Consideraciones metodológicas y resultados del modelo de pronóstico de la primera vuelta de las elecciones presidenciales colombianas 2018

Segunda versión

Cifras & Conceptos

4 de mayo de 2018

En estas notas describimos la segunda versión de la metodología de elaboración de los resultados del modelo de pronóstico de la primera vuelta de las elecciones presidenciales de 2018 en Colombia, elaborado por Cifras & Conceptos. Dado el debate público que ha tenido nuestro modelo, y luego de recibir una serie de preguntas y sugerencias, hemos decidido ampliar nuestras explicaciones con el fin de hacer lo más transparente posible el ejercicio.

1. Consideraciones preliminares

La primera consideración es que los modelos de pronóstico *no* son encuestas. La diferencia crucial es que, mientras las segundas no buscan pronosticar una variable, los primeros sí.

Las encuestas son una «foto» de una situación en un momento dado en el tiempo, la cual está sometida a error, por razones estadísticas. La idea fundamental de una encuesta es indagar qué piensa una población sobre un tema, sin necesidad de indagar a toda la población. Uno de los temas más arcanos para el gran público es cómo se puede saber qué es lo que piensan los colombianos cuando se les pregunta en una encuesta a solo 1.200 de ellos, por decir algo. Sin embargo, una de las fortalezas de la estadística es que nos dice que podemos *aproximar* la opinión de una *población* simplemente consultándole la opinión a una *muestra* bien definida. Cómo definir bien la muestra es un problema estadístico. Los principios estadísticos para definir bien una muestra han sido derivados de una rama de la estadística, conocida como *estadística clásica* o *frecuentista*, y bien merecen una mayor difusión y explicación entre el público consumidor de información proveniente de las encuestas.

Adicionalmente, vemos con desazón cómo, en el debate público, las encuestas son utilizadas como un pronóstico de lo que va a ocurrir en el futuro. Creemos que eso es incorrecto.

Parte del esfuerzo por enseñar a la opinión pública a «leer» las encuestas, fuera de explicar temas técnicos como la representatividad, los márgenes de error y otros similares, debe incluir la lección de que las encuestas no deben ser usadas como una descripción de lo que pueda suceder.

Un modelo de pronóstico, en cambio, es un instrumento estadístico cuyo propósito esencial es intentar estimaciones sobre las probabilidades futuras de un fenómeno. Nosotros hacemos una distinción entre *predecir* y *pronosticar*. *Predecir* es adivinar el futuro; *pronosticar* es describir un futuro probable, sujeto a incertidumbre. En ese orden de ideas, suponemos que es posible pronosticar, pero no predecir. Como señala Nate **?**, p. xv, la visión de que los estadísticos son adivinos «es una idea obstinadamente equivocada *[wrongheaded]* y bastante peligrosa» (en inglés en el original). Para evitar el «problema de la predicción» (que consiste en que «Amamos predecir cosas, y [en] que no somos muy buenos en ello», p. 13), **?**, p. 15 recomienda un «cambio actitudinal», representado por el teorema de Bayes, que obliga a «pensar distinto sobre nuestras ideas, y cómo testearlas. Debemos llegar a sentirnos más cómodos con la probabilidad y la incertidumbre. Debemos pensar más cuidadosamente sobre los supuestos y creencias que traemos a un problema» (en inglés en el original). Al hablar de pronóstico, debemos reconocer las condiciones de incertidumbre del proceso sobre el cual se quiere trabajar.

Un modelo de pronóstico no es una encuesta, pero puede nutrirse de ellas (y de otras fuentes de información). El modelo de pronóstico corresponde a un abordaje metodológico distinto. El modelo de pronóstico puede intentar hacer pronósticos a través del tiempo, como cuando me pregunto cuál va a ser la tasa de cambio mañana dada toda la información presente y pasada que tengo sobre la tasa de cambio (y quizás sobre otras variables), o puede hacer pronósticos sobre variables contemporáneas sobre las cuales no se puede hacer una medición directa, como cuando me pregunto cuánto puede pesar una persona cuando conozco con precisión sus hábitos alimenticios. Una forma bastante general de un modelo de pronóstico es el *modelo lineal generalizado* (MLG), cuya expresión formal es la siguiente (ver **?**, c. 15):

$$\mu = f(\text{lin}(x), [\text{parámetros}]) \tag{1}$$

$$y \sim \text{fdp}(\mu, [\text{parámetros}]) \tag{2}$$

donde μ es la tendencia central de la variable que queremos pronosticar (variable predicha), f es la función de vínculo inversa, $\text{lin}(\cdot)$ es la función lineal, x son las variables predictoras, y son los datos y «fdp» denota el nombre de una función de densidad de probabilidad cualquiera. Lo que dice la expresión (**??**) es que la tendencia central de los datos está matemáticamente descrita por una función lineal de las variables predictoras, que contiene unos parámetros (el problema es hallar esos parámetros). Lo que dice la expresión (**??**) es que los datos están distribuidos alrededor de la tendencia central μ de acuerdo con la función de densidad de probabilidad denominada acá «fdp», que contiene otros parámetros. En otras palabras, el modelo matemático no es exacto, sino que está sometido a error: es solo correcto en un sentido probabilístico.

Los principios estadísticos para construir un modelo de pronóstico usualmente han provenido de la estadística clásica o frecuentista, pero en los últimos años, debido al desarrollo de técnicas computacionales, también se han puesto en práctica principios estadísticos para construir un modelo de pronóstico que provienen de la estadística bayesiana. El principio fundamental de la estadística bayesiana es que la probabilidad de que un evento ocurra puede ser actualizada, en una forma matemáticamente precisa, en la medida en que más información o datos están disponibles.

Hemos decidido hacer un modelo de pronóstico para la primera vuelta de las elecciones presidenciales colombianas de 2018 principalmente por dos razones:

1. Hemos hecho, con bastante éxito, ejercicios de pronóstico electoral en el pasado. Cifras & Conceptos, en los últimos cuatro años, ha adelantado dos ejercicios en el ámbito nacional, orientados a pronosticar la composición del Senado en las elecciones de 2014 y 2018, y uno de carácter internacional, para pronosticar tres variables del voto latino para las elecciones presidenciales de Estados Unidos de 2016, en cinco estados (ver **?**). En los tres casos obtuvimos bastante, pero no total, éxito en nuestros pronósticos (a nuestros críticos les gusta concentrarse en nuestros fracasos, ignorando nuestra alta proporción de éxitos).

2. Creemos que, en la actualidad, las encuestas no están captando una realidad de la política colombiana, que consiste en que una parte de la intención de voto está guiada por consideraciones que, en este escrito, denominamos de *estructura política*. Estas consideraciones tienen que ver con que:

   > los candidatos al Congreso movilizan a sus estructuras partidistas para lograr grandes participaciones electorales (a veces sólo al Congreso, y otras veces también a la Segunda vuelta presidencial) (**?**, p. 7).

   Esto implica que, en las actuales condiciones, creemos que las encuestas tienen un sesgo, dado por el hecho de que no están captando adecuadamente el voto de estructura. El propósito, por lo tanto es incluir una estimación del voto de estructura. Para ello, es necesario explicitar cómo entendemos el proceso electoral colombiano y las hipótesis sobre las cuales se sustenta el modelo de pronóstico. Para comenzar, creemos que el resultado de una campaña electoral es incierto por su naturaleza y está afectado por diversas fuerzas, actores y eventos: la cultura política de la sociedad, la fortaleza de los partidos políticos, el contexto económico, el cubrimiento de los medios y, por supuesto, las propuestas, personalidad y desempeño de los candidatos. Todo esto determina en últimas el resultado final. Estos elementos se pueden organizar en dos grandes dimensiones: la estructura y la opinión. Ningún candidato podrá ganar solo con estructura, pero tampoco ninguno podrá hacerlo solo con opinión.

Es pertinente precisar que el modelo de pronóstico utiliza los resultados públicos de las encuestas como una de las varias fuentes de información, pero no es la única. También prevé

el levantamiento de información primaria en la forma de entrevistas semiestructuradas de alto nivel, y utiliza otra información secundaria en la forma de registros administrativos oficiales de la Registraduría Nacional del Estado Civil.

Es muy importante enfatizar que este modelo de pronóstico no refleja el interés de ningún candidato en particular. La independencia del estudio se garantiza por el hecho de que fue financiado 100 % con recursos de Cifras & Conceptos.

2. La analogía con los modelos bayesianos

En los modelos bayesianos:

1. La incertidumbre se expresa en términos de probabilidades. Por lo tanto, se requiere pensar probabilísticamente y presentar los resultados en un lenguaje que deje claro el nivel de incertidumbre de cada una de las afirmaciones. Cuando se *predice*, la afirmación es generalmente dicotómica y en términos absolutos: sucede o no sucede. Cuando se *pronostica* se enfatiza sobre el nivel de riesgo de no acertar que está implícito en el pronóstico.

2. Hay una *creencia previa*, expresada en términos probabilísticos; esa creencia se confronta con los *datos*; y esa confrontación produce, a través de la regla de Bayes, una *creencia posterior*, también expresada en términos probabilísticos.

3. El pronóstico es afectado por la ocurrencia de nuevos eventos, razón por la cual requiere actualizaciones permanentes, según van apareciendo nuevos datos. Es decir, el pronóstico está condicionado por la disponibilidad de información.

El modelo de pronóstico que construimos es de inspiración bayesiana. De una parte, la experiencia nos ha mostrado que, en muchos casos, el montaje de toda una batería de análisis con alto rigor estadístico no produce resultados sustancialmente mejores que procedimientos más simples.[1] De otra, la información para Colombia es escasa. Sin embargo, creemos que la esencia de un modelo bayesiano, que está expresada en el numeral 2 de la lista anterior, está contenida en nuestro modelo.

[1] **?**, pp. 129–130 lo ponen así:

> Mientras que los superpronosticadores ocasionalmente despliegan sus propios modelos matemáticos explícitos, o consultan los de otras personas, eso es raro. La gran mayoría de sus pronósticos son simplemente el producto de un pensamiento cuidadoso y de un juicio matizado. [...]
>
> Pero el hecho de que los superpronosticadores son casi uniformemente gente altamente numérica no es mera coincidencia. Una numericidad superior ayuda a los superpronosticadores, pero no porque eso les permite aprovechar modelos matemáticos arcanos que adivinan el futuro. La verdad es más simple, sutil y mucho más interesante (en inglés en el original).

Creencia previa. En nuestro caso, no tenemos una creencia previa expresada en términos probabilísticos, pero podemos asociar nuestra *creencia previa* con los resultados electorales del 11 de marzo, que eligieron al Congreso de la República. Obviamente, las elecciones al Congreso no son las elecciones a la presidencia, pero proveen una pista. La pregunta clave aquí es cómo los votos que se manifestaron en las elecciones de Congreso se transforman en votos a la presidencia de la República. En otras palabras, de acuerdo con ciertas reglas, asignamos los votos registrados en las elecciones al Congreso a los distintos candidatos presidenciales. La información para hacer esto la obtuvimos de los resultados de las elecciones al Congreso publicados por la Registraduría Nacional del Estado Civil y de entrevistas que realizamos con expertos.

Datos. Los *datos*, por su parte, los proveen las encuestas sobre intención de voto que regularmente están publicando las encuestadoras, no solo la nuestra. En otras palabras, nuestros datos están dados por la intención de voto en las encuestas, que es equiparable, en nuestro modelo, al voto de opinión.

Como fuentes para este ejercicio solo serán consideradas e incluidas las encuestas presentadas por firmas con alianzas formales con grandes medios de cobertura nacional. Las tenidos en cuenta para este ejercicio son:

- YanHaas, para Semana, La República, RCN y otros medios.
- Invamer, para Caracol Televisión, El Espectador y Blu Radio.
- Centro Nacional de Consultoría, para CM&.
- Guarumo, para El Tiempo.
- Cifras y Conceptos, para Caracol Radio y Red Más Noticias.
- Datexco, para W Radio.

Creencia posterior. Con estos *datos* revisamos nuestra *creencia previa* para producir una *creencia posterior*.

Los ponderadores. En una primera aproximación, ponderamos los votos de estructura y de opinión por los tamaños estimados de la población que votan de acuerdo con uno y otro criterio. En la medida en que en un modelo bayesiano hay más datos, la creencia posterior tiende a distanciarse más de la creencia previa, y a acercarse más a los «verdaderos» parámetros del sistema. Una pregunta clave acá es cómo la ponderación del voto de estructura va disminuyendo, en la medida en que la intención de voto captada en las encuestas va «incorporando» esa estructura.

Electos que apoyan al candidato de su partido	60 %
Electos que no apoyan al candidato de su partido	40 %
No electos que apoyan al candidato de su partido	30 %
No electos que no apoyan al candidato de su partido	20 %
Logo	80 %
Logo sin candidato	30 %
Alianza de logos	50 %

Cuadro 1: Tabla de endosos electorales

3. El modelo en detalle

El propósito de esta sección es que todo aquel que esté interesado pueda entender la lógica de construcción de nuestro modelo.

3.1. El cálculo de la intención de voto

La intención de voto por el candidato presidencial i es $v(i)$. Esa intención de voto se estima de la siguiente manera:

$$v(i) = \alpha(e)v(i,e) + \alpha(o)v(i,o)$$

donde $v(i)$ es la votación del candidato i; $\alpha(k)$ es la ponderación del factor k, donde k puede ser estructura (e) u opinión (o); y $v(i,k)$ es la votación del candidato presidencial i en el factor k.

3.2. El cálculo del voto de estructura

Para el cálculo del voto de estructura se tomaron los resultados electorales de las elecciones parlamentarias del 11 de marzo de todos los candidatos al Senado, elegidos o no. Se tomaron en consideración las votaciones iguales o superiores a 5.000 votos, incluyendo por defecto las votaciones asociadas con los logos de los partidos (no asociadas con ningún candidato). Sea $v(j)$ la votación del candidato (o partido) j en las pasadas elecciones de Congreso. Se supone que una porción β de esa votación se transfiere al candidato presidencial i: $v(i,j) = \beta(i,j)v(j)$. Para cada candidato al Senado j se calcula una transferencia o endoso de votos a su candidato presidencial i, de acuerdo con el cuadro **??**. Para definir la afiliación de cada candidato al Senado j a un candidato presidencial i se utilizó la información de expertos, recabada por medio de entrevistas. La única información que Cifras & Conceptos se abstiene de publicar es la asignación de los candidatos al Senado j a las categorías descritas en el cuadro **??**, para salvaguardar la reserva de la fuente. El voto de estructura para el candidato presidencial i es simplemente $\sum_j \beta(i,j)v(j)$, donde j puede ser un candidato al Senado o un partido político.

Los cálculos anteriores señalan que hubo 13.7 millones de personas que votaron por candidatos al Senado el pasado 11 de marzo. De estos, se supone que se transfieren a algún candidato presidencial 5.6 millones de votos. Esta es la magnitud inicial del voto de estructura. En posteriores actualizaciones, esta magnitud puede cambiar.

Aún hay un número plural de parlamentarios electos sobre los cuales no hay total certeza a cuál candidato a la presidencia van a adherir, y que, dependiendo de las dinámicas de las campañas y la proximidad a las elecciones, se irán precisando.

3.3. El cálculo del voto de opinión

El voto de opinión se asocia con la pregunta de intención de voto reflejada en las encuestas públicas de reconocidas firmas, realizadas a la fecha después de las elecciones parlamentarias del 11 de marzo de 2018. La intención de voto de cada candidato es el promedio ponderado de las intenciones de voto del candidato en cada una de las encuestas, resultado de aplicar cuatro ponderaciones, así:

1. La vejez de la encuesta, la cual hace referencia al número de días desde la realización de la encuesta y la fecha de las elecciones. Cada encuesta es ponderada por vejez de la siguiente manera: si la encuesta es la más reciente disponible, su ponderación es 1. Si no es la más reciente, entonces su ponderación es dr/dn, donde dr son los días que faltan para la elección en la encuesta más reciente, y dn son los días que faltan para la elección en la encuesta n.

2. El tamaño de la muestra adoptada en la aplicación de las encuestas realizadas, conforme a las fichas técnicas que acompañan los resultados de cada medición. Cada encuesta es ponderada por tamaño de la siguiente manera: si la muestra es la de mayor tamaño, su ponderación es 1 multiplicado por la vejez. Si no es la de mayor tamaño, entonces su ponderación es $dr/dn \times tn/tg$, donde tn es el tamaño de la muestra n, y tg es el tamaño de la muestra más grande.

3. El método de recolección (telefónico o presencial) dando mayor ponderación al segundo. A las encuestas telefónicas se les da una ponderación de 0,6 y a las encuestas presenciales se les da una ponderación de 1.

4. La dispersión geográfica de las unidades muestrales, reflejada a partir del número de municipios en los que se aplicaron las diferentes encuestas. Cada encuesta es ponderada por dispersión geográfica de la siguiente manera: si la muestra es la de mayor número de municipios, su ponderación es 1 multiplicado por la vejez. Si no es la de mayor número de municipios, entonces su ponderación es $dr/dn \times mn/mg$, donde mn es el número de municipios de la encuesta n, y mg es el número de municipios de la encuesta con mayor número de municipios.

3.4. El cálculo de las ponderaciones α

Las ponderaciones $\alpha(k)$ que utilizamos para el ejercicio son $\alpha(e) = 0,38$ y $\alpha(o) = 0,62$. Los ponderadores irán cambiando en la medida en que las elecciones se encuentran más próximas.

3.5. Los resultados finales

Con la combinación de los resultados de las dimensiones de opinión y estructura se construyen dos escenarios en los que se valoran de manera diferente las dos dimensiones, lo que se traduce en la probable brecha del resultado final de las elecciones para cada candidato. Un resumen de los resultados se ilustra a continuación:

Nombre	Rango	4 abril	17 abril	2 mayo
Iván Duque	Max	37,7 %	36,4 %	37,9 %
	Min	37,2 %	35,6 %	33,7 %
Gustavo Petro	Max	17,4 %	18,9 %	21,6 %
	Min	15,7 %	17,0 %	18,3 %
Sergio Fajardo	Max	13,3 %	12,4 %	14,2 %
	Min	12,8 %	12,3 %	11,3 %
Germán Vargas	Max	21,0 %	23,8 %	24,1 %
	Min	18,2 %	20,9 %	20,8 %
Humberto de la Calle	Max	6,3 %	5,1 %	5,1 %
	Min	5,9 %	4,8 %	3,3 %
Voto en blanco	Max	10,7 %	6,9 %	4,4 %
	Min	5,3 %	3,5 %	3,1 %

A continuación describimos las referencias y la bibliografía recomendada. Los asteriscos denotan los trabajos no citados en este artículo que igual incluimos como bibliografía recomendada.

Referencias y bibliografía recomendada

Cifras & Conceptos, «Un modelo de pronóstico del voto latino para las elecciones presidenciales de Estados Unidos de 2016», mimeo.

* Gelman, Andrew, John B. Carlin, Hal S. Stern, David D. Dunson, Aki Vehtari y Donald B. Rubin (2013), *Bayesian Data Analysis*, Third Edition, Boca Raton: CRC Press (Taylor and Francis Group, Chapman and Hall Book).

Kruschke, John K. (2011), *Doing Bayesian Data Analysis: A Tutorial with R, JAGS and Stan*, Second Edition, Amsterdam: Academic Press (Elsevier). Esta segunda edición de 2015.

Silver, Nate (2015), *The Signal and the Noise: Why So Many Predictions Fail—but Some Don't*, New York: Penguin Books.

Stephens–Davidowitz, Seth (2017), *Everybody Lies: Big Data, New Data, and What the Internet Can Tell Us About Who We Really Are*, New York: HarperCollins.

Tetlock, Philip E. y Dan Gardner (2015), *Superforecasting: The Art and Science of Prediction*, New York: Broadway Books.

Vargas, Camilo (2014), «Las expresiones del voto en Colombia: elecciones nacionales 2014», mimeo, disponible en https://moe.org.co/home/doc/moe_mre/2014/-Las %20expresiones %20del %20voto %20en %20Colombia-Revista %20Foro.pdf