Un modelo bayesiano para el pronóstico de los resultados de las elecciones presidenciales en Estados Unidos

Cifras y Conceptos*

5 de septiembre de 2016
Versión preliminar para comentarios

1. Introducción

En este documento se presenta una metodología bayesiana para el pronóstico de los resultados de las elecciones presidenciales del 8 de noviembre de 2016 para la población latina de Estados Unidos en los estados de Arizona, Colorado, Florida, Nevada y Nuevo México. Esta metodología ha sido desarrollada por Cifras y Conceptos para complementar el seguimiento que hace la cadena de noticias Univisión de las elecciones presidenciales.

Se supone que el parámetro de interés del estudio (en cada estado) es la proporción de votantes para cada uno de los tres candidatos más importantes en la contienda presidencial. Así mismo, se supone que el conjunto de los estados se puede denotar como $S = \{1, 2, 3, 4, 5\}$. Los resultados de las elecciones en el estado s se denotan como $\boldsymbol{\pi}_s = (\pi_{s1}, \pi_{s2}, \pi_{s3})$ con $s = 1, \ldots, 5$ indicando los cinco estados. Nótese que $\pi_{s1}, \pi_{s2}, \pi_{s3}$ denota el porcentaje de votos por los candidatos republicano, demócrata y libertario, respectivamente. A su vez, se supone que $\pi_{s1} + \pi_{s2} + \pi_{s3} = 1$.

La estimación del parámetro $\boldsymbol{\pi}_s$ se hará por medio del análisis bayesiano, que busca combinar diferentes fuentes de información para estimar el parámetro de interés. El uso de la estadística bayesiana permite actualizar los resultados a medida que se actualizan las fuentes de información. Para su implementación, se requiere resumir la información previa en forma

* Este artículo estuvo dirigido por Andrés Gutiérrez y Hanwen Zhang. El equipo de Cifras y Conceptos para este trabajo está compuesto por César Caballero, Daniel Castellanos, Juan Martín Clavijo, David Rodr'iguez, Mar'ia Paula Rojas, Andrea Mateus, Miguel Ángel León, Juan Sebastián Garcés, Carolina Parra, Ángela Ordóñez y Giovany Babativa.

de una distribución de probabilidad denominada distribución *a priori*, la cual se denota por $p(\boldsymbol{\pi}_s)$. La fuente de información que se actualizará constantemente se resume en la función de verosimilitud, denotada por $p(\mathbf{y}|\boldsymbol{\pi}_s)$. Estas dos fuentes de información se combinan por medio del teorema de Bayes, induciendo una nueva distribución denominada la distribución *a posteriori* (de donde se obtiene la estimación bayesiana del parámetro $\boldsymbol{\pi}_s$). Esta distribución se obtiene mediante el teorema de Bayes y está dada por:

$$p(\boldsymbol{\pi}_s|\mathbf{y}) \propto p(\boldsymbol{\pi}_s)p(\mathbf{y}|\boldsymbol{\pi}_s)$$

En este proyecto, la distribución *a priori* se obtendrá usando información histórica de votación de las últimas seis elecciones presidenciales en los cinco estados de interés. Esta información incluye fundamentalmente la información demográfica, los resultados de una encuesta post–electoral y las *exit polls*. Es decir, debemos aclarar que allí no se usan las encuestas electorales. Estos datos se analizan por medio de un modelo de regresión multinomial con efectos aleatorios por estados. Con este modelo se busca pronosticar los resultados de votación para el año 2016, de donde se obtiene la distribución *a priori*, la cual estará dada en forma de una distribución Dirichlet (puesto que los componentes del parámetro de interés corresponden a tres porcentajes que suman 100 %). A este lo denominamos el *modelo de estructura*.

Este modelo se construyó en varias etapas que es importante enfatizar: primero, se hizo un modelo para cada uno de los tres candidatos, y luego uno combinando los resultados de los tres. El modelo se contrastó contra los resultados de la elección de 2012, y la comparación con lo efectivamente sucedido muestra un capacidad explicativa muy fuerte. La siguiente etapa fue calcular de la misma forma el voto latino, lo cual también se realizó con buenos resultados. Con eso listo, se produce la distribución *a priori*, que es el pronóstico de estructura para 2016, y que, repetimos, *no utiliza* ninguna encuesta electoral de ese periodo.

La forma de la función de verosimilitud se obtiene del análisis de los resultados de encuestas por medio de una regresión multinivel con post–estratificación (llamado MRP, por sus siglas en inglés), y también tiene en cuenta variables que describen la situación económica y social del momento (variables de *contexto*, que describen la coyuntura que está por fuera del control de las campañas políticas), así como variables que describen la estrategia de campaña de cada candidato (variables de *estrategia*, que describen la coyuntura que está bajo el control de las campañas políticas). El centro de la información de este componente son las encuestas electorales, pero incluimos otra información, como financiación, registro final de votantes, agenda, cubrimiento mediático y redes sociales. También construimos y utilizamos una matriz de transición de la intención de voto de los ciudadanos que participaron en las primarias *(gross flow)*, la cual es un insumo importante para este segundo componente del modelo.

Específicamente, con el modelo MRP se obtiene una estimación cruda de la intención de voto para cada uno de los cinco estados. Luego, estas estimaciones se ajustan usando las variables de los componentes de contexto y estrategia por medio de una regresión multinomial. A esto lo llamamos el *modelo de contexto y estrategia*, que sí utiliza las encuestas electorales de 2016. Los resultados finales se expresarán en términos de una distribución multinomial —que corresponderá a la distribución de la verosimilitud— la cual, junto con la distribución Dirichlet —como distribución *a priori*—, induce una familia de distribución conjugada, con la cual se realizará la inferencia *a posteriori*. Los resultados se contrastan contra el agregado de encuestas mensualizado proveniente de dos fuentes: *RealClearPolitics* y *The Huffington Post*.

Una vez se tengan la distribución *a priori* y la función de verosimilitud, se aplicará la teoría bayesiana estándar para un modelo Dirichlet–multinomial y, de esta forma, obtener la estimación bayesiana de la intención de voto en los cinco estados. Cada vez que se disponga de los resultados de nuevas encuestas, se actualiza el modelo MRP y la función de verosimilitud. En la actualidad se dispone de cinco puntos de observación, y en noviembre se podrán alcanzar 12 más. De esta forma, se obtienen estimaciones de la intención de voto en diferentes puntos del tiempo.[1] El nivel de precisión de las estimaciones aumenta en la medida en que se vayan acumulando más observaciones. Se supone que cada punto tiene menos peso en el pronóstico final a medida que se acerca la fecha de las elecciones presidenciales (8 de noviembre de 2016). Es decir, las observaciones más recientes tienen más peso a la hora de determinar las estimaciones que las observaciones más antiguas. Estas estimaciones son producidas por medio de un modelo de decrecimiento exponencial para cumplir el objetivo final de pronosticar la intención de voto para el día de las elecciones. Finalmente, luego de tener el modelo general, basados en información cualitativa y de expertos, se harán ajustes por estado, en términos de pesos de las variables. Adicionalmente, el documento también explica brevemente el uso de los modelos de estimación de áreas pequeñas para estimar la intención de voto a nivel de condados.

Este documento se organiza de la siguiente forma: en la sección 2 se presenta el modelo de estructura, el cual hace uso de los datos históricos de las últimas seis elecciones presidenciales en los cinco estados de interés; en la sección 3 se presenta el modelo MRP; en la sección 4 se presenta el modelo de contexto-estrategia y evolución, el cual utiliza los datos de las encuestas, junto con variables que miden la coyuntura socioeconómica y la estrategia de campaña de los diferentes candidatos; en la sección 5 se presenta la forma de combinar los resultados de los modelos de estructura y de contexto–estrategia y evolución para obtener una estimación

[1] Note que cada punto en el tiempo corresponde a las estimaciones obtenidas de una encuesta particular.

bayesiana final de los resultados de las elecciones; en la sección 6, se presenta el modelo para la estimación de áreas pequeñas utilizado con el fin de pronosticar la intención de voto en los condados de los estados; finalmente, la sección 7 presenta los principales hallazgos y conclusiones.

2. Modelo de estructura

El modelo de estructura utiliza los datos correspondientes a los resultados de votación de elecciones presidenciales en los cinco estados entre 1992 y 2012. Puesto que la fuente de información es histórica, los resultados de este modelo serán utilizados para construir la distribución *a priori* del parámetro de interés $\boldsymbol{\pi}_s$ en el estado s. Dada la naturaleza del vector $\boldsymbol{\pi}_s$, la distribución *a priori* por excelencia es la distribución Dirichlet (en la sección 4 se discutirá con más detalles). Para conocer los parámetros de esta distribución, es necesario saber cuál es la estimación de $\boldsymbol{\pi}_s$ usando únicamente los datos históricos, esto es, se requiere conocer el pronóstico de $\boldsymbol{\pi}_s$ para el año 2016 usando el modelo de estructura.

Los datos disponibles para el ajuste del modelo de estructura se denotarán por $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_{30}$, donde cada $\mathbf{z}_i$ contiene el números de votos de los partidos Republicano, Demócrata e Independiente de la i-ésima observación. Esto es,

$$\mathbf{z}_i = (z_{i1}, z_{i2}, z_{i3}), \quad i = 1, \ldots, 30$$

donde z_{i1} denota el número de votos obtenidos por el candidato republicano en la i-ésima observación, z_{i2} el número de votos obtenidos por el candidato demócrata y z_{i3} el número de votos obtenido por el candidato independiente. Teniendo en cuenta la naturaleza de los datos, se tiene que la distribución de $\mathbf{z}_i$ es multinomial, esto es,

$$\mathbf{z}_i \sim multinomial(n_i, \boldsymbol{\xi}_i)$$

donde n_i representa el número total de votos, y $\boldsymbol{\xi}_i = (\xi_{s1}, \xi_{s2}, \xi_{s3})$ es el vector de probabilidades para los tres partidos con $\xi_{i1} + \xi_{i2} + \xi_{i3} = 1$.

La regresión multinomial busca relacionar el vector de probabilidades de la i-ésima observación $\boldsymbol{\xi}_i$ con Q variables explicativas $W_{i1}, W_{i2}, \ldots, W_{iQ}$. No existen restricciones sobre los valores que toman las variables explicativas, pero se debe garantizar que $\xi_{s1}, \xi_{s2}, \xi_{s3}$ sean todos escalares positivos menores que 1, y que además la suma de los tres escalares sea la unidad. La forma más común de formular el modelo es tomar una de las tres categorías como base y modelar el cociente entre las probabilidades de las otras con respecto a la primera por medio de la función logarítmica. En este caso, se tomó como categoría base la inducida por el candidato independiente.

En resumen, el modelo viene dado por:

$$\ln \frac{\xi_{ik}}{\xi_{i3}} = \eta_{ik}, \quad k = 1, 2$$

que es equivalente a:

$$\begin{aligned}
\xi_{i1} &= \frac{e^{\eta_{i1}}}{e^{\eta_{i1}} + e^{\eta_{i2}} + 1} \\
\xi_{i2} &= \frac{e^{\eta_{i2}}}{e^{\eta_{i1}} + e^{\eta_{i2}} + 1} \\
\xi_{i3} &= \frac{1}{e^{\eta_{i1}} + e^{\eta_{i2}} + 1}
\end{aligned}$$

Por otro lado, las variables explicativas entran al modelo siguiendo las expresiones:

$$\begin{aligned}
\eta_{i1} &= \mathbf{W}_i' \boldsymbol{\gamma}_1 \\
\eta_{i2} &= \mathbf{W}_i' \boldsymbol{\gamma}_2
\end{aligned}$$

para $i = 1, \ldots, 30$. El vector $\mathbf{W}_i$ contiene los valores de todas las Q variables explicativas en la observación i-ésima, y los vectores $\boldsymbol{\gamma}_1$ y $\boldsymbol{\gamma}_2$ denotan los vectores de coeficientes de regresión.

En el apéndice se encuentran los nombres de las variables utilizadas en el modelo, así como los coeficientes estimados. La estimación del modelo fue implementada en el software estadístico `R`, específicamente usando el paquete `rjags` (Plummer, 2016), que permite la integración del software `JAGS` (Plummer, 2003) con `R`. Los códigos computacionales están disponibles a petición de los lectores.

En el siguiente cuadro se presenta el pronóstico para las elecciones de 2016 usando el modelo de estructura. El número de votos fue obtenido usando los porcentajes pronosticados junto con la proyección de la variable `número de votos` para el año 2016. En el contexto de la estadística bayesiana, los siguientes porcentajes corresponden a las estimaciones *a priori* de los vectores de parámetros $\boldsymbol{\pi}_s$ con $s = 1, \ldots, 5$. En la sección 5 se describen los detalles para escoger los parámetros de las distribuciones *a priori*.

		Republicano	Demócrata	Libertario
Arizona	Porcentaje	49.99 %	49.55 %	0.46 %
	Número de votos	1,424,005	1,411,316	13,001
Colorado	Porcentaje	46.03 %	52.22 %	1.75 %
	Número de votos	1,367,493	1,551,332	52,000
Florida	Porcentaje	46.96 %	52.67 %	0.37 %
	Número de votos	4,951,140	5,552,772	39,224
Nevada	Porcentaje	43.44 %	55.44 %	1.12 %
	Número de votos	581,255	741,864	15,014
New Mexico	Porcentaje	38.44 %	58.82 %	2.74 %
	Número de votos	385,125	589,306	27,433

Cuadro 1: *Resultados del modelo de estructura.*

3. Modelo MRP

Para analizar los resultados de las encuestas se decidió usar una herramienta estadística conocida como *modelo de regresión multinivel con post-estratificación*, que es ampliamente utilizado en estudios electorales (Gelman y Hill, 2007; Wyatt, Dropp, Byrum y Dulin, 2016; Ghitza y Gelman, 2013; Park, Gelman y Bafumi, 2016). Este modelo utiliza principalmente dos fuentes de información:

1. *Encuestas a personas* que contienen la intención de voto de la persona entrevistada, así como información personal como el género, edad, nivel educativo y estado donde está registrada para votar, entre otras.

2. *Información agregada* a nivel de estados sobre el número total de personas en cada una de las posibles combinaciones de las variables de información personal.

El modelo de regresión multinivel con post-estratificación combina las dos anteriores fuentes de información para lograr la estimación de los parámetros de interés a nivel de estados. Dicho modelo está compuesto por dos partes: la primera consiste en el ajuste de un modelo de regresión multinivel usando la primera fuente de información y la segunda corresponde a la post-estratificación. A continuación se exponen los detalles.

3.1. Modelo de regresión multinivel

En las encuestas a personas, la variable de interés para analizar es la intención de voto de la persona. Aunque en este momento se cuenta con tres posibles candidatos de los partidos republicano, demócrata y libertario, hay personas que al momento de ser encuestadas manifiestan estar indecisas; es decir, personas que no saben por quién votar. De esta forma, los datos sobre la intención de voto recolectados en las encuestas corresponden a datos multinomiales con cuatro categorías: las correspondientes a los tres candidatos y una categoría de indecisos.

Por lo anterior, se define $\boldsymbol{\rho}_i = (p_{i1}, p_{i2}, p_{i3}, p_{i4})'$ como el vector que contiene la probabilidad de que la persona i-ésima vote por el candidato republicano, demócrata, libertario o indeciso, respectivamente. El objetivo de este modelo es relacionar los componentes de $\boldsymbol{\rho}_i$ con la información personal, $\mathbf{X}_i$. Esto se realiza de manera análoga al procedimiento descrito en la sección 2, tomando la probabilidad de una categoría como base, y modelando el logaritmo del cociente entre la probabilidad de las otras categorías con la base. Al tomar la categoría de indecisos como base, tenemos que el modelo está formulado como:

$$
\begin{aligned}
\ln \frac{\rho_{i1}}{\rho_{i4}} &= \mathbf{X}_i'\boldsymbol{\beta}_1 \\
\ln \frac{\rho_{i2}}{\rho_{i4}} &= \mathbf{X}_i'\boldsymbol{\beta}_2 \\
\ln \frac{\rho_{i3}}{\rho_{i4}} &= \mathbf{X}_i'\boldsymbol{\beta}_3
\end{aligned}
$$

Las variables que describen la información personal son:[2]

- Género, que toma dos posibles valores: hombre y mujer.
- Edad, que toma cuatro posibles valores: entre 18 y 34 años, entre 35 y 49 años, entre 50 y 64 años, y 65 años o más.
- Raza, que toma cuatro valores: blanca, hispana, negra y otra.
- Estado, que toma cinco valores: Arizona, Colorado, Florida, Nevada y Nuevo México.

De esta forma, el modelo de regresión multinomial viene dado por:

$$
\begin{aligned}
\ln \frac{\rho_{i1}}{\rho_{i4}} &= \beta_{0,1} + \beta_{l[i],1} \times Genero_i + \beta_{m[i],1} \times Edad_i + \beta_{r[i],1} \times Raza_i + \beta_{s[i],1} \times Estado_i \\
\ln \frac{\rho_{i2}}{\rho_{i4}} &= \beta_{0,2} + \beta_{l[i],2} \times Genero_i + \beta_{m[i],2} \times Edad_i + \beta_{r[i],2} \times Raza_i + \beta_{s[i],2} \times Estado_i \\
\ln \frac{\rho_{i3}}{\rho_{i4}} &= \beta_{0,3} + \beta_{l[i],3} \times Genero_i + \beta_{m[i],3} \times Edad_i + \beta_{r[i],3} \times Raza_i + \beta_{s[i],3} \times Estado_i
\end{aligned}
$$

[2] Estas variables pueden cambiar dependiendo de la definición definitiva de las preguntas de las encuestas.

para $i = 1, \ldots, n$, donde n denota el número total de personas entrevistadas. En las dos anteriores expresiones, los coeficientes β denotan los efectos de las variables Género, Edad y Raza sobre las probabilidades. Cada uno de estos efectos tiene varios subíndices puesto que su naturaleza es aleatoria. Por ejemplo, para el coeficiente $\beta_{m[i],1}$, $\beta_{m[i],2}$ o $\beta_{m[i],3}$, el subíndice i denota, naturalmente, el individuo; el subíndice m puede tomar valores de 1 hasta el 4 (pues la variable Edad es una variable categórica con cuatro categorías), dependiendo de la edad del individuo i; el subíndice después de la coma hace referencia a que el efecto de la variable Edad es diferente para modelar $\ln \frac{\rho_{i1}}{\rho_{i4}}$, $\ln \frac{\rho_{i2}}{\rho_{i4}}$ y $\ln \frac{\rho_{i3}}{\rho_{i4}}$. Lo anterior implica que en nuestro modelo de regresión multinomial, para una variable explicativa categórica con q categorías, se tienen $3q$ efectos diferentes, que corresponden a q efectos para modelar $\ln \frac{\rho_{i1}}{\rho_{i4}}$, q efectos para modelar $\ln \frac{\rho_{i2}}{\rho_{i4}}$ y q efectos para modelar $\ln \frac{\rho_{i3}}{\rho_{i4}}$.

Por otro lado, la caracterización de un modelo multinivel radica en que los efectos de las variables explicativas no son considerados parámetros fijos, sino variables aleatorias. De esta forma, tenemos que:

$$
\begin{aligned}
\beta_{l,k} &\sim normal(0, \sigma^2_{l,k}), \text{para } l = 1, 2 \\
\beta_{m,k} &\sim normal(0, \sigma^2_{m,k}), \text{para } m = 1, \ldots, 4 \\
\beta_{r,k} &\sim normal(0, \sigma^2_{r,k}), \text{para } r = 1, \ldots, 4 \\
\beta_{s,k} &\sim normal(0, \sigma^2_{s,k}), \text{para } s = 1, \ldots, 5
\end{aligned}
$$

para $k = 1, 2, 3$. La estimación de este modelo de regresión multinomial multinivel se lleva a cabo por medio del paquete `rjags` del software `R`.

3.2. Post-estratificación

Después de estimar el modelo de regresión multinomial multinivel, se puede calcular el vector de probabilidades $\hat{\boldsymbol{\rho}}$ para cualquier persona una vez se conozca su información personal en términos del género, edad, raza y estado. Ahora, como las variables que describen la información personal son variables categóricas, solo habrá un número determinado de posibles valores para $\hat{\boldsymbol{\rho}}$. En el caso de las anteriores variables, existen en total $2 \times 4 \times 4 = 32$ combinaciones para cada uno de los cinco estados, y habrá 32 posibles valores para el vector de probabilidades $\hat{\boldsymbol{\rho}}$ para cada estado s. Estos valores los denotaremos con $\hat{\boldsymbol{\rho}}_{s1}, \hat{\boldsymbol{\rho}}_{s2}, \ldots, \hat{\boldsymbol{\rho}}_{s32}$.

Ahora, para llevar a cabo la post-estratificación, se requiere conocer, para cada estado, cuántas personas hay en cada una de las 32 combinaciones de todos los cruces posibles de las variables Género, Edad y Raza. Esto es, se debe conocer el número de hombres blancos entre 18 y 34 años, el número de mujeres blancas entre 18 y 34 años, el número de hombres blancos entre 35 y 49 años, el número de mujeres blancas entre 35 y 49 años, etc. Estos números se denotarán como $n_{s1}, \ldots, n_{s32}$.[3]

Finalmente, la estimación de la intención de voto para el estado s viene dada por

$$\hat{\boldsymbol{\rho}}_s = \frac{\sum_{j=1}^{32} n_{s,j} \hat{\boldsymbol{\rho}}_{sj}}{\sum_{j=1}^{32} n_{s,j}}$$

Esto es, se utilizan los 32 posibles valores de $\hat{\boldsymbol{\rho}}_{sj}$, ponderando por el número de personas en los 32 posibles cruces de las variables Género, Edad y Raza.

Finalmente, se hace la observación de que cada vez que se cuente con nuevas encuestas, se ejecutará nuevamente el modelo MRP, y así en un determinado punto del tiempo se contará, para un estado determinado, con diferentes resultados de $\hat{\boldsymbol{\rho}}_s$.

4. Modelo de contexto–estrategia y evolución

En esta sección, se presenta el modelo de contexto–estrategia y evolución. Este modelo busca resolver las siguientes problemáticas:

1. Tener en cuenta las variables del componente del contexto y estrategia, tales como: índice de sufrimiento macroeconómico, imagen del presidente actual y gasto de los candidatos en las campañas electorales, entre otras.

2. Tener en cuenta la «vejez» de las encuestas, esto es, el tiempo entre el momento en que se realizó el levantamiento de la información de la encuestas y el día de las elecciones.

3. Corregir los resultados obtenidos del modelo MRP en cuanto a la categoría de «indecisos», puesto que, para el día de las elecciones, esta categoría desaparece automáticamente, y por consiguiente el pronóstico final debe contener solo tres categorías: republicano, demócrata y libertario.

Dado lo anterior, suponga que en un punto determinado del tiempo, para un estado determinado s, se cuenta con un número total de J encuestas, a las cuales se les ha aplicado el modelo MRP. Así, se tienen las estimaciones de intención de voto (incluyendo la categoría

[3] El segundo subíndice de $n_{s1}, \ldots, n_{s32}$ debe coincidir con el segundo subíndice de $\hat{\boldsymbol{\rho}}_{s1}, \hat{\boldsymbol{\rho}}_{s2}, \hat{\boldsymbol{\rho}}_{s32}$.

de indecisos). Usando estas estimaciones, junto con la proyección del número de votantes del estado s, se puede obtener el número de votos republicanos, demócratas, libertarios e indecisos. La configuración y la notación de estas variables es como se indica en la siguiente tabla:

ID de la encuesta	Republicano	Demócrata	Libertario	Indecisos
1	$u_{1,1}$	$u_{1,2}$	$u_{1,3}$	Ind_1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
j	$u_{j,1}$	$u_{j,2}$	$u_{j,3}$	Ind_j
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
J	$u_{J,1}$	$u_{J,2}$	$u_{J,3}$	Ind_J

Cuadro 2: *Número de votos estimados utilizando el modelo MRP aplicado a J encuestas en un determinado estado s (en la notación se omite el subíndice s).*

Dado que el objetivo es proyectar los resultados al día de la votación, en el cual el indeciso deberá decidirse por un candidato, el objetivo de esta sección se centra en relacionar la probabilidad de las categorías Republicano, Demócrata y Libertario con el número de personas indecisas, que se utilizará como una variable explicativa. De esta forma, para pronosticar los resultados del día de las elecciones, se toma el número de personas indecisas como cero, y el pronóstico final se obtiene solo para los tres candidatos. Asimismo, también se incluye la variable Vejez para poder pronosticar los resultados para el día de las elecciones fijando el valor de la variable Vejez en cero.

Dado lo anterior, se busca relacionar el números de votos de las J encuestas $u_{j,k}$ ($j = 1, \ldots, J$ y $k = 1, 2, 3$) con valores de variables del componente de contexto–estrategia, junto con la variable Vejez y el número de indecisos. Por la naturaleza de los datos, es natural suponer que el vector $(u_{j,1}, u_{j,2}, u_{j,3})'$ sigue una distribución multinomial con parámetro m (número proyectado de votantes para el año 2016) y vector de probabilidades $(\omega_{j,1}, \omega_{j,2}, \omega_{j,3})$. Tomando como categoría base el candidato libertario, el modelo queda expresado como:

$$\ln \frac{\omega_{j,1}}{\omega_{j,3}} = \mathbf{V}'_j \boldsymbol{\lambda}_1 + \delta_1^{(Vejez)} \times Vejez_j + \delta_1^{(Ind)} Ind_j$$
$$\ln \frac{\omega_{j,2}}{\omega_{j,3}} = \mathbf{V}'_j \boldsymbol{\lambda}_2 + \delta_2^{(Vejez)} \times Vejez_j + \delta_2^{(Ind)} Ind_j$$

para $j = 1, \ldots, J$. El vector $\mathbf{V}_j$ contiene los valores de las variables de contexto y estrategia en el momento que se llevó a cabo la encuesta j, los vectores $\boldsymbol{\lambda}_1$ y $\boldsymbol{\lambda}_2$ contienen los coeficientes de regresión. $Vejez_j$ describe el tiempo entre el momento en el cual se realizó la encuesta j y el día de las elecciones y Ind_j denota el número de personas indecisas, el cual fue obtenido con el modelo MRP descrito en la sección anterior.

Una vez ajustado el anterior modelo de regresión multinomial, podemos contestar a la pregunta: *¿Cómo hubieran sido los resultados de las J encuestas si estas hubiesen sido realizadas el día de las elecciones?* Lo anterior se puede responder calculando los valores ajustados de $\mathbf{u}_j = (u_{j,1}, u_{j,2}, u_{j,3})$, usando: (1) los valores proyectos de las variables $\mathbf{V}_j$ al día de las elecciones, (2) el valor cero para la variable «Vejez», y (3) el valor cero para la variable «Ind». Estos valores ajustados se denotarán con $\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_J$. En la siguiente sección, se discutirá la forma de combinar estos J pronósticos en uno solo, y la forma de incorporar este pronóstico en el análisis bayesiano.

5. Estructura bayesiana

En esta parte del documento, se explica con detalle cómo combinar los resultados del modelo de estructura y el modelo de contexto–estrategia y evolución para obtener la estimación bayesiana de los parámetros de interés.

5.1. Modelo de decaimiento exponencial

Tal como se comentó al final de la sección anterior, se cuenta con los valores de las J encuestas proyectadas para el día de las elecciones $\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_J$. Para combinarlas y producir un solo pronóstico se deben tener en cuenta las características propias de estas encuestas y darles a cada una un peso diferente, inducido por la dinámica cambiante de la opinión electoral a través del tiempo. Entre las características más importantes se encuentran el número de personas entrevistadas (n_j) y la vejez ($Vejez_j$) de cada encuesta.[4] La ponderación para cada encuesta está dada por:

$$w_j = n_j \times \tau^{Vejez_j}$$

en donde τ es una constante que toma valores entre 0 y 1, de tal forma que el peso de una encuesta más «vieja» sea menor que el de una encuesta más reciente. Por otro lado, una encuesta con un número mayor de personas entrevistadas tendrá mayor peso. Una vez calculados los pesos $w_1, \ldots, w_J$, se deben normalizar para asegurar que tengan una suma igual a uno. Esto es, los pesos finales $w_1^*, \ldots, w_J^*$ se calculan como:

$$w_j^* = \frac{w_j}{w_1 + \cdots + w_J}$$

[4] Otras características, como la reputación de la firma encuestadora y el número de aciertos históricos en elecciones pasadas, también pueden ser tenidas en cuenta.

Usando los anteriores pesos, el pronóstico para el día de elecciones se calcula como:

$$\hat{\mathbf{u}} = w_1^* \hat{\mathbf{u}}_1 + \cdots + w_J^* \hat{\mathbf{u}}_J$$

Otra forma de combinar estos J pronósticos es usar los pronósticos suavizados $\tilde{\mathbf{u}}_j$ en vez de usar los resultados directos $\hat{\mathbf{u}}_j$. Estos pronósticos suavizados se pueden obtener al ajustar los valores $\hat{\mathbf{u}}_j$ a un modelo de regresión multinomial con efectos aleatorios por candidatos. La ventaja de utilizar estos pronósticos suavizados es que permiten mostrar la tendencia de las J encuestas por cada candidato.

Finalmente, se debe tener en cuenta que el vector $\hat{\mathbf{u}}$ contiene el pronóstico del números de votos de los tres candidatos. Para conocer el porcentaje de votos, se debe dividir $\hat{\mathbf{u}}$ por el número proyectado de votantes para el año 2016.

Este pronóstico, que fue obtenido principalmente de datos de encuestas, será insumo para extraer la función de verosimilitud y combinarla con la distribución *a priori*, con el fin de obtener la estimación bayesiana. A continuación se discute cada una de estas distribuciones.

5.2. Distribución *a priori*

Dado que el vector de parámetros $\boldsymbol{\pi}_s = (\pi_{s1}, \pi_{s2}, \pi_{s3})$ (que describen el porcentaje de votos obtenidos por los tres candidatos en el estado s) tiene la característica de que cada $\pi_{sk} \in (0, 1)$, y $\pi_{s1} + \pi_{s2} + \pi_{s3} = 1$, la distribución por excelencia es la distribución Dirichlet, cuya función de densidad está dada por

$$p(\pi_{s1}, \pi_{s2}, \pi_{s3}) = \frac{\Gamma(\beta_{s1})\Gamma(\beta_{s2})\Gamma(\beta_{s3})}{\Gamma(\beta_{s1} + \beta_{s2} + \beta_{s3})} \pi_{s1}^{\beta_{s1}-1} \pi_{s2}^{\beta_{s2}-1} \pi_{s3}^{\beta_{s3}-1}$$

donde $\beta_{sk} > 0$ para $k = 1, 2, 3$.

En esta distribución, para escoger los valores de los parámetros $\beta_{s1}, \beta_{s2}, \beta_{s3}$, se debe tener en cuenta que:

- La suma de los parámetros $\beta_{s1} + \beta_{s2} + \beta_{s3}$ representa el «número de datos» de la información previa y aquí se ve reflejado el grado de incertidumbre sobre la información previa, puesto que la varianza de la distribución $p(\theta)$ es grande para valores grandes de $\beta_{s1} + \beta_{s2} + \beta_{s3}$.

- La esperanza de la distribución *a priori* de cada parámetro π_{sk} está dada por:

  $$\frac{\beta_{sk}}{\beta_{s1}+\beta_{s2}+\beta_{s3}}$$

  para $k = 1, 2, 3$. De esta forma, se deben escoger los valores de $\beta_{s1}, \beta_{s2}, \beta_{s3}$ de manera tal que las estimaciones *a priori* de los parámetros $\pi_{s1}, \pi_{s2}, \pi_{s3}$ correspondan a los pronósticos de la intención de voto para el año 2016 (obtenidos del modelo de estructura).

Dado lo anterior, para fijar el valor de $\beta_{s1}+\beta_{s2}+\beta_{s3}$ se debe tener en cuenta el número de datos representados en la función de verosimilitud y los pesos que se determinan de antemano para la distribución *a priori* y la función de verosimilitud. Para el caso de este proyecto, estos pesos son 20 % para el modelo de estructura, y 80 % para el modelo de contexto y estrategia. De esta forma, suponga que la función de verosimilitud (la cual resume los resultados del modelo de contexto–estrategia) representa 1.000 datos. Entonces el número de datos que representa el modelo de estructura será $1.000 \times 0.2/0.8 = 250$, esto es, $\beta_{s1} + \beta_{s2} + \beta_{s3} = 250$. Tomando como ejemplo el estado de Arizona, donde los resultados del pronóstico para 2016, usando el modelo de estructura, están dados por $(49.99\,\%, 49.55\,\%, 0.46\,\%)$, entonces los parámetros $\beta_{Arizona,k}$ serán calculados como

$$\begin{aligned}
\beta_{Arizona,1} &= 250 \times 0.4999 = 124.975\\
\beta_{Arizona,2} &= 250 \times 0.4955 = 123.875\\
\beta_{Arizona,3} &= 250 \times 0.0046 = 1.15
\end{aligned}$$

Esto es, la distribución *a priori* del vector de parámetros $\boldsymbol{\pi}_{Arizona}$ viene dada por

$$\boldsymbol{\pi}_{Arizona} \sim Dirichlet(124.975, 123.875, 1.15).$$

5.3. Función de verosimilitud

En la función de verosimilitud se busca resumir las estimaciones obtenidas del modelo de contexto–estrategia y evolución. Esta estimación puede ser representada por medio de la distribución multinomial cuya función de densidad está dada por:

$$p(\mathbf{y}_s|X, \pi_{s1}, \pi_{s2}, \pi_{s3}) \propto \pi_{s1}^{y_{s1}}\pi_{s2}^{y_{i2}}\pi_{s3}^{y_{s3}} \tag{1}$$

Por ejemplo, suponga que, para un estado s, los resultados del modelo de contexto–estrategia y evolución arrojan una estimación de intención de voto igual a $\hat{\mathbf{u}}_s = (0.53, 0.38, 0.09)$, la cual representa a 5.000 datos. Este resultado puede ser representado por medio de un vector de datos con distribución $multinomial(5.000, \boldsymbol{\pi}_s)$ y el vector de datos viene dado por $5.000 \times \hat{\mathbf{u}}_s$. Esto es, la función de verosimilitud viene dada por:

$$
\begin{aligned}
p(\mathbf{y}_s|\pi_{s1}, \pi_{s2}, \pi_{s3}) &\propto \pi_{s1}^{5000\times0.53}\pi_{s2}^{5000\times0.38}\pi_{s3}^{5000\times0.09}\\
&\propto \pi_{s1}^{2650}\pi_{s2}^{1900}\pi_{s3}^{450}
\end{aligned}
$$

En conclusión, si para un determinado estado s, el modelo MRP arroja una estimación de intención de voto de $\hat{\mathbf{u}}_s$ representando a M datos, entonces la función de verosimilitud viene dada por la función de densidad de una distribución $multinomial(M, \boldsymbol{\pi}_s)$ con $\mathbf{y} = M \times \hat{\mathbf{u}}_s$.

5.4. Distribución *a posteriori* y pronóstico final

En este punto, para cada uno de los cinco estados y para el vector de interés $\boldsymbol{\pi}_s = (\pi_{i1}, \pi_{i2}, \pi_{i3})$ se tiene identificada la distribución *a priori* en forma de una distribución Dirichlet, denotada como:

$$\boldsymbol{\pi}_s \sim Dirichlet(\beta_{s1}, \beta_{s2}, \beta_{s3})$$

También se tiene la función de verosimilitud en forma de una distribución multinomial, esto es,

$$\mathbf{y} \mid \boldsymbol{\pi}_s \sim multinomial(M, \boldsymbol{\pi}_s)$$

En la literatura, es bien conocido que, bajo este marco de referencia, la distribución *a posteriori* del parámetro de interés $\boldsymbol{\pi}_s$ viene dada por:

$$\boldsymbol{\pi}_s \mid \mathbf{y} \sim Dirichlet(\beta_{s1} + y_{s1}, \beta_{s2} + y_{s2}, \beta_{s3} + y_{s3})$$

De donde podemos obtener la estimación vía simulación de Monte Carlo de la siguiente forma:

1. Fijar el número de iteraciones G.

2. Simular G valores de la distribución $Dirichlet(\beta_{s1}+y_{s1}, \beta_{s2}+y_{s2}, \beta_{s3}+y_{s3})$. Denotamos estos valores simulados por

$$(\pi_{s1}^{(1)}, \pi_{s2}^{(1)}, \pi_{s3}^{(1)})$$

$$(\pi_{s1}^{(2)}, \pi_{s2}^{(2)}, \pi_{s3}^{(2)})$$

$$\vdots$$

$$(\pi_{s1}^{(G)}, \pi_{s2}^{(G)}, \pi_{s3}^{(G)})$$

3. La estimación bayesiana del vector de parámetros viene dada por $\hat{\boldsymbol{\pi}} = (\hat{\pi}_{s1}, \hat{\pi}_{s2}, \hat{\pi}_{s3})$ donde

$$\hat{\pi}_{sk} = \frac{1}{G}\sum_{g=1}^{G} \hat{\pi}_{sk}^{(g)}, \text{ para } k = 1, 2, 3.$$

   También se puede calcular un intervalo de credibilidad de $(1-\beta) \times 100\,\%$ para cada π_{sk} utilizando los percentiles muestrales de $\beta/2$ y $1-\beta/2$ de los valores simulados $\pi_{sk}^{(1)}, \pi_{sk}^{(2)}, \ldots, \pi_{sk}^{(G)}$

6. Estimación en áreas pequeñas (SAE)

En esta parte se describe el modelo de unidades de Fay-Herriot (*Basic Unit Level Model*) (Rao, 2003) que será útil para estimar la intención de voto a nivel de condados. De nuevo, al utilizar los datos de una encuesta, para cada persona entrevistada i, se conoce su intención de voto, que es una de las cuatro categorías: republicano, demócrata, libertario o «indeciso». El modelo de unidades busca asociar las probabilidades de estas categorías con variables auxiliares, además de efectos específicos de las áreas (que en nuestro caso corresponden a condados). La formulación del modelo es:

$$\begin{aligned}
\ln \frac{\rho_{ij1}}{\rho_{ij4}} &= \beta_{ij1} = \mathbf{U}_{ij}' \boldsymbol{\varrho}_1 + v_{j1} \\
\ln \frac{\rho_{ij2}}{\rho_{ij4}} &= \beta_{ij2} = \mathbf{U}_{ij}' \boldsymbol{\varrho}_2 + v_{j2} \\
\ln \frac{\rho_{ij3}}{\rho_{ij4}} &= \beta_{ij3} = \mathbf{U}_{ij}' \boldsymbol{\varrho}_3 + v_{j3}
\end{aligned}$$

donde $j = 1, \ldots, l$ con l denotando el número de condados y $i = 1, \ldots, N_j$, con N_j denotando el número de personas del condado j. $\mathbf{U}_{ij}$ contiene los valores de variables auxiliares y los vectores $\boldsymbol{\varrho}_1$, $\boldsymbol{\varrho}_2$ y $\boldsymbol{\varrho}_3$ contienen los coeficientes de regresión. v_{j1}, v_{j2} y v_{j3} denotan los efectos del condado j con $v_{jk} \sim normal(0, \sigma_{vk}^2)$ para $k = 1, 2, 3$.

Bajo el supuesto de que N_j es grande, se puede calcular la media del condado j como:

$$\begin{aligned}\bar{\beta}_{j1} &= \bar{\mathbf{U}}'_j \boldsymbol{\varrho}_1 + v_{j1}\\ \bar{\beta}_{j2} &= \bar{\mathbf{U}}'_j \boldsymbol{\varrho}_2 + v_{j2}\\ \bar{\beta}_{j3} &= \bar{\mathbf{U}}'_j \boldsymbol{\varrho}_3 + v_{j3}\end{aligned}$$

donde las variables $\bar{\mathbf{U}}'_j$ se calculando promedio los valores de $\mathbf{U}_{ij}$ sobre los individuos i. Finalmente, podemos calcular las intenciones de voto del condado j como:

$$\begin{aligned}\rho_{j1} &= \frac{e^{\bar{\beta}_{j1}}}{e^{\bar{\beta}_{j1}} + e^{\bar{\beta}_{j2}} + e^{\bar{\beta}_{j3}} + 1}\\ \rho_{j2} &= \frac{e^{\bar{\beta}_{j2}}}{e^{\bar{\beta}_{j1}} + e^{\bar{\beta}_{j2}} + e^{\bar{\beta}_{j3}} + 1}\\ \rho_{j3} &= \frac{e^{\bar{\beta}_{j3}}}{e^{\bar{\beta}_{j1}} + e^{\bar{\beta}_{j2}} + e^{\bar{\beta}_{j3}} + 1}\\ \rho_{j4} &= \frac{1}{e^{\bar{\beta}_{j1}} + e^{\bar{\beta}_{j2}} + e^{\bar{\beta}_{j3}} + 1}\end{aligned}$$

7. Conclusiones

Este documento ha presentado una metodología que busca modelar la intención de voto de los latinos en cada uno de los cinco estados de interés. Un supuesto fundamental de esta aproximación metodologógica es que la dinámica electoral es cambiante y está en función de las votaciones anteriores, así como de variables de contexto (macroeconómicas, de opinión y percepción) y variables de estrategia propias de las campañas.

Con base en lo anteriormente mencionado, se hace necesario usar un modelo avanzado, que dé cuenta de esta realidad compleja, que los pronósticos electorales simples (basados en la respuesta de los entrevistados) a veces desconocen.

Referencias

Gelman, A. & Hill, J. (2007), *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Analytical Methods for Social Research, Cambridge University Press. URL: https://books.google.com.co/books?id=lV3DIdV0F9AC

Ghitza, Y. & Gelman, A. (2013), “Deep Interactions with MRP: Election Turnout and Voting Patterns Among Small Electoral Subgroups”, *American Journal of Political Science*, 57(3), 762–776.

Park, D. K., Gelman, A. & Bafumi, J. (2016), “Bayesian Multilevel Estimation with Poststratication: State-Level Estimates from National Polls”, *Political Analysis*, 12, 375–385.

Plummer, M. (2003), “Jags: A program for analysis of bayesian graphical models using gibbs sampling”.

Plummer, M. (2016), *Rjags: Bayesian Graphical Models using MCMC*. R package version 4-6. URL: https://CRAN.R-project.org/package=rjags.

Rao, J. N. K. (2003), *Small Area Estimation*, Wiley.

Wyatt, J., Dropp, K., Byrum, T. & Dulin, A. (2016), “Estimating General Election Support for President Using Multilevel Regression & Poststratication”, *Morning Consult*, 16.

Apéndice

Variables y estimaciones del modelo de estructura

En la siguiente tabla se encuentran la descripción de las variables incluidas en el modelo de estructura, así como los coeficientes estimados.

Variable	Estado	Beta Republicano	Beta Demócrata
(Intercepto)	Todos	-1,670203000	0,182398100
PWE1	Todos	-28,332480000	-21,750670000
PBE1	Todos	0,436419600	1,034637000
HTO	Todos	15,637740000	12,370070000
WTO	Todos	2,920792000	2,445066000
FTO	Todos	-11,973550000	-21,484550000
PH4	Todos	0,000000044	0,000000355
BTCP	Todos	-4,217628000	-0,693693400
GOV	Todos	5,512242000	-1,579114000
HPE2	Todos	-2,926257000	-1,857822000
VLDEM	Todos	3,742284000	0,425380000
PWE2	Todos	2,953954000	5,396506000
PHM	Todos	-16,304500000	-6,352269000
PBE4	Todos	0,000000884	0,000000797
MTCP	Todos	-0,026098320	0,047491220
PBM	Todos	-5,565691000	-8,586905000
NPAV	Arizona	0,000001298	0,000000601
	Colorado	0,000001124	0,000000639
	Florida	0,000000258	0,000000188
	Nevada	0,000004418	0,000007363
	New Mexico	0,000006754	0,000005217
PLAV	Arizona	4,412798000	17,874520000
	Colorado	12,371750000	22,034700000
	Florida	15,343190000	14,299830000
	Nevada	-9,500050000	-33,742430000
	New Mexico	-3,678377000	4,286852000

Cuadro 3: Variables utilizadas en el modelo de estructura para la población general y los coeficientes estimados.

La definición de las variables es la siguiente:

PWE1: Porcentaje de personas blancas entre los 18 y los 34 años.

PBE1: Porcentaje de personas afrodescendientes entre los 18 y los 34 años.

HTO: Turn out de personas latinas.

WTO: Turn out de personas blancas.

FTO: Turn out de mujeres.

PH4: Porcentaje de personas latinas con 65 años o más.

BTCP: Total de personas afrodescendientes.

GOV: Afiliación partidista del gobernador del estado.

HPE2: Porcentaje de participación de los latinos en el electorado.

VLDEM: Porcentaje de voto latino demócrata.

PWE2: Porcentaje de personas blancas entre los 35 y los 49 años.

PHM: Porcentaje de latinos que son hombres.

PBE4: Porcentaje de personas afrodescendientes con 65 años o más.

MTCP: Número total de hombres.

PBM: Porcentaje de afrodescendientes que son hombres.

NPAV: Nmero de personas en edad para votar (efecto aleatorio).

PLAV: Porcentaje de latinos en edad para votar (efecto aleatorio).

Variables y estimaciones del modelo de contexto

En la siguiente tabla se encuentran la descripción de las variables incluidas en el modelo de contexto, así como los coeficientes estimados.

La definición de las variables es la siguiente:

SBPT: Porcentaje de menciones positivas de Trump en medios.

IMMIGRATIONT: Porcentaje de personas que consideran que el tema migratorio será mejor manejado por Trump.

OF: Favorabilidad de Obama.

RNONE: Número de personas registradas sin afiliación partidista.

CIN: Porcentaje de personas que no tienen opinión de Clinton.

CF: Favorabilidad Clinton.

ROTH: Número de personas registradas con un partido diferente al Demócrata o al Republicano.

Variable	Estado	Beta Republicano	Beta Demócrata
(Intercepto)	Todos	0,494634900	0,261381200
SBPT	Todos	-7,709151000	-5,237373000
IMMIGRATIONT	Todos	-5,954350000	-4,827742000
OF	Todos	1,402496000	-0,174276300
RNONE	Todos	0,000002296	0,000001740
CIN	Todos	-0,471379100	0,487771600
CF	Todos	-2,653335000	-2,955432000
ROTH	Todos	0,000007577	0,000005918
CMAC	Todos	-0,000000002	-0,000000003
TIN	Todos	-1,725943000	-3,904541000
RDEM	Todos	-0,000000192	0,000000164
RTOTAL	Arizona	0,000002109	0,000002324
	Colorado	0,000002490	0,000002058
	Florida	-0,000000102	-0,000000225
	Nevada	0,000002393	0,000001445
	New Mexico	0,000009124	0,000008130
ISM	Arizona	-124,803300000	-130,417600000
	Colorado	-26,234520000	-11,103600000
	Florida	92,808210000	103,310700000
	Nevada	114,471700000	117,262400000
	New Mexico	-14,796040000	-15,088250000

Cuadro 4: Variables utilizadas en el modelo de contexto para la población general y los coeficientes estimados.

CMAC: Gasto promedio mensual de las menciones de Clinton.

TIN: Porcentaje de personas que no tienen opinión de Trump.

RDEM: Número de personas registradas que se consideran demócratas.

RTOTAL: Número total de personas registradas para votar (efecto aleatorio).

ISM: Índice de sufrimiento macroeconómico (efecto aleatorio).

Variables y estimaciones del modelo de estrategia

En la siguiente tabla se encuentran la descripción de las variables incluidas en el modelo de estrategia, así como los coeficientes estimados.

La definición de las variables es la siguiente:

DIFFIGT: Diferencia entre el dinero recaudado y el dinero gastado por la campaa de Trump.

DIFFD: Diferencia absoluta entre los desembolsos de la campaña de Clinton y la de Trump.

Variable	Estado	Beta Republicano	Beta Demócrata
(Intercepto)	Todos	0,535232900	-0,863748200
DIFFIGT	Todos	0,000000099	0,000000181
DIFFD	Todos	-0,000000075	-0,000000071
GFH	Todos	3,068926000	3,260631000
SPOT_T	Todos	0,000769432	0,000714438
DT	Todos	0,000000044	0,000000045
DIFFR	Todos	0,000000008	0,000000007
VT	Todos	-0,056745780	-0,042922680
GFT	Todos	14,838480000	13,522200000
PRIMDEM	Arizona	-0,000005732	-0,000002410
	Colorado	-0,000020665	-0,000008087
	Florida	-0,000001775	-0,000000820
	Nevada	-0,000164781	-0,000029478
	New Mexico	-0,000008220	-0,000000509

Cuadro 5: Variables utilizadas en el modelo de estrategia para la población general y los coeficientes estimados.

GFH: Probabilidad de votar por Clinton teniendo en cuenta por quién votó en las primarias.

SPOT_T: Número de pautas publicitarias emitidas de la campaña de Trump.

DT: Total de desembolsos hechos por la campaña de Trump.

DIFFR: Diferencia absoluta entre el recaudo de la campaña de Clinton y la de Trump.

VT: Acumulado de visitas de Trump a cada estado.

GFT: Probabilidad de votar por Trump teniendo en cuenta por quién votó en las primarias.

PRIMDEM: Total número de votos por el partido demócrata en las primarias (efecto aleatorio).