Universidad de
SanAndrés

Universidad de San Andrés

Departamento de Economía

Licenciatura en Economía

Google Trends: una nueva herramienta para la predicción económica.

Aplicaciones para complementar el análisis econométrico tradicional

Autor: José Morán

Legajo: 20151

Mentor: Walter Sosa Escudero

Victoria, Buenos Aires, Marzo 2016

Contenido

1. Introducción

El crecimiento exponencial de las ciencias informáticas, como los datos disponibles sobre los que estas se basan, ha llevado a una revolución informática en la última década. La inmensa cantidad de datos disponibles en Internet ha revolucionado no solo las comunicaciones sino todas las ciencias que se basan en el análisis e interpretación de datos. Antes de la difusión de internet, uno de los problemas centrales de un econometrista consistía en la recopilación de datos de los cuales partir un análisis. Hoy en día esa problemática parece haber disminuido con la gran cantidad de información de público conocimiento que un econometrista puede obtener online. Sin embargo la dificultad ya no se basa en la obtención de datos, sino justamente en la filtración de aquellos que resulten relevantes.

Una nueva disciplina informática como el análisis de Big Data lidia con la difícil tarea de estudiar estas nuevas bases de datos. Estas cuentan con cientos de miles de registros para cuantificar un fenómeno en particular, filtrando el ruido de aquellos otros miles de datos irrelevantes para el análisis seleccionado. Los nuevos avances en la econometría han visto esto como una oportunidad para complementar el análisis tradicional si uno sabe bien dónde mirar. Con esto, los economistas se han volcado en estudiar al internet como un nuevo canal donde los usuarios expresan sus preferencias y desarrollan comportamientos que pueden ser medibles y teorizados en modelos microeconómicos. Esta nueva rama de la econometría muchas veces llamada *nowcasting* busca complementar el análisis econométrico tradicional con estas nuevas variables de información para aprovechar la disponibilidad instantánea de los datos para acortar los tiempos de predicción.

El objetivo de este trabajo consiste en estudiar cómo se ha desarrollado esta nueva rama de la econometría y analizar las potencialidades de su aplicación a nuestro contexto microeconómico. Ante una literatura nueva y con escasos reviews es fundamental un estudio sobre su desarrollo y expansión a modo de encontrar convenciones y métodos comunes para estructurarla como una nueva disciplina. La primera parte de este trabajo consiste en el orden y sugestión de esta estructura. En la segunda parte me enfoco en la posible aplicación de esta metodología para un caso concreto. Particularmente para nuestro país, un análisis de este tipo puede ser de suma utilidad debido a la crisis de información pública. Sin contar con acceso a bases de datos e índices confiables, es de gran interés para un economista en Argentina desarrollar canales de información alternativos, transparentes e instantáneos. Basando el análisis en las señales que los usuarios realizan cada vez que interactúan en la web intento explicar comportamientos microeconómicos con datos que de otra forma no estarían disponibles.

En la sección 2 se analiza una de las herramientas introducidas por la compañía Google Inc para cuantificar las búsquedas de los usuarios: Google Trends. Se estudian las ventajas que estas nuevas variables trajeron al estudio de predicciones y los trabajos econométricos que se basaron en este nuevo instrumental. En las sección 3 se analiza el paper precursor de Varian & Choi 2009 que desarrolló el primer modelo econométrico para introducir esta nueva variable para complementar los métodos de predicción tradicionales. En la sección 4 aparece el estudio de otra herramienta de Google: Google Flu Trends, que ha inspirado trabajos posteriores de predicciones en el campo de la epidemiología. En la sección 5 repasamos papers que han sido influenciados por estos trabajos precursores y muestran el gran desarrollo de esta nueva disciplina como una alternativa al estudio econométrico tradicional. Por último en la sección 6 planteo una aplicación para el contexto argentino: explicar la variabilidad de la cotización del dólar paralelo utilizando como variable explicativa las búsquedas en internet relacionadas con el dólar.

2. Google Insights for Search & Google Trends

Hoy en día las redes sociales son unos de los principales medios de expresión y comunicación. Con la popularización de los buscadores de internet y redes sociales como Facebook o Twitter, se multiplicó la cantidad de información disponible en la red. Esta revolución tecnológica nos permite acceder a millones de datos diariamente de manera gratuita e instantánea. La constante interacción entre los usuarios e internet nos permite analizar de cierta forma el comportamiento conjunto de los individuos y su declaración de preferencias a través de su comportamiento en internet.

Esto fue lo que intentó capturar Google cuando lanzó su herramienta Google Trends en mayo de 2006. Al ser el buscador con mayor tráfico de internet Google cuenta con millones de datos sobre las búsquedas de los usuarios. Google no revela el tráfico total por término de búsqueda ya que su negocio consiste en la venta de publicidad y de hacerlo perdería su ventaja competitiva. Pero con Google Trends desarrolló un índice de búsquedas relativo cuyo objetivo fue determinar que palabra o *query* tenía un mayor interés para los usuarios, destacando una tendencia por ciertos términos.

Este índice (de ahora en adelante *query index***)** mostraba el volumen de búsqueda total de un término específico (*query*) sobre el volumen total de búsquedas en un mismo período de tiempo. De esta forma se obtenía un query share, o sea la participación de un término específico dentro del volumen total.

$$\text{Query Share} = \frac{\text{Volumen Total por Query en tiempo: t}}{\text{Volumen Total de Búsquedas en tiempo: t}}$$

Luego el índice era normalizado para tener una media de cero y una desviación de uno. Con esto descartaban efectos por volumen ya que lo importante no era el volumen total de búsquedas de un término específico sino la importancia relativa de ese término dentro del total de búsquedas para una población. Al normalizar el índice es posible comparar poblaciones de distintos tamaños sin que la mayor cantidad de búsquedas impacte en la importancia relativa de un término elegido. Por ejemplo: es probable que en Estados Unidos haya mayor cantidad de búsquedas para el término "bar" que en Escocia. Esto no quiere decir que haya una mayor preferencia en los Estados Unidos por los bares que en Escocia. Al normalizar las muestras podemos comparar los resultados sin que el mayor volumen de usuarios pese en las preferencias finales.

Google Trends descarta también aquellos términos que no lleguen a un piso determinado de búsquedas para eliminar la variación total de la muestra. Por lo tanto los términos por debajo de cierto tráfico no son tenidos en cuenta y no arrojan resultados en caso de buscar su query share. Las búsquedas con connotación pornográfica o con faltas de ortografía tampoco son consideradas.

En agosto de 2008 Google publicó una extensión de su herramienta Google Trends, llamada Google Insight for Search. Esta mejora orientada a los usuarios de Adworks (estadísticas para sitios de internet) buscaba realizar un análisis más minucioso de las búsquedas de los usuarios. Si entendemos el negocio de Google como la venta de publicidad segmentada por palabras entonces una aplicación que les permita a los usuarios analizar que palabras tienen mayor interés direccionaría eficazmente la compra de publicidad.

Esta actualización trajo ciertas mejoras que permitieron un análisis más profundo acerca de las búsquedas de los usuarios. Se introdujo una segmentación de los resultados de búsqueda por zona geográfica, desagregando el análisis y permitiendo obtener datos a nivel nacional, estatal y municipal en los Estados Unidos. Esto fue un avance significativo ya que mientras más acotada sea la muestra a nivel espacial más representativo será el análisis para esa población.

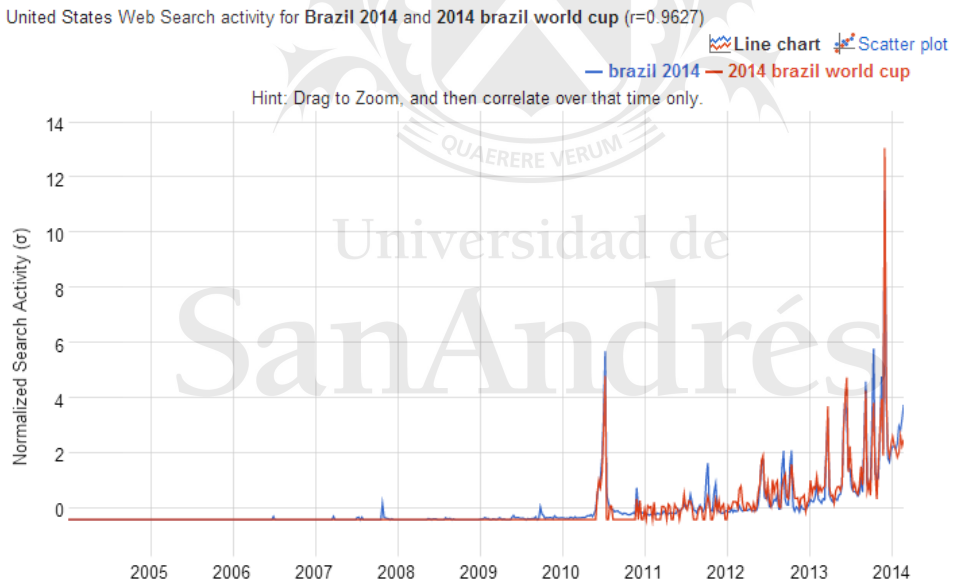
Además con Insights for Search Google comenzó con la categorización de los queries similares agrupándolos dentro de un mismo tema en común. Por ejemplo la palabra "wheel" o "clutch" pertenecen a la subcategoría "automotive parts" que a su vez pertenece a la categoría "automotive". De esta forma Google permite realizar un análisis por categorías sin necesidad de ir sumando cada palabra que esté relacionada con esa categoría. Este catálogo de palabras es un trabajo que Google actualiza diariamente, introduciendo nuevas palabras a cada categoría para volver más significativa la muestra. A su vez la nueva herramienta permitió descargar los resultados obtenidos del índice a un formato de CSV ampliando las posibilidades del análisis. De esta forma uno puede descargar los volúmenes de búsqueda relativos por palabra o por categoría y realizar estudios comparativos entre las tendencias de distintos términos.

En septiembre de 2012 Google fusionó ambas herramientas creando la versión definitiva de Trends tal como se encuentra disponible hoy en día, donde se pueden descargar los resultados obtenidos, graficar las tendencias por palabra y categoría y ver los resultados por región geográfica. Se amplió la categorización de palabras extendiéndola a nuevos idiomas y países. Se continuó trabajando en la desagregación a nivel temporal y espacial. Los datos obtenidos son semanales y se pueden obtener datos desde el 1ro de enero de 2004. Para Estados Unidos se pueden obtener datos sobre el interés por una determinada palabra a nivel municipal, estatal o nacional. Para el resto de los países el mayor nivel de detalle que podemos obtener es a nivel provincial/ regional, incluida la Argentina.

Google Correlate

La nueva versión de Trends trajo una herramienta de correlación llamada Google Correlate. Con esta nueva aplicación podemos buscar una palabra y automáticamente el sistema nos devuelve las 10 primeras palabras con el índice de correlación de Pearson más alto con el término escogido. Por ejemplo si buscamos la frase "Brazil 2014" (búsqueda realizada el 28 de agosto de 2014) los resultados que obtenemos para Estados Unidos son los siguientes:

Cuadro 1. Resultados de Búsqueda "Brazil 2014" y los 10 términos con mayor correlación para esa búsqueda. Fecha de búsqueda agosto 2014.
FUENTE: http://www.google.com/trends/correlate/



Cuadro 2. Búsquedas Normalizadas de "Brazil 2014" y "2014 Brazil World Cup" de 2005 a 2014.
FUENTE: http://www.google.com/trends/correlate/

Esto nos demuestra que hay una alta correlación entre Brasil 2014 y la Copa del Mundo de Brasil 2014. Por lo tanto el nivel de actividad de búsqueda de las personas interesadas en Brasil 2014 es muy similar a aquellas interesadas por la Copa del Mundo realizada en Brasil y aquellos términos que tengan relación con ella. Google Correlate realiza el análisis contrario a Google Trends ya que al ingresar una serie de datos busca aquellos términos correlacionados con esta tendencia.

3. Una primera modelización y posible aplicación: Predicting the Present with Google Trends

A mediados del año 2009 el economista Hal Varian junto a Hyunyoung Choi utilizaron la herramienta analítica Google Trends para predecir comportamientos de masas. En su paper **"Predicting the Present with Google Trends"** utilizan las búsquedas de Google como un nuevo método de predicción comparándolo con los métodos econométricos tradicionales para medir la actividad económica.

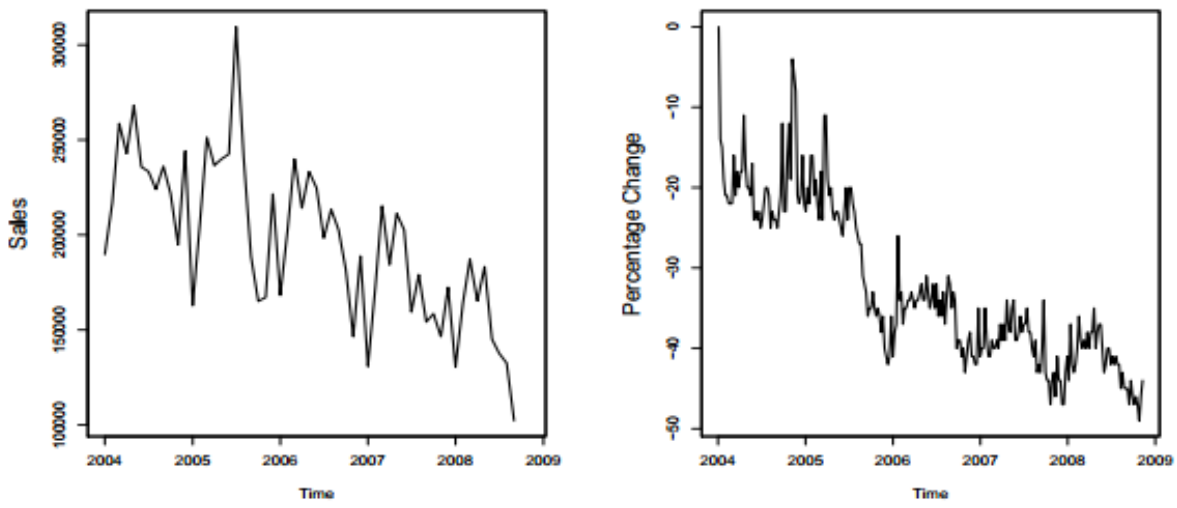
Su hipótesis consiste en que las búsquedas de los usuarios en internet puede estar correlacionado con ciertas preferencias cuantificables, por ejemplo el volumen de ventas de Ford o los precios en el mercado inmobiliario. Se basan en el supuesto que para la toma de ciertas decisiones (comprar un auto, una casa) los usuarios previamente deben investigar acerca del producto que quieren comprar y esto lo realizan a través de internet. Por lo tanto Google Trends sería una herramienta para cuantificar está declaración de preferencias y podría estar relacionado con el resultado final de la toma de decisión.

La principal motivación de los autores para utilizar esta herramienta se debe a la disponibilidad instantánea de datos con la que cuenta el programa. Al tener datos semanales e instantáneos sobre los términos buscados por los usuarios, los autores cuentan con una ventaja para su modelo de predicción ya que los modelos estimativos convencionales trabajan con datos que tienen cierto retraso o *lag*. Por ejemplo el reporte mensual de ventas que publica Automotive News[1] recién está disponible una semana después de terminado el mes. Al realizar un modelo de estimación lineal utilizando información del mes pasado recién se obtienen los datos una semana después de comenzado el mes a analizar. Para los autores, una de las ventajas principales de su modelo se debe justamente a que los datos provistos por Google pueden ayudar a reducir esa brecha ya que están disponibles sin este desfasaje.

La hipótesis de su paper consiste en que las búsquedas de los usuarios al principio del mes pueden ayudar a predecir el resultado de ventas de ese mismo mes cuando se publiquen al mes siguiente. Por lo tanto esta nueva variable puede ayudar a "predecir el presente".

En primer lugar los autores grafican una serie con los volúmenes de ventas mensuales de Ford y lo comparan contra los resultados de búsqueda de la query "Ford", tal como se observa en los gráficos debajo.

[1] . http://www.autonews.com/section/datacenter?cciid=internal-aninside-dcnav#axzz2mtRQcocM



(a) Ford Monthly Sales (b) Ford Google Trends

Cuadro 3. Volumen de ventas Mensuales de Ford (a) y Resultados de Búsqueda de "Ford" (b). Fuente: Varian & Choi (2009)

A simple vista se puede observar que ambas tienen una marcada estacionalidad y una tendencia a la baja en los últimos años. Esto puede deberse al comportamiento pro-cíclico de la industria automotriz, con una fuerte caída de ventas a fines de 2008 coincidiendo con la crisis económica internacional. Pero la intuición de Varian y Choi parece ser correcta al afirmar que las búsquedas para el término "Ford" pueden ser entendidas como una declaración de preferencias y un paso previo a la toma de decisión de comprar un auto Ford. Por eso la naturaleza de la correlación entre las ventas y los resultados de las búsquedas.

En la siguiente sección analizaremos el modelo utilizado por los autores y algunos ejemplos adicionales donde este razonamiento puede ser aplicado.

El Modelo

La metodología que utilizan los autores es bastante simple y se basa en los modelos básicos de predicción: modelos estacionales autoregresivos (Seasonal AR Model) con información de un mes anterior y de doce meses anteriores. Su enfoque consiste en comparar dos modelos y ver cuál de ellos ajusta mejor y tiene menor promedio de error.

En primer lugar tienen el modelo 0 donde utilizan solamente dos variables: las ventas del mes anterior (yt-1) y las ventas del año anterior (yt-12)

Modelo 0: $\log (yt) \approx \log (y_{t-1}) + \log (y_{t-12}) + e_t$;

Donde la variable e_t es el término de error.

Luego agregan al modelo inicial el termino de búsqueda "**Ford**" durante la primer semana del mes. Esta nueva variable la denominan $x_t^{(1)}$

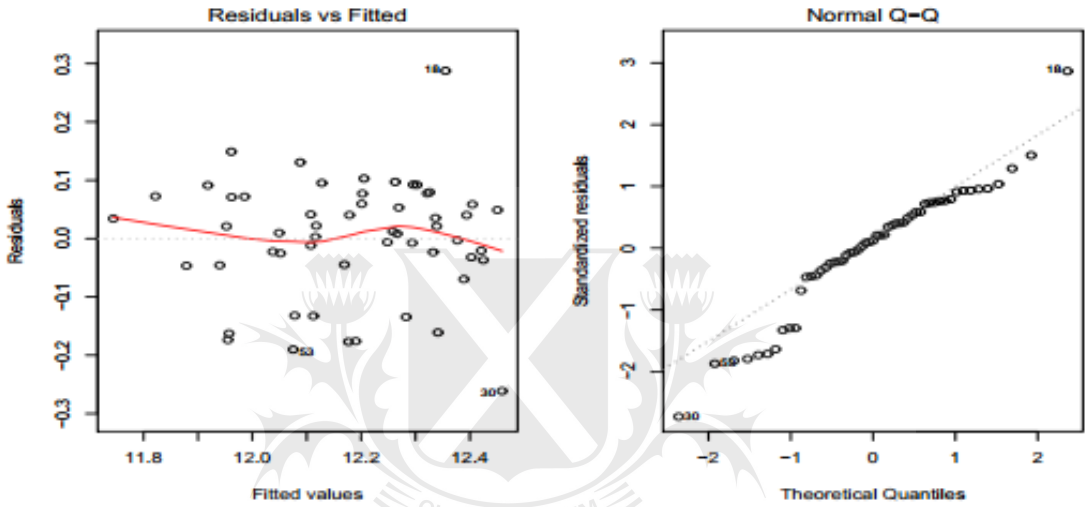
Modelo 1: $\log (yt) \approx \log (y_{t-1}) + \log (y_{t-12}) + x_t^{(1)} + e_t$

El resultado de la regresión del modelo propuesto es el siguiente:

$$\log(y_t) \quad = \quad 2.312 + 0.114 \cdot \log(y_{t-1}) + 0.709 \cdot \log(y_{t-12}) + 0.006 \cdot x_t^{(1)}$$

Donde el coeficiente 0.006 de la variable query: "**Ford**" podemos entenderlo como una elasticidad, un aumento del 1% en las búsquedas por Ford se pueden traducir en un aumento del 0,6% de las ventas.

Abajo se pueden encontrar los resultados de los gráficos de dispersión de las regresiones realizadas en el paper.



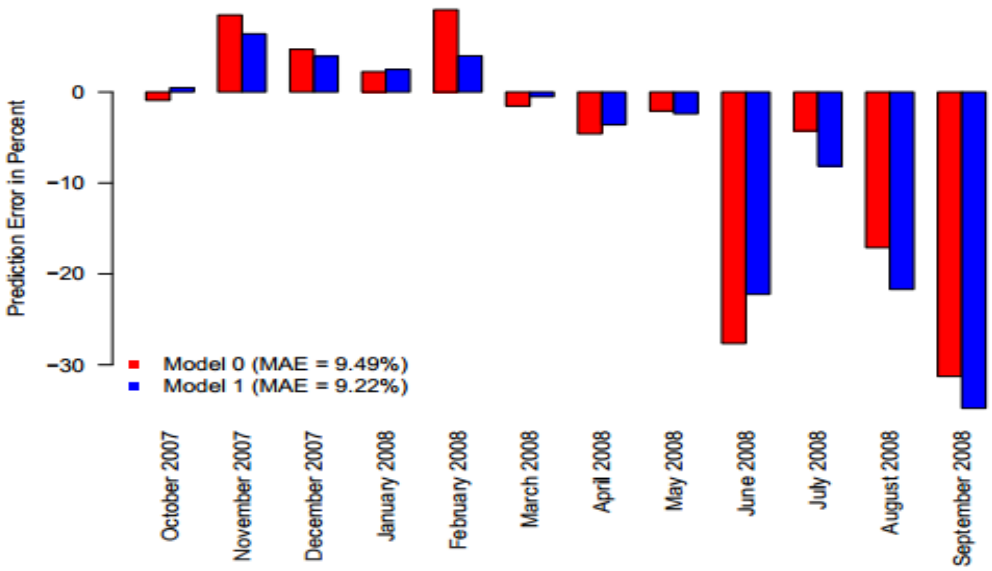
Cuadro 4. A la izquierda el Plot de los Residuos del Modelo vs los estimados. A la derecha grafico Q-Q plot. Fuente: Varian & Choi (2009)

Al obtener un outlier en el dato 18 (julio 2005) Varian y Choi analizan este caso donde las ventas se disparan. Ese mes coincide con un descuento promocional para los empleados que produjo un incremento en las ventas. Para contrarrestar este efecto agregan una variable *dummy* para el mes de julio 2005. El resultado de la nueva regresión es el siguiente:

$$\log(y_t) \quad = \quad 2.007 + 0.105 \cdot \log(y_{t-1}) + 0.737 \cdot \log(y_{t-12}) + 0.005 \cdot x_t^{(1)} + 0.324 \cdot \mathrm{I}(\text{July } 2005).$$

Se puede observar que el coeficiente de correlación del query **Ford** bajó a 0,005 ya que al no considerar la variable *dummy* para el incremento promocional de julio estaban sobreestimando la importancia de las búsquedas de internet. La introducción de la nueva *dummy* también ayuda a reducir el error medio absoluto del modelo de estimación haciendo que ajuste de mejor manera a los resultados de ventas.

Al calcular el error absoluto medio (MAE por sus siglas en ingles) de estimación los autores llegan a la conclusión que este es 3% menor para todo el periodo de estimación para el modelo que incluye la variable de búsqueda de Ford.



Cuadro 5. Gráfico comparativo de error absoluto medio (MAE) del modelo 0 (sin Google Trends) y Modelo 1 (agregando las variables de búsqueda "Ford")
Fuente: Varian & Choi (2009)

En el gráfico 5 observamos en el error de predicción que el modelo 1 ajusta mejor teniendo un MAE menor para la mayoría de los meses.

Automóviles

Continuando con esa metodología los autores realizan otros ejemplos para analizar la eficiencia del modelo propuesto. Utilizan los reportes de ventas mensuales del US Census Bureau [2] que dividen a las industrias en distintas categorías de acuerdo al North American Industry Classification System (NAICS)[3]. En este caso en vez de utilizar el query index de un término específico, utilizan el volumen de búsqueda para una categoría que sea comparable con la categorización del NAICS.

Al igual que en el caso anterior, los reportes de ventas de un mes publicados por el Departamento de Censos de Estados Unidos solamente están disponibles de una a dos semanas después de terminado el mes en cuestión. Aquí Varian & Choi introducen el query index de cada categoría especifica de las primeras dos semanas del mes para correlacionarlo con el nivel de ventas a final del mes para aquella categoría compatible del NAICS.

Para la categoría automotores del NAICS utilizan diversas sub categorías de Google Trends que se correlacionan con las ventas de automotores: Camiones y Camionetas (Trucks & SUV´s), Seguros de autos (Auto Insurance) y Motocicletas (Motorcycles). Proponen dos modelos comparativos, uno que cuente solamente con las búsquedas de la subcategoría Camiones y Camionetas y otro modelo que incluya además de esa sub categoría las de Motocicletas y Seguros para Autos.

[2] http://www.census.gov/retail/
[3] http://www.census.gov/eos/www/naics/

En este ejemplo llegan al mismo resultado que en el análisis anterior ya que los modelos que incluyen los índices de búsqueda por categoría tienen un menor MAE que el modelo clásico basado solamente en el historial de ventas. A su vez, el modelo que incluye más de una subcategoría de Trends ajusta mejor que aquel que incluye solamente la categoría Camiones y Camionetas (Trucks & SUV´s). En este caso podemos observar que a mayor nivel de desagregación de las búsquedas de los usuarios, mayor información vamos a tener sobre sus preferencias de búsquedas al comienzo de mes y esto puede traducirse a una mayor decisión de compra al final del mes.

Mercado Inmobiliario

En el mismo paper realizan un análisis interesante sobre una aplicación del modelo al mercado inmobiliario. Parten del supuesto que la compra o venta de una propiedad requiere una investigación previa con mayor anticipación de lo que nos llevaría la transacción de cualquier otro bien de consumo. Con esto en mente consideran que las búsquedas de Google sobre el mercado inmobiliario pueden ser una buena señal de la intención de comprar o vender propiedades y esta variable puede ayudar a predecir el resultado final al terminar el mes.

Continuando con datos a nivel de categorías los autores analizan las subcategorías que componen la categoría "Real State" (mercado Inmobiliario) y analizan cual es la subcategoría con mayor correlación al nivel de ventas. Como referencia continúan usando las estadísticas del Departamento de Censos de Estados Unidos y el US Department of Housing and Urban Development.[4] Las subcategorías que conforman "Real State" son: Real Estate Agencies (Google Category Id: 96), Rental Listings & Referrals (Id 378), Property Management (Id 425), Home Inspections & Appraisal (Id 463), Home Insurance (Id 465), Home Financing (Id 466).

El modelo que utilizan para comparar es similar al modelo descripto anteriormente pero quitan la variable "12 meses anteriores" para ajustar por estacionalidad, dejando un modelo que correlaciona las ventas del corriente mes con lo sucedido en el mes inmediatamente anterior: y_{t-1}

El modelo estimativo al que llegan los autores está dado por la siguiente ecuación:

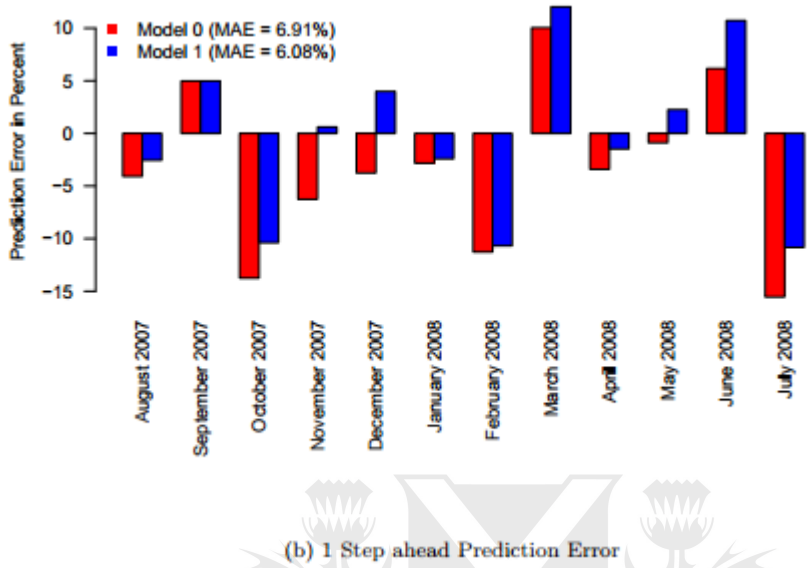
$$\text{Model 1: } \log(y_t) = 5.795 + 0.871 \cdot \log(y_{t-1}) - 0.005 \cdot x^{(1)}_{378,t} + 0.005 x^{(2)}_{96,t} - 0.391 \cdot \text{Avg Price}_t(2.5)$$

La variable $x_{378,\,t}$ $^{(1)}$ representa las búsquedas dentro de la categoría Rental Listings & Referrals, que se refiere a los avisos para ofertar y publicar hogares en venta, que puede entenderse como una *proxy* para la oferta de inmuebles. La variable $X_{96,\,t}$ $^{(2)}$ corresponde a la categoría Real State Agencies que son las búsquedas relacionadas a las inmobiliarias que puede ser interpretada como una señal para demandar un inmueble. Los autores utilizan esta última variable en su modelo predictivo ya que según su investigación es la categoría que mejor estima el nivel de ventas.

Si analizamos los signos de los coeficientes estimados por el modelo podemos inferir cierta lógica en el comportamiento del mercado inmobiliario: las cantidades de búsquedas relacionadas a la publicación y oferta de nuevos inmuebles están negativamente correlacionadas al nivel de ventas. Mientras que un aumento de las *queries* sobre agencias inmobiliarias están positivamente relacionadas con las ventas, indicando un deseo de los usuarios de comprar una propiedad.

[4] http://www.census.gov/construction/nrs/

En un nivel general podemos observar que el modelo propuesto por los autores estima con menor error que los modelos convencionales al obtener un promedio de error menor, tal como indica el grafico debajo:



(b) 1 Step ahead Prediction Error

Cuadro 6. Gráfico con MAE del Modelo 0 (sin Google Trends) y Modelo 1 (agregando la variables de Búsquedas relacionadas a "Real State")
Fuente: Varian & Choi (2009)

Turismo

Bajo el supuesto que internet es usado comúnmente para planear viajes, Varian & Choi utilizan las búsquedas de destinos turísticos para predecir las visitas efectivas a esos lugares.

Los datos de este nuevo modelo provienen del Hong Kong Tourism Board[5] que publica mensualmente las estadísticas de los visitantes por país, lugar de residencia, modo de transporte, modo de entrada y otros criterios. Utilizan datos por país desde enero de 2004 hasta agosto 2008 para realizar su análisis. Agregan también una variable dummy para los Juegos Olimpicos de Beijing desde el 8 al 24 de agosto de 2008 para ajustar las grandes diferencias en las visitas, ya que durante ese período el tráfico a Hong Kong bajó más de lo normal.

La nueva variable que proponen los autores es el número de búsquedas de la query "Hong Kong", que es una subcategoría dentro de Google Trends dentro de "Destinos de Vacaciones". Para esta variable obtienen datos de los siguientes países de origen: Estados Unidos, Gran Bretaña, Canadá, Francia, Italia, Australia, Alemania, Japón e India. En conjunto estos 9 países representan el 19% del total de visitas a Hong Kong durante el período examinado.[6]

El modelo estimativo al que llegan es el siguiente:

[5] http://partnernet.hktourismboard.com
[6] Varian & Choi, pg 16.

$$\log(y_{i,t}) = 2.412 + 0.059 \cdot \log(y_{i,t-1}) + \beta_{i,12} \cdot \log(y_{i,t-12}) \times \text{Country}_i$$
$$+ \delta_i \cdot \text{Beijing} \times \text{Country}_i + 0.001 \cdot x_{i,t}^{(2)} + 0.001 \cdot x_{i,t}^{(3)} + e_{i,t},\ e_{i,t} \sim N(0, 0.09^2)$$

Del cual sacan las siguientes conclusiones:

- Las visitas del mes anterior $(y_{i,t-1})$ y del año anterior $(y_{i,t-12})$ estan positivamente correlacionadas con las visitas del mes corriente.
- Las búsquedas de Google de la query "Hong Kong" en la segunda $x_{i,t}^{(2)}$ y tercer $x_{i,t}^{(3)}$ semana del mes corriente estan positivamente relacionadas con las visitas del mes corriente.
- Durante las Olimpiadas de Beijing las visitas a Hong Kong disminuyeron .

Al analizar las varianzas en la siguiente tabla muestran que la mayoría de la varianza esta explicada por las variables "históricas" (visitas en t-1 y t-12) y que la contribución de la variable Google Trends es estadísticamente significativa.[7]

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
log(y1)	1	234.07	234.07	29,220.86	< 2.2e-16	***
Country	8	5.82	0.73	90.74	< 2.2e-16	***
log(y12)	1	9.02	9.02	1,126.49	< 2.2e-16	***
$x_{i,t}^{(2)}$	1	0.44	0.44	54.34	1.13E-12	***
$x_{i,t}^{(3)}$	1	0.03	0.03	3.87	0.049813	*
Beijing	1	0.41	0.41	51.23	4.53E-12	***
Country:log(y12)	8	0.23	0.03	3.59	0.000504	***
Country:Beijing	8	0.14	0.02	2.12	0.033388	*
Residuals	366	2.93	0.01			

Table 2.2: Estimates from Model (2.6)
Note: Signif. codes: '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.10

Cuadro 7. Tabla con estimados para el modelo con la variable adicional Búsquedas de "Hong-Kong". Fuente: Varian & Choi (2009)

Su modelo estimativo se ajusta con un buen nivel obteniendo un R^2 de 0.9875 y podemos observarlo gráficamente en los siguientes cuadros.

[7] Varian & Choi, pg 16.

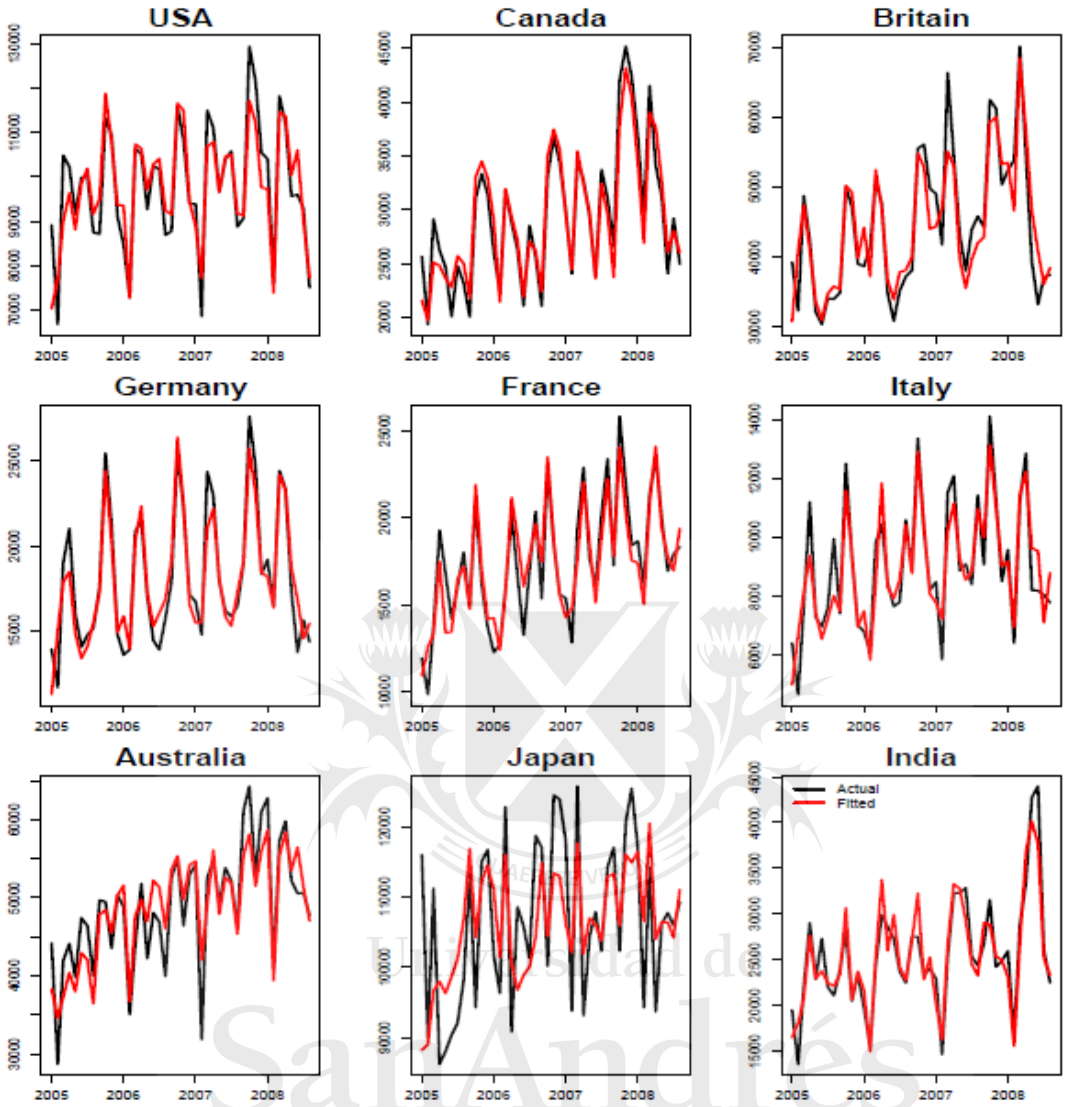


Figure 2.9: Visitors Statistics and Fitted by Country

Cuadro 8. Estadísticas de Visitas y Estimaciónes por País.
Fuente: Varian & Choi (2009)

Lo novedoso de este nuevo enfoque no es suplantar los modelos estacionales autoregresivos clásicos donde solo la historia pesa, sino suplementar estos modelos con una nueva herramienta de recolección de "datos instantáneos". Demostraron que se puede utilizar estos datos para ayudar a predecir comportamientos presentes disminuyendo los errores de estimación. Sin embargo su aporte más valeroso es introducirnos en una nueva herramienta versátil y relevante que nos permite aprovechar la infinidad de datos volcados en internet y canalizarlos como una declaración de preferencias. Esta propuesta a utilizar datos no convencionales y analizarlos para realizar inferencias en el comportamiento microeconómico es su gran aporte ya que abre una puerta a un nuevo método de análisis.

Esta invitación se puede observar en trabajos posteriores que se desprenden de este paper original y buscan aplicar la metodología de Varian & Choi a diversos campos de las ciencias del

comportamiento. Surge entonces una rama de la econometría que se puede entender como la microeconomía de Big Data.

4. Google Flu Trends

En 2004 Johnson, Wagner, Hogan, Chapman, Olszewski, Dowling y Barnas [8] testearon una intuición que ya se venía manifestando en internet: existe cierta correlación entre el comportamiento de los usuarios en internet y la ocurrencia de ciertos eventos. En su caso analizaron la correlación entre el acceso a páginas de salud relacionadas con la gripe y los casos de gripe reportados por el Centro de Control y Prevención de Enfermedades de Estados Unidos (CDC por sus siglas en ingles). El supuesto que buscaban probar los autores era que las búsquedas de información relacionada a la gripe se podían interpretar como una señal de un posible caso de infección. Aunque no toda persona que busca información sobre la gripe está realmente enferma, a nivel agregado existe una correlación entre este comportamiento de búsqueda y los casos detectados.

Su análisis se basó en analizar el Log de entrada de un sitio de internet www.healthlink.com/ para cuantificar el número de usuarios en Estados Unidos que accedieron a artículos relacionados con la gripe. Para esto tomaron 17 artículos y cuantificaron el número de visitas que tuvo cada artículo por semana y armaron así una serie de tiempo. Al correlacionar esa serie de tiempo con el reporte semanal del CDC de visitas a los médicos por causas relacionadas con la gripe, encontraron que existía una correlación positiva entre ambos. A mayor número de visitas a artículos relacionados con la gripe mayor número de casos reportados. Aún con datos acotados a un solo sitio a nivel nacional los autores fueron precursores en introducir y tratar de cuantificar este concepto de correlación entre las búsquedas en internet y los casos de gripe.

Basándose en la idea introducida por el paper anterior, Eysenbach en 2006[9] expandió esta metodología para obtener un canal de recolección de datos más amplio, analizando las palabras buscadas en Google relacionadas con la gripe. Esta nueva disciplina introducida por el autor como *Infodemiology* (Information Epidemiology) buscaba detectar ciertos términos relacionados con enfermedades virales y su correlación con los casos de infección reportados por la Agencia de Salud Pública de Canadá[10]. Al no contar con el total de datos de búsqueda en los servidores de Google, optó por crear una campaña en Google, comprando ciertos términos de búsqueda como "Flu" o "Flu Symptoms". Cuando un usuario realizaba esa búsqueda aparecía un aviso publicitario (también contratado por Eysenbach) que decía "Do you have the flu? Fever, Chest, discomfort, Weakness, Aches, Headache, Cough" y al hacer click llevaba a un sitio de educación para el paciente. De esta forma Google le otorgaba ciertas métricas para su campaña que le permitía cuantificar el número de búsquedas de los términos relacionados con la gripe y la cantidad de clicks en su anuncio publicitario. Con esos datos el autor correlacionó el número de clicks y búsquedas en Google con el número de casos reportados por gripe. Encontró de esa forma una correlación positiva entre las búsquedas de Google y los clicks en el aviso publicitario y los casos reportados según la Agencia de Salud Pública de Canadá. Todos los coeficientes fueron significativos para un nivel de P<.001. La variable clicks tiene una correlación del 0.91 con los

[8] Analysis of Web Access Logs for Surveillance of Influenza, Johnson et al, Medinfo 2004.

[9] 'Infodemiology: tracking flu-related searches on the web for syndromic surveillance', Eysenbach, G. (2006), American Medical Informatics Association.

[10] http://www.phac-aspc.gc.ca/fluwatch/index-eng.php

casos de gripe ocurridos una semana posterior. Para los casos ocurridos esa misma semana el coeficiente baja a 0.88. Sin embargo podemos observar la estrecha relación que existe entre ambas variables, confirmando la suposición de Eysenbach.

Este trabajo inspiró un paper que sería fundamental para el desarrollo de esta nueva disciplina ya que por primera vez se utilizaron datos de búsquedas de Google para predecir brotes de epidémicos.

En noviembre de 2008 un grupo de economistas de Google junto con el Centro de Control y Prevención de Enfermedades de Estados Unidos (CDC) publicaron un paper sobre cómo usar las búsquedas de Google para estimar casos de epidemias gripales en Estados Unidos. Este trabajo llevó al desarrollo por parte de Google de una herramienta sofisticada para estimar y predecir enfermedades, hoy conocido como Google Flu Trends[11].

El trabajo de Ginsberg J, Mohebbi MH, Patel RS, Brilliant L, Smolinski MS, (Google Inc) y Brammer L (Centers for Disease Control and Prevention, CDC) formalizó una metodología que permitía correlacionar las búsquedas de ciertos términos relacionados a enfermedades gripales (Influenza) con el número de casos de visitas al médico debido a esas mismas enfermedades en una región determinada. El tema tomó relevancia a mediados de 2009 con el brote de la llamada Gripe A o Influenza H1N1 que tuvo una propagación mundial.

Lo novedoso de la propuesta de los autores fue que al utilizar las búsquedas de Google para estimar y predecir brotes epidémicos obtenemos información diaria que nos permite realizar un diagnóstico sin los habituales retrasos de las fuentes de información tradicionales. El CDC publica datos semanales sobre el número de visitas médicas relacionadas a enfermedades virales con un retraso de 1 a 2 semanas. [12] Por lo tanto al analizar el comportamiento de los usuarios de búsquedas en Google sobre "salud" podemos mejorar la detección temprana de epidemias y tomar acciones rápidas para prevenir su expansión. Las búsquedas de los usuarios en Google son una señal de las tendencias de salud a nivel mundial, siempre y cuando contemos con un gran número de datos para que sean significativos.

Los autores toman resultados regionales de Estados Unidos sobre búsquedas relacionadas a enfermedades virales y correlacionan los resultados a aquellos obtenidos por el CDC. Su investigación transcurre en una etapa preliminar de Google Trends por lo cual no contaban con las categorización de palabras que hoy nos ofrece la herramienta. Procedieron entonces a seleccionar mediante un proceso de iteración aquellas palabras que mejor correlacionaban con las visitas a los Hospitales por cuestiones virales.

Proceso de autoselección de palabras.

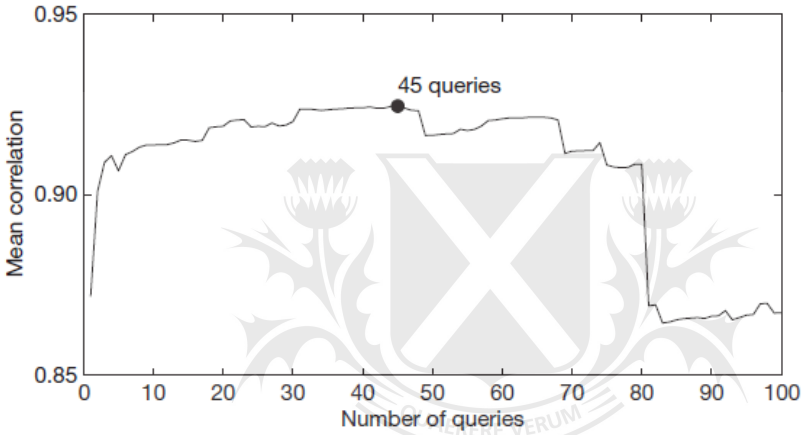
Los autores evaluaron 9 veces cada búsqueda (una por región) utilizando la metodología four cross –validation folds por región, obteniendo 36 correlaciones entre la búsqueda candidata y los porcentajes de visitas relacionadas con la gripe. Luego aplicaron la transformación Z de Fisher a cada correlación y tomaron el promedio de las 36 correlaciones Z transformadas.

[11] http://www.google.org/flutrends/es/#ES

[12] Detecting Influenza epidemics using search engine query data, Ginsberg et al, Nature Mg, Feb 2009.

Seleccionaron aquel conjunto de palabras con la mayor correlación (calculada con el método descripto arriba) y definieron esas búsquedas como búsquedas relacionadas con la gripe[13].

Luego de realizar un ranking de aquellos términos relacionados con la gripe era necesario definir qué cantidad de términos era necesario agregar para obtener la mayor correlación entre el modelo estimado y los resultados del CDC. Para ello utilizaron un modelo con una sola variable definida como "términos relacionados con la gripe" donde agruparon aquellas palabras con mayor correlación. Al regresar sobre 1152 observaciones (128 semanas X 9 regiones) obtuvieron que la mayor correlación se obtiene con 45 términos. Como podemos observar en el gráfico de abajo al exceder los 80 términos la correlación se dispersa debido a la aleatoriedad de los mismos. (Dentro de esa selección de 80 se encuentran términos como "Oscar Nominations" que a simple vista no parecen estar correlacionados con la gripe)



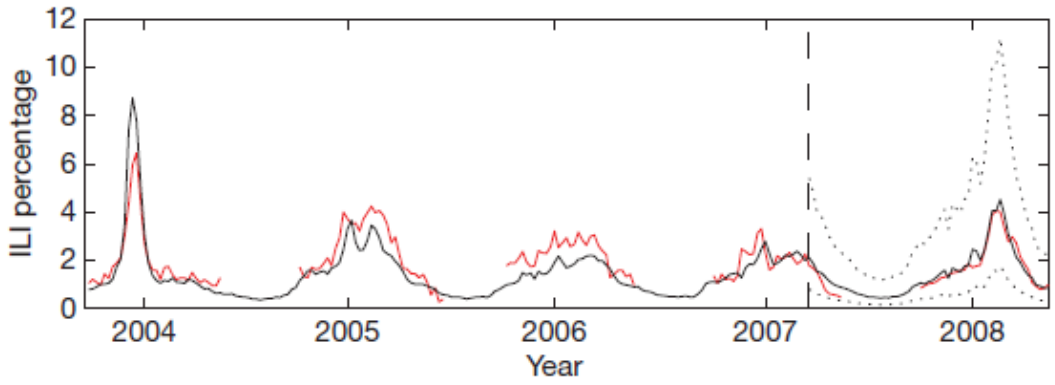
Cuadro 9. Correlación media de acuerdo al número de queries tomadas. El máximo se produce en 45. Fuente: Ginsberg et al (2009)

Utilizando los términos con mayor correlación como variable explicativa estimaron un modelo lineal para los porcentajes de visitas semanales por gripe entre 2003 y 2007. El modelo obtiene buenos resultados de ajuste con una correlación media de 0.90 (min=0.80, max=0.96 en las 9 regiones).[14]

De esta forma, utilizando las búsquedas de Google relacionadas con la gripe pudieron estimar de forma consistente el porcentaje de visitas relacionadas con la gripe 1 a 2 semanas antes de que se obtengan los reportes del CDC.

En el cuadro de abajo podemos observar las estimaciones (en negro) contra los resultados obtenidos por el CDC (en rojo). Las líneas punteadas indican intervalos con predicción al 95%. Las regiones incluyen Nueva York, Nueva Jersey y Pennsylvania.

[13] Idem Ver "Methods"
[14] Idem pg 1013



Cuadro 10. Estimaciónes del modelo en negro vs resultados reportados por CDC, en rojo.
Fuente: Ginsberg et al (2009)

En conclusión las búsquedas de los usuarios en la Web pueden ser utilizadas para estimar las visitas a los médicos por casos de gripe, que puede entenderse como una proxy de los casos detectados. La ventaja es que al obtener y procesar los datos de forma rápida se puede prever 1 a 2 semanas antes del reporte del CDC. Esta prevención es fundamental en el caso de la epidemiología ya que mientras antes se detecte y concientice sobre la expansión de enfermedades contagiosas mejor será la contención de la epidemia.

La limitación del modelo proviene que en caso de un brote epidémico o de un miedo asociado a una epidemia los usuarios pueden entrar en una histeria y realizar más búsquedas relacionadas a esta enfermedad. De esta forma se estarían sobreestimando la expansión porque puede haber muchos usuarios buscando estos términos cuando no presenten síntomas de la enfermedad.

Testeando Google Flu Trends

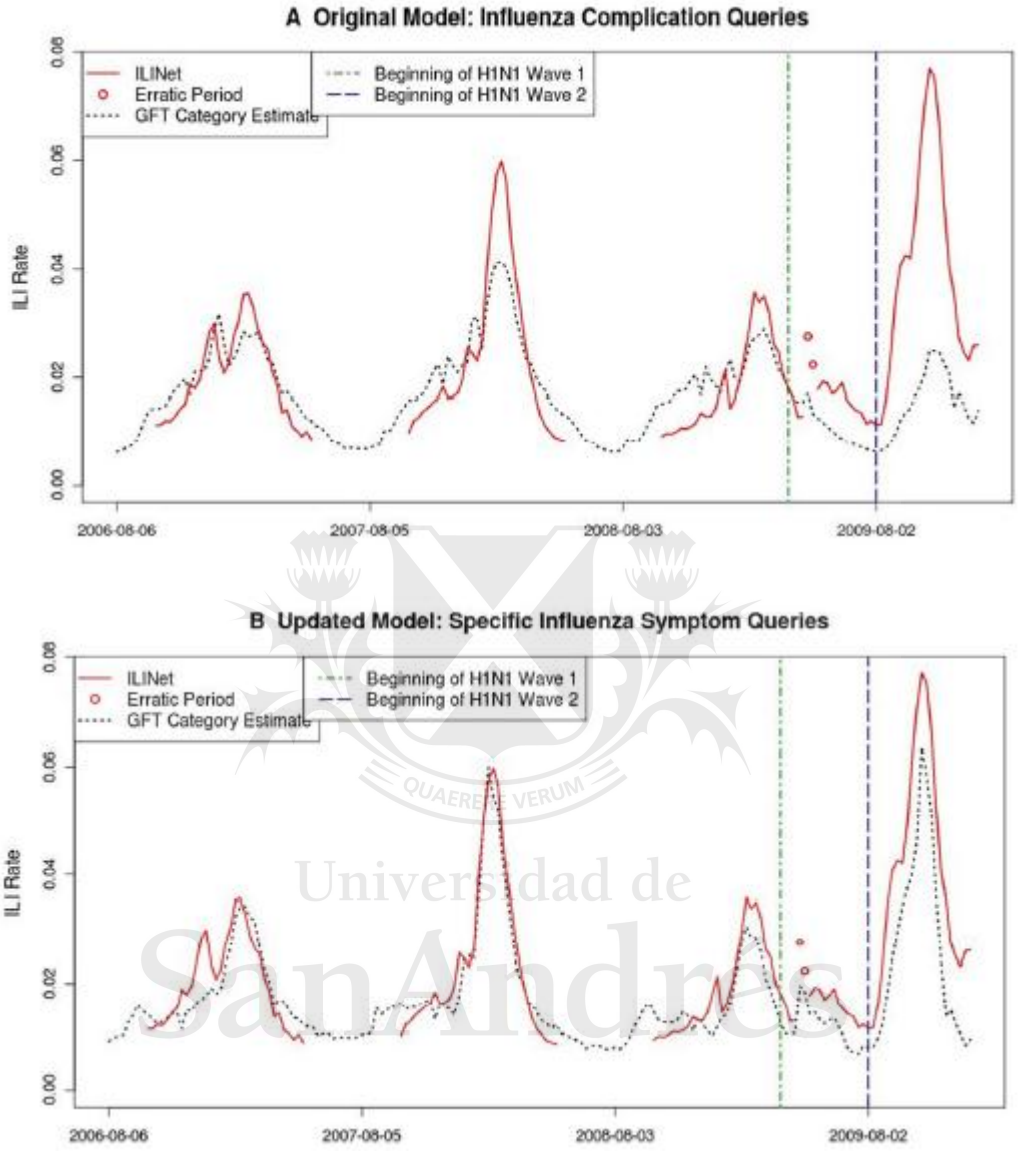
Con el brote epidémico de Gripe H1N1 en el año 2009, fue necesario actualizar el modelo predictivo ya que surgió una nueva cepa del virus y cambió la terminología de búsqueda de los usuarios. En el modelo original se advertía que una de las limitaciones de la herramienta era que si cambiaban las palabras con las que los usuarios buscaban en Google podía llegar a pasar desapercibidos ciertas búsquedas relacionadas con la gripe. Por lo tanto ante un nuevo virus era necesario recalibrar el modelo incorporando los nuevos términos.

Los economistas de Google hicieron esto extendiendo los datos utilizados para clasificar los nuevos términos relacionados con la gripe. Agregaron a su serie de datos originales aquellas búsquedas de los primeros meses de 2009. El modelo actualizado contuvo 160 términos relacionados con la gripe versus los 40 del modelo original. Aunque contiene cuatro veces más términos, estos contienen un cuarto del volumen de búsquedas ya que se incluyeron términos más específicos con la nueva cepa del virus.

Para probar el modelo los autores dividieron en dos los periodos de comparación: una etapa pre epidemia H1N1 (de septiembre de 2003 a marzo de 2009) y la etapa de la epidemia mundial H1N1 (marzo a diciembre de 2009). En el periodo de epidemia separan a su vez dos olas de brote: la primera ola de marzo a agosto y la segunda de agosto a diciembre. Para su análisis no tienen en cuenta la semana del 27 de abril al 3 de mayo ya que en esa semana se disparó la atención de los medios hacia la epidemia y esto incitó a los usuarios a buscar más términos relacionados con el nuevo virus aun cuando no estaban afectados por tal.

Los gráficos muestran como ajustan ambos modelos y podemos observar menor discrepancias con el nuevo modelo actualizado. Esto se debe a que la versión original subestimó la epidemia del 2009 ya que no contó con esa historia para calibrar sus términos de búsqueda por lo tanto estaba omitiendo nuevos términos que surgieron a partir del brote y eran una señal de la expansión de la enfermedad. Con la integración de nuevos términos mejor relacionados al virus H1N1 se obtuvo una mayor correlación y capacidad de predicción de los casos de infección. Durante la primera ola de la epidemia el modelo actualizado tenía una correlación de 0.945 mientras que el modelo anterior solo de 0.29. Ya en la segunda ola el modelo original obtuvo un 0.916 contra un 0,985 del actualizado.

Este nuevo "entrenamiento" al incluir la historia de búsqueda del 2009 mejoró la capacidad predictiva ya que incorpora la evolución de la terminología relacionada con la nueva cepa del virus. Esto nos demuestra que para no volverse obsoleta esta herramienta debe actualizarse periódicamente.



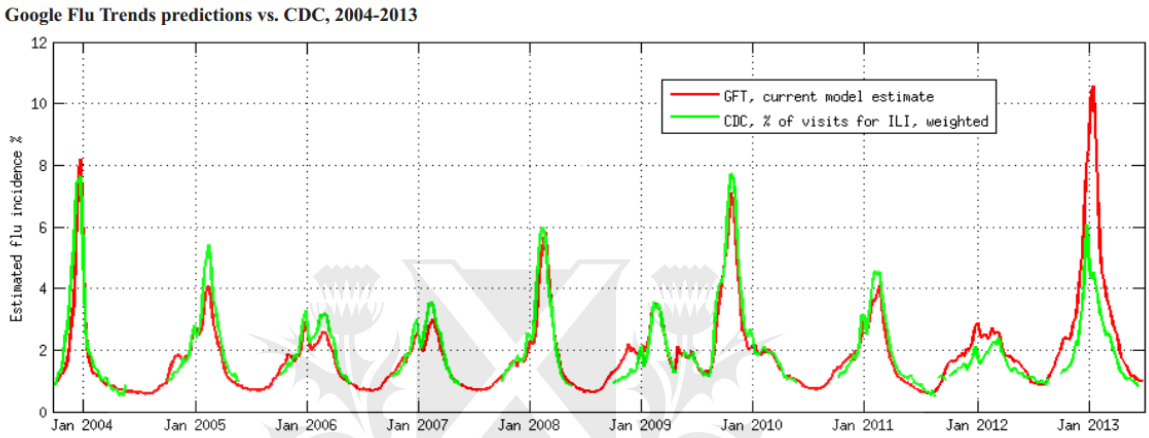
Cuadro 11. Grafico A Estimaciones del modelo original (2009) vs Casos reportados de Gripe. Gráfico B Estimaciones del Modelo Actualizado (2011) vs Casos reportados de Gripe. Fuente: Cook, Conrad, Fowlkes & Mohebbi (2011) Google Inc & CDC.

A principios de 2013 la revista Nature publicó un artículo analizando las predicciones de Google Flu y como esta herramienta sobreestimó el brote de gripe de la temporada 2012 -2013 en Estados Unidos. [15] El equipo de Google publicó un paper para explicar este error de estimación y actualizaron el modelo de acuerdo a las nuevas tendencias de búsqueda relacionadas con la

[15] "When Google got Flu Wrong", **Nature Journal of Science**, 13 Feb 2013. http://www.nature.com/news/when-google-got-flu-wrong-1.12413

gripe. En su trabajo "Google Disease Trends: An update" [16] los investigadores de Google explican la sobreestimación de su modelo con respecto a lo reportado por el CDC.

En el siguiente cuadro podemos ver como se ha ajustado el modelo con respecto a los casos reportados por gripe, desde 2004 a 2013. Desde el ajuste realizado en 2008 hasta la temporada 2012-2013 el error absoluto medio durante todas las estimaciones semanales fue de 0,30 puntos porcentuales. Sin embargo durante la temporada 2012- 2013 la sobreestimación pasó a 6,04 pp. Este desajuste se puede observar en el aumento exponencial que comienza a mediados de 2012 y llega a su cima en enero de 2013.



Cuadro 12. Predicciones Google FLU Trends vs CDC, 2004 a 2013.
Fuente: Copeland, Romano, Zhang, Hecht, Zigmond y Stefansen (Google Inc. 2013).

Al igual que el modelo actualizado en 2009, el algoritmo utilizado para la estimación es susceptible al sesgo cuando se disparan ciertas búsquedas específicas de un término relacionado con la gripe en un período de tiempo muy corto. Desde 2009 el modelo no se actualizó porque las estimaciones se ajustaban a los casos reportados, eliminando aquellos picos de búsqueda fomentados por una alta cobertura mediática. Sin embargo estos picos solo eran detectados si duraban unos pocos días; el modelo no se encontraba preparado para detectar picos de búsqueda sostenidos a lo largo del tiempo que se debían más a una histeria colectiva alimentada por los medios que a casos reales. Por eso los autores introdujeron en su paper de 2013 lo que ellos llaman "spike detectors" para identificar picos de búsquedas sostenidos que no están relacionados con posibles casos de infección sino alentados por la histeria colectiva. De esta forma el sistema recibe series de tiempo de búsquedas relacionadas con la gripe y valida si las últimas mediciones están dentro de las expectativas del modelo de acuerdo a la varianza histórica. Se calibra el modelo constantemente eliminando aquellos términos que están más relacionados a la cobertura mediática e introduciendo términos más específicos que utilizarían las personas que potencialmente pueden haber contraído la gripe.

Su conclusión es que para que el modelo ajuste correctamente con los casos reportados es necesario re-calibrar cada temporada aquellos términos clasificados como "ILI-Related" (relacionados con la gripe), eliminando aquellos que se hayan hecho populares por una mayor

[16] "Google Disease Trends: an Update", Copeland, Romano Et Al. Google Inc. 2013. http://research.google.com/pubs/pub41763.html

cobertura en los medios e introducir nuevos términos que tengan más relación con el tipo de gripe en esa temporada.

Al igual que Varian & Choi el mayor logro de este trabajo no son los resultados a los que llegan sino la intuición e inspiración que incorporan al análisis micro. Se trata de utilizar una nueva variable con datos casi instantáneos para suplementar los métodos de estimación tradicionales basados en la historia, a modo de ahorrar el tiempo que toman estos métodos en recolectar y analizar su información.

5. Aplicaciones Económicas

Inspirados en el trabajo original de Ginsberg et Al 2009[17] Askitas & Zimmerman buscaron demostrar que la información proveniente de Google puede ser utilizada para predecir comportamientos económicos y compararlo con las fuentes de medición tradicionales. En su paper "Google Econometrics and Unemployment Forecasting"[18] los autores correlacionan ciertas búsquedas de Google en Alemania con el índice de desempleo reportado mensualmente por la Agencia Federal de Empleo de Alemania.[19] Este índice de desempleo se reporta a final de cada mes y se utilizan datos recolectados en la última quincena del mes anterior y la primera quincena del mes corriente para reportar el índice de desempleo del mes actual. Al igual que en el trabajo de Varian & Choi, al utilizar datos semanales de búsquedas en Google se evita este *lag* al utilizar datos casi instantáneos.

Para evitar el desfasaje con los datos tradicionales reportados por la agencia de desempleo los autores compilan las búsquedas semanales en dos quincenas y correlacionan las búsquedas de la segunda quincena del mes anterior y la primera del mes corriente con los datos oficiales reportados a final del mes. Además al separarlo por quincenas los autores buscan evaluar que quincena de datos ajusta mejor con el desempleo ya que las búsquedas realizadas en las primeras dos semanas del mes ($W12_{M}$) van a estar influenciadas por el índice anunciado del mes anterior (UM_{-1}) que se realiza a inicios del mes en curso. Las búsquedas de las últimas dos semanas del mes anterior ($W34_{M-1}$) se ven menos impactadas por este anuncio ya que se ven afectadas solamente por el anuncio del desempleo de dos meses atrás (UM_{-2})

Para esto eligen cuatro términos relacionados con el desempleo y testean cuál de ellos se ajusta mejor al índice reportado al final del mes. Estos cuatro términos de búsqueda son:

1. "Arbeitsamt" o "Arbeitsagentur" (Oficina o Agencia de Desempleo)
2. "Arbeitslosenquote" (Índice de desempleo)
3. "Personalberater" o "Personalberatung" (Consultor Personal)
4. "Stepstone" o "Jobworld" o "Jobscout" o "Meinestadt" o "meine Stadt" o "Monster Jobs" o "Monster de" "JobboerseStepsone" (Portales de empleo más populares de Alemania)

El primer término (K1) está relacionado a personas que buscan contactarse o han contactado con la oficina de desempleo, siendo una señal de ingreso al desempleo. El segundo término (K2) hace referencia directa al desempleo. Con este tercer término K3 los autores buscan la señal de aquellos trabajadores altamente calificados que temen despidos y compañías que buscan prepararse también para despidos y restructuraciones de personal. Con el cuarto término (k4) buscan un indicador de las actividades relacionadas con la búsqueda de empleo y tomar esto como una señal de una salida del desempleo.

Luego de seleccionar los términos que para ellos serán buenos indicadores del desempleo, construyen una serie temporal con datos bisemanales de 2004 hasta 2009 utilizando ECM (Error

[17] Detecting Influenza epidemics using search engine query data, Ginsberg et al, Nature Mg, Feb 2009
[18] Google Econometrics and Unemployment Forecasting, Askitas & Zimmerman, IZA DP N° 4201, Jun 2009
[19] http://statistik.arbeitsagentur.de/

Correction Model) para realizar las correlaciones entre las búsquedas de los usuarios y el índice de desempleo oficial. Para cada quincena (W34 $_{M-1}$; W12 $_{M)}$ estiman dos modelos: uno con dos variables explicativas (dos términos de búsqueda) y otro modelo con múltiples variables (más de dos términos de búsqueda). Al regresar estos modelos utilizan el Criterio de Información Bayesiano (BIC) para seleccionar aquel modelo que mejor ajuste con el índice de desempleo oficial. Como se puede observar en la tabla de abajo, el modelo que logra esto es aquel donde se utilizan los datos de búsqueda para los términos k1 y k4 en las últimas dos semanas del mes anterior al estudiado (W34 $_{M-1)}$.

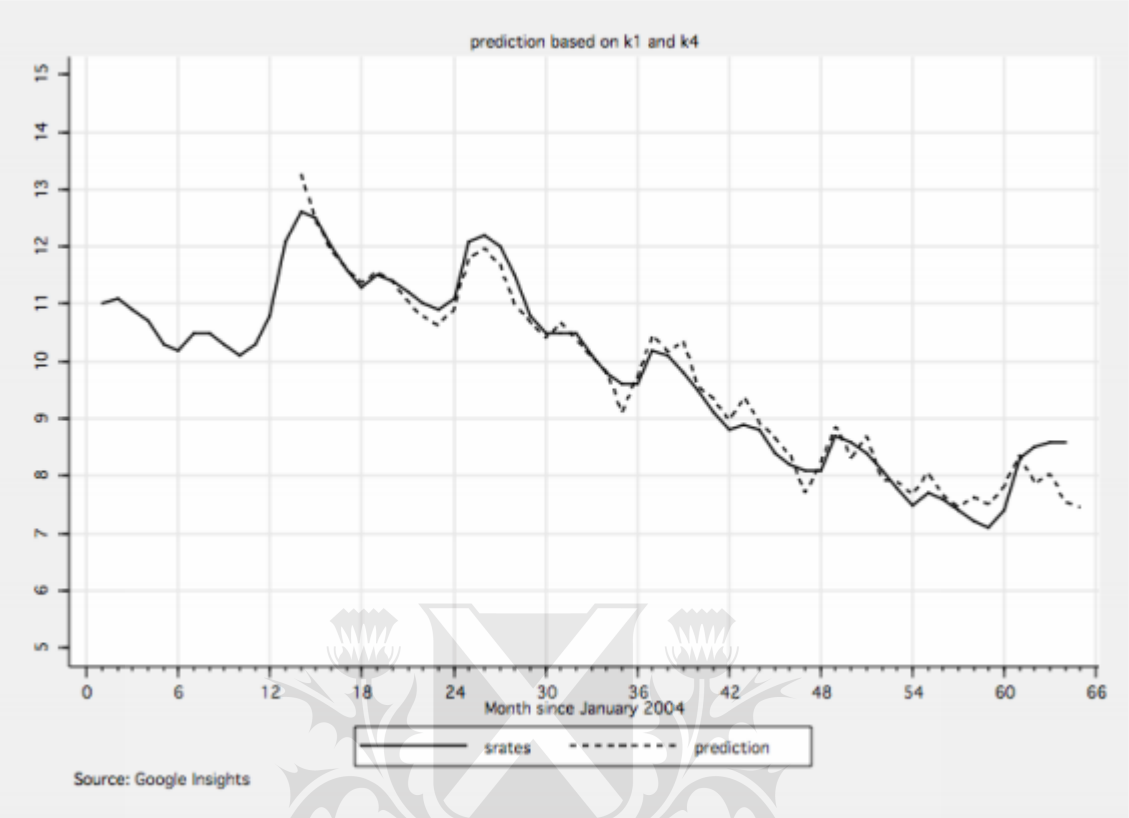
Table 3. Models with two variables involving activity in weeks 3,4

	w34k1_2 b/t	w34k1_3 b/t	w34k1_4 b/t	w34k2_3 b/t	w34k2_4 b/t	w34k3_4 b/t
L12.srates	-0.156 (-1.95)	-0.378*** (-4.01)	-0.406*** (-6.65)	-0.414*** (-4.51)	-0.429*** (-4.95)	-0.455*** (-6.81)
L13.w34k1	0.000 (0.04)	-0.021* (-2.11)	0.024*** (6.65)			
LS12.w34k1	-0.045*** (-4.80)	-0.039*** (-4.08)	0.016* (2.34)			
L13.w34k2	0.014** (2.95)			0.004 (0.75)	0.012*** (3.66)	
LS12.w34k2	0.032*** (4.71)			0.021* (2.48)	0.013* (2.45)	
L13.w34k3		0.031*** (4.32)		0.020*** (3.58)		0.017*** (6.41)
LS12.w34k3		0.017 (1.58)		-0.010 (-0.83)		0.009 (1.50)
L13.w34k4			-0.031*** (-11.41)		-0.028*** (-10.72)	-0.025*** (-11.79)
LS12.w34k4			-0.014** (-3.46)		-0.010* (-2.19)	-0.013** (-3.40)
_cons	-0.918 (-1.06)	1.600 (1.41)	5.503*** (6.98)	0.482 (0.58)	6.911*** (6.61)	5.977*** (7.41)
N	51	51	51	51	51	51
AIC	102.860	103.445	40.470	112.394	72.194	46.150
BIC	114.451	115.036	52.061	123.985	83.785	57.741
Log Lik.	-45.430	-45.722	-14.235	-50.197	-30.097	-17.075
R^2	0.692	0.688	0.909	0.628	0.831	0.899

Ln and Sn are the nth monthly lag and difference operators respectively. . The variable naming convention is as follows: w12=first monthly half, w34=second monthly half; k1, k2, k3, k4 are the keywords defined in Section 2. A model which is denoted by eg w12ki_j is one involving the two activity variables in the first monthly halves i and j whereas w34ki_j_l is a model with 3 keywords i, j and l in the second monthly halves. The variable srates is the seasonal unemployment rates. Finally the significance stars mean: * p<0.05; ** p<0.01; *** p<0.001

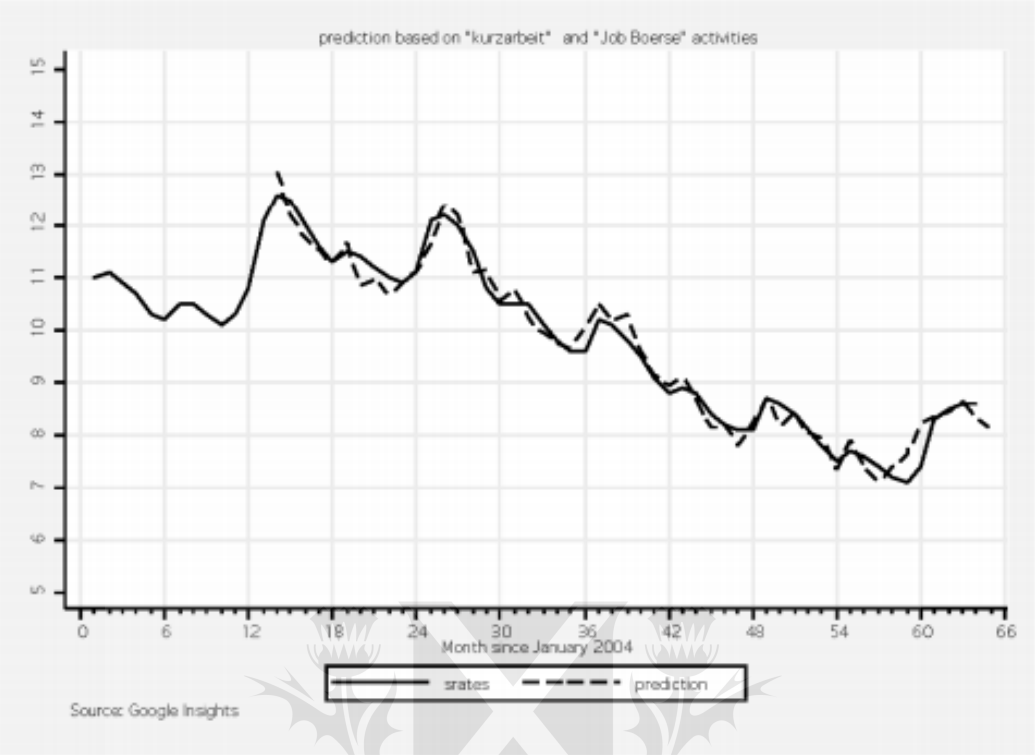
Cuadro 13. Estadísticos para modelos con dos Variables (dos términos de búsqueda)
Fuente: Askitas & Zimmerman (2009)

También podemos observar que los coeficientes para el k1 ("Agencia u Oficina de Desempleo") son positivos, confirmando la relación positiva que intuían los autores tomando este tipo de búsqueda como una señal de entrada al desempleo. En cambio los coeficientes para k4 (Portales de Empleos de Alemania) son negativos, siendo la actividad de búsqueda de trabajo un indicio de una salida del desempleo. En el cuadro de abajo se puede observar gráficamente como el modelo estimado se ajusta o "predice" el índice de desempleo.



Cuadro 14. Predicciones vs Índice de Desempleo en Alemania.
Fuente: Askitas & Zimmerman (2009)

En líneas generales la predicción acompaña a la tendencia del desempleo observándose una discrepancia solo al final de diciembre de 2008 (Mes 59-60). La explicación que ofrecen los autores a este aumento del desempleo que el modelo no anticipó se debe a que en diciembre de 2008 el gobierno anunció la implementación de un beneficio para la jornada reducida (short time working) pasando el período de 6 a 18 y finalmente 24 meses, reduciendo los costos laborales para esta modalidad. Por lo tanto el creciente interés por este empleo con una jornada de menor duración que el modelo no captaba puede haber contribuido a una baja en el desempleo. Para testear esta hipótesis reemplazaron el primer término de su modelo (k1) por el término "Kurzarbeit" (Short-time work, Jornada Reducida). Los resultados que obtuvieron se pueden observar en la siguiente serie.



Cuadro 15. Predicción del Modelo con *queries* "Kurzarbeit" (trabajo a corto plazo) y "JobBoerse" (portal de búsquedas laborales) vs Índice de Desempleo.
Fuente: Askitas & Zimmerman (2009)

Se puede notar que la discrepancia a final de enero de 2009 desaparece y el modelo parece ajustar mejor, sin embargo los autores notan la limitación que puede tener este arreglo en el largo plazo ya que la política de jornada laboral reducida puede quedar obsoleta con el correr del tiempo por lo tanto dejaría de ser un buen indicador.

En conclusión los autores prueban estadísticamente que hay una correlación entre el comportamiento de búsqueda de los usuarios y un fenómeno económico medible tradicionalmente como el desempleo. Al utilizar esta nueva modalidad de información tan masiva e instantánea se logra complementar los modelos predictivos con datos que el análisis económico tradicional no ha tenido en cuenta hasta el momento. Tener en cuenta esta nueva metodología podría ser sumamente útil en momentos de crisis cuando los flujos tradicionales de información son demasiado lentos para tomar una decisión rápida y coherente.

Predicting Initial Claims for Unemployment Benefits

Continuando con la metodología introducida en su paper original, Varian & Choi también utilizan la nueva herramienta de Google para correlacionar las búsquedas de trabajo en el servidor con el reporte de solicitudes de beneficios de desempleo (Initial Jobless Claims) que realiza el US Department of Labor semanalmente[20]. Este reporte mide la cantidad de personas que han

[20] http://www.dol.gov/opa/media/press/eta/ui/current.htm

aplicado para beneficios de desempleo y es considerado un buen indicador del mercado laboral[21]. Las solicitudes semanales se publican a nivel nacional y estatal 5 días después del fin de esa semana. Al igual que Askitas & Zimmerman (2009) los autores ven la oportunidad de utilizar las búsquedas de ciertas categorías de Google Trends relacionadas con el desempleo para obtener un indicador que estime estas solicitudes de desempleados de manera inmediata, ahorrando así la semana de lag que tardan las fuentes de información oficiales en reportar.

Para realizar esto utilizan las series de tiempo de Google Trends "Jobs" & "Welfare & Unemployment". Estas series indican la cantidad de *query share* que tiene cada categoría de búsqueda, siendo "Jobs" la categoría que agrupa aquellos términos relacionados con la búsqueda de trabajo, y "Welfare & Unemployment" la categoría relacionada con el desempleo y las solicitudes de subsidios, beneficios para desempleados.

Como los Estados Unidos entró en recesión en diciembre de 2007, de acuerdo al NBER[22], los autores prueban sus modelos con una serie de datos a largo plazo y otra a corto plazo. En el largo plazo utilizan una serie de datos desde 2004 bajo el supuesto que no hubo un cambio estructural en los últimos 5 años. En el corto plazo solamente utilizan datos de diciembre de 2007 hasta julio 2009 tomando solo el periodo de recesión asumiendo que existen diferencias estructurales y hay estacionalidad en el mercado de trabajo.[23]

Utilizan ARIMA y AR (1) para sus regresiones armando dos modelos:

Modelo de Base: $\log(y_t) = \text{Intercept} + \phi \log(y_{t-1}) + e_t$ donde y_t representa la serie de tiempo para las "Initial Claims".

Modelo Alternativo: $\log(y_t) = \text{Intercept} + \alpha \text{Jobs}_t + \beta \text{Welfare}_t + \phi \log(y_{t-1}) + e_t$ se agregan Jobs_t y Welfare_t que representan las series de tiempo para las categorías de Google Trends.

Lo que intentan probar es si al introducir las búsquedas de los usuarios el modelo autoregresivo mejora la predicción. En la siguiente tabla se muestran los resultados de las regresiones y podemos observar que al introducir las series de Google Trends el modelo mejora significativamente. Esto se evidencia en la reducción del error absoluto medio (MAE) tanto para las series de largo plazo como de corto. En la serie de largo plazo el MAE se reduce de 3,24% a 2,73% mejorando en un 15,74%. En la serie de corto plazo pasa de 3,10% a 2,7%, una disminución de 12,9%. Además los coeficientes obtenidos tanto para "Jobs" como para "Welfare" son positivos, indicando que a medida que aumentan las búsquedas de Google de términos relacionados al desempleo, también aumentan las solicitudes para beneficios por desempleo.

[21] Predicting Initial Claims for Unemployment Benefits, Varian & Choi, Julio 2009 pg 1.

[22] http://www.nber.org/cycles/recessions.html

[23] Varian & Choi 2009, pg 2

		Baseline Model				Alternative Model					
		Intercept	ϕ	σ	MAE	Intercept	ϕ	Jobs	Welfare	σ	MAE
LT	Est	0.1269	0.9902	0.0443	3.24%	1.6498	0.8727		0.0014	0.0429	2.73%
	SE	0.1618	0.0126			0.3754	0.029		0.0003		
ST	Est	0.2174	0.9839	0.0432	3.10%	1.792	0.8632	0.0014	0.0010	0.0398	2.70%
	SE	0.2632	0.0202			0.5541	0.0427	0.0006	0.0004		

Table 2: Summary of Long Term Models and Short Term Models

Cuadro 16. Resumen de Estadísticos Modelos a corto y largo plazo.
Fuente: Varian & Choi (2009) "Predicting Initial Claims for Unemployment Benefits".

En conclusión, Varian & Choi comprueban nuevamente que al introducir esta nueva variable de Google Trends a los análisis tradicionales económicos se complementan los modelos predictivos obteniendo predicciones con menor error.

How Google Search Forecast Housing Prices. Wu & Brynjolfsonn (2014)

En este paper preliminar los autores buscan probar si las búsquedas de ciertos términos en Google pueden ayudar a predecir las ventas en el mercado inmobiliario. Su intuición consiste en que antes de realizar una compra o venta de un inmueble el público necesita investigar a través de internet, siendo este comportamiento una señal de una futura transacción en el mercado inmobiliario.

Continuando con la metodología propuesta por Varian & Choi, proponen complementar los modelos predictivos tradicionales con los términos de búsquedas relacionados a inmuebles que realizan los usuarios en Google. Para esto utilizan búsquedas de Google a nivel estatal de Estados Unidos desde el 1Q de 2004 al 3Q del 2011 y lo correlacionan con el House Price Index (HPI) que compila la Office of Federal Housing Enterprise Oversight[24] y el reporte de ventas trimestrales de la National Association of Realtors[25], utilizando un modelo AR (1). En primer lugar buscan seleccionar cuales son los categorías de búsquedas dentro de Google con mayor correlación con las ventas y utilizan el periodo del 1Q 2006 al 4Q 2008 para encontrar estos términos. Como criterio de selección utilizan aquellas categorías de búsqueda cuya predicción del nivel de ventas contenga el menor MAE. Encuentran dos categorías que ajustan bien con la serie histórica de ventas del mercado inmobiliario; estas son "Real State Listings" y "Real State Agency."

Una vez seleccionadas las categorías de Google Trends con las cuales realizar las predicciones proponen analizar la correlación entre las búsquedas pasadas con el nivel de ventas presentes y luego intentan predecir el nivel de ventas futuras.

En una primera prueba comparan el modelo de predicción original que tiene en cuenta solamente las ventas pasadas contra el modelo que agrega al modelo base los ratios de búsquedas relacionadas con el mercado inmobiliario (dentro de las categorías "Real State Listings" y "Real State Agency."). En el gráfico debajo (figure 3) se muestran las diferencias del MAE entre el modelo alternativo y el modelo base. Cuando los valores están por encima de cero quiere decir que el MAE del modelo base es mayor que aquel del modelo complementado por el historial de

[24] www.ofheo.gov
[25] http://www.realtor.org/research-and-statistics

búsquedas. En promedio el MAE disminuye en un 2.3% (0.174 vs 0.170) al introducir estas nuevas variables.

Luego testean si las búsquedas pueden complementar la predicción de las ventas futuras. En este caso la historia que nos proporciona la búsqueda complementa mejor el análisis ya que los datos tradicionales (ventas pasadas) para predecir ventas futuras tienen un *lag* de dos cuatrimestres, mientras que los resultados de Google están disponibles al instante. La mejora en la predicción es mayor, obteniendo una reducción de un 7.1% del MAE en comparación con el modelo base. (0.172 vs 0.185). La intuición de los autores respecto a esta mejora se debe a que la decisión de comprar o vender un inmueble generalmente toma más de un trimestre por lo tanto es esperable que las búsquedas realizadas hace varios meses recién se relacionen con una venta en el presente. Además puede que las ventas en el trimestre anterior estén correlacionadas con las ventas presentes pero esa tendencia puede irse diluyendo cuando se compara con ventas futuras, dos trimestres por delante. Por estas razones es útil y valedero agregar al modelo de predicción tradicional las búsquedas relacionadas con el mercado inmobiliario tanto pasadas como presentes.

Figure 3: Y-axis indicates the average difference in MAE between the baseline model (Equation 1) and the model that uses search indices (Equation 2). We use predictions from the first quarter of 2009 to the third quarter of 2011. When the dots are above the zero line, the baseline MAE is worse than the MAE from the model that uses search.

Cuadro 17. Diferencia promedio en MAE entre el modelo de base y modelo alternativo con índices de búsqueda. Cuando los puntos están por arriba de cero el MAE del modelo base es peor que el MAE del modelo propuesto.

Fuente: Wu & Brynjolfsonn (2014)

Para reforzar su modelo los autores también lo comparan contra un modelo de forecast que es utilizado y consultado por los expertos en el mercado inmobiliario publicado. El National

Association of Realtors (NAR) publica trimestralmente un forecast sobre las ventas de inmuebles a nivel nacional para Estados Unidos. Wu & Brynjolfsonn comparan este forecast con sus predicciones desde el 2Q de 2009 al 3Q de 2011 tomando un total de 10 trimestres. Aunque sus predicciones sobre el nivel de ventas presentes son mejores que las del NAR la diferencia no es estadísticamente significativa. Sin embargo sus predicciones sobre el nivel de ventas futuras si lo son. En este caso el MAE para el NAR es de 0.110 mientras que el de su modelo con índices de búsqueda es de 0.084, una mejora del 23.6%.

Finalmente los autores extienden su análisis para el mercado de artículos para el hogar, bajo la intuición de que una venta en el mercado inmobiliario lleva a una demanda futura de artículos domésticos. Por lo tanto puede existir una relación entre las ventas de casas y las búsquedas en internet de artículos para el hogar. Para demostrar esto invierten la variable explicativa, siendo en este caso las ventas de casas las que llevarían a un aumento en los índices de búsqueda de artículos para el hogar. En la tabla de abajo (cuadro 18) se pueden observar los resultados de las correlaciones entre ambas variables. En la columna 1 y 2 vemos que las correlaciones no son significativas mientras que en la columna 3 si lo son. Si entendemos que la compra de artículos para el hogar no se realiza inmediatamente después de la compra de la casa entonces es esperable que exista un *lag* entre la compra y la búsqueda de artículos para el hogar.

Table 5. Linear Regression on Search Terms Related to Home Appliances and the Volume of Housing Sales				
Dependent Var. Search Terms related to Home Appliances	Search Terms on Home Appliances (quarterly)	Search Terms on Home Appliances (quarterly)	Search Terms on Home Appliances (quarterly)	Search Terms on Home Appliances (quarterly)
	(1)	(2)	(3)	(4)
	Fixed effect	Fixed effect	Fixed effect	Fixed effect
Home Sale $_t$	-.054 (.0001)			0.188 (0.0004)
Home Sale $_{t-1}$		-.020 (.0001)		-0.627 (0.393)
Home Sale $_{t-2}$			.590** (.3)	1.140*** (0.427)
Obs.	254	203	152	152
Controls	Quarters	Quarter	Quarters	Quarters
States	51	51	51	51
*p<.1, **p<.05, ***p<.001, Huber-White robust standard errors are shown in parentheses				

Cuadro 18. Regresiones para términos de búsqueda relacionados con productos del hogar y el volumen de ventas de inmuebles.
Fuente: Wu & Brynjolfsonn (2014)

En conclusión los autores demuestran que al introducir el comportamiento de búsqueda de los usuarios en Google como variable explicativa, se puede complementar los modelos predictivos mejorando sus errores de estimación.

6. Una Posible aplicación local: "DÓLAR BLUE"

En la siguiente sección analizamos si la metodología propuesta por Varian & Choi podría resultar útil para una aplicación local para la Argentina. Al igual que en varios países Google Trends cuenta con varias categorías formadas por determinados términos de búsqueda relacionados a un tema en particular. En Argentina adicionalmente está dividida la región por provincias permitiendo desagregar el interés de los usuarios a nivel provincial. Una pregunta relevante en nuestro contexto económico es si Google Trends puede ser usado para representar el interés de los usuarios por un tema principal de la economía argentina: el mercado paralelo de divisas. ¿Puede utilizarse entonces el comportamiento de búsquedas de Google para predecir por ejemplo el valor del dólar paralelo?

Al tratarse justamente de una cotización paralela y sin contar con un sitio oficial muchas veces los usuarios necesitan utilizar el buscador como un canal para encontrar la cotización promedio del dólar que se opera por afuera del mercado formal de divisas. Tomando esta búsqueda como una declaración de preferencias hacia la compra o venta de dólares podemos testear si esta señal se condice con el precio de la moneda en el mercado paralelo. Esto representa una oportunidad de captar y cuantificar el interés de los personas por las divisas extranjeras ayudando así a predecir cuál será su cotización en el corto plazo.

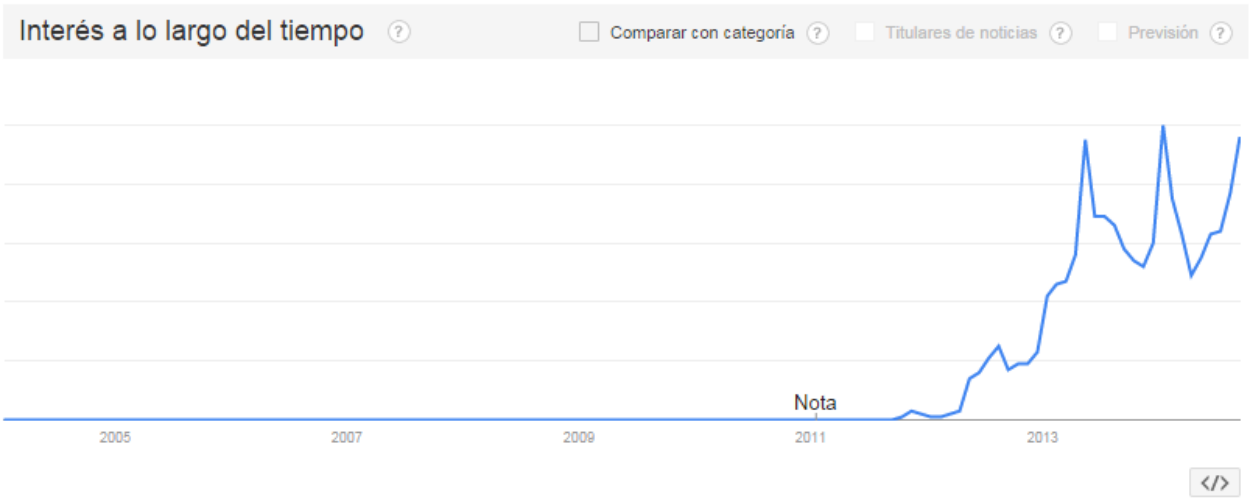
Otra ventaja adicional de la propuesta se trata de que la herramienta ya cuente con una categoría de búsqueda que contiene términos relacionados a las divisas extranjeras, la categoría "Divisas y Cambio de Moneda Extranjera". Sin embargo esta categoría no contiene un mayor grado de desegregación y no puede atribuirse los términos relacionados a cada moneda en particular. Por lo tanto sería un ejercicio interesante buscar las palabras relacionadas a la búsqueda de cotización del dólar en el mercado informal, por ejemplo "dólar blue", "dólar paralelo" etc. El índice de búsquedas de estos términos puede ayudarnos a explicar la volatilidad del tipo de cambio no oficial y predecir cuál sería su valor en el corto plazo.

Tomando la cotización del Dólar informal que publica Ámbito Financiero diariamente[26] podemos construir una serie histórica de la cotización para analizar si se correlaciona con los resultados de búsquedas en Google de la categoría "Divisas y Moneda Extranjera , el término "Dólar" y el término "Dólar Blue". Los datos relevantes tomados constan de octubre de 2011 hasta septiembre 2015. Aunque Google cuenta con información histórica desde el 2004, el término "dólar blue" tomó relevancia a partir de octubre de 2011 luego de la introducción del cepo en el mercado cambiario. Esto produjo un desdoblamiento del tipo de cambio generando una brecha entre el mercado formal e informal de moneda extranjera.

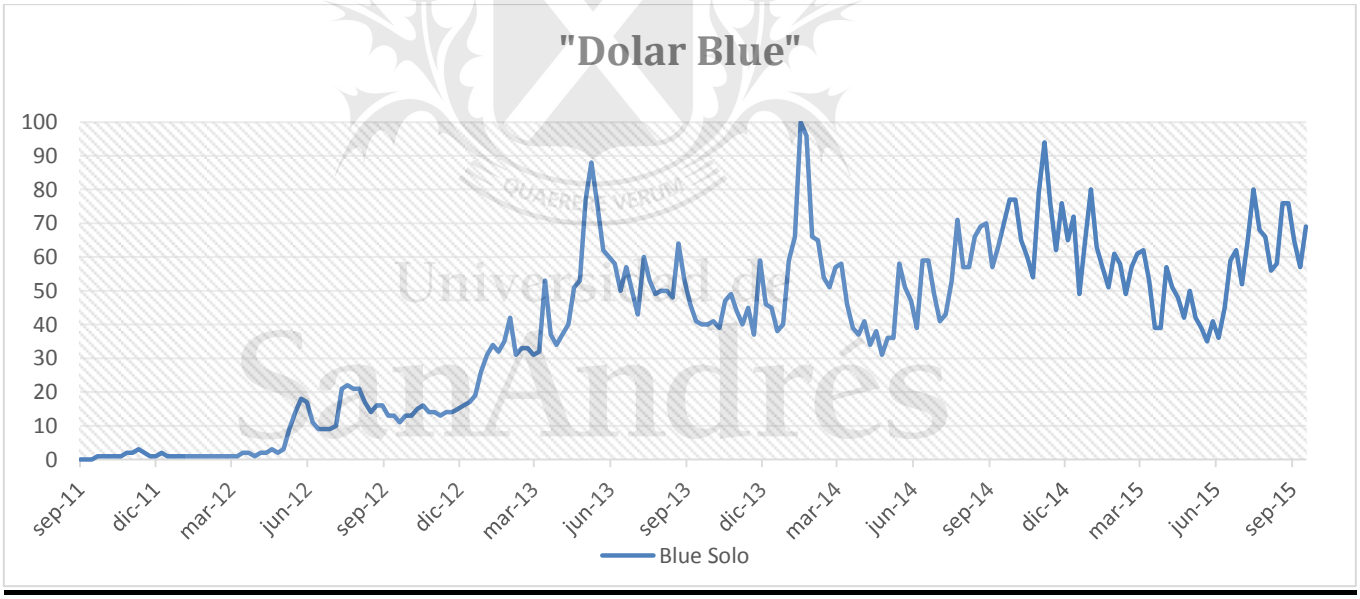
En el cuadro 19 podemos observar el nivel de búsqueda por el término "Dólar Blue" y vemos que el interés se dispara en octubre de 2011 y de ahí incrementa exponencialmente. A partir de entonces el valor de esta moneda informal tomó mayor relevancia para la mayoría del público ya que prácticamente se imposibilito el acceso al Dólar a la cotización oficial.

En los cuadros 20 y 21 se puede observar la tendencia alcista tanto en la cotización como en las búsquedas de "Dólar Blue". Las dos series tienen también picos en junio 2013, enero 2014, noviembre 2014 y agosto- septiembre 2015 mostrando un comportamiento estacional similar. El análisis preliminar de estos gráficos muestra indicios de una posible correlación entre ambas series.

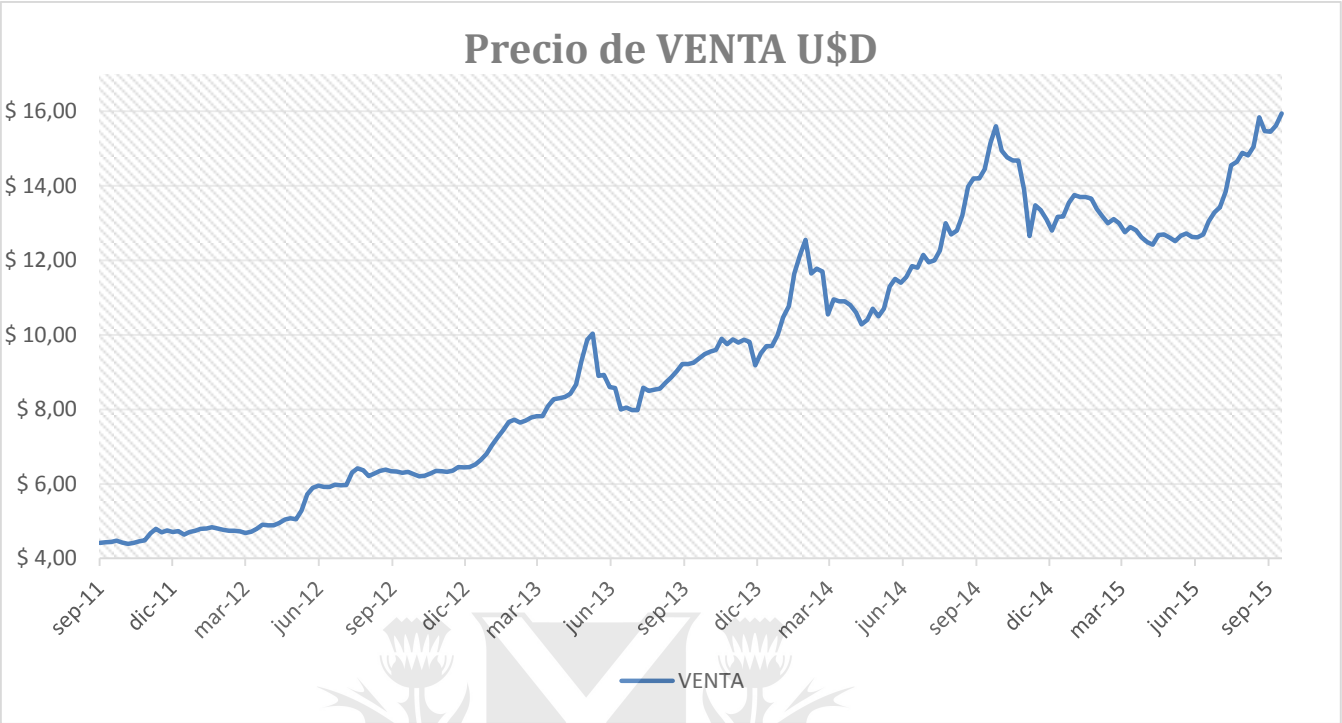
[26] http://www.ambito.com/economia/mercados/monedas/dolar/info/?ric=ARSB=



Cuadro 19: Gráfico del Índice de búsquedas del término "Dólar Blue" desde 2005 a la fecha. Fuente: Google Trends.



Cuadro 20: Grafico del Índice de búsquedas del término "Dólar Blue" desde 2011 a la fecha. Fuente: Google Trends.



Cuadro 21: Cotización Precio de Venta Dólar informal o Blue de Octubre 2011 a Septiembre 2015. Fuente: Ámbito Financiero.

Un ejercicio para testear la correlación podría ser intentar regresar la variable búsquedas de "Dólar Blue" sobre la cotización informal histórica. En el cuadro 22 se observan los resultados de la regresión y el estimador lineal encontrado. Si vemos el R^2 de 0,754 se puede afirmar que gran parte de la variación de la cotización del dólar informal puede ser explicada a través de una regresión lineal del índice de búsquedas del dólar informal. Por lo tanto el coeficiente β de 0,1194 puede ser entendido como una elasticidad entre la búsqueda de cotización en internet y el precio de venta del dólar al día siguiente. Un aumento del 1% en las búsquedas de internet llevaría a un aumento del 0,12% del precio del dólar blue. Estadísticamente tenemos evidencia significativa para rechazar la hipótesis nula, que las búsquedas de internet no guardan relación alguna con el precio del dólar informal. El P-Valor se aproxima a cero por lo que la probabilidad de estar en un escenario de no relación entre las variables es muy baja.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,86814631
R Square	0,75367801
Adjusted R Square	0,75250505
Standard Error	1,70658421
Observations	212

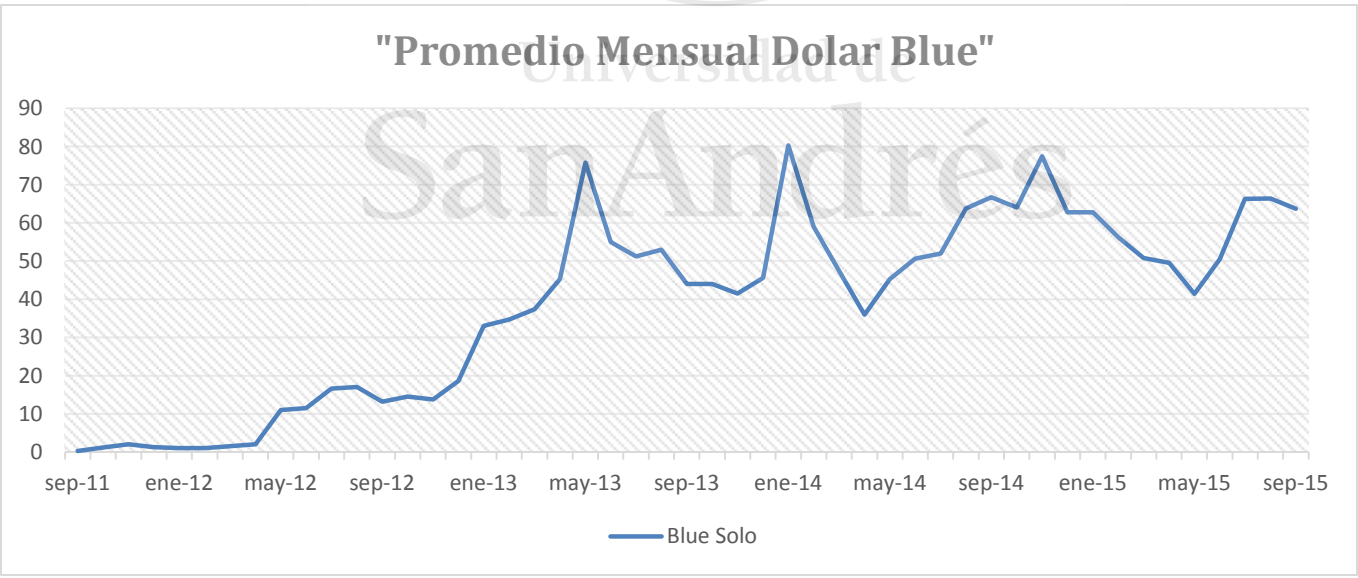
ANOVA

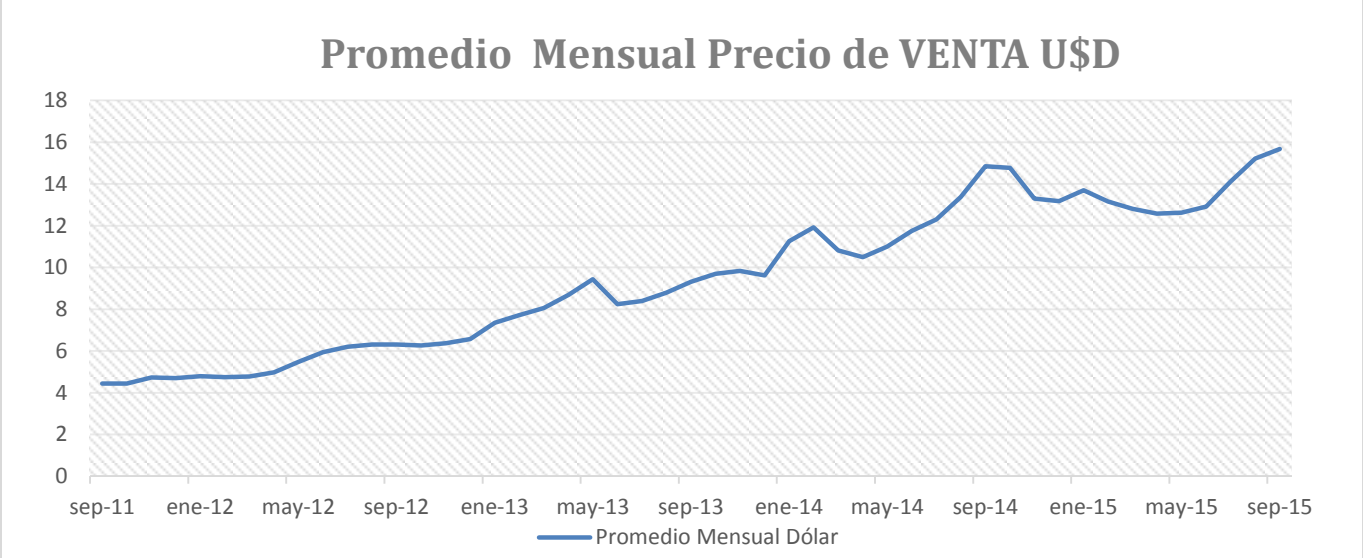
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	1871,360253	1871,36025	642,542643	8,1075E-66
Residual	210	611,6102296	2,91242966		
Total	211	2482,970483			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95,0%*	*Upper 95,0%*
Intercept	4,82471879	0,216128054	22,3234268	2,3646E-57	4,3986602	5,25077739	4,3986602	5,25077739
X Variable 1	0,11943924	0,0047119	25,3484249	8,1075E-66	0,11015056	0,12872793	0,11015056	0,12872793

Cuadro 22: Resultados estadísticos de la regresión lineal entre X: Índices de búsqueda de Dólar Blue en Argentina e Y: Serie de cotización de Dólar informal para la venta en Argentina. Fuente: Ámbito Financiero y Google Trends.

Habiendo observado que existe evidencia para suponer una correlación entre ambas variables podemos profundizar el análisis intentando encontrar una tendencia en las series. En el cuadro 23 y 24 se toma solo el promedio mensual de cada serie a modo de quitar la estacionalidad que viene dada a principio y a fin de cada mes.



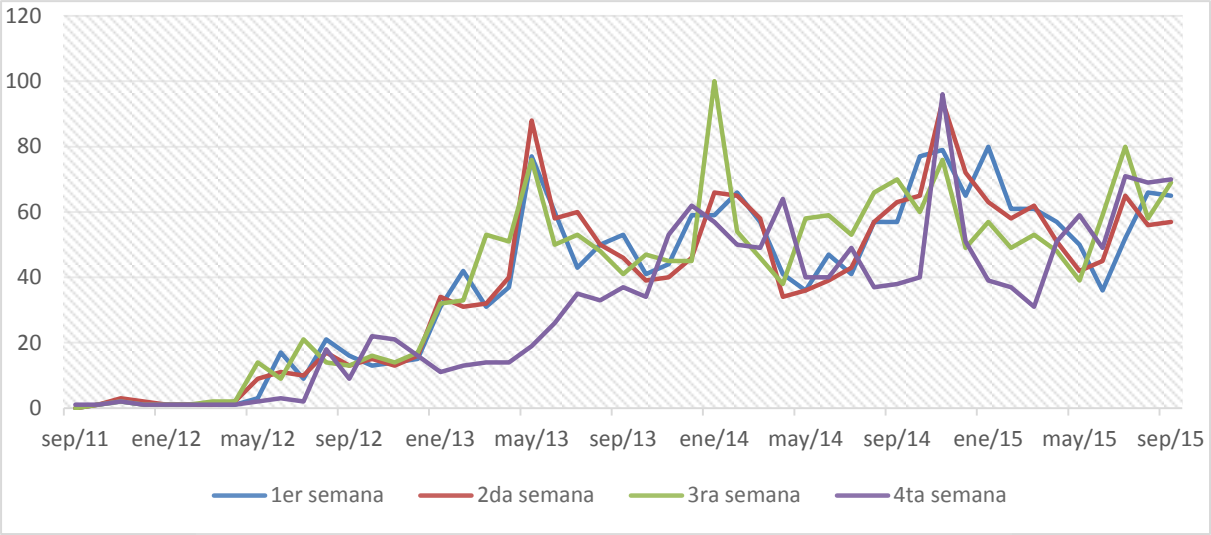


Cuadro 23 y 24: Promedio mensual de búsquedas de Dólar Blue y cotización histórica del Dólar informal. Fuente: Google Trends y Ámbito Financiero

Observando los promedios podemos detectar una tendencia a la suba en ambas series y observar con mayor claridad los picos en junio 2013, enero y noviembre 2014 y septiembre 2015 que continua en alza. Los picos abruptos pueden deberse a variables exógenas no contempladas en este modelo acotado. Por ejemplo períodos post elecciones donde el dólar puede sufrir ataques especulativos, principios de año donde el dólar paralelo tiende a subir por los meses de vacaciones donde se demanda más moneda extranjera, cobro de aguinaldo, shocks de política monetaria como mayor rigidez en el cepo cambiario o una operación de Mercado Abierto del BCRA para bajar la cotización paralela al ofrecer divisas oficiales.

Por último si separamos la serie por semana del mes y las graficamos por separado se puede identificar los ciclos mensuales del modelo. En el grafico 25 se encuentran las búsquedas agregadas por semana. Aun cuando las 3 primeras semanas del mes parecen contar con un nivel similar, la última semana del mes representa a lo largo del gráfico niveles más bajos de búsqueda de Dólar Blue en Google. Esto podría deberse a que a fin de mes se cuenta con menos dinero y las compras de dólar para ahorro se deben haber realizado en las semanas anteriores. Esta disparidad genera un ciclo mensual en la serie el cual pudimos observar en los gráficos anteriores.

Cuadro 25: Serie de Búsqueda de Dólar Blue desagregada por semana del mes Fuente: Google Trends



Con este análisis preliminar el buscador de Google podría ser un buen indicador sobre el interés de una moneda informal de la cual no existen canales oficiales de información y el usuario que busca comprar o vender dólares primero realiza una consulta en Google para encontrar algún índice de cotización confiable.

Un trabajo posterior podría ser proponer un modelo econométrico similar al que utilizan Varian & Choi para analizar las correlaciones entre ambas series y estudiar si agregando la variable de búsquedas en Google, mejora la capacidad de predicción del modelo.

7. Conclusiones

Un mundo ideal para un econometrista sería aquel donde tuviese los datos relevantes disponibles de manera inmediata y poder predecir en el corto plazo. Aun cuando un escenario así estaría más cercano a la futurología y especulación, el desarrollo de internet y las redes sociales parecen haberlo acercado al menos a una posibilidad no tan lejana. Hoy en día internet y las redes sociales han democratizado la información, descentralizando su administración llevándola al alcance del público.

El problema ya no parece encontrarse en la falta de datos sino en la abundancia de ellos. Y es aquí donde la econometría como disciplina tiene una gran oportunidad de crecimiento. Si el avance de esta disciplina se enfoca en la detección de datos relevantes y quitar el ruido en el medio de este universo de datos, entonces hay un gran futuro para la econometría moderna.

Lo precursor de Varian & Choi fue justamente este acercamiento del análisis tradicional con las nuevas herramientas de análisis de bases de datos. Invitar a los usuarios a realizar sus propias predicciones o correlaciones, complementar modelos tradicionales con nuevos datos relevantes e instantáneos o simplemente sugerir otra perspectiva de análisis a un estudio particular. El análisis econométrico puede flexibilizarse utilizando declaraciones de preferencias tan simples y directas como una búsqueda de internet.

Sin embargo contar con demasiados datos irrelevantes puede sesgar el análisis al igual que contar con insuficientes datos. Hemos observado que tomar datos brutos de internet puede ser un buen indicador pero no el único. Aun siendo una declaración de preferencias, es imposible conocer el incentivo detrás de esa búsqueda. Es por eso que esta nueva disponibilidad debe usarse con criterio, para complementar y no sustituir los modelos clásicos de análisis de estadísticas.

La economía como ciencia puede obtener grandes resultados de este tipo de herramientas. Su crecimiento parece trascender las fronteras tradicionales para aventurarse en otras disciplinas como por ejemplo la epidemiología, donde un pronóstico responsable puede ayudar a prevenir epidemias y mejorar la salud general. O mejorar políticas públicas como analizar el desempleo tomando nuevos indicadores provenientes directamente del público como puede ser una búsqueda de empleo en un portal de trabajo.

Hemos observado la diversa literatura que se ha escrito sobre estas técnicas y aunque su crecimiento es evidenciable todavía no se desarrollado como una disciplina propia de la economía. Para ello es necesaria una estructura metodológica que asiente parámetros y convenciones para sistematizar este tipo de análisis. Ordenar la literatura existente es un punto de partida para la expansión de esta disciplina y será un gran desafío para transformar estos diversos estudios aislados en una escuela econométrica a futuro. El aprovechamiento de estos nuevos datos puede ayudarnos a entender mejor el mundo moderno. Esta nueva era de información instantánea parece habernos acercado a cuantificar y detectar comportamientos sociales. Debemos estar preparados para acompañar este crecimiento con una escuela de pensamiento estructurada, donde la nueva literatura pueda florecer. Donde los nuevos trabajos traigan mayores modelos y aspiren al orden en esta inmensidad de datos.

8. Bibliografía

Askitas, N., & Zimmermann, K. F. (2009). 'Google econometrics and unemployment forecasting'. *German Council for Social and Economic Data (RatSWD) Research Notes*, (41).

Butler, D. (2013). 'When Google got flu wrong.' *Nature*, *494*(7436), 155.

Carneiro, H. A., & Mylonakis, E. (2009). 'Google trends: a web-based tool for real-time surveillance of disease outbreaks.' *Clinical infectious diseases*,*49*(10), 1557-1564.

Choi, H., & Varian, H. (2009).'Predicting initial claims for unemployment benefits.'*Google Inc* 1-5.

Choi, H., & Varian, H. (2012). 'Predicting the present with google trends.' *Economic Record*, *88*(s1), 2-9.

Cook, S., Conrad, C., Fowlkes, A. L., & Mohebbi, M. H. (2011). 'Assessing Google flu trends performance in the United States during the 2009 influenza virus A (H1N1) pandemic' *PloS one*, *6*(8), e23610.

Copeland, P., Romano, R., Zhang, T., Hecht, G., Zigmond, D., & Stefansen, C. (2013). 'Google disease trends: an update' *Nature*, *457*, 1012-1014.

Cowpertwait, P. S., & Metcalfe, A. V. (2009). *Introductory time series with R*. Springer Science & Business Media.

Eysenbach, G. (2006). 'Infodemiology: tracking flu-related searches on the web for syndromic surveillance'. In *AMIA Annual Symposium Proceedings* (Vol. 2006, p. 244). American Medical Informatics Association.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). 'Detecting influenza epidemics using search engine query data.' *Nature*, *457*(7232), 1012-1014.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (p. 6). New York: springer.

Johnson, H. A., Wagner, M. M., Hogan, W. R., Chapman, W., Olszewski, R. T., Dowling, J., & Barnas, G. (2004). 'Analysis of Web access logs for surveillance of influenza.' *Stud Health Technol Inform*, *107*(Pt 2), 1202-6.

Mohebbi, M., Vanderkam, D., Kodysh, J., Schonberger, R., Choi, H., & Kumar, S. (2011). 'Google correlate whitepaper'. *Web document: correlate. googlelabs. com/whitepaper. Pdf*

Seifter, A., Schwarzwalder, A., Geis, K., & Aucott, J. (2010). 'The utility of "Google Trends" for epidemiological research: Lyme disease as an example.' *Geospatial Health*, *4*(2), 135-137.

Varian, H. R. (2014). 'Big data: New tricks for econometrics'. *The Journal of Economic Perspectives*, 3-27.

Wilson, N., Mason, K., Tobias, M., Peacey, M., Huang, Q. S., & Baker, M. (2008). 'Interpreting Google flu trends data for pandemic H1N1 influenza: the New Zealand experience'. *Euro surveillance: bulletin européen sur les maladies transmissibles= European communicable disease bulletin*, *14*(44), 429-433.

Wu, L., & Brynjolfsson, E. (2014). 'The future of prediction: How Google searches foreshadow housing prices and sales'. In *Economic Analysis of the Digital Economy*. University of Chicago Press.