carrière-Swallow, Y., y Labbé, F. (2013). "Nowcasting with google trends in an emerging market." *Journal of Forecasting*, *32*(4), 289–298.

Chang, J., y Río, A. D. (2013). "Google trends: Predicción del nivel de empleo agregado en perú usando datos en tiempo real, 2005-2011." *Documento de Trabajo, Banco Central de Reserva del Perú*.

Choi, H., y Varian, H. R. (2009a). "Predicting the present with google trends." *Technical Report, Google Inc*.

Choi, H., y Varian, H. R. (2009b). "Using internet seach data as economic indicators." *Techincal Report, Google Inc*.

D'Amuri, F., y Marcucci, J. (2010). "Google it! forecasting the us unemployment rate with a google job search index." *SSRN*.

Diebold, F. X. (2005). "Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of the diebold-mariano tests." *Journal of Business & Economic Statistics*, *33*(1), 1–24.

Diebold, F. X., y Mariano, R. S. (1995). "Comparing predictive accuracy." *Journal of Business & Economic Statistics*, *20*(1), 134–144.

Friedman, M. (1968). "The role of monetary policy." *The American Economic Review*, *58*(1), 1–17.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., y Brilliant, L. (2009). "Detecting influenza epidemics using search engine query data." *Nature*, *457*, 1012–1014.

Granger, C. W. J. (1969). "Investigating causal relations by econometric models and cross-spectral methods." *Econometrica*, *37*(3), 424–438.

Gujarati, D. N. (1995). *Basic econometrics*. McGraw-Hill International Editions.

Guzman, G. (2011). "Internet search behavior as an economic forecasting tool: the case of inflation expectations." *The Journal of Economic and Social Measurement*, *36*(3), 119–167.

Huang, H., y Penna, N. D. (2009). "Constructing consumer sentiment index for us using google searches." *University of Alberta, Department of Economics*.

Kapetanios, G., Shin, Y., y Snell, A. (2003). "Testing for a unit root in the nonlinear star framework." *Journal of Econometrics*, *112*(2), 359–379.

Khraief, N., Shahbaz, M. Q., Heshmati, A., y Azam, M. A. (2015). "Are unemployment rates in oecd countries stationary? evidence from univariate and panel unit root tests." *IZA Discussion Paper No. 9571*.

OECD. (2016). "Main economic indicators - complete database." *Main Economic Indicators (database) - http://stats.oecd.org/index.aspx?queryid=36324*.

Pavlicek, J., y Kristoufek, L. (2015). "Nowcasting unemployment rates with google searches: Evidence from the visegrad group countries." *PloS one*, *10*(5).

Polgreen, P. M., Chen, Y., Pennock, D. M., y Nelson, F. D. (2008). "Using internet searches for influenza surveillance." *Clinical Infectious Diseases*, *47*(11), 1443–1448.

Smith, P. (2016). "Google's midas touch: Predicting uk unemployment with internet search data." *Journal of Forecasting*, *35*(3), 263–284.

Stephens-Davidowitz, S., y Varian, H. R. (2015). "A hands-on guide to google data." *Technical Report, Google Inc*.

Suhoy, T. (2009). "Query indices and a 2008 downturn: Israeli data." *Bank of Israel Discussion Paper 2009.06*.

Toda, H. Y., y Yamamoto, T. (1995). "Statistical inference in vector autoregressions with possibly integrated processes." *Journal of Econometrics*, *66*(1), 225–250.

Tuhkuri, J. (2015). "Google searches predict unemployment: How far, when, and how much?"

Varian, H. R. (2014). "Big data: New tricks for econometrics." *The Journal of Economic Perspectives*, *28*(2), 3–27.

Vicente, M. R., López-Menéndez, A. J., y Pérez, R. (2015). "Forecasting unemployment with internet search data: Does it help to improve predictions when job destruction is skyrocketing?" *Technological Forecasting and Social Change*, *92*, 132–139.

Yilanci, V. (2008). "Are unemployment rates nonstationary or nonlinear? evidence from 19 oecd countries." *Economics Bulletin*, *3*(47), 1–5.