¿Manipulación de datos en la elección?

Por
Jorge A. López, Ph.D.

Muchos de nosotros aprovechamos la última tecnología y seguimos las elecciones del domingo a través de un método nuevo: las actualizaciones de los votos que el PREP puso por Internet. Con lo que el IFE no contaba, es que además de informar, las actualizaciones mostradas también nos daban valiosos datos que pueden ser examinados para determinar su credibilidad. Un análisis somero muestra que lo más probable es que los datos del PREP hayan recibido vida artificial de algún algoritmo computacional.

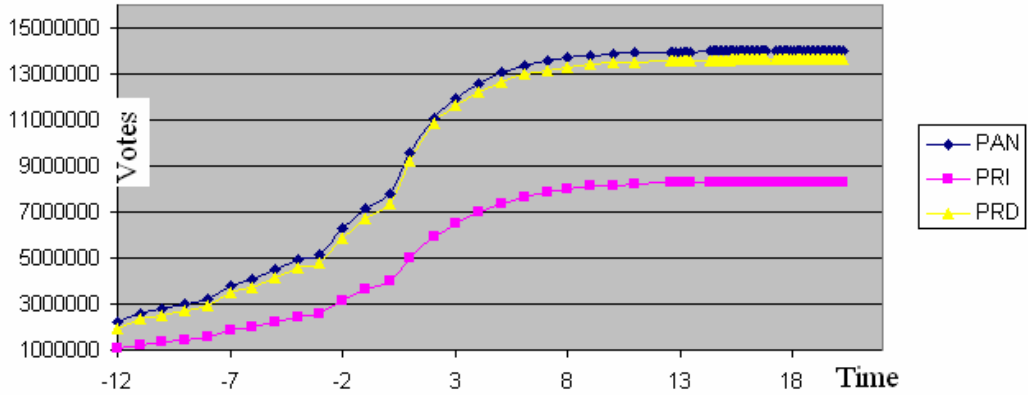
Lo que los navegantes de la red vimos ese domingo, fue un inicio fortísimo de Calderón de más de 4% de ventaja sobre AMLO que empezó a disminuir en porcentajes en el transcurso de la tarde. La tendencia decreciente continuó y, alrededor de medianoche, muchos nos fuimos a la cama con la convicción de que empatarían a las 3:00 AM del lunes, y que AMLO llevaría una ventaja de un punto porcentual para cuando nos despertáramos. La gran sorpresa matutina fue que la tendencia había cambiado en la noche y Calderón aparecía con una pequeña pero invariante ventaja de aproximadamente 1%; esto nos mandó a algunos de nosotros, profesores de física, precisamente a lo que hacemos para ganarnos la vida: análisis de datos.

Para el lunes en la tarde los primeros datos de las actualizaciones del PREP empezaron a aparecer en los *blogs* y *chat-rooms*, y las insinuaciones de manipulación empezaron a tomar forma. El Dr. Luis Mochan, físico de la UNAM, e incontables *surfers* anónimos ayudaron a completar la película.

Los Datos

Después de recabar datos de varias fuentes independientes y confirmar su confiabilidad, la primera preocupación apareció al graficar la tendencia de los votos en función del tiempo (medido respecto a las 00:00 del lunes). Las tres curvas correspondientes a los tres candidatos principales tenían un comportamiento sorprendente: presentaban una ventaja constante del PAN sobre el PRD como se muestra en la figura.

Este efecto de *espejo* entre las dos curvas no debería existir ya que los votos que



estaban llegando arribaban de diferentes partes del país donde el apoyo de los candidatos variaba por márgenes enormes.

El "Cuchareo"

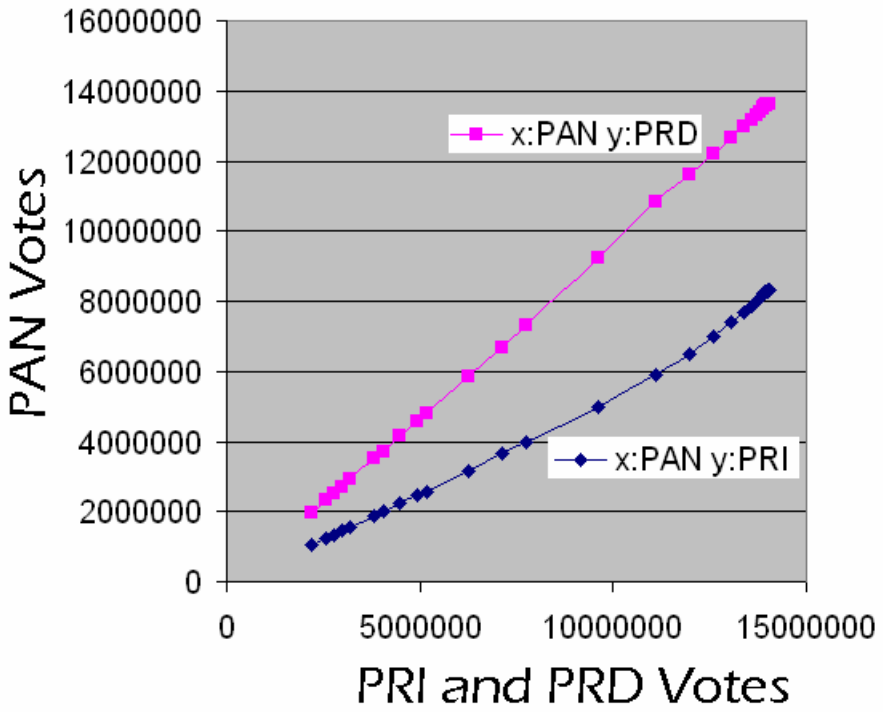
La pregunta inmediata fue: ¿como cuantificar esta anormalidad? La respuesta obvia es por medio de una prueba, por ejemplo el coeficiente de Pearson que mide la correlación del producto de los momentos, y que es usado desde ciencias sociales hasta ingeniería. El coeficiente se define como

$$r = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 \sum (y-\bar{y})^2}}$$

el cual, en español, determina si dos variables varían juntas; un coeficiente de cero significa independencia total de las variables, y un valor de uno significa dependencia total. No fue muy sorprendente encontrar que el coeficiente de Pearson para los votos PAN-PRD tenía un valor de 0.999974! [Para poner este valor en perspectiva, valores mayores de 0.80 son tomados generalmente —por ejemplo, por la NASA—como indicación de confiabilidad.] Las correlaciones de las otras curvas fueron 0.998205 (PAN-PRI) y 0.998196 (PRI-PRD); era obvio que el control había sido establecido sobre la relación PAN-PRD y –aún más importante— eso demostraba que los datos del PREP, si no eran falsos, al menos habían sido modificados, o *cuchareados* para ponerlo en términos de la jerga política actual.

El Algoritmo

Ya descubierta la existencia de una relación, la siguiente pregunta fue: ¿Qué tipo de relación fue impuesta sobre los votos artificiales? Como las curvas estaban extremadamente paralelas, uno podría esperar una relación lineal entre las tendencias de votación; esto fue confirmado por la siguiente gráfica.



La gráfica muestra, en el eje y, el número de votos que el PAN tenía al tiempo en el que el PRI y PRD tenían los votos de la coordenada x. La relación lineal es obvia. Un ajuste rápido a la línea PAN-PRD nos da la ecuación lineal

$$Y_{PAN}=279926.7904 + 1.008060294\ X_{PRD}.$$

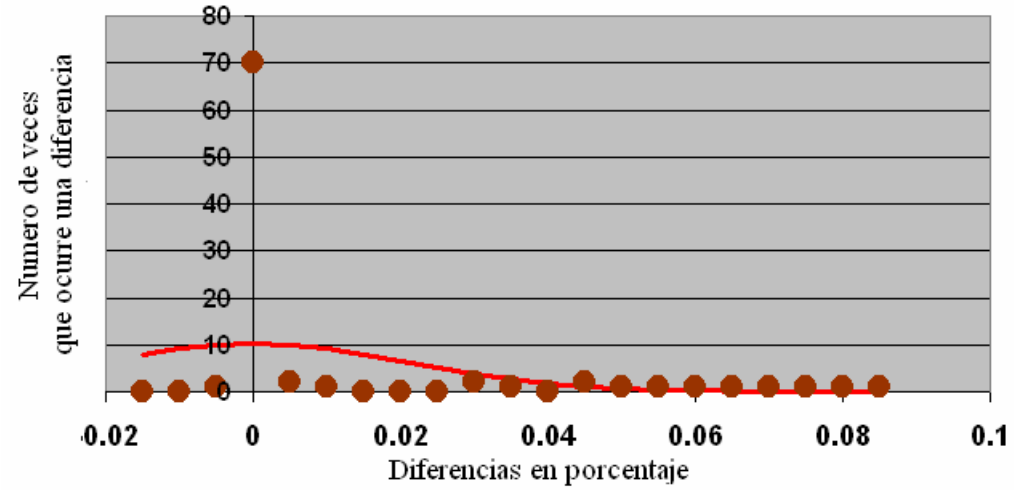
Cabe notar que la ventaja final de Calderón reportada por el PREP fue de 257,532 votos, después de que redujeron el reporte inicial de 402,708, lo cual está muy de acuerdo a la intersección de la ecuación, de 279,926, que correspondería a la ventaja inicial otorgada por el *software* al PAN.

Certidumbre en el ajuste

Ante esta evidencia, cualquier estadístico diría que la existencia de una relación lineal entre dos variables, por buena que esta sea, no es evidencia de manipuleo de datos. Lo cual nos llevó a pensar en la siguiente pregunta: ¿Podría existir una relación lineal tan perfecta en datos reales de una votación que debieran de contener un cierto grado de aleatoriedad introducido por la mezcla geográfica?

La respuesta esta vez se debe obtener de un estudio de las desviaciones que los datos tienen respecto al ajuste lineal. Muestreos reales siempre tienen desviaciones pequeñas de la tendencia principal, y se sabe que estas desviaciones siguen un patrón conocido como distribución de "*Gauss*", también llamada "Normal" por su ocurrencia tan frecuente en procesos naturales.

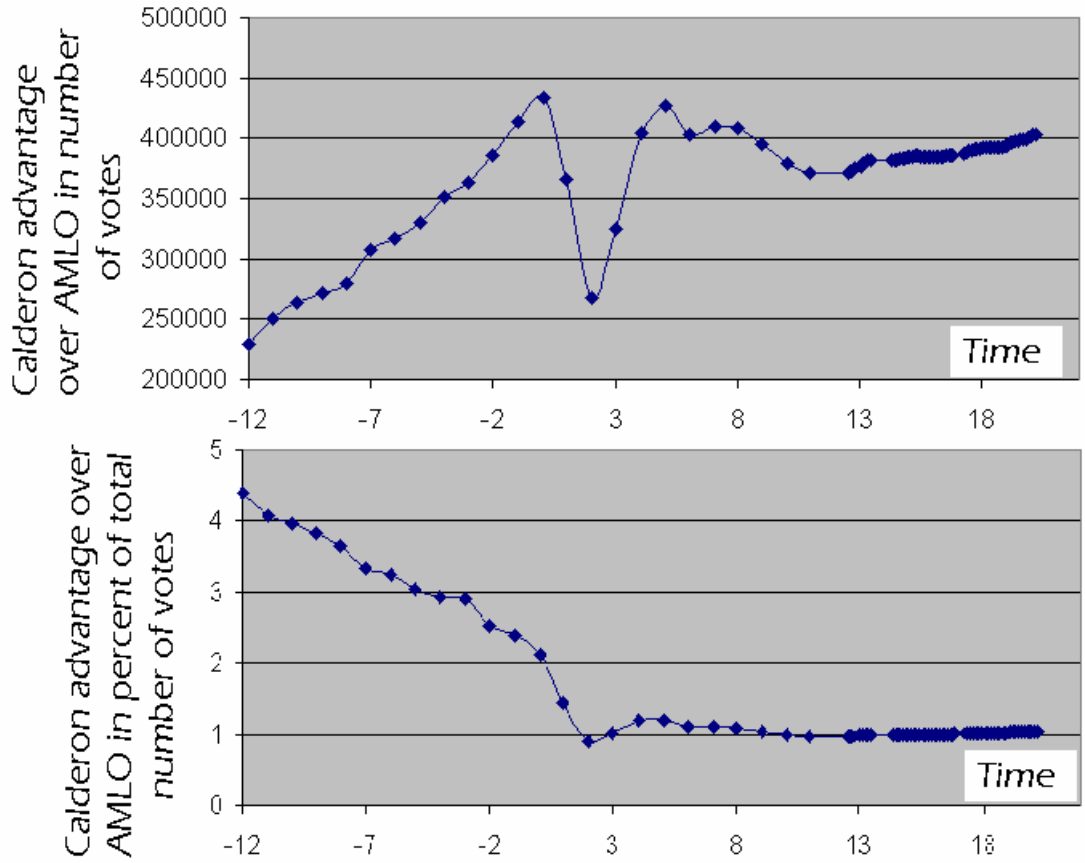
Viendo las diferencias entre los datos de la curva PAN-PRD de la gráfica previa, y la expresión analítica, uno puede obtener una distribución de estas diferencias y mostrarlas en una gráfica de frecuencia como la siguiente.



Usando los puntos para representar el número de veces que ocurre una cierta diferencia entre los datos y el ajuste, y dibujando con la curva roja la distribución normal esperada para tal muestra, es claro que los datos no siguen una distribución Gaussiana. El hecho de que la diferencia que ocurre con mayor frecuencia sea la de cero, es una indicación clara que los datos fueron producidos artificialmente y no tienen posibilidad de pasar como datos originados en un muestreo como el de una votación.

El Plan

¿Cuál era, entonces, el plan que siguieron los controladores del PREP? Esta duda se puede disipar viendo la variación en tiempo de la diferencia de votos de los dos contendientes principales. Las siguientes gráficas muestran la evolución en tiempo de esa diferencia tanto en número de votos como en porcentaje del total.



Y con esto se puede identificar el plan fácilmente. Aparentemente el algoritmo que operó en el "*conteo*" del PREP fue programado para dar a Calderón una abundante ventaja tempranera, y así forjar la impresión de triunfo y presionar al IFE a que lo declarara ganador en la conferencia de las 8:00 PM. Como el "conteo rápido" que hizo el IFE -independiente del PREP- bajo la supervisión de cinco científicos, no ratificó la ventaja ficticia, el anuncio del ganador se pospuso para la conferencia de las 11:00 PM.

Pero debido a que la ventaja de Calderón (en porcentaje) continuaba erosionándose, el anuncio fue pospuesto de nuevo, y el programa aparentemente entró en un segundo modo de operación en el que la caída de Calderón se aceleró. [Fue en ese momento que muchos de nosotros nos fuimos a la cama convencidos del triunfo de AMLO para la mañana.] Pero luego, alrededor de las 3:00 AM, el código entró en un tercer modo de operación y constriñó la diferencia Calderón-AMLO aproximadamente a 1% y la mantuvo constante en el resto del proceso. Por increíble que esto parezca, la relación

lineal entre los votos de los candidatos se mantuvo intacta durante todas estas maniobras turbulentas.

¿Y ahora qué?

Después de haber discutido este análisis con varios (muchos) colegas científicos de EU y México, no he encontrado uno que presente objeciones a las conclusiones. Tomando esto como una comprobación tácita de que los datos de PREP son artificiales, son muchas las dudas que surgen. ¿Cuál fue entonces el resultado real de la votación? ¿Quién manipuló el PREP para que construyera estos resultados? Y –tal vez más importante– ¿se usó este sistema computacional en alguna –cualquier- parte del recuento de votos?

Dos moralejas son las que se desprenden del presente ejercicio: 1) hay que tener cuidado con los métodos electrónicos de votación, y 2) nunca hay que subestimar el poder de las matemáticas.