Documentos CEDE

ISSN 1657-7191 Edición electrónica.

Hechos y palabras: la realidad colombiana vista a través de la prensa escrita

Juan Manuel Caicedo
Alejandro Gaviria
Javier Moreno

46

NOVIEMBRE DE 2011

Universidad de los Andes
Facultad de Economía

CEDE
Centro de Estudios
sobre Desarrollo Económico

Serie Documentos Cede, 2011-46
ISSN **1657-7191 Edición electrónica.**

Noviembre de 2011

Calle 19A No. 1 – 37 Este, Bloque W.
Bogotá, D. C., Colombia
Teléfonos: 3394949- 3394999, extensiones 2400, 2049, 3233
infocede@uniandes.edu.co
http://economia.uniandes.edu.co

Ediciones Uniandes
Carrera 1ª Este No. 19 – 27, edificio Aulas 6, A. A. 4976
Bogotá, D. C., Colombia
Teléfonos: 3394949- 3394999, extensión 2133, Fax: extensión 2158
infeduni@uniandes.edu.co

Edición y prensa digital:
Cadena S.A. • Bogotá
Calle 17 A Nº 68 - 92
Tel: 57(4) 405 02 00 Ext. 307
Bogotá, D. C., Colombia
www.cadena.com.co

Impreso en Colombia – *Printed in Colombia*

Universidad de los Andes
Facultad de Economía

CEDE
Centro de Estudios
sobre Desarrollo Económico

Hechos y palabras: la realidad colombiana vista a través de la prensa escrita

Juan Manuel Caicedo, Alejandro Gaviria y Javier Moreno[1]

Noviembre de 2011

Resumen

Este artículo presenta la primera aplicación al estudio de la realidad colombiana de *culturomics*, una nueva metodología de investigación de las ciencias sociales que describe tendencias culturales, sociales y lingüísticas con base en el análisis cuantitativo de textos digitalizados. El artículo usa la totalidad de las noticias y opiniones publicadas durante los últimos veinte años en tres medios escritos de circulación nacional con el propósito de describir las trayectorias de algunos fenómenos socioeconómicos y políticos: la corrupción, la división de poderes, el conflicto, el optimismo económico, etc. Más allá de los hallazgos concretos, este artículo muestra de qué manera puede usarse la metodología propuesta para describir la cambiante realidad de un país en desarrollo.

Palabras clave: Colombia, periódicos, análisis de textos, cambio institucional y *culturomics.*

Clasificación JEL: O54 y Z10.

[1] Juan Manuel Caicedo está cursando sus estudios de posgrado en Carnegie Mellon (juan@cavorite.com). Alejandro Gaviria es profesor de la Facultad de economía de la Universidad de los Andes en Bogotá, Colombia (agaviria@uniandes.edu.co) y Javier Moreno es un postdoctoral fellow en la Universidad de Waterloo (bluelephant@gmail.com)

Words and facts: an analysis of Colombian reality through the written news media

Juan Manuel Caicedo, Alejandro Gaviria y Javier Moreno

November 2011

Abstract

This paper presents the first application to Colombia of culturomics, a novel research method in the social sciences that describes cultural, social and linguistics trends based on the quantitative analysis of millions of digitized texts. The paper uses all news and editorials published during the past two decades in three national periodicals in order to analyze the trajectories of various phenomena of interest: corruption, economic optimism, the relative power of the judicial branch, etc. Beyond the specific findings, this paper showcases how to use culturomics to describe the changing reality of a developing country.

Key words: Colombia, newspapers, institutional change and culturomics.

JEL classification: O54 and Z10.

1. Introducción

Este artículo presenta la primera aplicación a la realidad colombiana de *culturomics*, una nueva metodología de investigación de las ciencias sociales que describe tendencias culturales, sociales y lingüísticas con base en el análisis cuantitativo de textos digitalizados. En principio, la mayor o menor frecuencia de aparición de ciertas palabras o expresiones en millones de textos digitalizados revela cambios relevantes en la cultura, la sociedad o el lenguaje. Dicho de otra manera, los textos escritos, analizados de manera integral, leídos exhaustivamente, contienen información útil sobre ciertos aspectos de la realidad.

El artículo utiliza la totalidad de las noticias y comentarios publicados durante los últimos veinte años en tres medios escritos de circulación nacional: *El Tiempo*, *Semana* y *Dinero*. En números redondos, este trabajo analiza más de dos millones de artículos que contienen, en conjunto, aproximadamente seiscientos millones de palabras. Los cambios en la aparición de ciertas palabras, *desempleo*, *recesión*, *corrupción*, *magistrados*, entre otras, ayudan a entender algunos aspectos de la realidad contemporánea colombiana. Contar palabras permite contar historias.

Este trabajo es una de las primeras aplicaciones de *culturomics* basada enteramente en publicaciones periódicas. Las aplicaciones más conocidas y difundidas hacen uso de libros publicados en el transcurso de varias décadas e incluso de siglos. Este trabajo, por el contrario, utiliza artículos publicados en periódicos en un período más breve. Por lo tanto, se hace hincapié no en los cambios culturales de larga duración, sino en cambios institucionales y sociales de corto y mediano plazo. Además, este artículo es probablemente la primera aplicación de *culturomics* a una realidad local, a un período específico en un país particular. Las aplicaciones anteriores han sido transnacionales, han abarcado una realidad más amplia, al menos geográfica y socialmente.

El artículo muestra que algunos fenómenos económicos, el desempleo y el crecimiento de la economía, entre otros, son descritos o seguidos adecuadamente por los cambios en las menciones a las palabras correspondientes: *desempleo* y *recesión* en este caso. Muestra igualmente que las frecuencias de aparición de *verano* e *invierno* siguen de cerca las fluctuaciones de la temperatura del océano Pacífico. Revela también que, desde una perspectiva de mediano plazo, la aparición de la palabra *corrupción* no ha crecido, la sigla *FARC* se registra frecuentemente junto con el vocablo *secuestros* y la palabra *magistrados* se ha encontrado recientemente un mayor número de veces que *congresistas*.

Pero más que los aspectos de fondo, la contribución de este artículo es metodológica. El artículo describe una base de datos, presenta una metodología de análisis y muestra el potencial de esta última mediante una serie de ejemplos. La estructura del artículo es la siguiente. La sección 2 presenta los antecedentes y repasa brevemente la literatura relevante. La sección 3 describe los datos. La sección 4 compara, para algunos fenómenos socioeconómicos, el comportamiento de los indicadores con el

comportamiento de la frecuencia en que aparecían las palabras correspondientes. La sección 5 utiliza frecuencias de palabras para estudiar varios fenómenos de difícil medición. El propósito de esta sección, la más polémica del artículo, es mostrar que la metodología en cuestión permite el desarrollo de un nuevo tipo de indicadores en las ciencias sociales. Finalmente, la sección 6 presenta algunas ideas para futuras investigaciones.

2. Motivación y antecedentes

Culturomics es el análisis cuantitativo de tendencias culturales, sociales y lingüísticas con base en libros, periódicos y otros textos digitales disponibles en Internet o en medios similares. Este tipo de análisis usa millones de páginas de texto para estudiar la evolución de patrones culturales y para identificar cambios significativos en la opinión pública. En opinión de Michel *et al.* (2011), el análisis cuantitativo de textos constituye un nuevo método de análisis en las ciencias sociales, cuya virtud radica no en el estudio minucioso de algunos textos seminales —la estrategia tradicional de las ciencias sociales—, sino en la lectura automatizada de millones de textos de diversa calidad y trascendencia. *Culturomics* compensa con volumen su falta de discernimiento; es un método de fuerza bruta.

Michel *et al.* (2011) usan un corpus de más de cinco millones de libros en inglés (el 4% de los libros publicados en este idioma en todos los tiempos) con el propósito analizar, entre otras cosas, la evolución de la gramática de la lengua inglesa, el auge y la caída de la reputación política, científica y artística y algunos eventos de censura en contra de artistas judíos. Muchas otras aplicaciones son posibles. Este tipo de análisis permitiría estudiar, por ejemplo, la cambiante popularidad de algunas teorías científicas (la teoría de la evolución), de ciertas ideologías (el marxismo) o incluso de varias formas de pensamiento (los sesgos étnicos o raciales).

En general, las fluctuaciones en el uso de ciertas palabras contiene información relevante sobre el mundo del lenguaje y las ideas, sobre la realidad exterior y sobre lo que ha ocurrido (y está ocurriendo) en la mente de los hombres. Algunos ejemplos bastan para ilustrar este tipo de análisis. El gráfico 1 muestra la frecuencia de la preposición *a* en dos formas distintas de escritura: una con acento y otra sin acento[2]. La preposición acentuada (*á*) era la más utilizada durante el siglo XIX, pero la preposición sin acento (*a*) pasó a convertirse, durante la primera mitad del siglo XX, en la norma de uso general. La transición fue rápida, tomó aparentemente menos de una década. En teoría, la existencia de las academias de la lengua hace que algunos cambios ortográficos sean mucho más rápidos en el español que en el inglés (Michel *et al.*, 2011).

[2] Como se explicará más adelante, la frecuencia se calcula como el número de ocurrencias de la palabra en cuestión (a) en el universo de textos disponibles en un año dado dividido por el número total de palabras en los mismos textos en el mismo año. El gráfico puede reproducirse fácilmente en http://ngrams.googlelabs.com/.

El gráfico 2 muestra, para todo el siglo XX, la frecuencia de la palabra *marxismo* en el corpus de libros en español y de la expresión equivalente, *marxism*, en el corpus de libros en inglés. En ambos idiomas, la frecuencia aumenta casi de manera continua entre 1920 y 1980. En español comienza a disminuir en 1980; en inglés, unos pocos años más tarde. La evolución es similar en ambos casos, pero la frecuencia es mucho mayor en los textos publicados en español. El gráfico 3 repite el análisis para la palabra *neoliberalismo*. El auge comienza lentamente en los años ochenta, toma fuerza en los años noventa y empieza a revertirse a partir del año 2002, coincidiendo paradójicamente con la recuperación de la economía mundial y de las economías latinoamericanas que habían aplicado, años atrás, las recetas neoliberales. Finalmente, el gráfico 4 muestra los cambios en la influencia de Francia y los Estados Unidos en las letras hispanas. Francia dominó hasta finales del siglo XIX y los Estados Unidos comenzaron a consolidar su dominio en la segunda mitad del siglo anterior. Francia es el pasado, los Estados Unidos el presente, pero no necesariamente el futuro.

Gráfico 1.
Frecuencia de las preposiciones *a* y *á*

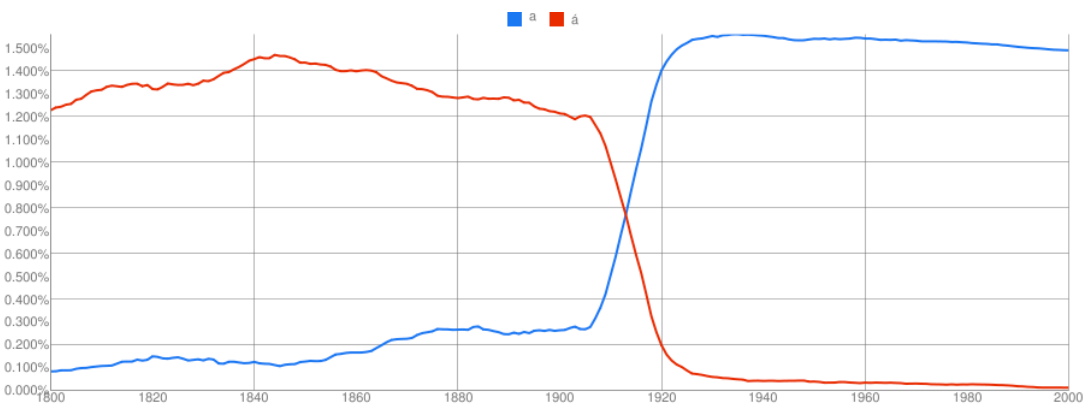


Gráfico 2.
***Marxismo* en el siglo *xx*: inglés y español**

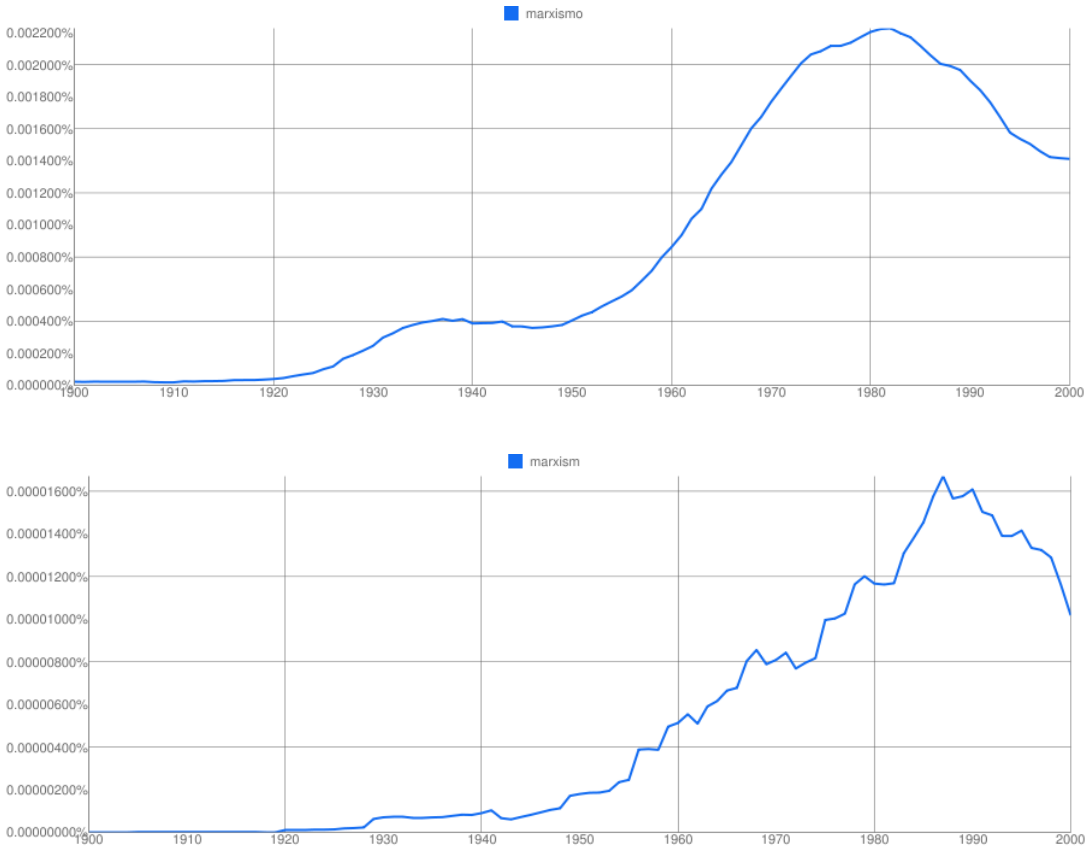
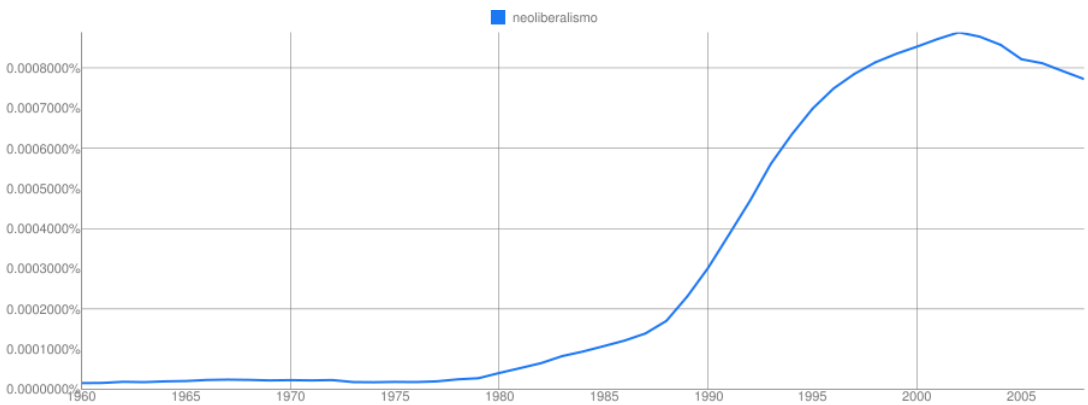
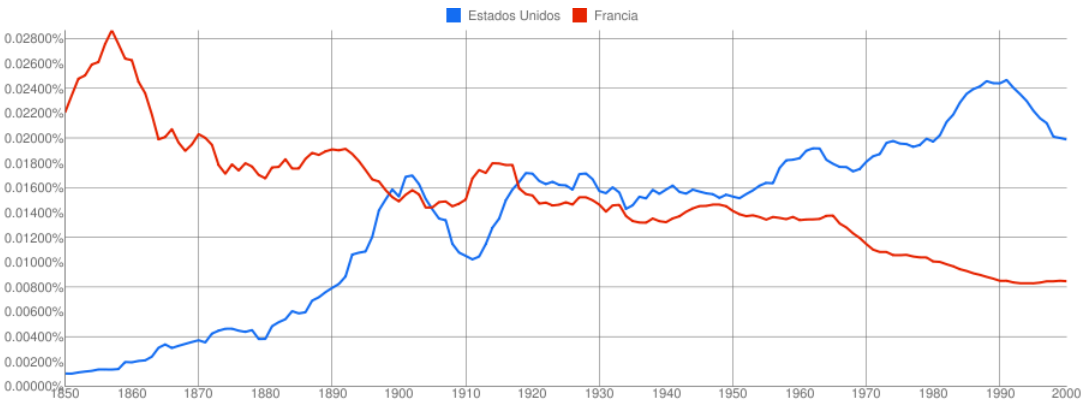


Gráfico 3.
Auge del uso del término *neoliberalismo* en los años noventa



Gráfica 4.
Influencias foráneas en el mundo hispano: Francia y Estados Unidos



Recientemente *culturomics* ha sido usado para estudiar el origen y el desarrollo de algunas ideas fundamentales de la economía, la sociología y la ciencia política. Ravallion (2011) muestra, por ejemplo, que a comienzos de los años sesenta las palabras *pobreza*, *desigualdad* y *crecimiento económico* comenzaron a ganar popularidad. Este cambio refleja un cambio intelectual de fondo, una reconceptualización de la idea del desarrollo económico: el desarrollo comenzó a ser visto como un problema tratable, no sólo como un reto intelectual, sino también como una responsabilidad inaplazable de la comunidad internacional.

Michel *et al.* (2011) y Ravallion (2011) utilizan las mismas herramientas, el conteo de palabras en un corpus de cinco millones de libros y 360,000 millones de palabras, para estudiar la importancia de algunas enfermedades infecciosas. Con la notable excepción del sida y la hepatitis, las enfermedades infecciosas perdieron participación a partir de la segunda mitad del siglo XX. Ravallion anota, además, que las menciones al sida superan, en todo momento, su impacto sobre la morbilidad y la mortalidad; resultado que ilustra un hecho fundamental: el análisis de textos revela aspectos relevantes de la realidad, pero la revelación contiene muchas veces un sesgo cultural, está sesgada por las creencias, las teorías, las opiniones y las modas de cada momento. *Culturomics* es el estudio realidad a través del filtro de la cultura.

Aquí cabe una aclaración: Michel *et al.* (2011) excluyen intencionalmente cualquier tipo de publicación periódica. Esta decisión, aunque inexplicada en el artículo, no es arbitraria. Los libros, por su estilo de producción y desarrollo, ofrecen una perspectiva decantada de la cultura, desconectada de las fluctuaciones bruscas de la opinión pública. Al restringir el análisis a los libros, se reduce el ruido y se gana en claridad global y capacidad generalizadora, pero, por lo mismo, se pierde especificidad: los libros no dan cuenta de la manera como la sociedad responde e interpreta el flujo de información, imperfecto y a veces incluso contradictorio, que se produce a diario. Para explorar fenómenos sociales sensibles a una información que cambia y se adapta en tiempo real, conviene utilizar más bien archivos de noticias.

Antes del lanzamiento de la base de datos de libros digitalizados de Google que popularizó el tipo de análisis descrito anteriormente, Glaeser y Goldin (2002) examinaron la frecuencia de aparición de las palabras *corrupción* y *fraude* en *The New York Times* y un conjunto de diarios regionales con la idea de construir un indicador de la trayectoria de la corrupción en los Estados Unidos. Su análisis muestra que la frecuencia de las palabras mencionadas aumentó durante la primera parte del siglo XIX y disminuyó súbitamente después de 1870. En opinión de los autores, esta trayectoria replica la evolución de la corrupción en los Estados Unidos a pesar de los sesgos mediáticos ya mencionados[3]. En síntesis, la palabra escrita puede dar cuenta de la trayectoria de algunos fenómenos sociales de difícil medición.

Más recientemente, Leetaru (2011) utilizó un archivo de treinta años de noticias recopiladas por servicios de inteligencia de los Estados Unidos e Inglaterra para medir, mediante un análisis automatizado del tono de los artículos (positivo o negativo y en qué grado), la opinión global sobre eventos como la llamada *primavera árabe* o la *guerra en los Balcanes*. El análisis muestra que las crisis políticas son usualmente precedidas por una caída significativa en el tono de los artículos. Con métodos automatizados de geoposición de textos, Leetaru pudo producir además una estimación aproximada de la localización de Osama Bin Laden antes de su muerte. El autor propone el uso de estas metodologías para predecir eventos de importancia global de manera similar a como Bollen *et al.* (2011) anticipan movimientos del mercado de valores mediante un análisis de frecuencias del caudal de Twitter. Con todo, estos resultados evidencian el potencial descriptivo de *culturomics*.

La dificultad de este tipo de análisis radica entonces en la interpretación; en la necesidad de discernir, para el fenómeno en cuestión, qué tanto corresponde la cambiante frecuencia de las palabras a una faceta real y qué tanto a una distorsión mediática. La distorsión depende, en general, del fenómeno estudiado, del momento histórico y de las publicaciones. Cuando existen cifras objetivas, como ocurre para algunos fenómenos económicos, la comparación de la ocurrencia de las palabras y la realidad del fenómeno bajo estudio da algunas pistas sobre la los sesgos culturales y de opinión. Cuando no existen cifras objetivas, ambos aspectos son difíciles de separar; los gráficos dicen tanto de los vaivenes realidad como de los ciclos de la cultura y la opinión.

Este artículo utiliza un archivo de noticias para estudiar la realidad, la opinión, la cultura y la economía de Colombia durante los últimos veinte años. El análisis es más sugestivo que definitivo. Plantea muchos interrogantes, revela algunos sesgos y sugiere algunos temas de investigación.

[3] En este caso, como lo reconocen Glaeser y Goldin (2002), el indicador propuesto, basado en las menciones de prensa, tiene un problema adicional: la aparición en prensa, esto es, el reporte escrito de la corrupción, puede afectar directamente el fenómeno que se está tratando de medir. La prensa no sólo refleja, también puede influir sobre la realidad del fenómeno analizado.

3. Base de datos y cálculo de frecuencias

El corpus de noticias utilizado contiene todos los artículos publicados en las versiones electrónicas del periódico *El Tiempo* y de las revistas *Semana* y *Dinero*. Los archivos de noticias tienen una cobertura temporal distinta. El archivo de *El Tiempo* comienza en 1991, el de *Semana*, en 1980 y el de *Dinero*, en 1993. Los tres archivos se extienden hasta el 31 de julio del 2011, fecha de corte del análisis. Las tres publicaciones mencionadas cuentan con los archivos electrónicos de noticias más antiguos y completos de los medios impresos colombianos. Al menos en contenido, estas publicaciones pueden considerarse representativas de la prensa escrita de circulación nacional.

El método usado para la construcción de los archivos de noticias es simple. Primero un programa recorre los sitios web de las publicaciones seleccionadas y descarga la totalidad de los artículos. Luego el mismo programa elimina los elementos adicionales (barras de navegación, enlaces, imágenes, anuncios publicitarios, etc.) y almacena una versión simplificada de cada artículo. Finalmente el programa descarta los artículos (más de 100,000) con el mismo título y el mismo contenido. El análisis final está basado en un archivo depurado que contiene una sola copia de los artículos.

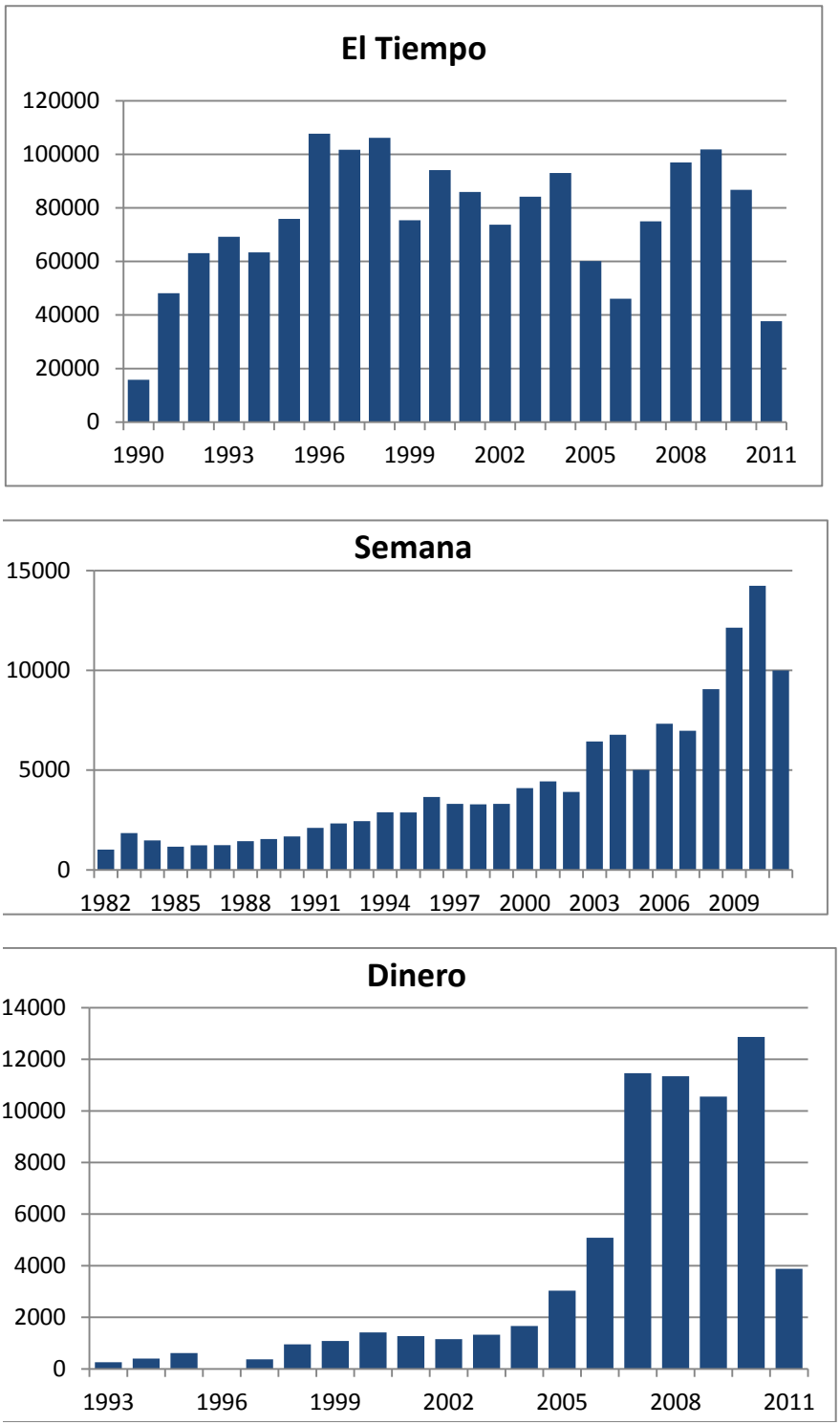
El archivo analizado contiene casi dos millones artículos. En números redondos, el 90.0% de los artículos proviene de *El Tiempo*, el 6.5% de *Semana* y el 3.5% de *Dinero*. El gráfico 5 muestra que el número de artículos varía de manera sustancial de un año al siguiente. El número de artículos de *Semana* y *Dinero* aumentó de manera considerable después del 2005, como consecuencia de la introducción de blogs y de artículos informativos que no hacen parte de las ediciones impresas. El número de artículos de *El Tiempo* no cambió grandemente entre 1993 y 2010, con la excepción de un bache (inexplicado) en los años 2004 y 2005. En los años ochenta, la muestra sólo contiene unos pocos artículos de la revista *Semana*: menos de 2,000 anuales en promedio.

Una vez depurado el archivo de noticias, el título, la fecha de publicación y el texto completo de cada uno de los artículos fueron identificados y almacenados en registros separados. Las letras mayúsculas fueron convertidas a minúsculas y la totalidad del texto fue dividido en *n-gramas*. Un *n-grama* es una secuencia de *n* palabras consecutivas dentro de un texto determinado. Así, por ejemplo, la división de un texto en *1-gramas* arroja un listado de todas las cadenas de caracteres separadas por espacios o signos de puntuación, incluidas las palabras (*partido* o *Colombia*), los números (1984 y 8,000) y otras expresiones (como *M-19* o *F1*). La división del mismo texto en *2-gramas* arroja secuencias tales como *derechos humanos* o *nueva constitución*. La división en *3-gramas* muestra secuencias tales como *5 a 0* o *Valle del Cauca*.

La frecuencia de aparición se calcula como el cociente entre el número de ocurrencias de un *n-grama* en todos los artículos publicados durante un mes dado en una de las tres publicaciones analizadas y el número total de *1-gramas* publicados durante

el mismo mes en la misma publicación. Los *n-gramas* que aparecen menos de diez veces en un mes fueron excluidos del análisis.

Gráfico 5.
Cantidad de artículos por publicación



El corpus contiene más de 600 millones de 1-*gramas*. La distribución por publicación de los *1-gramas* es similar, pero no idéntica, a la distribución de los artículos: el 86.5% está en *El Tiempo*, el 9.5% en *Semana* y el 4.0% restante en *Dinero*. La participación de *Semana* y *Dinero* es mayor en la distribución de *1-gramas* que en la de artículos, habida cuenta de la mayor longitud de los artículos publicados en estos medios de circulación semanal o quincenal con relación a aquellos publicados en *El Tiempo*, un medio de circulación diaria.

El análisis de las secciones siguientes está limitado al período 1992-2011. Con anterioridad a 1992, el corpus incluye apenas unos pocos artículos, provenientes en su gran mayoría de la revista *Semana*. En el período estudiado, por el contario, el número de artículos permite analizar los cambios en las frecuencias de palabras de escaza aparición (*bonanza*, *desempleo*, *sequía*, etc.). El período de análisis coincide con los primeros veinte años de la Constitución Política de Colombia. Aunque el período fue escogido por razones pragmáticas, asociadas a la disponibilidad de información, tiene también un sentido o significado histórico.

Como se dijo anteriormente, las menciones a las palabras de interés (*corrupción*, por ejemplo) están siempre normalizadas por el número total de palabras o *1-gramas* en la totalidad del archivo de noticias. En principio, el aumento en el número de artículos, como resultado, por ejemplo, de la inauguración de contenidos virtuales, no es un problema: la frecuencia de la palabra *corrupción*, para usar el mismo ejemplo, no tiene por qué aumentar simplemente si aumenta el número de artículos o el volumen de información.

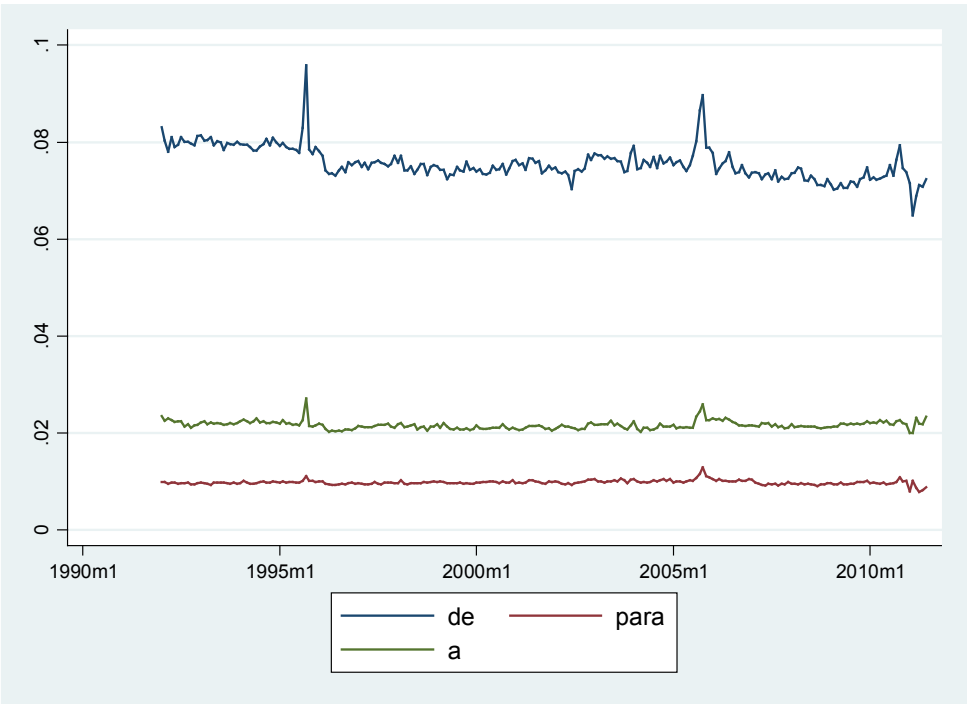
El gráfico 6 muestra la frecuencia de las preposiciones más comunes del idioma español: *a*, *de* y *para*. Estas frecuencias no deberían cambiar de un año al siguiente a pesar del aumento en el número total de artículos publicados; por lo tanto, la existencia de cambios abruptos o de tendencias bien definidas podrían indicar la presencia de sesgos o problemas en los archivos de noticias. El gráfico muestra que los cambios son marginales, asociados probablemente a distorsiones aleatorias y a algunos baches en el archivo de noticias (en septiembre de 1995, septiembre del 2005 y octubre del 2010). Este resultado descarta, en principio, la presencia de grandes errores de construcción o programación.

Si el contenido de una publicación cambia de manera sustancial, por ejemplo, si la publicación se concentra en la información internacional o deportiva, los cambios en las frecuencias darían, en teoría, una idea equivocada sobre la trayectoria de ciertos fenómenos: una disminución de la frecuencia de la palabra *corrupción* obedecería no tanto a una disminución del fenómeno bajo análisis, como a una reducción de su cobertura mediática asociada, a su vez, a los cambios en el contenido de la publicación. En general, el sesgo derivado de los cambios en el contenido de las publicaciones periódicas puede atenuarse mediante los cambios en la normalización de las series. Las menciones a las palabras de interés pueden dividirse ya no por el número total de 1-*gramas*, sino por el número total de apariciones de algunas palabras genéricas que capturan indirectamente las posibles variaciones de importancia o

contenido. Las apariciones de *corrupción* podrían dividirse por las apariciones de *política*, *gobierno* o *presidente*, palabras que reflejan, en términos generales, la importancia del cubrimiento local dentro del contenido general del periódico en cada momento del tiempo. En general, los cambios en la normalización no afectaron los resultados de manera significativa.

Finalmente, el análisis siguiente depende en buena medida de la comparación de series de tiempo. La comparación está basada en la inspección visual y en el simple cálculo de correlaciones. Una comparación más sofisticada podría usar, por ejemplo, los análisis de supervivencia comunes en biología (Jones y Crowley, 1989) o los indicadores de bondad de ajuste basados en conceptos de entropía (Cowell *et al.*, 2011). La sofisticación metodológica, sin embargo, no necesariamente resulta más informativa.

Gráfico 6.
Frecuencias de algunas preposiciones



4. Algunos ejemplos: realidades y palabras

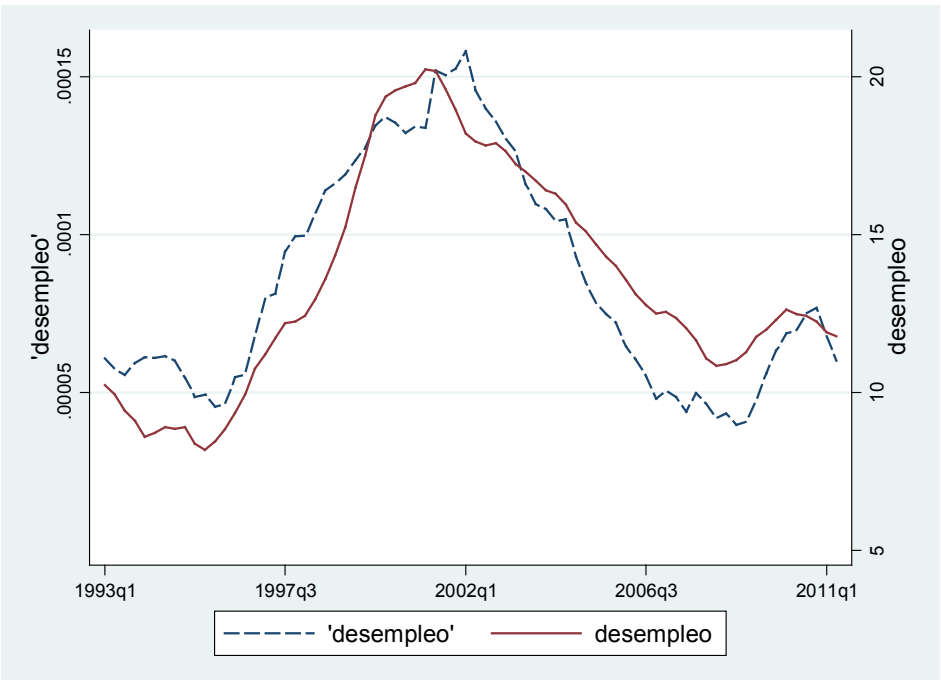
Esta sección presenta cinco ejemplos de fenómenos distintos que comparten una misma característica: todos son medidos de manera sistemática mediante indicadores conocidos y probados. El análisis propuesto compara, en todos los casos, la evolución de dos series: el indicador del fenómeno estudiado (desempleo, por ejemplo) y la frecuencia de la palabra correspondiente (*desempleo*, en este caso). La comparación entre indicadores y frecuencias revela, por una parte, la pertinencia

de la metodología propuesta y, por otra, la magnitud y dirección de algunos sesgos mediáticos. En síntesis, el análisis permite entender de qué manera la realidad es reflejada (y al mismo tiempo distorsionada) por la prensa escrita[4].

4.1. Desempleo

El gráfico 7 muestra la tasa trimestral de desempleo de las siete principales ciudades de Colombia y la frecuencia de la palabra *desempleo* en el archivo de noticias del diario *El Tiempo* correspondiente al mismo trimestre. Las series abarcan el período comprendido entre el primer trimestre de 1993 y el segundo del 2011. Ambas series fueron filtradas con base en un promedio móvil de un año (cuatro trimestres). La tasa de desempleo se limita a las siete principales ciudades con el fin de asegurar la comparabilidad de los datos a lo largo del período. El análisis no cambia en absoluto si el archivo de *El Tiempo* se complementa con archivos de *Semana* y *Dinero*: el grueso de las noticias sobre el tema en cuestión proviene de *El Tiempo*.

Gráfico 7.
Desempleo y *desempleo* en *El Tiempo*



La correlación entre ambas series es evidente. El coeficiente de correlación es de 0.90 en todo el período. La tasa de desempleo y la frecuencia de la palabra *desempleo* crecieron a un ritmo similar durante la crisis de finales de los años

[4] Una versión del buscador está disponible en http://ngrams.cavorite.com.

noventa, pero el descenso de ambas series fue distinto. La frecuencia (una medida del interés mediático) descendió más rápidamente que la tasa de desempleo (una medida objetiva de la desocupación). La inercia de la realidad fue aparentemente mayor que la inercia del interés mediático. Dicho de otra manera, el fenómeno del desempleo fue más duradero que las noticias y comentarios al respecto.

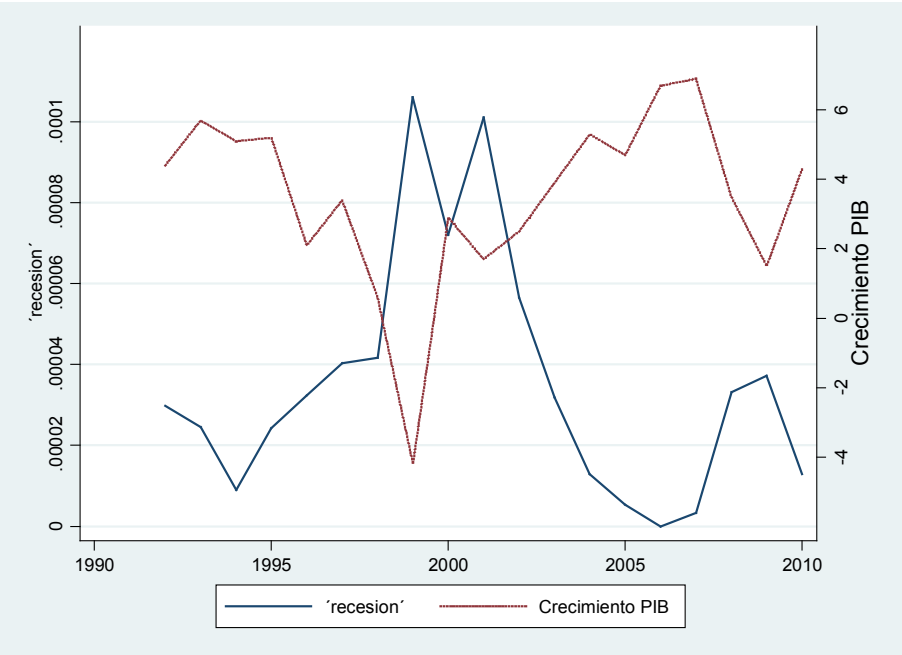
En el 2008, coincidiendo con la crisis internacional y con el aumento del desempleo interno, el interés mediático revivió nuevamente. La frecuencia de la palabra *desempleo* aumentó de manera desproporcionada entre el 2009 y el 2010. La prensa reaccionó de más frente al repunte del desempleo, podría decirse. En términos más generales, la disminución injustificada en el interés mediático después de la crisis de los años noventa y el aumento desproporcionado después de la crisis internacional del año 2008, sugieren que los medios escritos son más sensibles al agravamiento de un problema social que a su persistencia. La prensa escrita perdió el interés en un problema duradero y acuciante. Sólo cuando la situación empeoró, los periodistas y comentaristas volvieron a ocuparse del tema.

En suma, dos hechos merecen resaltarse: la alta correlación entre las dos series y las respuestas asimétricas de la prensa escrita: el olvido relativo de un problema persistente y la reacción abrupta ante su empeoramiento.

4.2. Recesión

El gráfico 8 muestra simultáneamente la tasa anual de crecimiento económico y la frecuencia de la palabra *recesión* en el diario *El Tiempo*. Las series cubren el período comprendido entre los años 1992 y el 2010. La frecuencia corresponde a los promedios móviles de doce meses. La serie de crecimiento corresponde, por su parte, a la tasa anual de crecimiento del producto interno bruto (PIB).

Gráfico 8.
Crecimiento del PIB y *recesión* en *El Tiempo*



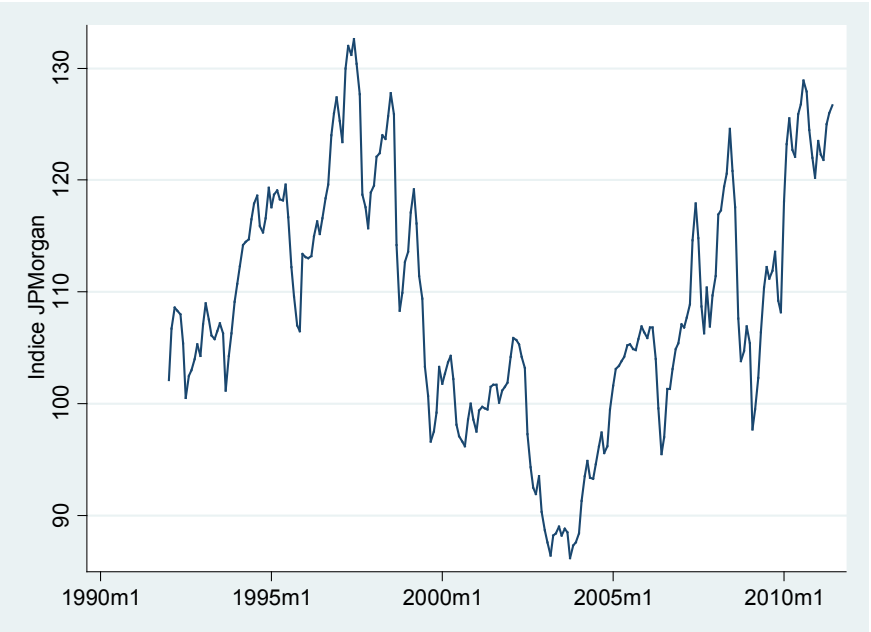
Las conclusiones de este ejemplo son similares a las del ejemplo anterior. Como en el caso del desempleo, la correlación de las dos series es evidente. La frecuencia de *recesión* aumentó cuando cayó la tasa de crecimiento y viceversa. La correlación es de -0.82. Sólo hay una discordancia notable: la leve desaceleración económica del año 2002 estuvo acompañada de un aumento desproporcionado de la frecuencia. Pero, en términos generales, el comportamiento de ambas series es similar.

El anexo muestra un gráfico complementario, que relaciona la frecuencia de la palabra *recesión*, ya no con la tasa de crecimiento económico, sino con una variable idéntica pero inversa: *10 – tasa de crecimiento anual*. Este gráfico permite observar más claramente el movimiento conjunto de ambas series. La caída de la frecuencia coincide con la recuperación de la economía. A diferencia de lo ocurrido en el ejemplo anterior, en el cual el desempleo cayó más lentamente que la frecuencia de la palabra con que se designa, en este ejemplo la frecuencia de *recesión* cayó a un ritmo similar al de la recuperación económica. Mientras que el conteo de palabras no capturó plenamente la trayectoria asimétrica de la tasa de desempleo (empeoramiento súbito y mejoramiento lento), sí parece capturar la trayectoria más simétrica de la tasa de crecimiento económico (empeoramiento y mejoramiento de duraciones semejantes).

4.3. Revaluación

El gráfico 9 presenta el índice de tasa de cambio real calculado por el banco J.P. Morgan[5]. El gráfico permite apreciar las fluctuaciones de la tasa de cambio real: el peso se depreció durante la crisis de los años noventa, se apreció años más tarde durante la recuperación económica, se volvió a depreciar durante la crisis financiera del 2008 y se apreció finalmente en los últimos dos años de crecimiento acelerado. Los ciclos fueron pronunciados, no muy distintos a los experimentados por otros países latinoamericanos. El comportamiento del real brasileño, por ejemplo, fue muy similar.

Gráfico 9.
Índice de tasa de cambio real (ITCR) de J.P. Morgan

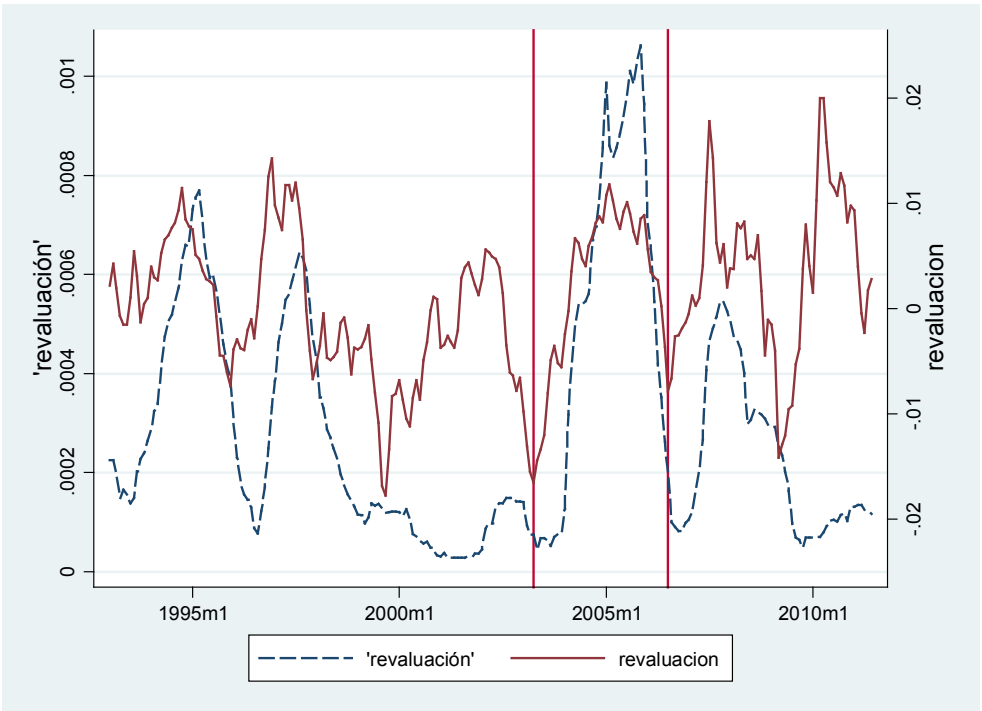


El gráfico 10 muestra simultáneamente el cambio porcentual mes a mes en el ITCR y la frecuencia de la palabra *revaluación* en el archivo de *El Tiempo*. El período de análisis es el mismo del ejemplo anterior, va de 1992 hasta el 2011. Los datos corresponden a los promedios móviles anuales (12 meses). El gráfico distingue tres períodos: 1993(1)-2003(4), 2003(5)-2006(7) y 2006(7)-2011(6). El comovimiento de las series analizadas, la revaluación real y la frecuencia de la palabra *revaluación*, fue diferente en cada uno de los períodos señalados.

5 A diferencia del índice calculado por el Banco de la República (ITCR), el índice usado aumenta cuando el peso colombiano se valoriza (o el dólar se desvaloriza) y disminuye en caso contrario. Las conclusiones del análisis no cambian en absoluto si se usa el ITCR.

El coeficiente de correlación entre la revaluación y *revaluación* es alto, cercano al 0.5 entre 1993 y el 2011. Pero el valor del coeficiente cambió de manera sustancial a lo largo del período de análisis: fue de 0.50 entre 1993 y el 2003, de 0.90 entre el 2003 y el 2006, y de 0.11 entre el 2001 y el 2007. En el primer período (1993-2003), hubo tres eventos de revaluación. Los dos primeros, ambos anteriores a la crisis de finales de los años noventa, estuvieron acompañados de un aumento moderado en la frecuencia. Por el contrario, el último evento posterior a la crisis y de menor magnitud no suscitó un cambio sustancial en la frecuencia; no mereció mayor atención por parte de la prensa.

Gráfico 10.
Revaluación y *revaluación* en *El Tiempo*



En el segundo período (2003-2007), la revaluación generó mucho mayor interés mediático: la frecuencia aumentó de manera sustancial, mucho más rápidamente que la revaluación misma. El mayor interés mediático pudo haber estado impulsado por la respuesta oficial a las presiones de los exportadores e industriales. Entre el 2003 y el 2007, el Gobierno del entonces presidente Uribe estableció una serie de subsidios a los exportadores y trató fallidamente de fijar un piso para la tasa de cambio en diciembre del 2004[6]. Aparentemente el activismo oficial se tradujo en

[6] Las presiones del ejecutivo sobre el Banco de la República fueron entonces un secreto a voces. Véase, por ejemplo: "Los ex codirectores del Banco, Carlos Caballero Argáez, Salomón Kalmanovitz y Sergio Clavijo, así como el decano de la Facultad de Economía de la Universidad de Los Andes y ex director de Planeación Nacional, Juan Carlos Echeverry cuestionaron públicamente las presiones que ejerció el Ejecutivo ese 20 de diciembre de 2004, insinuando que decretaría la emergencia económica e impondría un control de cambio en Colombia" (http://www.primerapagina.com.co/MostrarDocumentoPublico.aspx?id=1113575).

más noticias y comentarios sobre la revaluación. Sea como sea, el interés mediático por la revaluación creció decididamente durante este período.

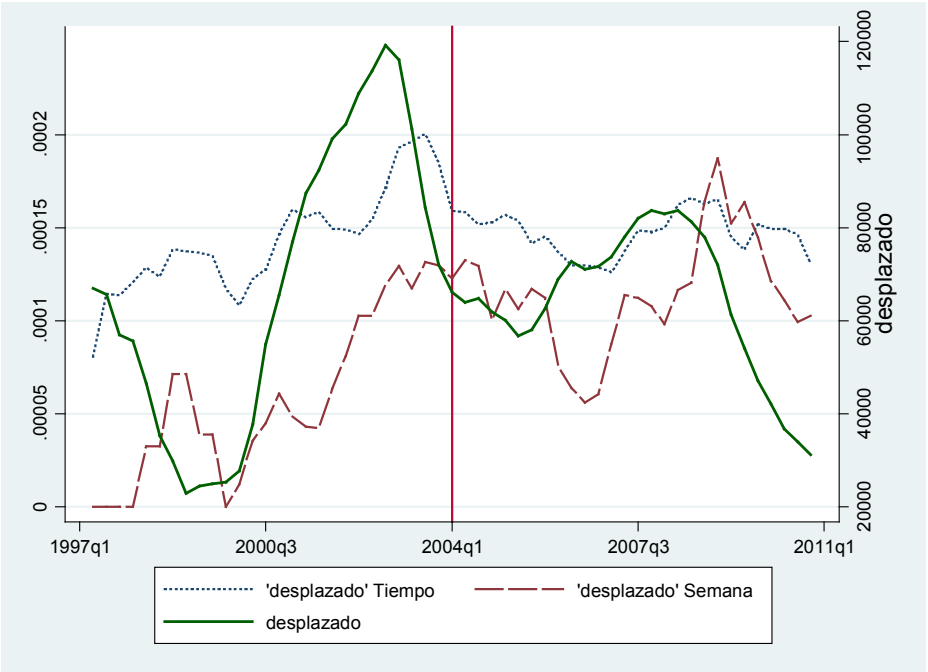
Pero el interés mediático en la revaluación parece haberse desvanecido durante los últimos años. En el tercer período (2007-2011), la revaluación ha pasado casi desapercibida. La frecuencia apenas aumentó a pesar del fuerte aumento en el ITCR. De manera especulativa, uno podría decir que hay menos noticias porque el Gobierno ha hecho menos, pero también que el Gobierno ha hecho menos porque hay menos noticias. Usualmente la prensa alimenta las preocupaciones del Gobierno y viceversa. En teoría, este tipo de retroalimentación positiva abre la posibilidad de equilibrios múltiples: unos de obsesión mediática y otros de desatención o indiferencia. El hecho cierto es que el mismo fenómeno fue primero cubierto obsesivamente y después casi olvidado por completo.

4.4. Desplazados

El gráfico 11 estudia el cubrimiento del desplazamiento forzado por la prensa escrita colombiana. El gráfico muestra el flujo mensual de desplazados según los registros oficiales de la oficina de Acción Social y la frecuencia de la palabra *desplazado* (y sus variantes[7]) tanto en el diario *El Tiempo* como en la revista *Semana*. Las series fueron suavizadas con base en promedios móviles de doce meses. El período de análisis va de enero de 1997 a junio del 2011. Antes de 1997, el número de desplazados era despreciable según los registros oficiales.

[7] Las palabras buscadas fueron desplazado, desplazada, desplazados y desplazamiento. El análisis suma las apariciones individuales de cada una de las palabras referidas.

Gráfico 11.
Desplazados y *desplazados* en *El Tiempo*



El análisis está dividido en dos períodos: antes y después de enero del 2004, esto es, antes y después de la sentencia de la Corte Constitucional que ordenó al Gobierno, entre otras cosas, priorizar la atención de emergencia a los desplazados y garantizar su acceso a los servicios sociales básicos. Entre 1997 y el 2004, el flujo de desplazados aumentó sustancialmente, pasó de 20,000 a comienzos del período, a 120,000 en los años intermedios y a 60,000 al final del 2003. Este aumento estuvo acompañado de un incremento en la frecuencia de aparición la palabra *desplazado* y sus variantes. Aparentemente la prensa escrita reaccionó al incremento en el flujo de personas desplazadas por la violencia. La reacción fue menor, si se quiere, que en los ejemplos anteriores, pero fue notable en todo caso.

Después del 2004, el comovimiento de las series ha sido menos evidente. Las fluctuaciones en la frecuencia poco tuvieron que ver con las fluctuaciones en el flujo de desplazados. Además, la caída del flujo posterior al 2007 no estuvo acompañada de una caída consustancial en la frecuencia. Probablemente porque el *stock* de desplazados siguió creciendo a pesar de la disminución en el flujo o porque la aplicación de la sentencia de la Corte Constitucional fue una fuente ocasional de noticias, en parte por las polémicas constantes entre el Gobierno y la misma Corte.

Antes del 2004, el coeficiente de correlación entre el flujo de desplazados y la frecuencia de la palabra *desplazado* (y sus variantes) fue de 0.60 tanto para *El Tiempo* como para *Semana*. Después del 2004, el coeficiente de correlación cayó a 0.15 para *El Tiempo* y a -0.08 para *Semana*. El coeficiente de correlación de las dos

series de frecuencias de *desplazado* fue de 0.69 para todo el período, de 0.88 para el período inicial (antes del 2004) y de 0.64 para el período final (después del 2004). En general, *El Tiempo* y *Semana* tuvieron un cubrimiento similar del fenómeno en cuestión: el promedio de las frecuencias fue semejante y los patrones temporales fueron también parecidos.

En suma, los mayores flujos de desplazados sí suscitaron el interés de la prensa nacional. El interés creció rápidamente con el aumento de los flujos y no cayó con su disminución. Aparentemente el cubrimiento respondió tanto a los flujos como a la cantidad total de desplazados.

Clima: *El Niño* y *La Niña*

Las noticias de prensa capturan la dinámica del clima y pueden ser usadas como una medida indirecta del impacto de algunos fenómenos climáticos globales. El gráfico 12 presenta, a la izquierda, la temperatura del Pacífico ecuatorial en la llamada zona 3.4[8] y, a la derecha, la frecuencia de las palabras *sequía* y *verano* en los archivos de *El Tiempo*. Los datos corresponden a los promedios móviles de doce meses. Ambas figuras muestran, mediante líneas verticales, las últimas cuatro apariciones de fenómeno El Niño, un calentamiento atípico de la temperatura del océano Pacífico[9].

Los aumentos de la temperatura del Pacífico afectan los patrones de lluvia y generan períodos prolongados de sequía, con consecuencias conocidas sobre las cosechas, el volumen de los embalses, etc. El gráfico muestra que la frecuencia de aparición de las palabras sequía y verano aumentó de manera notable durante los períodos en que ocurrió el fenómeno El Niño. El aumento fue especialmente notorio en dos momentos: a finales de los años noventa y al final del período de análisis, en el año 2010.

El gráfico 13 repite el análisis anterior, ya no para los aumentos de temperatura sino para las disminuciones, esto es, ya no para los períodos de El Niño sino de La Niña[10]. Los resultados son similares. La figura de la izquierda ilustra los períodos de caída en la temperatura, la de la derecha, la frecuencia de las palabras *inundación* e *invierno*. De nuevo, los aumentos de la frecuencia de las palabras en cuestión coincidieron con la llegada de La Niña. El aumento al final del período de análisis es particularmente notable, refleja el mucho mayor impacto del último evento de La Niña, ocurrido en la primera mitad del año 2011. El gráfico sugiere la existencia de un evento extremo en comparación con los eventos precedentes.

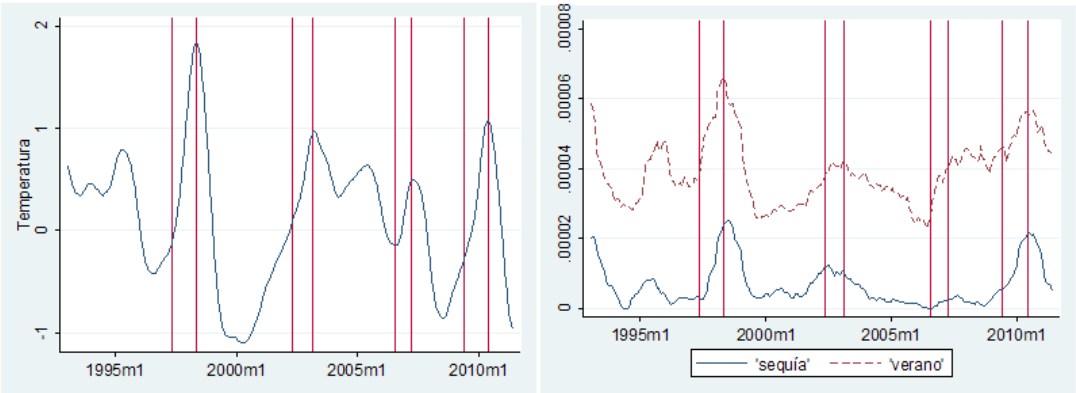
---

8 La Zona 3.4 está ubicada entre la latitud 5 N y 5 S y entre el meridiano 170 y el 120. Usualmente la temperatura de esta zona se usa para monitorear la presencia de los fenómenos del Niño y la Niña.

9 Ver, por ejemplo, http://es.wikipedia.org/wiki/El_Niño.

10 Ver http://es.wikipedia.org/wiki/La_Niña_clima.

Gráfico 12:
El Niño, *sequía* y *verano* en *El Tiempo*



El gráfico 14 muestra los comovimientos entre la temperatura del Pacífico ecuatorial y la frecuencia de la palabra *sequía*. La relación es evidente, sorprendente incluso. El coeficiente de correlación es de 0.61 para todo el período. Si la frecuencia de *sequía* da una idea indirecta sobre el impacto de las distorsiones climáticas —a mayor impacto, más noticias—, el gráfico 6 sugiere que el impacto de El Niño no ha aumentado durante las últimas dos décadas. Todo lo contrario. La pendiente de la línea es negativa y significativamente diferente de cero. El evento del 2010 tuvo un impacto considerable, pero no implica, por sí solo, un agravamiento de los efectos económicos, sociales y ambientales de las sequias.

El gráfico 15 repite el análisis anterior para la frecuencia de la palabra *inundación*. El coeficiente de correlación es de nuevo alto, de –0.54 para todo el período. Las conclusiones son en este caso opuestas a las del caso anterior. La pendiente de la línea de regresión es positiva y estadísticamente significativa. Los datos parecen consistentes con la idea de un agravamiento gradual, no espectacular pero sí notable. Esta conclusión depende, sin embargo, del evento extremo del 2011 y no debería considerase como definitiva.

Gráfico 13:
La Niña, *inundación* e *invierno* en *El Tiempo*

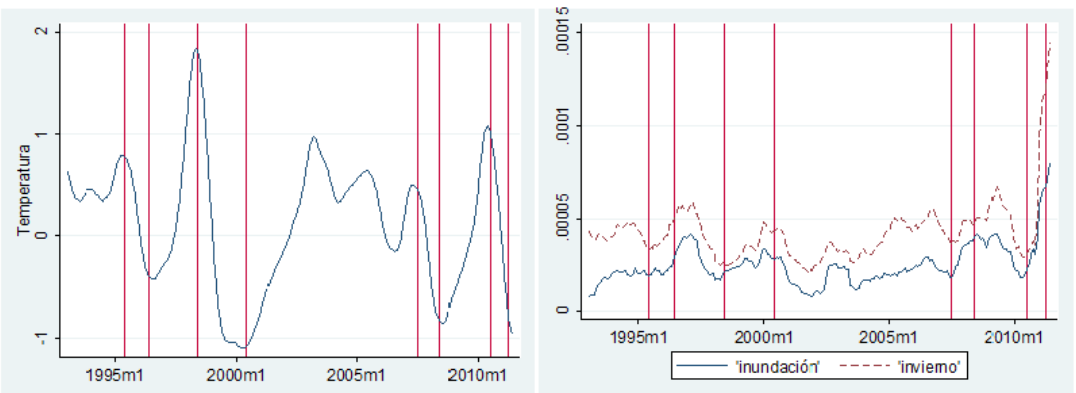


Gráfico 14.
Temperatura y *sequía*

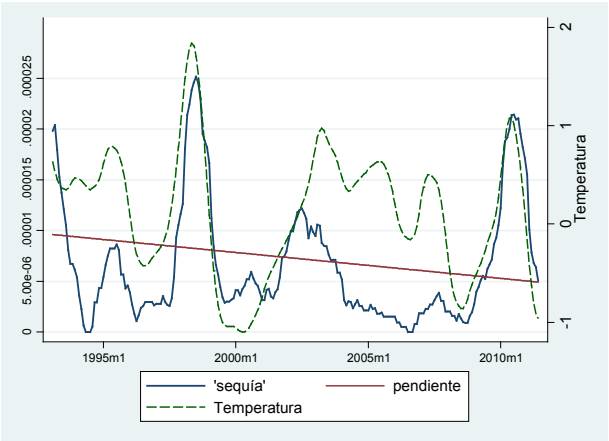
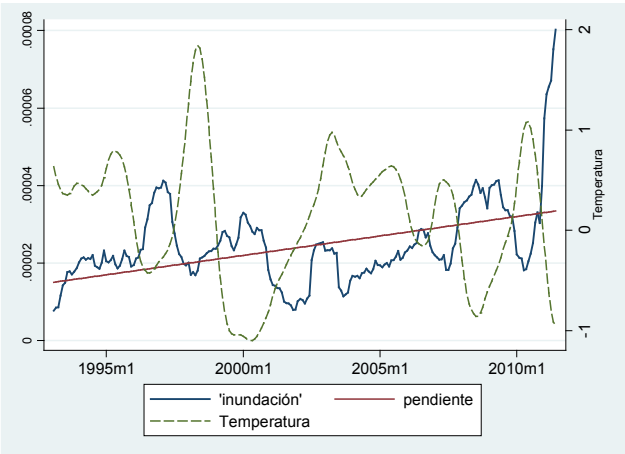


Gráfico 15.
Temperatura e *inundación*



En suma, las frecuencias de las palabras estudiadas dan una idea del impacto general de los eventos climáticos durante las dos últimas décadas. El análisis no es concluyente. La evidencia no sugiere un empeoramiento de las sequias, pero sí de las inundaciones. Sin embargo, los resultados dependen de los eventos extremos de finales del período. Sea lo que sea, la frecuencia de noticias es una forma útil de medir, al menos preliminarmente, el impacto de los eventos climáticos.

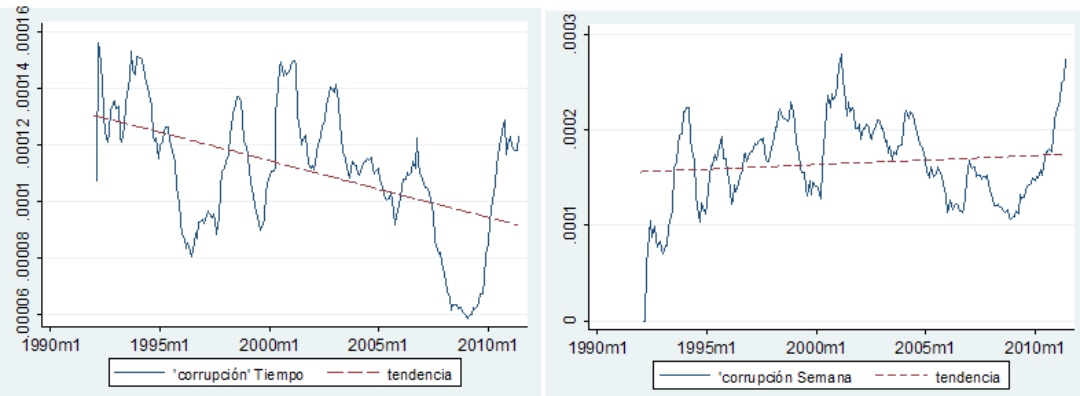
5. Veinte años, cinco historias: realidades como palabras

Esta sección presenta cinco ejemplos, cinco estudios de caso que ejemplifican el uso de *culturomics* como metodología relevante de cuantificación de fenómenos que, por su misma naturaleza, son difíciles de cuantificar. Los ejemplos estudian varios temas centrales de la realidad colombiana de los últimos 20 años: la corrupción, la guerra, el optimismo económico y el equilibrio de poderes (los congresistas frente a los jueces y el presidente frente a los mandatarios locales). Dadas las dificultades obvias de medición, los temas estudiados no han sido cuantificados de manera sistemática. Ninguno de ellos cuenta con indicadores conocidos y respetados. Las comparaciones entre indicadores y frecuencias no son por lo tanto posibles. Las frecuencias son, en este caso, los indicadores: la forma imperfecta de cuantificar los cambios y las tendencias de la corrupción, el conflicto, el entusiasmo y la distribución del poder.

5.1. Corrupción

El gráfico 16 muestra, para el período comprendido entre enero de 1992 y julio del 2011, la evolución de la frecuencia de la palabra *corrupción* y sus accidentes[11]. Las series mostradas corresponden a los promedios móviles de doce meses: inicialmente se calcularon las frecuencias mensuales y seguidamente los promedios móviles anuales. El gráfico presenta, separadamente, las frecuencias correspondientes al diario *El Tiempo* y a la revista *Semana*. En la revista *Dinero*, el número de noticias es relativamente menor y las apariciones de la palabra *corrupción* son casi despreciables.

Grafico 16:
Corrupción* en *El Tiempo* y *Semana



Ambas figuras cuentan una historia similar. Ambas revelan, por ejemplo, grandes fluctuaciones alrededor de una tendencia más o menos horizontal. En *El Tiempo*

[11] El análisis muestra la frecuencia conjunta de las palabras corrupción, corrupta, corruptas, corrupto y corruptos. Las conclusiones del análisis no cambian en absoluto si se incluyen otras palabras relacionadas como desfalco, peculado, robo al erario, etc.

(figura de la izquierda), la tendencia es negativa; en *Semana* (figura de la derecha), es positiva. Pero más allá de estas diferencias, el gráfico sugiere, en esencia, una considerable inercia de la corrupción: los escándalos ocurren cada cierto tiempo pero no parece existir una tendencia clara. En suma, la corrupción ha sido fluctuante en el corto plazo pero constante desde una perspectiva de más largo plazo[12]. Todo cambia y todo sigue igual.

El conteo de noticias, opiniones y comentarios no es un indicador perfecto de la corrupción. Este indicador está sesgado por los eventos más costosos o por algunos casos que concentran, por razones muchas veces fortuitas, la atención de la opinión pública. En algunas coyunturas específicas, el indicador propuesto recoge los sesgos ideológicos o los intereses políticos de los directores y editores de los medios de comunicación estudiados. En fin, los cuestionamientos abundan. Pero este tipo de análisis no debería descartarse fácilmente. En cierta medida, equivale a simple un ejercicio memorístico —contar para recordar—, a una forma de contrarrestar los juicios impresionistas del presente con los juicios del pasado, de comparar la indignación de hoy con la de ayer.

Como se dijo en la sección 2, un indicador similar fue usado por Goldin y Glaeser (2001) para estudiar la estudiar la evolución de la corrupción en los Estados Unidos desde un horizonte de largo plazo. Más recientemente, Goel, Nelson y Naretta (2011) usaron la frecuencia de búsqueda de la palabra *corrupción* en Internet para hacer comparaciones entre países. Los indicadores tradicionales de corrupción están basados en opiniones, las cuales, en la mayoría de los casos, están influenciadas por el cubrimiento de la prensa. Los indicadores aquí propuestos están basados en la intensidad del cubrimiento, en la idea de que la cambiante realidad de un fenómeno complejo puede cuantificarse, en cierta medida al menos, con base en su cubrimiento mediático.

Volviendo al gráfico 16, hay un hecho peculiar que merece un comentario aparte. En ambas figuras, tanto en la de *El Tiempo* como en la de *Semana*, la frecuencia de la palabra corrupción cayó de manera notable entre finales del 2005 e inicios del 2010, y aumentó seguidamente de manera también notable. Aparentemente los medios analizados se desentendieron de la corrupción durante buena parte del segundo período del ex presidente Uribe (2006-10) y luego, como si tuvieran que ponerse al día, volvieron a preocuparse por el tema con una intensidad renovada. Después de una calma de varios años, vino la tempestad mediática de los meses recientes.

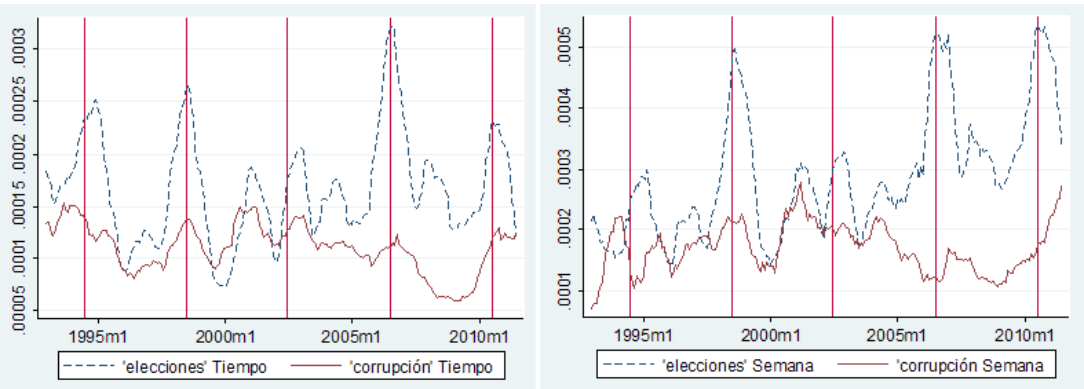
---

[12] En cada momento del tiempo, abrumados por los eventos de la coyuntura, los comentaristas políticos tienden a percibir la corrupción presente como la peor en mucho tiempo, tienden, en otras palabras, a confundir las fluctuaciones con la pendiente. En diciembre de 1997, un reconocido periodista escribió: "nunca antes el país había presenciado tan impresionante sucesión de hechos escandalosos. Trátese de peculados o desfalcos en entidades del Estado; de narcomicos en el Congreso; de testaferratos o de simple venalidad administrativa, el panorama de la corrupción en Colombia es francamente desolador". En octubre del 2011, un periodista distinto escribió lo mismo: "Lo que se robó en Colombia en los últimos años no tiene antecedentes y no es que fuéramos el paraíso anti-corrupción". El presentismo domina las opiniones sobre la corrupción

Las razones de este comportamiento no son fáciles de precisar. Pero el gráfico 17 da algunas pistas al respecto. El gráfico muestra conjuntamente las frecuencias de las palabras *corrupción* y *elecciones,* y señala las fechas de las elecciones presidenciales ocurridas durante el período de análisis. Claramente la frecuencia de *corrupción* aumentó cíclicamente en los meses previos y posteriores a las elecciones presidenciales: subió y cayó coordinadamente con la frecuencia de *elecciones*. Esta regularidad mediática tuvo una excepción notable: las elecciones del año 2006, precisamente las únicas elecciones de todo el período en las cuales el presidente en ejercicio fue candidato[13].

Durante los meses que antecedieron y sucedieron las elecciones presidenciales del 2006, las noticias, comentarios y opiniones sobre la corrupción (una medida indirecta de la intensificación de las denuncias y los debates al respecto) no aumentaron de manera sustancial como lo habían hecho en el pasado durante períodos similares. Pero en las elecciones del 2010, ya con el presidente en ejercicio por fuera de la contienda, todo pareció volver a la normalidad: la "corrupción" creció sustancialmente antes y después de las elecciones. En apariencia las denuncias y debates que se habían postergado salieron a flote súbitamente. En suma, más que un aumento permanente de la corrupción, el crecimiento súbito de la frecuencia noticiosa al final del período de análisis podría indicar una suerte de actualización, de desfogue.

Gráfico 17.
corrupción* y *elecciones



Pero más allá de los ciclos y las fluctuaciones temporales, los datos sugieren que la corrupción permaneció más o menos constante durante los últimos veinte años. Al menos, la frecuencia de *corrupción* no muestra una tendencia clara, ni positiva ni negativa. Las variaciones fueron muchas, pero la tendencia no cambió notablemente.

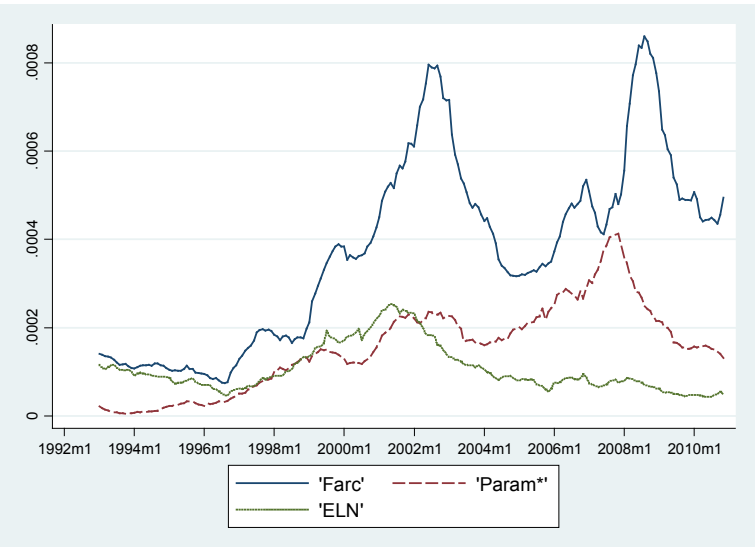
[13] En El Tiempo el coeficiente de correlación entre corrupción y elecciones fue de 0.60 antes del 2002 y de 0.35 después. En Semana fue de 0.36 y de -0.07 respetivamente.

5.2. Conflicto

El conflicto colombiano concentró la atención de los medios de comunicación durante los últimos 20 años. El fortalecimiento de los distintos grupos armados durante la primera mitad de los años noventa, las posteriores negociaciones con las FARC, la subsecuente ofensiva militar, los acuerdos con los paramilitares y los rescates de los secuestrados, produjeron muchas noticias, comentarios y editoriales de prensa. En teoría, al menos, las frecuencias de aparición de las palabras *FARC*,*ELN* y *paramilitares* ilustran la manera como los medios de comunicación dieron cuenta de la cambiante realidad del conflicto colombiano. Las palabras permiten, en suma, tomarle el pulso a la obsesión mediática con el conflicto.

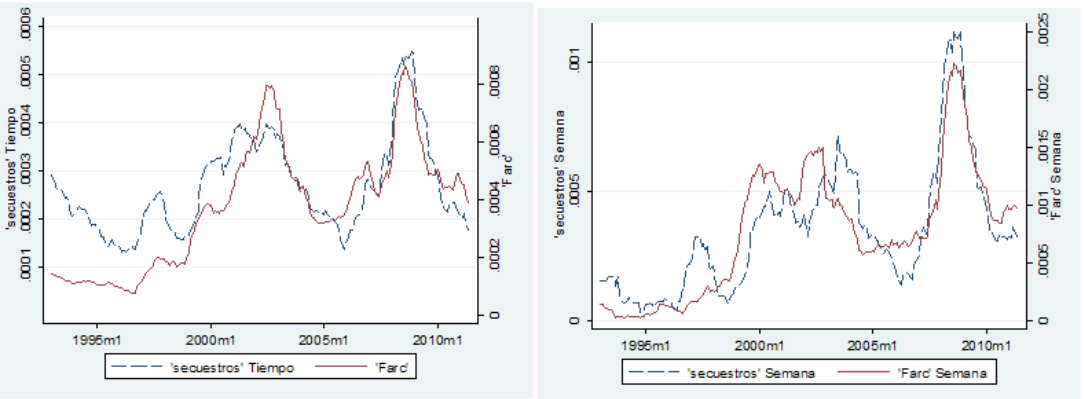
El gráfico 18 muestra la frecuencia mensual de las palabras *FARC*, *ELN* y *paramilitares* para el período 1992-2011[14]. El análisis corresponde en este caso a los archivos del diario *El Tiempo* (el análisis conjunto de los tres medios disponibles es casi idéntico). La frecuencia de la sigla *FARC* supera ampliamente, en más de diez veces, las frecuencias de palabras como *desempleo* y *corrupción*. Supera incluso las de expresiones genéricas como *congreso* y *elecciones*. Las frecuencias de las palabras *ELN* y *paramilitares* son relativamente menores, pero no despreciables. En general la importancia mediática del conflicto fue enorme. Las FARC tuvieron dos momentos de ebullición mediática (en el 2002 y el 2009), los paramilitares uno (en el 2008) y el ELN otro en el 2001. La desaparición de la frecuencia de aparición de *ELN* fue gradual y continua; la de *paramilitares*, mucho más abrupta.

Gráfico 18.
FARC, ELN* y *paramilitares* en *El Tiempo



[14] La frecuencia de *paralimitare*s corresponde a la frecuencia de sus accidentes y del mismo término en inglés: *paramilitar, paramilitares, paramilitaries, paramilitarism, paramilitarismo, paramilitarizado y paramilitary.*

Gráfico 19. *FARC* y *secuestros*



El gráfico 19 da algunas pistas sobre las causas de los grandes altibajos en el cubrimiento mediático de la guerrilla de las FARC. El gráfico muestra, tanto para *El Tiempo* como para *Semana*, el cambio mensual de las frecuencias de las palabras *FARC* y *secuestros*. Las coincidencias son enormes. Ambas series se mueven de manera casi sincrónica. El coeficiente de correlación es de 0.78 de para *El Tiempo* y de 0.87 para *Semana*. La evidencia indica que la visibilidad de las FARC estuvo asociada esencialmente al tema del secuestro. Los rescates y las liberaciones, en particular, parecen haber generado todo tipo de noticias, reacciones y comentarios que, en conjunto, aumentaron de manera sustancial la visibilidad mediática de este grupo. De nuevo, el gráfico sugiere que los secuestros (y los secuestrados) garantizaron a las FARC una gran visibilidad a pesar de su debilitamiento militar. La fórmula *FARC* = *secuestros* resume adecuadamente esta historia mediática.

En resumen, los secuestros de las FARC fueron el tema predominante en el cubrimiento del conflicto colombiano, El cubrimiento tuvo dos o tres momentos de ebullición, pero, en general, el interés fue sostenido por al menos una década.

5.3. Bonanza

El gráfico 20 muestra la frecuencia de las palabras *bonanza* y *boom* en el archivo de *El Tiempo*: las conclusiones no cambian si se incluyen los otros medios disponibles. En principio, esta serie mide, de manera indirecta, el entusiasmo colectivo ante las buenas noticias económicas originadas, por ejemplo, en un descubrimiento petrolero o minero, o en un aumento sustancial en los precios de los principales productos de exportación. Los datos sugieren que el mayor entusiasmo colectivo de las últimas dos décadas ocurrió entre 1993 y 1995 como consecuencia de los hallazgos petroleros de Cusiana y Cupiaga. La prensa reaccionó mucho más fuertemente ante el descubrimiento de un nuevo yacimiento que ante los altos precios del petróleo y del carbón de los últimos años. Este resultado sugiere, en últimas, la existencia de una realidad sociológica relevante (de un sentimiento colectivo de abundancia, en este caso) que pudo haber incidido sobre las decisiones públicas y privadas.

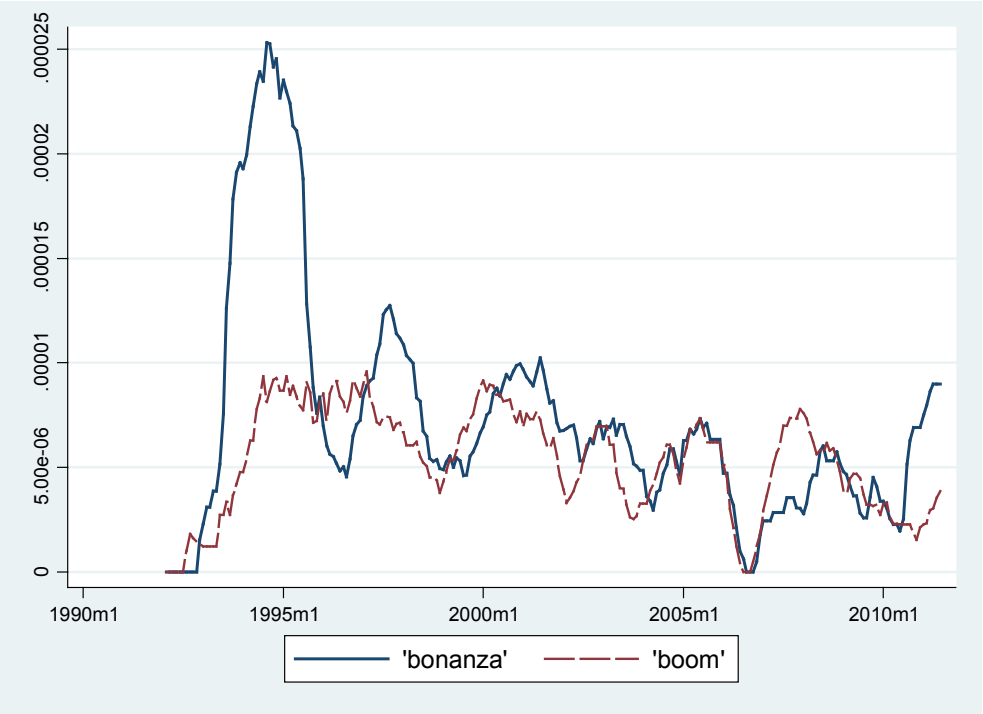
En teoría, los descubrimientos petroleros crearon una sensación de abundancia de recursos y de ausencia de restricciones, impulsaron un auge en el consumo público y privado y pudieron, incluso, haber sembrado la semilla de la crisis de finales de los años noventa, la peor en la historia moderna del país (Echeverry, 1996). Más allá de las consecuencias, la evidencia sugiere que en la primera mitad de los años noventa, más que en cualquier otro momento de la últimas dos décadas, la idea de una bonanza o de un *boom* económico capturó la imaginación de mucha gente. La prensa puede a veces servir de termómetro del entusiasmo colectivo.

5.4. División de poderes

La Constitución de 1991 redefinió la estructura de poder en Colombia. Formalmente, la descentralización le otorgó mayor poder a los departamentos y a los municipios. De la misma manera, la independencia del Banco de la República y la creación de la Corte Constitucional le restaron poder a la rama ejecutiva. Pero los cambios institucionales no siempre tienen consecuencias reales. La estructura del poder no sólo depende de la constitución o de las instituciones formales. Otras factores, económicos y sociológicos, pueden ser determinantes.

La frecuencia de aparición de algunas palabras puede dar alguna idea sobre los cambios reales (no formales) en la estructura de poder. Por ejemplo, si la frecuencia de las palabras *alcaldía* y *gobernación* aumenta con respecto a la de *presidencia*, podría hablarse de un mayor protagonismo político de los poderes territoriales o de una mayor visibilidad de los mandatarios locales y, por lo tanto, de un profundización efectiva de la descentralización que trasciende los meros cambios institucionales. Asimismo, si la frecuencia de la palabra *magistrado* (y sus accidentes) aumenta con relación a la de la palabra *congresista* (y sus accidentes), podría hablarse de una transferencia de poder hacia el poder judicial.

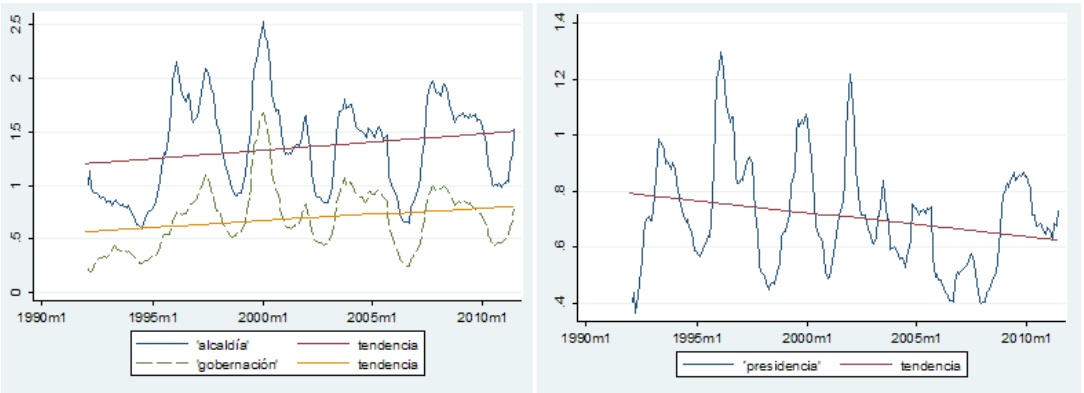
Gráfico 20. *Bonanza* y *boom* en *El Tiempo*



El gráfico 21 muestra, para el período 1992-2011, la frecuencia de las palabras *alcaldía*, *gobernación* y *presidencia*. Los datos corresponden al diario *El Tiempo*. Las series fueron normalizadas con base en las menciones a la palabra *elecciones* con el propósito de corregir por los ciclos electorales: la frecuencia de las palabras en cuestión tiende a aumentar, por razones obvias, durante los períodos de elecciones. Los resultados muestran, por una parte, un aumento tendencial en las frecuencias de *alcaldía* y *gobernación* y, por otra, una disminución en la frecuencia de *presidencia*. Las pendientes son estadísticamente significativas en cada una de las tres gráficas.

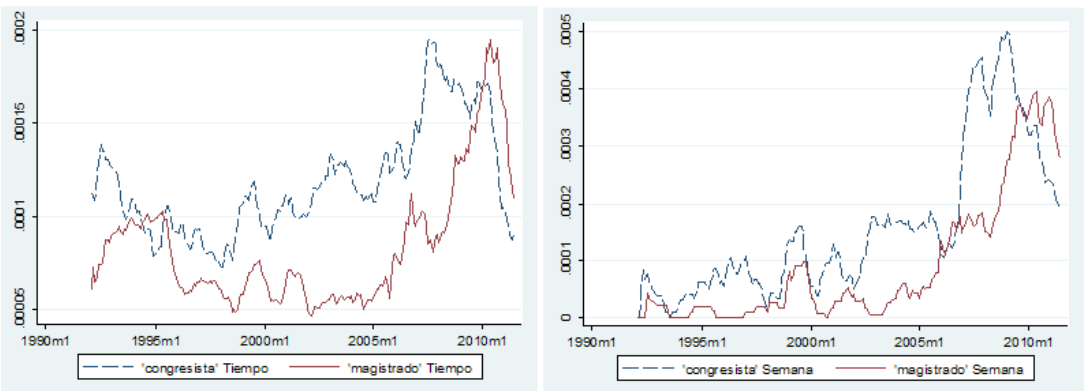
Este resultado sugiere, en últimas, que la descentralización sí vino acompañada de una mayor visibilidad mediática de los centros regionales de poder. La mayor visibilidad puede indicar, a su vez, una transferencia real de poder de la nación a las regiones o puede simplemente mostrar un mayor interés de la prensa nacional por la suerte de los municipios y departamentos. Sea lo que sea, el gráfico brinda información, en principio relevante, sobre una faceta no estudiada de la descentralización.

Gráfica 21. *Alcaldía, gobernación* y *presidencia* en *El Tiempo*



El gráfico 22 muestra la frecuencia de las palabras *magistrados* y *congresistas*. La serie de la izquierda corresponde a *El Tiempo*, la de la derecha a *Semana*. Ambas series cuentan una historia similar: un aumento sostenido de la frecuencia de aparición de magistrado y sus accidentes desde mediados de la década anterior. En el 2010, por primera vez durante las últimas dos décadas, la frecuencia de *magistrado* superó la de *congresista*. El mayor protagonismo mediático de los magistrados probablemente tuvo mucho que ver con los escándalos de la parapolítica y de las interceptaciones telefónicas. Pero puede también reflejar un cambio estructural, no asociado solamente a una coyuntura específica: la mayor injerencia de los magistrados en las decisiones públicas.

Gráfico 22. *Magistrados* y *congresistas*



En los últimos meses, ambas series cayeron abruptamente, pero, al mismo tiempo, la preeminencia mediática de los magistrados se mantuvo. En general, el resultado sugiere un cambio significativo de la estructura de poder en Colombia[15].

[15] En una entrevista publicada en El Espectador (18/12/2010) el abogado y columnista Yesid Reyes hizo una interesante observación sobre la vida pública de su padre, el presidente de la Corte Suprema, Alfonso Reyes Echandía, inmolado en la toma y retoma del Palacio de Justicia: "la exposición de mi padre a la prensa en el año 1985, cuando era el presidente de la corporación, fue mínima. No tengo idea de cuántas veces saldría en la prensa pero en todo caso no fueron más de tres o cuatro: dos de ellas antes de morir, durante la toma del Palacio".

6. Conclusiones

Este artículo presenta un análisis preliminar de algunos aspectos de la realidad colombiana basado en el conteo de palabras en tres periódicos de circulación nacional. El análisis tiene una dificultad obvia: no es neutral; la descripción necesariamente incorpora los sesgos de los editores y comentaristas de los periódicos bajo escrutinio. La sección 4 muestra, sin embargo, que el análisis describe adecuadamente la cambiante realidad de algunos fenómenos socioeconómicos. La sección 5 muestra, de otro lado, que el análisis permite describir la dinámica de otros fenómenos que, por su misma naturaleza, son difíciles de medir o cuantificar. Las descripciones no son definitivas, pero plantean preguntas interesantes, sugieren hipótesis no triviales y pueden servir de punto de partida para investigaciones posteriores.

En esencia, este artículo describe una base de datos mediante una serie de ejemplos que, en conjunto, dan algunas luces sobre las transformaciones económicas y sociales ocurridas en Colombia durante los últimos veinte años. Pero el objetivo es más ilustrativo que descriptivo, más de forma que de fondo. Más que medir o incluso explicar algunos fenómenos socioeconómicos, el artículo quiere mostrar la utilidad de una metodología novedosa, de una nueva herramienta de investigación de las ciencias sociales.

Este es el primer artículo de *culturomics* para Colombia. Pero no será el último. Algunas ideas sobre posibles investigaciones o análisis posteriores son obvias. Temáticamente, los trabajos futuros podrían retomar algunos de los temas aquí planteados: la corrupción, el cubrimiento periodístico de las políticas económicas, los determinantes del cubrimiento del conflicto, etc. Podrían también explorar algunos temas adicionales: las relaciones internacionales, las percepciones de inseguridad, el cubrimiento relativo de las distintas regiones, el papel del Banco de la República, etc.

Metodológicamente, las posibilidades de investigaciones futuras son variadas. Valdría la pena, por ejemplo, estudiar la coexistencia de dos o más palabras en los artículos y comentarios de prensa. Este tipo de enfoque permitiría ir más allá del simple análisis de series de tiempo y brindaría información relevante sobre relaciones causales entre las variables de interés. Por ejemplo, valdría la pena conocer la medida en que las palabras *regalías* y *corrupción* (o *crisis* y *pobreza* o *salario mínimo* e *inflación*) vienen juntas en los artículos estudiados. En otras palabras, del simple análisis univariado se podría pasar al multivariado.

También sería útil estudiar el tono de la información. El conteo no discrimina entre la cobertura positiva o negativa, mucho menos entre las muchas posibles variaciones en el tono de las noticias y comentarios. Convendría, por ejemplo, analizar el tono de la cobertura mediática de una institución determinada (el Banco de la República), de una figura política (el presidente Uribe) o de un país (Venezuela). Convendría, en últimas, complementar el análisis de frecuencias con información sobre el sentido y tono de la cobertura.

Finalmente, valdría la pena estudiar las diferencias entre noticias y opinión. Las noticias y las opiniones pueden reflejar la realidad de manera diferente y podrían estudiarse separadamente. Este tipo de análisis daría algunas luces sobre los sesgos de los medios y los cambios en la opinión publicada. Por último, esta metodología podría combinarse con encuestas de opinión con el fin de analizar las interacciones, no siempre obvias, entre opinión pública y publicada.

Referencias

Cowell, F. A., E. Flachaire y S. Bandyopadhyay (2011), “Inequality, Entropy and Goodness of Fit”, Document de Travail n°2011-23, Groupement de Recherche en Economie Quantitative d'Aix-Marseille - UMR-CNRS 6579, Ecole des Hautes études en Sciences Sociales.

Echeverry, J. C. (1996), "The Fall in Colombian savings during the 1990s. Theory and evidence", en *Borradores de economía 3593*, Banco de la República.

Glaeser, E. L. y C. Goldin (2006), "Corruption and Reform: Introduction", en Corruption and Reform: *Lessons from America's Economic History*, National Bureau of Economic Research, pp. 2-22.

Jones, M. P. y J. Crowley (1989, marzo), “A General Class of Nonparametric Tests for Survival Analysis”, en *Biometrics*, vol. 45, núm. 1, pp. 157-170.

Michel, J. B., Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, The Google Books Team, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. A. Nowak y E. L. Aiden (2010, diciembre), “Quantitative Analysis of Culture Using Millions of Digitized Books”, *Science*, 331, 176 (2011).

Goel, R. K., M. A. Nelson y M. A. Naretta (2011, septiembre), “The internet as an indicator of corruption awareness”, en *European Journal of Political Economy*, en imprenta.

Anexo

Gráfico A1. ***10 – tasa de crecimiento del PIB y recesión en El Tiempo***