Agregación de encuestas en 2018: Un modelo bayesiano en México

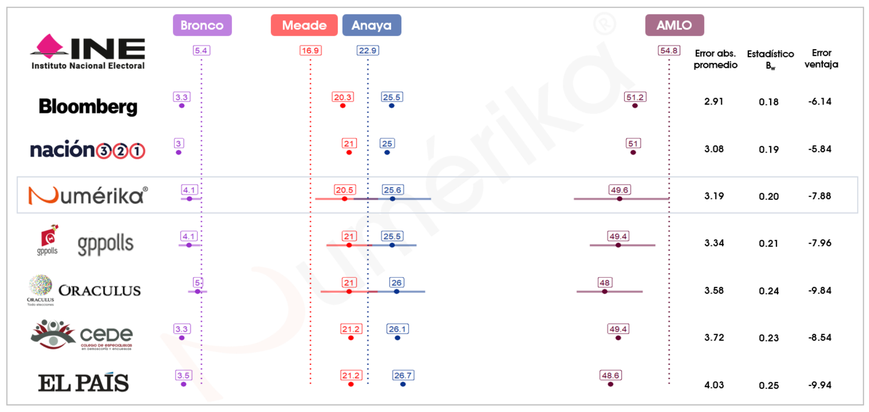
Fernando Antonio Zepeda Herrera (fazepher@gmail.com) *Numérika*

Resumen

En los procesos electorales de México, así como en otros países, diferentes actores realizan encuestas con el objetivo de obtener información sobre el estado de las preferencias de los votantes. Las estimaciones publicadas generalmente son muy distintas entre sí, por lo que la opinión pública puede beneficiarse de métodos que combinen las diferentes fuentes de información.

Estos modelos de agregación de encuestas, conocidos como *Poll of Polls*, han cobrado mayor relevancia, particularmente desde el pronóstico de Nate Silver en 2008 para la elección presidencial en EUA. Sin embargo, en el caso de México, no había habido muchos modelos públicos de este tipo hasta la elección presidencial de 2018.

Presento aquí, el modelo que desarrollé en Numérika, basado principalmente en el trabajo de modelos multinomiales dinámicos jerárquicos de Cargnoni, Müller y West. Este *Poll of Polls* y 6 más fueron recopilados por el sitio Oráculus, lo que introdujo la agregación modelada de encuestas a la discusión pública en México.



Estimación puntual y límites de los intervalos de probabilidad al 95% para las preferencias efectivas entre los 4 candidatos de la elección presidencial. Por cuestiones de redondeo, las sumas pueden no dar 100%

La línea punteada indica el porcentaje efectivo de los cómputos distritales del INE. El error promedio considera las desviaciones absolutas para los 4 candidatos, el estadístico B_w fue propuesto por Arzheimer y Evans en 2014 para sistemas multipartidistas, mientras que el error ventaja es la diferencia en la estimación del margen entre los dos punteros. Las tres medidas son mejores entre más cercanas sean al cero.

Supuestos

Considerando el objetivo de agregación de encuestas, necesitábamos un modelo que contemplara los siguientes supuestos generales en un proceso electoral:

- En cada punto en el tiempo existen unas *preferencias subyacentes* entre los votantes que determinarían el resultado de la elección, *si esta se realizara en ese momento*.
- Cada encuesta busca estimar dichas preferencias pero, debido a errores muestrales y no muestrales, el dato obtenido es una desviación de las mismas.
- Dicha desviación depende, primordialmente, de la casa encuestadora que realizó el ejercicio demoscópico a través de dos fuentes independientes:
  - Las posibles propensiones, constantes en el tiempo, a sobre o subestimar a cada candidato. Es decir, pueden existir *sesgos constantes de casa*.
  - Asimismo, *cada casa encuestadora tiene una volatilidad común entre sus encuestas*, proveniente de todos los errores muestrales y no muestrales que no pueden clasificarse como parte de los sesgos constantes.

Modelo

Modelaremos series de tiempo multinomiales:

$$Y_e \sim Mult(\theta_e, n_e)$$

Cada encuesta tiene asociada una *medición latente* η_e relacionada con las probabilidades multinomiales θ_e mediante la transformación logística aditiva:

$$\eta_{e,i} = ln\left(\frac{\theta_{e,i}}{\theta_{e,I}}\right) \quad \forall\, i = 1, \ldots, I-1,$$

donde I es el número de candidatos. La medición está centrada en una media que depende del nivel "real" del sistema en la semana de la encuesta y de los posibles sesgos constantes de la casa encuestadora; su varianza depende exclusivamente de un parámetro de dispersión de la casa encuestadora:

$$\eta_e \sim N(\mu_{t[e]} + \gamma_{c[e]}, V_{c[e]}).$$

La evolución del sistema depende del último punto en el tiempo, teniendo como origen una media inicial m:

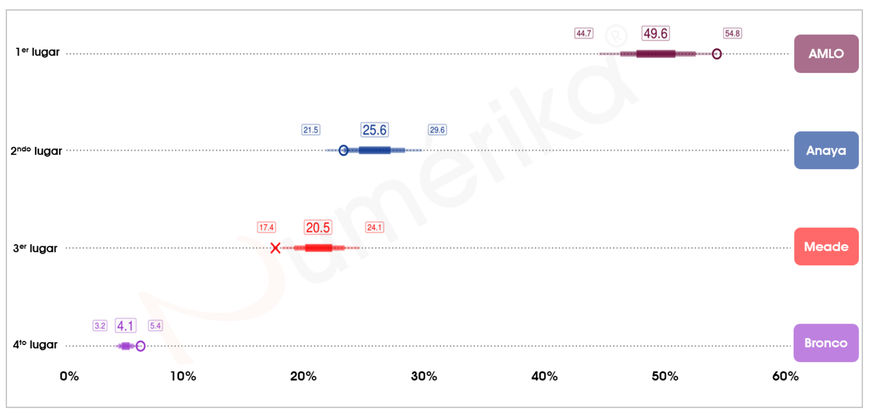
$$\mu_t \sim N(\mu_{t-1}, W), \qquad \mu_0 \sim N(m, W).$$

Se requiere completar el modelo asignando distribuciones iniciales a los parámetros. Para no imputar, *a priori*, un sesgo determinado a ninguna casa encuestadora, se decidió que las distribuciones de los sesgos fueran normales centradas en 0:

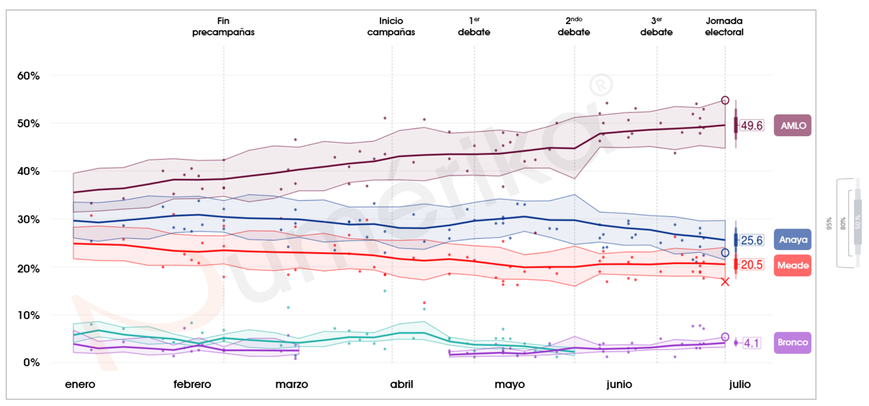
$$\gamma_c \sim N(0, \Omega_c) \quad \forall\, c = 1, \ldots, C,$$

donde C es el número de casas encuestadoras en muestra.

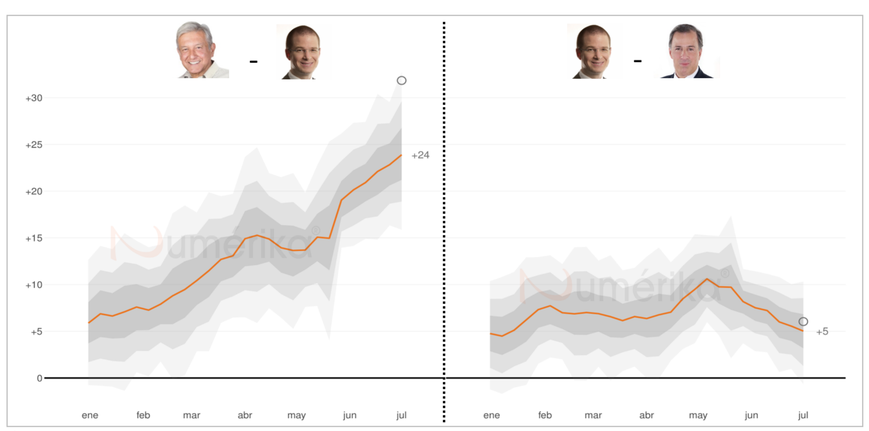
Principales resultados



Estimación puntual y límites de los intervalos de probabilidad al 95% para las preferencias efectivas entre los 4 candidatos de la elección presidencial. Por cuestiones de redondeo, la suma no da 100%
Los segmentos de diferente grosor representan intervalos centrales de probabilidad al 50%, 80% y 95% como se muestra en el diagrama a la derecha. Un círculo representa que el intervalo contuvo al verdadero valor obtenido por el candidato en la jornada electoral, mientras que un tache significa que la votación obtenida se encontró fuera del intervalo.



Estimaciones individuales de encuestas públicas para la proporción efectiva de votos entre los 4 candidatos de la elección presidencial. Evolución semanal de las preferencias modeladas.
Los últimos datos se presentan como estimaciones puntuales e intervalos centrales de probabilidad al 50%, 80% y 95%, como se muestra en el diagrama del costado derecho.
Un círculo representa que el intervalo contuvo al verdadero valor obtenido por el candidato en la jornada electoral, mientras que un tache significa que la votación obtenida se encontró fuera del intervalo.



Última estimación puntual de la diferencia en puntos porcentuales de las preferencias efectivas entre cada par de candidatos (margen C1 – C2).
Evolución de la estimación del margen, la línea central indica la estimación puntual y las bandas son los intervalos centrales de probabilidad al 50%, 80% y 95%. El círculo representa que el intervalo contuvo al verdadero margen entre los candidatos.

Retos, lecciones, comentarios...

El proyecto ofreció varios retos y lecciones que considero valioso compartir, más allá del desempeño específico del modelo. Los invito a platicar conmigo sobre los que más les interesen:

- Modelar incertidumbre en varianzas es difícil, más cuando se tiene *small data*.
  - Se consideraron 58 encuestas en 6 meses, en EUA a 2 años de la elección ya hay en la base de *Fivethirtyeight* 99 encuestas en los últimos 5 meses sobre las primarias.
- Las circunstancias alejan el modelo "del pizarrón":
  - ¿Qué hacer cuando el Bronco no junta las firmas y las encuestadoras lo sacan de sus estimaciones? ¿Cómo reintroducirlo después de la decisión del TEPJF? ¿Cómo reaccionar ante la declinación de Margarita Zavala? ¿Cómo hacer esto "en vivo"?
- *All models are wrong... some are useful*:
  - ¿Cuál es la unidad de tiempo adecuada? ¿Todas las casas encuestadoras deben considerarse insesgadas? ¿Cómo incorporar información adicional?
- ¿Cómo determinar las distribuciones iniciales bajo el paradigma bayesiano?
  - Las medias *a priori* fueron 35 % AMLO, 30 % Anaya, 25 % Meade, 5 % Zavala y 5 % Bronco. Para las varianzas fueron Wishart inversas de 6 grados de libertad y matrices de escala diagonales con elementos iguales a 0.05.
- Hacia el futuro:
  - ¿Cómo contribuir a mejorar la calidad de las encuestas? ¿Para qué son útiles las agregaciones y para qué no? ¿Cómo mejorar la precisión del modelo?...

Referencia principal

Cargnoni, Claudia, Peter Müller y Mike West. 1997. "Bayesian Forecasting of Multinomial Time Series through Conditionally Gaussian Dynamic Models". *Journal of the American Statistical Association* 92 (438): 640-647.

Para más información visitar www.numerika.mx o contactarme vía correo electrónico o Twitter.