La ley de Benford y su aplicabilidad en el análisis forense de resultados electorales

Gonzalo Castañeda*

Resumen: En este artículo se analiza la viabilidad de la ley de Benford para el estudio forense de la detección de fraudes electorales. De acuerdo con esta ley, los dígitos iniciales de un conjunto de números siguen una distribución logarítmica cuando los datos no han sido perturbados. La generación de esta *ley* en datos socioeconómicos y de otra índole depende de la presencia de un doble proceso aleatorio: eventos de una distribución y distribuciones de probabilidad elegidas de un conjunto reducido. A partir de un modelo basado en agentes se muestra que esta *ley* no ofrece una prueba robusta para distinguir entre elecciones limpias y las que han sido manipuladas. En dicho modelo las preferencias partidistas se modifican a partir del contagio social, y los parámetros son calibrados con datos de las elecciones mexicanas de 2006.

Palabras clave: ley de Benford, análisis de fraude electoral, modelos basados en agentes, preferencias políticas, elecciones mexicanas de 2006.

Benford's Law and its Applicability in the Forensic Analysis of Electoral Results

Abstract: This article analyses the viability of Benford's Law in the forensic study of the detection of electoral frauds. According to this law, the initial digits of a set of numbers follow a logarithmic distribution as long as the data has not been disturbed. Implementation of this *law* in socioeconomic data —or in other types of data— depends on the presence of a two-fold randomized process: events of a distribution, and probability distributions selected from a reduced group. Through agent-based models, it is demonstrated that this *law* does not offer a robust test to distinguish between clean elections and elections that have been manipulated. In the model-in-question, partisan preferences are modified by means of social transmission and the parameters are calibrated with data from the Mexican 2006 elections.

Keywords: Benford's Law, analysis of electoral fraud, agent-based models, political preferences, Mexican 2006 elections.

---

*Gonzalo Castañeda es profesor-investigador del Centro de Estudios Económicos de El Colegio de México. Camino al Ajusco 20, Pedregal de Santa Teresa, México D.F., 10740, Apartado Postal 20671. Tel. (52 55) 54 49 30 00, fax (52 55) 56 45 04 64. Correo electrónico: sociomatica@hotmail.com.

Artículo recibido en julio de 2009 y aceptado para su publicación en octubre de 2010.

Introducción

Al percatarse de que las páginas de los primeros dígitos en las tablas de logaritmos estaban más desgastadas que las páginas de los últimos dígitos, el astrónomo y matemático Simon Newcomb descubrió, en 1881, que los dígitos iniciales significativos de los números (*i.e.* excluyendo el cero) no se distribuían de manera uniforme. Dado que estas tablas eran utilizadas por científicos de diferentes disciplinas, Newcomb conjeturó que este fenómeno debía estar presente en bases de datos provenientes de distintos ámbitos de la vida. Pero fue en 1938, cuando el físico Frank Benford redescubrió el fenómeno en 20 muestras de diferentes fuentes, que se aportó evidencia rigurosa sobre la presencia recurrente de la distribución logarítmica de los dígitos (Hill, 1998). Entre las bases de datos que mostraban esta frecuencia relativa se encontraban las siguientes: cuentas de electricidad, área de los ríos, peso atómico de los elementos químicos, números de los inmuebles en las calles, número de habitantes en las poblaciones, estadísticas de la liga americana de beisbol, número de defunciones en desastres.

Una aplicación importante de la ley de Benford (o ley del primer dígito) se encuentra en la detección de fraudes fiscales (Nigrini, 1996; Durtschi *et al*., 2004), por lo que varios analistas han sugerido que esta prueba forense podría utilizarse para detectar la posibilidad de manipulación en otros tipos de datos socioeconómicos (Varian, 1972; Diekmann, 2004). La inclusión de datos falsos en un conjunto de números suele llevarse a cabo mediante una distribución uniforme y, por lo tanto, el comportamiento anómalo en este conjunto puede detectarse al comparar la frecuencia empírica de estos números con la distribución teórica asociada con la ley de Benford.

En años recientes algunos autores, liderados por Walter Mebane, han empleado esta metodología para detectar fraudes electorales. Por lo general, este procedimiento estudia el primer o segundo dígito inicial en la cuenta de votos recibidos por cada candidato, ya sea en la casilla o en las secciones electorales. Ejemplos de estos estudios forenses se presentan en Mebane (2006a, 2006b, 2007a, 2007b, 2007c y 2008), y Pericchi y Torres (2004).

En este artículo se cuestiona la aplicabilidad de la ley de Benford como prueba forense de los resultados electorales. Este cuestionamiento obedece a la dificultad de saber si, efectivamente, las decisiones de voto producen una distribución logarítmica en los primeros dígitos de las cuentas de las diferentes casillas o distritos. Para descifrar la naturaleza estocástica de las campañas electorales se utiliza un modelo basado en agentes (ABM, por sus

siglas en inglés), ya que los procesos electorales son caracterizados como sistemas adaptables complejos. En otras palabras, el modelo computacional aquí descrito plantea que las preferencias partidistas se ven condicionadas por la información local y global, y que la interacción social de los individuos modifica, a su vez, los clústeres partidistas y los sondeos nacionales de opinión pública. De esta forma las campañas virtuales que se simulan con un ABM permiten generar datos artificiales para calcular las frecuencias relativas de los dígitos iniciales significativos y así contrastarlas con la ley de Benford.

Además de esta introducción, el artículo se divide en cuatro secciones. En la segunda se formaliza el concepto de la ley de Benford y se presentan estudios que aplican este tipo de análisis forense a datos electorales mexicanos. En la tercera sección se hace una breve síntesis de las premisas y los mecanismos del ABM utilizado, lo que de entrada es ya una contribución original para el estudio de preferencias partidistas durante una campaña electoral. En la cuarta sección se presentan simulaciones en las que la validación (rechazo) de la ley de Benford no es del todo consistente con la presencia de elecciones virtuales limpias (trucadas). En cambio, en las conclusiones se plantea que las simulaciones con ABM y sus patrones emergentes sí pueden ser empleados como una prueba forense alternativa en un análisis electoral.

La ley de Benford, sus limitaciones y aplicación en el caso mexicano

Resulta muy sencillo establecer la distribución teórica para el dígito ubicado en la *k*-ésima posición (de izquierda a derecha) de números generados de acuerdo con un cierto proceso estadístico. En particular, la frecuencia relativa que caracteriza la ley de Benford para el primer dígito significativo (1-BL) se describe de la siguiente manera:

$$\text{Prob}\,(d_1) = \log_{10}\left(1 + \frac{1}{d_1}\right), \quad d_1 = 1, 2, 3, \ldots, 9 \qquad (1)$$

de este modo, el dígito 1 tiene una probabilidad de 0.301 mientras que el dígito 9 tiene una probabilidad de sólo 0.0458. Las probabilidades restantes se presentan en el cuadro A.1 del apéndice A, en el que también se describen las distribuciones de la ley de Benford para diferentes dígitos.

Asimismo, la distribución teórica que caracteriza la ley de Benford para el segundo dígito está dada por la siguiente expresión:

$$\text{Prob}(d_2) = \sum_{k=1}^{9} \log_{10}\left(1 + \frac{1}{10k + d_2}\right), \quad d_2 = 0, 1, 2, \ldots\ldots, 9 \tag{2}$$

cabe notar que para la distribución 2-BL existe una probabilidad positiva para el dígito 0, que es igual a 0.11968, dado que el cero sí puede presentarse en la segunda posición inicial de un número.

Finalmente, la ley del dígito-significativo con que se generaliza la ley de Benford en términos de una densidad conjunta de los dígitos en las primeras *k* posiciones iniciales se define de la siguiente manera (Hill, 1995):

$$\text{Prob}(D_1 = d_1, \ldots\ldots, D_k = d_k) = \log_{10}\left[1 + \left(\sum_{i=1}^{k} d_i \bullet 10^{k-i}\right)^{-1}\right] \tag{3}$$

por ejemplo, Prob ($D_1 = 1, D_2 = 2, D_3 = 9$) = $\log_{10}$ [1 + (1 x 10^{3-1} + 2 x 10^{3-2} + 9 x 10^{3-3})$^{-1}$] = $\log_{10}$ [1 + (129)$^{-1}$] ≈ 0.00335.

También se puede demostrar que la distribución del *k*-ésimo dígito significativo (D_k) se aproxima muy rápidamente a una distribución uniforme cuando la *k*-ésima posición se mueve hacia la derecha (Zhipeng, Ling y Huajia, 2004). Por medio de histogramas es fácil visualizar la presencia de distribuciones no-uniformes para la primera y segunda posición pero no así para las demás posiciones. De hecho, cuando k ≥ 3 las distribuciones asociadas tienen un valor medio muy cercano a 4.5 y una varianza cercana a 8.25, valores que coinciden con los de una distribución uniforme.

Limitaciones de la ley

Hill (1995) presenta una explicación bastante convincente de por qué es muy común observar la ley de Benford en tablas numéricas generadas con fenómenos muy diversos. De acuerdo con este autor, la distribución logarítmica se establece cuando el conjunto de números analizados proviene de un proceso aleatorio generado con una *mezcla estadística*. Es decir, una tabla de números sigue la ley de Benford cuando éstos son producidos a partir de

realizaciones aleatorias de un conjunto pequeño de distribuciones diferentes que, a su vez, son elegidas aleatoriamente.[1] A manera de ejemplo, este escenario prevalece en una tabla recopilada con números que aparecen en un periódico, ya que estos números provienen de diferentes fuentes de datos: estadísticas deportivas, información financiera, fechas de eventos, precios de productos anunciados.

En consecuencia, para que la ley de Benford prevalezca en datos electorales la decisión de votar tiene que ser explicada en términos de una colección de funciones de distribución. Por ejemplo, un voto emitido en determinada casilla debe provenir, en ciertas ocasiones, de un votante duro; en otras, de la decisión de un individuo con un compromiso partidista débil; mientras que otros más serían emitidos por independientes. En cualquiera de estos tres casos la decisión de votar por un candidato en particular está determinada por una distribución de probabilidad específica. De esta forma, el tipo de votante y la decisión que adopta son el resultado de un doble proceso de aleatorización. No obstante, no existe garantía de que el patrón emergente observado en los resultados electorales reales efectivamente se produzca a partir de este tipo de caracterizaciones.

Algunos autores han sugerido que las leyes de dígitos-significativos no son relevantes para datos electorales. Por ejemplo, Taylor (2005), al igual que otros politólogos cuantitativos, plantea que la decisión de votar puede describirse a partir de un esquema de lotería. De esta manera, la probabilidad que tiene un individuo de votar por un candidato se define en términos de un modelo multinomial y, por ende, existen diferentes distribuciones dependiendo de variables sociodemográficas, ideología y otros elementos que ayudan a categorizar a la población de votantes; por lo tanto, la cuenta total de votos es la suma de una serie de realizaciones de eventos aleatorios. Para el caso de Venezuela, Taylor muestra que la distribución simulada de los dígitos-significativos a partir de un modelo multinomial que incluye exclusivamente variables que reflejan un proceso electoral limpio es similar a la distribución observada en el referéndum real bajo estudio, pero discrepa de la referencia teórica establecida por Benford.

Sin embargo, cabe resaltar que no es válido simular una elección limpia por medio de una multinomial en la que las regresiones no son controladas por variables que indican el efecto distorsionador de un fraude. La omisión de estas variables como factores explicativos puede sesgar los parámetros

[1] Para la explicación de un proceso estocástico alternativo véase Pietronero *et al*. (2001).

estimados y, por lógica, la simulación que procede de este modelo no refleja adecuadamente las condiciones que prevalecen en una elección nítida. Asimismo, dicho modelo probabilístico no incorpora la *mezcla estadística* que conduce a la ley de Benford. Como bien lo apunta Mebane (2006a), esta ley requiere un conjunto pequeño de distribuciones, en contraposición al enorme conjunto que se genera con una multinomial.

Una segunda limitante de la aplicabilidad de la ley de Benford para datos electorales tiene que ver con el hecho de que la frecuencia de ciertos dígitos se ve restringida por determinadas reglas institucionales, por lo que dichos datos violan la 1-BL por construcción. Por ejemplo, cuando una casilla tiene como mucho 750 boletas, como en el caso mexicano, las cuentas de votos para un candidato que inician con los dígitos 8 y 9 tienen menor probabilidad de aparición que aquellas cuentas que inician con los dígitos restantes; en otras palabras, las cuentas con 800 y 900 votos son descartados por definición.

Como sugiere Brady (2005), un artefacto similar se introduce cuando las secciones electorales se diseñan de tal forma que incluyen aproximadamente el mismo número de votantes potenciales. Este escenario permite, por ejemplo, que en una contienda bipartidista cerrada haya muchas secciones en las que las votaciones presenten participaciones de 50 por ciento. En consecuencia, con secciones electorales de tamaño similar existen muchas cuentas en las que cada candidato recibe un número de votos que empieza con un determinado dígito; por ejemplo, con 1 cuando el total de votos posibles es de 350 000 (=162 500 x 2). A raíz de estas dos complicaciones se aconseja usar la 2-BL como la referencia teórica comparable con las distribuciones empíricas.

Algunas aplicaciones para las elecciones mexicanas de 2006

Diferentes autores llevaron a cabo pruebas forenses con la metodología de Benford para detectar fraude en las elecciones que tuvieron lugar en México el 6 de julio de 2006.[2] Un ejemplo es Mansilla (s. f.), quien compara las

[2] Los partidos registrados para esta contienda electoral fueron los siguientes: Partido Acción Nacional (PAN), cuyo candidato era Felipe Calderón; Alianza por México, una coalición entre el Partido Revolucionario Institucional (PRI) y el Partido Verde Ecologista de México y cuyo candidato era Roberto Madrazo; la Coalición por el Bien de Todos que unió al Partido de la Revolución Democrática (PRD) con el Partido del Trabajo y Convergencia Democrática a través de la nominación de Andrés Manuel López-Obrador (AMLO); el Partido Nueva Alianza (Panal) que tenía

distribuciones empíricas del primer dígito inicial para los votos obtenidos por Calderón y AMLO respecto a la distribución teórica de la 1-BL. Como se mencionó, esta prueba no es válida debido a los sesgos que surgen de determinadas reglas institucionales de la contienda. A pesar de ello el autor sostiene que, de acuerdo con los datos a nivel casilla que se obtuvieron de una muestra del conteo del PREP (Programa de Resultados Electorales Preliminares), existe evidencia estadística que rechaza la validez de la ley de Benford para el caso mexicano, por lo que los resultados electorales quedan en entredicho.[3]

En un análisis mucho más riguroso, Mebane (2006a) pone a prueba la 2-BL para las elecciones presidenciales de México. Las cuentas de votos para los cinco partidos (o coaliciones) se definen, en este ejercicio, a nivel de la casilla y de la sección electoral. El autor presenta la bondad de ajuste de la distribución empírica para los 32 estados y para los datos globales (*i.e.* frecuencias observadas en el ámbito nacional). Con 95 por ciento de confianza, el autor encuentra que la ley de Benford se rechaza en los datos globales y en muchos de los estados.[4]

En otro artículo, Mebane (2007a) estudia los resultados de las elecciones de presidente, senadores y diputados de 2006 usando las cuentas por secciones. En su análisis encuentra muchas inconsistencias con la 2-BL en los votos para los candidatos del PAN y el PRD en municipalidades cuyo alcalde tiene la misma filiación política. Para este autor, dicho resultado pone en evidencia que la maquinaria política del partido localmente dominante operó en la fabricación de votos; no obstante, este patrón no se observa en municipalidades gobernadas por el PRI. Asimismo, Mebane sostiene que las anomalías detectadas en la prueba 2-BL para los partidos que no resultaron ser competitivos podrían ser producto de la intimidación o del voto estratégico. En un estudio más, Gutiérrez y Calderón (2006) presentan pruebas

como candidato a Roberto Campa y el Partido Alternativo Socialdemócrata (PAS) que postuló a Patricia Mercado

[3] Un estudio similar fue realizado por la compañía consultora AC Nielsen (2006), aunque en este caso se consideraron las 128 771 casillas procesadas en el PREP y también se calculó la distribución de los votos recibidos por Madrazo, de nueva cuenta la 1-BL es rechazada.

[4] Mebane sostiene que la prueba 2-BL no es válida a nivel de la casilla cuando existe un problema de "división aproximadamente similar con remanentes" (REDWL, por sus siglas en inglés). Esto es, la mayoría de los votos son emitidos en un número específico de casillas y se distribuyen equitativamente, mientras que los votos remanentes se esparcen en las urnas restantes. Este factor introduce un sesgo en la 2-BL empírica.

basadas en la 2-BL para resultados por sección analizados a nivel distrital. Los autores detectan 47 distritos con severas *anomalías*.

En contraste, Pliego (2007) argumenta que la violación a la ley de Benford no es un buen indicador de fraude electoral. En su trabajo, Pliego analiza exclusivamente las casillas en las que el tribunal electoral decidió recontar los votos. Por lo tanto, Pliego plantea que la 2-BL puede ser considerada como una prueba potente sólo cuando la desviación observada en la frecuencia empírica respecto a la distribución teórica se incrementa conforme aumenta el número de irregularidades detectadas en el recuento. De acuerdo con la Ji-cuadrada calculada para 12 de los 15 distritos electorales involucrados (usando cuentas a nivel de casilla y de sección), no se observa patrón alguno entre el valor de esta estadística y las modificaciones en los votos obtenidos. De aquí que la validez de la ley de Benford sea descartada. Cabe aclarar que una limitante del enfoque de Pliego tiene que ver con que la desviación respecto a la 2-BL puede deberse a la presencia de intimidación del voto y no a inconsistencias aritméticas en las actas de escrutinio, en cuyo caso no se esperaría detectar un patrón en las Ji-cuadradas obtenidas de una muestra que incluye casillas seleccionadas exclusivamente por irregularidades en las actas.

Metodología para la validación teórica de la ley de Benford en elecciones limpias

Con el propósito de corroborar que una competencia electoral nítida produce la 2-BL, Mebane (2006a) simula un proceso de votación en secciones electorales de igual tamaño en donde el resultado de los comicios se genera con un conjunto pequeño de distribuciones seleccionadas al azar. Para modelar el mecanismo de votación se suponen tres tipos de individuos (prefieren a la oposición, al gobierno establecido, seleccionan al azar), pero la decisión a favor de un candidato se efectúa con un cierto error cuya frecuencia varía de una sección a otra pero es constante al interior de las mismas. Aunque cada sección tiene el mismo número de votantes, la proporción de cada tipo varía entre secciones al definirse en términos de una distribución uniforme. En consecuencia, los votos recibidos por los distintos candidatos dependen de dos procesos aleatorios: *1)* la selección del tipo de individuo que emite el voto en una sección en particular, y *2)* la realización del voto efectuado por los individuos de cada tipo.

Por medio de una simulación de Monte-Carlo este autor encuentra que a partir de este mecanismo sencillo de votación se producen resultados electorales que satisfacen la 2-BL para diferentes especificaciones de los parámetros. Sin embargo, en el citado artículo no se ofrece evidencia empírica alguna de que los supuestos planteados caracterizan adecuadamente un escenario real de formación de preferencias y votación. No basta con suponer que detrás de un proceso electoral prevalece una *mezcla estadística* que replica una 2-BL en los resultados del conteo de votos. En aras del realismo metodológico también es necesario validar que dicha mezcla es producto de un proceso estocástico de formación de opiniones como el que se observa a lo largo de una campaña real.

Por lo tanto, en este artículo se sugiere que una metodología más sólida para validar la relevancia de la ley de Benford en el análisis forense de datos electorales requiere simular una campaña a través de un modelo basado en agentes. Un primer paso en esta dirección consiste en construir un modelo computacional cuyos parámetros y condiciones iniciales sean calibrados directamente con datos reales o a partir de la bondad de ajuste de los datos artificiales a una serie de sondeos de opinión realizados a lo largo de la campaña. Posteriormente, en una segunda etapa, se calcula la distribución para el segundo dígito inicial de los datos electorales que se generan en una campaña simulada en donde el conteo final se lleva a cabo con transparencia. De esta forma, la relevancia de la ley de Benford no se rechaza cuando la distribución generada con las cuentas de los votos artificiales obtenidos por cada candidato es estadísticamente cercana a la 2-BL teórica. Se dice que la distribución logarítmica es robusta cuando se pude *hacer crecer* independientemente de las realizaciones que producen las distintas variables aleatorias incluidas en el modelo.

La descripción de las campañas electorales mediante un ABM

A partir de la evidencia empírica que indica que las redes de discusión política son muy importantes para explicar la formación de opiniones, Castañeda e Ibarra (2011) desarrollan un ABM para las elecciones presidenciales de 2006. En este modelo computacional se establece un mecanismo de contagio social en el que cada agente está sujeto a la influencia de información local y global. Simultáneamente, el modelo permite combinar diversos elementos institucionales de los procesos electorales, como debates, campañas

negativas y sesgos inducidos por la televisión u otros medios masivos de comunicación, con los incentivos de los votantes, a través del voto estratégico y el costo-beneficio de la decisión de votar.

Una campaña electoral es considerada un sistema adaptable complejo, ya que las posiciones de individuos y partidos afectan las encuestas de opinión y éstas, a su vez, inciden en la formación de opiniones y estrategias partidistas. Estos sistemas son analizados a partir de modelos computacionales en los que es posible incorporar agentes heterogéneos y mecanismos de decisión condicionados por la interacción local y el contexto social. A diferencia de los modelos matemáticos tradicionales, la simulación mediante modelos basados en agentes permite explicar comportamientos agregados que provienen de dinámicas no lineales causadas por la retroalimentación entre agentes y entre éstos y el entorno de adaptación.[5]

El ABM descrito en este artículo enfatiza el contagio social entre votantes potenciales por lo que se estructura a partir de un autómata celular en el que una retícula caracteriza el espacio geográfico/social de interacción.[6] La regla de transición de cada agente (o célula) hace que la variable de estado (intención de voto) varíe, esencialmente, en función de opiniones mayoritarias locales (red de discusión política) y globales (encuestas nacionales). De esta forma, la preferencia partidista de un porcentaje de agentes activados aleatoriamente en cada periodo del modelo computacional (o día de campaña) puede cambiar por efecto del contagio social.

Las preferencias políticas iniciales de los ciudadanos y la mayoría de los parámetros del modelo son calibrados con datos agregados de encuestas electorales y con datos panel que le dan seguimiento a los cambios en las preferencias políticas de los encuestados.[7] En particular, el sembrado inicial de preferencias en las distintas regiones del país hace uso del Estudio Panel

[5] Castañeda (2009) presenta una revisión analítica de la literatura sobre sociomática, es decir, sobre la explicación de fenómenos socioeconómicos a través de la teoría de la complejidad y los ABM.

[6] El ABM fue construido con *NetLogo* versión 4.0.3 autorizado por Uri Wilensky en 1999, http://ccl.northwestern.edu/netlogo/. Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL.), el cual puede solicitarse en la siguiente dirección de correo electrónico: sociomatica@hotmail.com.

[7] Para la serie de preferencias agregadas se empleó la *encuesta sobre encuestas* calculada por el Centro de Investigación para el Desarrollo A.C. (CIDAC), que utiliza un conjunto de encuestas periódicas realizadas por diferentes compañías (Mitofsky y Gea-Isa) y periódicos nacionales (*Reforma*, *El Universal* y *Milenio*): http://www.cidac.org/es/modules.php?name=Content&pa=showpage&pid=98. [Consultado el 15 de febrero de 2008].

México 2006 organizado por MIT y el periódico *Reforma*, el cual ofrece tres oleadas de entrevistas (octubre 2005, abril-mayo 2006 y julio 2006) con 2 400 entrevistados y cerca de cien preguntas relacionadas con la intención del voto, información sociodemográfica y opiniones de temas económicos y políticos.[8]

La configuración del modelo computacional

La descripción detallada del modelo, bajo la premisa de un proceso electoral limpio, se presenta en el artículo antes referido. Sin embargo, en este apartado y en el apéndice B se hace una breve exposición del mismo de tal manera que el lector pueda entender cómo se simula la formación de preferencias políticas durante una campaña electoral. El ABM en cuestión está diseñado con módulos específicos que se activan en diferentes periodos según la cronología real de la campaña por la presidencia de México. Por ende, se plantea un proceso que dura 240 días (correspondiente a ocho meses, noviembre-junio); encuestas nacionales que se levantan mensualmente y en el día posterior a dos debates, éstos tienen una amplia cobertura y se efectúan en los periodos 180 y 220; un sesgo-TV que opera de manera continua hasta el día 200; escándalos políticos que son lanzados en 2 por ciento de los días que dura la campaña; un voto estratégico que es activado diez días antes de la elección y una decisión costo-beneficio sobre votar o abstenerse en el último día de cada corrida.

El entorno de interacción social se representa a través de una retícula 120 x 120 con fronteras y vecindades de tipo Moore (*i.e.* un agente tiene a lo más ocho vecinos con los que puede interactuar). El espacio geográfico está dividido en 16 zonas electorales, cada una de las cuales se identifica con un conjunto de estados de la república, de tal forma que todas las zonas corresponden, aproximadamente, a regiones del país que tienen el mismo número de votantes registrados en el padrón electoral real (véase el cuadro A.2 en

[8] Los cuestionarios para cada oleada, el conjunto de datos, la metodología y algunos estudios relacionados se encuentran disponibles en el portal de la 2006-MPS: http://web.mit.edu/polisci/research/mexico06 Proyecto dirigido por las siguientes personas (en orden alfabético): Andy Baker, Kathleen Bruhn, Roderic Camp, Wayne Cornelius, Jorge Domínguez, Kenneth Green, Joseph Klesner, Chappell Lawson (investigador principal), Beatriz Magaloni, James McCann, Alejandro Moreno, Alejandro Poiré y David Shirk; bajo el apoyo financiero de la National Science Foundation (SES-0517971) y el periódico *Reforma*; el trabajo de campo fue llevado a cabo por el equipo de investigación y encuestas del periódico bajo la dirección de Alejandro Moreno.

el apéndice A). Por otra parte, las zonas tienen una dimensión 30 x 30 y están conformadas por nueve distritos electorales de igual tamaño (10 x 10). Cabe mencionar que es, precisamente, en estos distritos donde los agentes ejercen su voto.

Mientras que las preferencias políticas (Calderón, Madrazo, AMLO, otros e indecisos) de cada agente se siembran aleatoriamente de acuerdo con la participación partidaria a nivel estatal observada en los datos de la encuesta panel, los valores de otras variables se siembran al inicio de cada corrida a partir de criterios definidos en el ámbito nacional. Entre estas variables se encuentran las siguientes: compromiso ideológico (voto-duro, voto-débil e indecisos), exposición a la televisión (o proclividad a ser influido), atributos sociodemográficos (género, edad, escolaridad, ingreso y religión), confianza en la equidad del proceso electoral.[9]

Módulos de contagio local y global

Durante los primeros 30 días, la información local es la única fuente de contagio en el modelo. Éste es posible cuando un agente que ha sido activado en la simulación establece comunicación con uno de sus interlocutores políticos. Los interlocutores son parte de la red de discusión política si pertenecen a la vecindad del agente y superan un umbral de similitud, es decir, si comparten al menos un número determinado de atributos sociodemográficos. Una vez que la comunicación se produce, el interlocutor elegido al azar logra incidir en la opinión del agente activado cuando la opinión del primero coincide con la expresión mayoritaria de la vecindad.

Posteriormente, al hacerse pública la primera encuesta de opinión, el contagio se produce cuando el interlocutor elegido al azar entre los vecinos del agente tiene una preferencia política que coincide con la opinión mayoritaria reflejada en la encuesta nacional. Estas encuestas se levantan con 10 por ciento de los agentes de cada una de las zonas electorales, por lo que se puede afirmar que los sondeos son representativos del sentir nacional. El observador especifica en la interfaz del programa el porcentaje de individuos que pueden ser activados en cada periodo,[10] aunque la influencia solamente es viable en la medida en que el agente activado no sea un votante duro.

[9] De estas variables, el compromiso ideológico se obtiene de datos de encuestas, la percepción de equidad se calibra mediante un método indirecto y las demás son definidas por el observador en la interfaz del programa.

[10] Este porcentaje y otros que se describen a continuación son calibrados indirectamente.

Módulo de campañas negativas

En 2 por ciento de los días definidos al azar, el partido en el segundo lugar de la contienda, de acuerdo con las encuestas, lanza un escándalo político en contra del candidato que lidera las encuestas. Si el observador especifica en la interfaz que el partido atacante sigue una estrategia agresiva, el escándalo puede ser de gran repercusión (o de amplio espectro) y por ello el atacante puede hacerse de 6 por ciento de los adherentes no duros del líder con una probabilidad de 80 por ciento. No obstante, también existe el riesgo de alienar al mismo porcentaje de sus adherentes cuando éstos desacreditan el uso de dichas tácticas. En contraste, cuando la estrategia se define como moderada, la posibilidad de ganar/perder adeptos una vez propagado el escándalo disminuye a sólo 2 por ciento de los votantes.

Módulo de sesgo-TV

La imagen de un candidato puede ser promovida de manera sostenida a través de medios electrónicos por parte de figuras públicas, noticieros, grupos de cabildeo y gerentes de campaña. Por esta razón, el ABM incluye un módulo que especifica que un partido tiene la posibilidad de atraer por este mecanismo a nuevos adeptos cada día de la campaña hasta 40 días antes de la elección debido a la suspensión de este tipo de publicidad. Periodo que coincide con la fecha establecida en el *acuerdo de neutralidad* firmado por todos los partidos al inicio de la campaña. Este acuerdo también prohibe la transmisión de spots promoviendo programas sociales y obras públicas.

El modelo supone que el sesgo-TV favorece al PAN debido al apoyo indiscriminado dado por Fox a Calderón en sus discursos púbicos y a las acusaciones del PRD sobre la injerencia de las cámaras empresariales. El observador especifica en la interfaz el porcentaje de individuos que están sujetos a la influencia de los mensajes de los medios. Esta propensión se limita a los agentes que no son votantes duros y que tienen una exposición a la TV relativamente alta. A diferencia del módulo de interacción local, en el que el contagio se presenta dentro de las vecindades, el cambio de preferencias políticas a través de este módulo se produce de manera esparcida en el territorio nacional. De tal manera que el sesgo-TV permite introducir agentes con preferencias políticas antagónicas en clústeres de individuos que mantienen cierta afinidad política.

Módulo de debates

En cada uno de los debates existe un ganador con una probabilidad que es especificada por el observador. En el primer debate, el candidato victorioso

se elige entre los dos contendientes principales (dado que AMLO no participó), y entre los tres candidatos importantes en el segundo debate. La victoria ofrece a los partidos la posibilidad de conseguir el apoyo de ciudadanos indecisos y votantes débiles. El porcentaje de individuos que pueden modificar su opinión en estas circunstancias también se especifica en la interfaz del programa.

Módulo de voto-estratégico

Cuando la primera opción política de un agente no corresponde a ninguno de los dos candidatos que van a la cabeza en la encuesta previa al día de las elecciones es posible que la intención de voto se modifique a favor de su segunda opción. Esta última se define, en el modelo, con el candidato que tiene las preferencias mayoritarias en la vecindad del agente. El ABM también incorpora la posibilidad de que los adherentes de Calderón vayan con Madrazo y viceversa en un escenario en que AMLO es considerado *un riesgo para el país*, mientras que los votantes indecisos activados en este módulo seleccionan al azar entre Calderón y Madrazo. De nueva cuenta el porcentaje de votantes tácticos se define en la interfaz del programa.

Módulo de participación

Mediante un análisis costo-beneficio los ciudadanos deciden participar o abstenerse de votar el día de la elección. Mientras que el costo de votar es definido por el observador, el beneficio se determina usando una probabilidad heurística que mide la posibilidad de ganar que tiene el candidato preferido por el agente. Esta probabilidad se estima al multiplicar la proporción de las intenciones de voto recibidas por su candidato en las encuestas nacionales, por el porcentaje de individuos que lo apoyan en la vecindad y por un índice binario que indica si el agente confía o no en la elección. Obviamente, los votantes duros siempre ejercen su voto.

Método de calibración

El porcentaje de votantes duros para cada partido se obtiene de una encuesta preelectoral conducida por el periódico *Reforma* entre los meses de enero y junio de 2006.[11] De acuerdo con esta encuesta, 59 por ciento de los

[11] Más detalles en Moreno y Méndez (2007, cuadro 1).

individuos entrevistados declaró una identidad partidaria; de este porcentaje, los votantes-duros se distribuyeron de la siguiente forma: PAN (38%), PRD (40%), PRI (48%). Por falta de información, el valor correspondiente de votantes duros para Otros (42%) es simplemente un promedio porcentual de los tres partidos principales.

Algunos parámetros que son definidos por el observador en la interfaz del programa también pueden ser calibrados indirectamente. Mediante un procedimiento de optimización no lineal (*hill-climbing*) se estiman los parámetros cuyos valores difícilmente se pueden obtener directamente de las encuestas. Con este propósito se especifica una función de adaptación que mide el ajuste de los datos simulados a los datos reales. Esta función se define como un error cuadrático medio, en el que cada error se calcula con la diferencia relativa entre la participación de la intención de voto de cada partido según las encuestas de opinión reales y la participación estimada con los datos que se generan artificialmente.[12]

Mecanismos de fraude electoral

En los módulos previos del modelo se representa un proceso electoral limpio dado que la decisión individual de voto es respetada; es decir, la votación por un candidato en específico no es inducida con amenazas o premios otorgados por autoridades o partidos, ni el conteo final de votos es alterado por funcionarios electorales. En contraste, en un ABM que simula un proceso fraudulento se supone que las decisiones de los agentes son manipuladas el día de la elección. En particular, el módulo de fraude plantea que existen varios distritos electorales *controlados por la maquinaria panista*. En estos distritos los operadores de Calderón pueden modificar a discreción cierto porcentaje de votos por medio de diversos mecanismos, porcentaje que el observador determina en la interfaz del modelo. Por lo tanto, el conteo final anunciado por las autoridades electorales oculta el hecho de que algunos de los votos obtenidos por el PAN fueron inducidos con *regalos* o alterados por funcionarios distritales o federales.

[12] En el cuadro A.3 del apéndice A se presentan los valores de parámetros calibrados a partir de este criterio, los cuales son utilizados para llevar a cabo las corridas en los escenarios de elecciones limpias y trucadas.

Los mecanismos de fraude implementados en el modelo consideran diferentes procedimientos para especificar en qué zonas electorales el voto puede ser manipulado a favor de Calderón: *1)* "autoridades": cuando en la zona electoral existe al menos un gobernador panista en el momento de la elección; *2)* "enclaves": cuando el PAN tiene una mayoría relativa en las preferencias políticas de la zona de acuerdo con los sondeos levantados al inicio de la campaña; *3)* "inconsistencias": cuando, en los datos reales, los distritos electorales de la zona tienen un gran número de actas que presentan irregularidades. De acuerdo con estos criterios, el cuadro A.4 del apéndice A muestra las zonas electorales que tienen el potencial de producir un escenario fraudulento.

Asimismo, el modelo supone que en los procedimientos de autoridades y enclaves la posibilidad de un fraude por parte de los operadores del PAN requiere superar dos filtros. En primer término, la zona electoral tiene que estar dominada por el PAN (ya sea por la presencia de gobernantes afiliados al partido o por la mayoría relativa de las preferencias panistas). En segundo término, las preferencias iniciales por Calderón tienen que superar un umbral de relevancia partidista. En otras palabras, para que una zona electoral pueda ser considerada como un reducto del PAN se requiere que el apoyo partidista no esté muy pulverizado.

En una zona de *control-panista* la fabricación de votos puede ser de dos tipos: "intercambio de votos" o "creación de votos". En el primer método, un porcentaje *f* de los votos no panistas (M) seleccionado al azar se modifican de tal forma que Calderón incrementa su cuenta en *f* x M votos en cada uno de los *distritos capturados*. En el segundo método, M corresponde al número de individuos registrados en el distrito que prefirieron no votar el día de la elección, de tal forma que las elevadas tasas de participación en los *distritos capturados* se deben a la fabricación de votos a favor de Calderón.

¿Es válida la 2-BL en los datos artificiales de una campaña electoral simulada?

Resulta indudable que los individuos emiten su voto en cada uno de los distritos electorales en función de la ubicación espacial de las preferencias e ideologías. De igual forma, la heterogeneidad en el voto se debe a la aleatoriedad inherente a las diferentes fuentes de influencia que inciden en el individuo (debates, contagio local y global, sesgo-TV) y a los incentivos

geográficamente diferenciados que condicionan sus decisiones (costo-beneficio de la participación, voto estratégico). En consecuencia, es razonable pensar que la cuenta de votos a favor de un candidato en los diferentes distritos electorales sea el resultado de una *mezcla de estadísticas*, aunque no es evidente que dicha mezcla tenga los atributos necesarios para producir la 2-BL. Debido a la complejidad de la interacción entre las diferentes variables aleatorias del modelo y a la dificultad de caracterizar matemáticamente el proceso estocástico, en esta sección se emplea un ABM de campañas electorales para simular la distribución del segundo dígito inicial.

La ley de Benford en una campaña virtual limpia

Una vez calibrado el modelo computacional con datos de la campaña presidencial mexicana de 2006 se llevan a cabo diez corridas suponiendo un proceso electoral limpio y, en cada caso, se compara la distribución simulada con la 2-BL teórica. En el cuadro 1 se presentan las diferentes Ji-cuadradas de Pearson con nueve grados de libertad que se calcularon para cada una de las diez simulaciones. Esta estadística se obtiene para los votos recibidos por cada uno de los principales candidatos de la contienda a nivel del distrito electoral (cien agentes) y las frecuencias relativas correspondientes se estiman de los datos artificiales generados en los 144 distritos de la retícula del ABM, como se indica en la siguiente expresión:

$$X^2_{2\,BL} = \sum_{i=0}^{9} \frac{(sd_i - sd \bullet pBL_{2i})^2}{sd \bullet pBL_{2i}} \qquad (4)$$

en donde sd_i *es el número de distritos con el número i en el segundo dígito inicial en la* cuenta de los votos recibidos por el candidato C en determinado distrito; $sd = \Sigma sd_i$ es el número total de distritos que tienen un segundo dígito inicial en la cuenta de votos para el candidato C (*i.e.* distritos en los que el candidato C obtuvo cuentas de votos de un dígito son descartados), pBL_{2i} es el valor teórico para el número i de acuerdo con 2-BL.

De acuerdo con los resultados presentados en el cuadro 1 existen tres corridas en donde los votos obtenidos para uno de los candidatos no obedece la ley de Benford dado que el valor crítico para la Ji-cuadrada es de 16.9 en una prueba con un nivel de confianza de 95 por ciento. Por lo tanto, la ley de Ben-

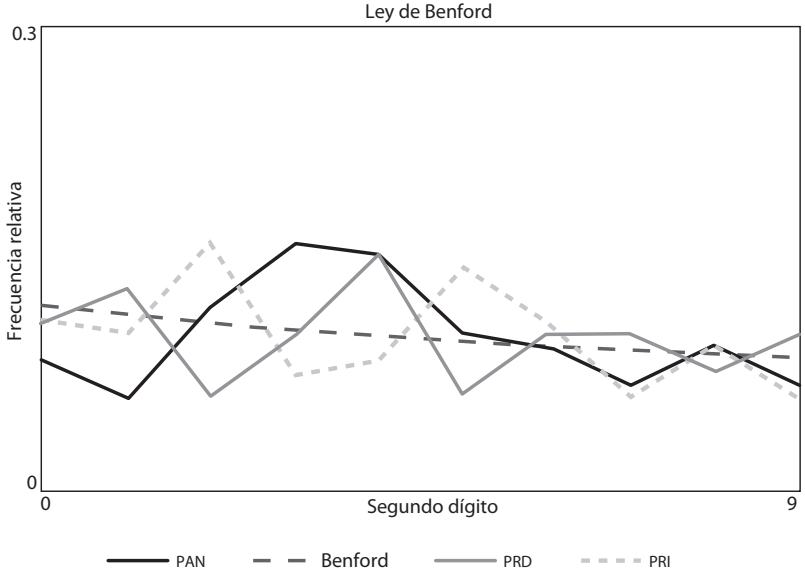
ford no parece ser un fenómeno que siempre está presente en los resultados de procesos electorales limpios. La gráfica 1 describe la 2-BL teórica (línea Benford) y las frecuencias relativas empíricas para los tres contendientes principales de la elección presidencial (PAN = Calderón, PRD = AMLO, PRI = Madrazo) en un contexto en el que se suponen elecciones nítidas.

CUADRO 1. Ji-cuadrada en un proceso electoral limpio

	1	2	3	4	5	6	7	8	9	10
Calderón	12.826	10.603	13.342	4.516	8.826	11.979	12.607	17.368*	5.468	12.484
AMLO	11.979	13.132	21.053*	6.356	4.85	11.595	12.885	7.421	5.685	9.449
Madrazo	14.934	9.837	13.58	7.982	4.278	10.749	19.018*	6.107	5.276	9.583

Fuente: Elaboración propia. *La 2-BL se rechaza con una confianza de 95%.

GRÁFICA 1. La 2-BL teórica y las frecuencias relativas simuladas en un proceso electoral limpio (con datos de la corrida 10 descrita en el cuadro 1)



Fuente: Elaboración propia.

La ley de Benford en el caso de un ABM con fraude en el conteo de votos

Una prueba forense que utiliza la ley de Benford es válida cuando la ley es robusta, es decir, cuando la ley se mantiene en ciertas condiciones (*i.e.* una elección nítida) pero no en otras (*i.e.* elecciones fraudulentas). En el apartado anterior se mostró que la 2-BL no siempre se presenta en una elección limpia, el siguiente paso consiste en analizar si este tipo de distribución se rechaza la mayoría de las veces en elecciones en las que un porcentaje de los votos son fabricados. Por esta razón, se simulan elecciones fraudulentas a través de un ABM que utiliza el procedimiento de *autoridades* para determinar los *distritos capturados* por la maquinaria panista y el mecanismo de *cambio de votos* para alterar las actas.

Se corren diez simulaciones con las que se estiman las distribuciones artificiales del segundo dígito inicial, las cuales se comparan con la distribución de la 2-BL teórica. El cuadro 2 presenta los promedios de fraude que se simulan (6.41, 14.40, 20.52 y 11.93%). En los dos primeros casos la fabricación de votos se produce sólo en las zonas con gobierno panista que tienen un control regional mayor de 20 por ciento, mientras que en los dos últimos casos el fraude se esparce por toda la retícula.

Cabe notar que la 2-BL se rechaza para las cuentas de un candidato en sólo cuatro corridas cuando el fraude es de alrededor de 6.41 por ciento de los votos emitidos, y para una sola corrida cuando el nivel de fraude es de 14.40 por ciento. En consecuencia, el grado de rechazo no es muy diferente al encontrado en los resultados del cuadro 1, en donde las elecciones son limpias por construcción. Estas simulaciones indican que una prueba forense para la detección de fraude basada en la 2-BL no es muy poderosa y, por ende, el rechazo de la 2-BL que se encontró en los diferentes estudios de los datos electorales mexicanos dista de ser una evidencia concluyente sobre la manipulación del voto.

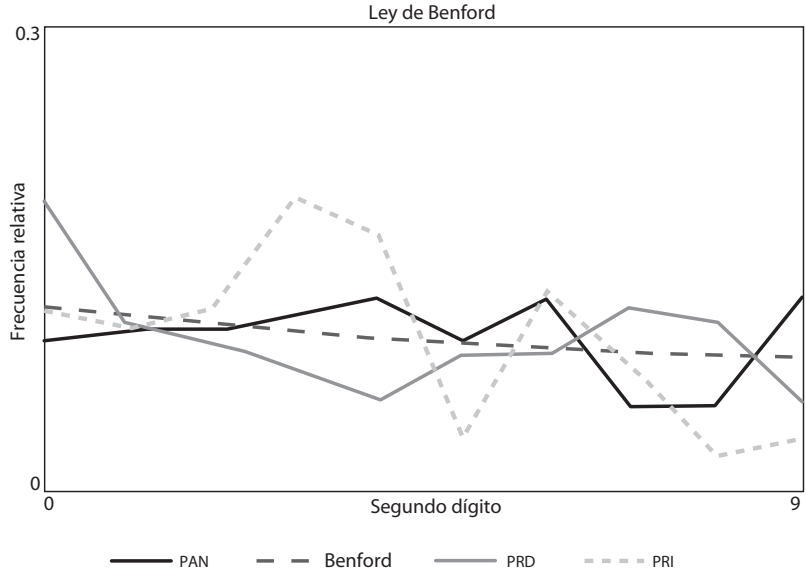
Sin embargo, se podría argumentar que esta prueba se desempeña mejor cuando se aplica a un nivel menos agregado, de tal forma que los datos analizados no mezclan zonas electorales con datos fabricados con aquellos que reflejan las decisiones verdaderas de los individuos. Esto se puede observar en el cuadro 2, cuando el fraude se esparce en toda la retícula y alrededor de 20.52 por ciento de los votos ha sido alterado. En este escenario, ocho de diez corridas rechazan la 2-BL para al menos un candidato; asimismo, este número de rechazos se incrementa cuando el grado de fraude es aún mayor (los resultados no se presentan aquí). En particular, en la gráfica 2 se presenta una

CUADRO 2 . Ji-cuadrada con elecciones fraudulentas simuladas (procedimiento de autoridades)

Corrida	1	2	3	4	5	6	7	8	9	10
Fraude (%)	6.41	6.45	6.42	6.47	6.49	6.5	6.4	6.26	6.48	6.25
Calderón	18.332*	6.334	8.281	15.515	26.048*	7.539	5.503	4.93	7.491	6.059
AMLO	11.712	7.588	8.087	11.547	3.999	11.387	5.802	8.753	11.904	18.274*
Madrazo	5.573	20.059*	15.425	7.232	5.695	5.049	11.993	16.831	13.906	10
Fraude (%)	14.72	14.08	14.5	14.22	14.48	14.5	14.48	14.47	14.01	14.58
Calderón	14.505	13.519	11.33	9.613	8.255	5.635	10.018	9.05	13.542	5.502
AMLO	7.893	14.002	13.965	5.11	5.857	6.504	6.437	5.395	11.828	5.516
Madrazo	10.195	5.86	20.075*	6.737	9.327	6.619	10.654	5.808	5.528	13.257
Fraude (%)	20.82**	20.04**	20.58**	20.82**	20.59**	20.67**	19.97**	20.23**	20.67**	20.82**
Calderón	23.574*	10.759	9.663	16.305	5.378	12.38	13.877	11.518	5.718	16.526
AMLO	22.641*	13.602	8.575	6.188	3.201	10.893	8.264	7.072	16.981*	9.777
Madrazo	16.028	20.13*	20.558*	15.273	19.967*	23.14*	17.942*	18.307*	21.741*	6.97
Fraude (%)	11.76**	12.02**	11.99**	11.85**	11.84**	11.98**	11.87**	12.01**	12.01**	12.01**
Calderón	10.524	12.109	26.354*	13.901	6.12	24.966*	9.81	2.839	6.673	9.456
AMLO	2.948	10.163	10.485	11.749	12.319	3.167	3.993	15.799	12.637	7.467
Madrazo	14.31	15.307	10.515	21.175*	16.644	18.176*	6.34	11.307	19.138*	9.069

Fuente: elaboración propia. *La 2-BL se rechaza con una confianza de 95%. **Incluye fraude en todas las zonas electorales de la retícula.

GRÁFICA 2. Ley de Benford con elecciones fraudulentas
(cuadro 2, corrida 3, fraude: 20.58% esparcido en todas las zonas)



Fuente: Elaboración propia.

desviación significativa de la frecuencia simulada respecto a la 2-BL teórica para el caso de Madrazo (PRI).

A pesar de que existen algunos escenarios en los que la 2-BL se rechaza, la mayoría de las veces cuando opera el fraude, pero no cuando se trata de elecciones limpias, la confiabilidad de esta prueba es limitada. Para que se pueda detectar la fabricación de votos, el fraude tiene que ser masivo y concentrado en regiones muy específicas, de tal forma que distritos *nítidos* (secciones electorales o casillas) no se mezclen con distritos *sucios*. Por ejemplo, en un escenario en donde el promedio de fraude es de alrededor de 11.93 por ciento de los votos emitidos, la prueba es muy débil ya que sólo en cuatro de diez corridas la 2-BL se rechaza aun cuando por construcción todas las zonas electorales presentan fraude.

Por otra parte, la debilidad de esta prueba se agudiza en un ABM que introduce un módulo que produce errores aritméticos neutrales en el conteo de votos. El cuadro 3 muestra un escenario con alrededor de 25.52 por ciento errores de conteo esparcidos en toda la retícula. En este escenario tres corridas rechazan la 2-BL aun cuando el proceso electoral sea limpio. Cabe

CUADRO 3. Ley de Benford con errores aritméticos

	1	2	3	4	5	6	7	8	9	10
Errores (%)	25.72	25.36	25.77	25.77	25.78	24.96	25.25	26	25.74	24.82
Calderón	6.582	11.398	15.713	2.381	13.661	10.965	11.252	5.271	10.849	16.053
AMLO	6.109	17.339*	6.033	12.873	8.355	26.078*	6.195	20.459*	8.321	6.516
Madrazo	6.543	9.36	6.716	6.861	7.367	4.076	10.121	8.211	7.724	12.239
Errores (%)	85.69	87.26	86.77	86.82	86.93	85.6	85.85	83.46	86.75	84.92
Calderón	7.413	5.876	16.859	3.562	22.089*	10.089	13.804	3.224	2.469	3.849
AMLO	18.81*	17.003*	14.01	12.867	11.359	10.594	9.646	9.218	5.69	5.629
Madrazo	7.97	19.814*	9.439	9.064	5.48	13.638	29.221*	10.375	16.772	22.226*

Fuente: Elaboración propia. *La 2-BL se rechaza con una confianza de 95%.

también notar que con errores que rondan alrededor de 86 por ciento de los votos emitidos, la distribución de preferencias políticas es más cercana a la distribución uniforme y, por ello, el rechazo de la 2-BL ocurre en 50 por ciento de las corridas.

Conclusiones

La presunción de fraude aparece recurrentemente en las contiendas electorales de países democráticos. El poder en disputa, los intereses en juego y las formas diferentes de concebir el mundo dan pauta a que determinados grupos y partidos intenten, en ocasiones, asegurar el triunfo por medio de la manipulación del voto. La existencia de fraude, o al menos la percepción de que éste ha ocurrido, es más común en países cuyos sistemas democráticos no están del todo consolidados. Si bien los actores políticos en estas sociedades han aceptado dirimir sus disputas a través de organismos gubernamentales elegidos por las mayorías, su compromiso con la democracia suele ser débil y la desconfianza hacia los grupos antagónicos es profunda. Dicho

entorno hace que los triunfos electorales de un partido sean, con frecuencia, cuestionados por los otros partidos argumentando la presencia de fraude en los comicios.

Las acusaciones de fraude son, por lo general, expresadas al calor de la contienda y utilizando argumentos poco fundamentados. De aquí la necesidad de desarrollar nuevas pruebas forenses de análisis electoral que permitan a los partidos y a los tribunales electorales tener criterios más objetivos para exponer alegatos y dirimir disputas. Tradicionalmente, las pruebas forenses se basan en el análisis estadístico de los resultados numéricos, ya sea a través de la detección de patrones anómalos en la distribución de los votos (o errores registrados) y en el comportamiento dinámico del conteo de actas, o bien mediante el establecimiento de asociaciones con variables que no deberían incidir en la decisión del voto en una contienda limpia.

En una reciente serie de artículos Walter Mebane Jr., investigador de la Universidad de Michigan, ha propuesto el uso de la ley de Benford para la detección del fraude electoral. De acuerdo con esta ley los dígitos iniciales de un conjunto de números (*i.e.* los votos obtenidos por los diferentes candidatos en cada casilla, sección o distrito) deben seguir una distribución logarítmica cuando los datos no han sido manipulados; sin embargo, en este artículo se muestra que dicha prueba no es una herramienta forense robusta. Mediante un modelo basado en agentes, calibrado con datos de la elección mexicana de 2006, se llevan a cabo simulaciones de Monte Carlo en las que la violación a la ley de Benford no permite distinguir entre procesos electorales limpios y trucados.

Por otra parte, el ABM de campañas virtuales utilizado en este artículo, en el que las preferencias partidistas se modifican a partir del contagio social, parece ser una herramienta más atractiva para los estudios forenses. Este tipo de ejercicio se lleva a cabo en Castañeda e Ibarra (2010), quienes comparan las distribuciones de datos artificiales y datos reales a través de una estadística no paramétrica. En este trabajo se observa que la distribución artificial generada a través de una contienda limpia replica mucho mejor la distribución real de las participaciones de los tres partidos principales que la distribución artificial obtenida en un escenario en el que un porcentaje de los votos han sido manipulados en el conteo final. Aunque los resultados del estudio sólo pueden descartar la presencia de un fraude masivo del orden de 5-6 por ciento de los votos emitidos, la técnica es promisoria para detectar fraudes más refinados a partir de modelos que presenten un mayor detalle y niveles de calibración más precisos.

Referencias bibliográficas

AC Nielsen (2006), "¿La Ley de Benford se puede aplicar a los votos en las casillas?", presentación en power point.

Brady, Henry (2005), "Comments on Benford's Law and the Venezuelan Election", manuscrito, Berkeley, University of California.

Castañeda, Gonzalo (2009), "Sociomática: El estudio de los sistemas adaptables complejos en el entorno socioeconómico", *El Trimestre Económico*, 76 (1), pp. 5-64.

Castañeda, Gonzalo e Ignacio Ibarra (2010), "Detección de fraude con modelos basados en agentes: Las elecciones mexicanas de 2006", *Perfiles Latinoamericanos*, 18 (36), julio-diciembre, pp. 43-69.

_____ (2011), "Campañas, redes de discusión y volatilidad de las preferencias políticas: Un análisis de las elecciones mexicanas de 2006", *Foro Internacional* 203 (LI), enero-marzo.

Diekmann, Andreas (2004), "Not the First Digit! Using Benford's Law to Detect Fraudulent Scientific Data", working paper, Swiss Federal Institute of Technology.

Durtschi, Cindy, William Hillison y Carl Pacini (2004), "The Effective Use of Benford's Law to Assist in Detecting Fraud in Accounting Data", *Journal of Forensic Accounting*, 5, pp. 17-34.

Gutiérrez, Luis H. y Emiliano Calderón (2006), "La Ley de Benford para el segundo dígito y los resultados electorales en México", UAM-Iztapalapa/ Facultad de Ciencias-UNAM.

Hill, Theodore P. (1995), "A Statistical Derivation of the Significant-Digit Law", *Statistical Science*, 10 (4), pp. 354-363.

_____ (1998), "The First Digit Phenomenon", *American Scientist*, 86 (4), julio-agosto.

Mansilla, R. (s.f.), "Análisis de los resultados electorales a partir de la Ley de Benford", Centro de Investigaciones Interdisciplinarias en Ciencias y Humanidades (CEIICH)-UNAM.

Mebane, Walter Jr. (2006a), "Election Forensics: Vote Counts and Benford's Law", working paper, Department of Political Science, University of Michigan.

_____ (2006b), "Election Forensics: The Second-digit Benford's Law Test and Recent American Presidential Elections", working paper, Department of Political Science, University of Michigan.

_____ (2007a), "Election Forensics: Statistics, Recounts and Fraud", wor-

king paper, Department of Political Science, University of Michigan.
_____ (2007b), "Statistics for Digits", working paper, Department of Political Science, University of Michigan.
_____ (2007c), "Evaluating Voting Systems to Improve and Verify Accuracy", working paper, Department of Political Science, University of Michigan.
_____ (2008), "Election Forensics: Outlier and Digit Tests in America and Russia", working paper, Department of Political Science, University of Michigan.
Moreno, Alejandro y Patricia Méndez (2007), "La identificación partidista en las elecciones presidenciales de 2000 y 2006 en México", *Política y Gobierno* XIV (1), pp. 43-75.
Nigrini, Mark (1996), "A Taxpayer Compliance Application of Benford's Law", *Journal of the American Taxation Association*, 18, pp. 72-91.
Pericchi, Luis Raúl y David Torres (2004), "La Ley de Newcomb-Benford y sus aplicaciones al referéndum revocatorio en Venezuela", *Reporte técnico no-definitivo*, 2ª versión, octubre, manuscrito, Universidad de Puerto Rico y Universidad Simón Bolívar.
Pietronero, E., E. Tossati, V. Tossati y A. Vespignani (2001), "Explaining the Uneven Distribution of Numbers in Nature: The Laws of Benford and Zipf", *Physica* A 293, pp. 297-304.
Pliego, Fernando (2007), "El mito del fraude electoral en México", México, Editorial Pax.
Taylor, Jonathan (2005), "Too Many Ties? An Empirical Analysis of the Venezuelan Referendum Counts", manuscrito, Stanford University.
Varian, Hall (1972), "Benford's Law", *American Statistics*, 23, pp. 65-66.
Zhipeng, Li, Cong Ling y Wang Huajia (2004), "Discussion on Benford's Law and its Applications"; http://arxiv.org/abs/math/0408057.

Apéndice A

CUADRO A.1. Distribuciones de la ley de Benford para la k-ésima posición inicial

Dígito (di)	$P(d_1)$	$P(d_2)$	$p(d_3)$	$p(d_4)$
0		0.11968	0.10178	0.10018
1	0.30103	0.11389	0.10138	0.10014
2	0.17609	0.10882	0.10097	0.10010
3	0.12494	0.10433	0.10057	0.10006
4	0.09691	0.10031	0.10018	0.10002
5	0.07918	0.09668	0.09979	0.09998
6	0.06695	0.09337	0.09940	0.09994
7	0.05799	0.09035	0.09902	0.09990
8	0.05115	0.08757	0.09864	0.09986
9	0.04576	0.08500	0.09827	0.09982

Fuente: Dieckmann (2004).

CUADRO A.2. Distribución de preferencias por zona electoral (porcentajes)

Región	Estados	Calderón (PAN, azul)	AMLO (PRD, amarillo)	Madrazo (PRI, rojo)	Otros (gris)	Indecisos (naranja)	Proporción de ciudadanos empadronados
1	Baja California y Sonora	34	17	32	2	15	0.0525
2	Sinaloa, Colima, Dgo. y Zacatecas	29	10	45	2	14	0.058
3	Chihuahua y Coahuila	18	18	38	6	20	0.073
4	Nuevo León y Tamaulipas	28	18	28	4	22	0.059
5	Nayarit y Michoacán	17	36	30	5	12	0.067
6	Jalisco	34	15	33	5	13	0.059
7	Guanajuato y Aguascalientes	45	12	23	7	13	0.06
8	Querétaro, San Luis Potosí e Hidalgo	32	23	23	2	20	0.05
9	Estado de México	18	40	25	5	12	0.065
10	Estado de México	18	40	25	5	12	0.065
11	Distrito Federal y Tlaxcala	17	59	8	7	9	0.055
12	Distrito Federal y Tlaxcala	17	59	8	7	9	0.055
13	Guerrero y Oaxaca	12	35	35	6	12	0.063
14	Puebla y Morelos	29	27	21	6	17	0.065
15	Veracruz	31	41	13	9	6	0.069
16	Tabasco, Campeche, Chiapas y Yucatán	13	34	38	5	10	0.079

Fuente: Cálculos propios con datos de la primera oleada del Estudio Panel para México 2006 y del IFE (ciudadanos empadronados por estado). *Notas:* Las participaciones de las preferencias por zonas se calcularon con los promedios ponderados de las participaciones de los estados incluidos en cada región, en donde los ponderadores están dados por la proporción de ciudadanos empadronados en cada estado. Debido a que por construcción cada región tiene aproximadamente el mismo número de ciudadanos empadronados (columna 8), las participaciones en las regiones (9 y 10) y (11 y 12) se duplican, de tal forma que las grandes poblaciones del Estado de México y el Distrito Federal están adecuadamente representadas en el espacio geográfico.

CUADRO A.3. Parámetros calibrados indirectamente

Parámetro	Valor
Contagio	20
Debate	10
Voto-estrat.	5
Sesgo	0.08
Equitativa	50
Costo	0.01
Crisis	No
Escándalo	Sí
Tipo-PAN	Moderado
Tipo-PRI	Moderado
Tipo-PRD	Agresivo

Fuente: Valores calculados en Castañeda e Ibarra (2011). *Nota: Contagio* se refiere al porcentaje de agentes que son activados en cada periodo para un posible contagio, ya sea local o global; *debate* se refiere al porcentaje de agentes que pueden ser influidos por los resultados de los debates presidenciales; *voto-estrat*, se refiere al porcentaje de individuos que utilizan un comportamiento estratégico al momento de votar en caso de ser necesario; *sesgo* se refiere a la proporción de agentes que pueden ser influidos por actores externos en caso de estar muy expuestos a los medios; *equitativa* se refiere al porcentaje de agentes que consideran que las elecciones son justas; *costo* se refiere a los costos de acudir a las urnas; *crisis* se refiere al uso del voto estratégico bajo la premisa de que la presidencia de AMLO puede dañar al país; *escándalo* se refiere al uso de campañas negativas para atraer votantes, en estas circunstancias las estrategias de campaña para cada partido pueden ser agresivas o moderadas.

CUADRO A.4. Zonas electorales *capturadas* de acuerdo con diferentes criterios

Zona	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Autoridades	sí					sí	sí	sí						sí		sí***
Enclaves*	sí			sí		sí	sí	sí						sí		
Inconsis-tencias**	7			1		4			2	2	5	5				8

Fuente: Cálculos propios con datos de Gobierno Legítimo de México, http://www.amlo.org.mx. *Fuente: Estudio Panel México 2006, las zonas electorales se definen en el cuadro A.2. **Ranking de acuerdo con el número más alto de inconsistencias en las actas de escrutinio. ***Esta zona no se incluye cuando el control regional se establece en 20 por ciento.

Apéndice B

Con el objetivo de que el lector tenga un mejor entendimiento sobre la forma en que opera el modelo computacional utilizado en este artículo se describe a continuación el protocolo ODD (*Overview, Design and Detail*) correspondiente.

Panóramica

(i) *Propósito*. El objetivo de este modelo es describir el proceso de formación de preferencias partidistas a lo largo de una campaña electoral concebida como un sistema adaptable complejo. Con este enfoque la intención de voto se ve influida por procesos de interacción social, los cuales se representan a través de influencias de carácter local (redes de discusión política) y global (encuestas nacionales). Con este ABM se pretende mostrar la relevancia que tienen la interacción social y los factores institucionales en la volatilidad de las preferencias individuales y en la distribución espacial del voto. Los factores institucionales considerados en el modelo son los siguientes: debates, campañas negativas y sesgos mediáticos.

(ii) *Entidades, variables de estado y escala temporal*. En relación con las entidades el ABM considera que los sitios de la retícula de un autómata celular describen a individuos, que pueden ser de dos tipos: votantes débiles (*i.e.* susceptibles de cambiar preferencias) o votantes duros (*i.e.* adheridos permanentemente a una preferencia política). Estos individuos se insertan en un entorno geográfico-social, por lo que forman parte de una vecindad (comunidad) ubicada en una zona geográfica del país y de una red social de discusión política al interior de su vecindad.

Las variables de estado de cada individuo son las siguientes: preferencia partidista (Calderón, Madrazo, AMLO, otros e indecisos), que puede cambiar en cualquier día de la campaña una vez que el agente ha sido activado; atributos sociodemográficos (ingreso, religión, edad, sexo y escolaridad); exposición a la TV, y confianza en las elecciones. Estas tres últimas variables permanecen fijas a lo largo de la campaña.

Cada periodo (o tic) de la corrida representa un día de la campaña, por lo que ésta dura hasta el día 240 en que los individuos votan y se hace el conteo electoral. Las preferencias agregadas de los votantes se describen a través de una encuesta de carácter nacional que se levanta cada 30 días y un

día después de los debates (periodos 180 y 220). El sesgo-TV deja de operar en el periodo 200 y el voto estratégico se produce diez periodos antes del día de la elección.

(iii) *Procesos, activación y cronología*. En cada periodo de la campaña un porcentaje de individuos es elegido al azar (*i.e.* activación asincrónica aleatoria), por lo que en caso de ser votantes débiles tienen la posibilidad de cambiar sus preferencias partidistas. Los procesos que determinan el cambio de preferencia partidista tienen que ver con las siguientes reglas de transición: interacción social, debate, escándalo político y voto estratégico. Asimismo, existe un proceso adicional que determina si el individuo opta por participar el día de las elecciones. Cada uno de estos procesos se describe en el texto. La interacción social opera todos los días de la campaña pero la influencia de las encuestas se inicia a partir del periodo en que se levanta la primera encuesta (día 30). El escándalo político (campaña negativa) surge exclusivamente en 2 por ciento de los días de la campaña, mientras que la incidencia de los debates y el voto estratégico sobre las preferencias se dan exclusivamente los días arriba referidos.

Diseño de conceptos

(i) *Emergencia y adaptación*. En cada periodo de la campaña y en el día de las elecciones se produce una distribución espacial de los votos que se puede describir a través de diferentes funciones. Por ejemplo: las participaciones que los partidos obtienen en los votos escrutados en los distintos distritos electorales o la frecuencia del diferencial de votos recibidos entre el primer y el segundo lugar de cada distrito. Estos patrones son producto de la interacción social que los individuos tienen en sus redes de discusión política, del proceso coevolutivo preferencias → encuestas → preferencias, y de los factores institucionales antes referidos.

(ii) *Objetivos, aprendizaje y predicción*. Los individuos no deciden por quién votar en función de algún criterio racionalista, por lo que su preferencia partidista obedece a factores ideológicos (votante duro) o bien a esquemas de imitación o contagio social (votante débil). Sin embargo, el agente sí hace uso de un análisis estratégico en dos circunstancias: cuando su candidato favorito se ha descarrilado según las encuestas y por ello se inclina por su mejor opción de entre los candidatos que tienen la posibilidad de ganar, y cuando toma la decisión de votar sopesando el costo de

acudir a las urnas con el beneficio esperado de ganar, el cual se define a partir de la heurística descrita en el texto.

(ii) *Percepción*. Los individuos son conscientes de las preferencias partidistas de cada uno de los integrantes de su red de discusión política y de las preferencias agregadas reflejadas en las encuestas levantadas en el ámbito nacional y difundidas ampliamente por los medios de comunicación. Con esta información se percatan de las posibles discrepancias entre sus preferencias personales y las del resto de la población. Asimismo, su exposición a la televisión los vuelve susceptibles a las influencias de los medios de comunicación y de las campañas publicitarias de amplia cobertura.

(iii) *Interacción*. Los individuos forman parte de un área geográfica del país y de una red de discusión política, por lo que el contagio social sólo es posible entre individuos que tienen un alto grado de similitud sociodemográfica y se comunican al momento de ser activados. En consecuencia, las preferencias de los individuos están sujetas a las presiones sociales de su entorno cuando éste tiene opiniones diferentes a las de su interlocutor y estas últimas son avaladas por las encuestas nacionales o el sentir mayoritario de la red de discusión.

(iv) *Estocasticidad*. Existen varios elementos de carácter aleatorio en el modelo cuya realización se produce en distintos periodos de la corrida: el sembrado inicial con que se establece una caracterización descriptiva de las participaciones partidistas observadas en las distintas zonas geográficas del país; el procedimiento utilizado para el levantamiento de la encuesta en el ámbito nacional; la activación asincrónica de los individuos seleccionados en cada periodo para el contagio social; el periodo en que surgen los escándalos políticos por parte de los candidatos que se encuentran en el segundo lugar de las encuestas; la selección de individuos que pueden cambiar sus preferencias por efecto de las campañas negativas, los debates y la exposición a la televisión.

Detalles

(i) *Inicialización*. En la interfaz del programa el observador puede optar por un sembrado aleatorio de las preferencias partidistas en los sitios de la retícula que refleje los promedios nacionales de participación obtenidos con los datos de la primera oleada de la encuesta panel. También es posible que la representatividad de las preferencias partidistas se exprese a nivel de las 16 zonas geográficas en las que se divide la retícula.

(ii) *Insumos*. El modelo utiliza datos de la encuesta panel de *Reforma*-MIT para calibrar el sembrado inicial, datos de encuestas para determinar el porcentaje de votantes duros por partido y datos de *promedios de encuestas periódicas* sobre preferencias agregadas para calibrar indirectamente parámetros del modelo no estimados con encuestas. Por último, el modelo se valida de forma empírica al comparar los datos artificiales con los datos reales a nivel distrito electoral según el cómputo oficial del Instituto Federal Electoral. Este proceso de validación se lleva a cabo a partir de una prueba no paramétrica (*Kolmogorov-Smirnov*) en la que se analiza si los datos de ambas fuentes provienen de la misma distribución teórica.

(iii) *Submodelos*. Los detalles particulares de los procesos con los que se pueden modificar las preferencias y con los que se decide votar o abstenerse se describen en el cuerpo central de este artículo, mientras que la justificación teórica/empírica de los mismos se presenta en Castañeda e Ibarra (2011).