UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE ESTUDIOS SUPERIORES ACATLÁN

Análisis estadístico de la encuesta de seguimiento diario Milenio GEA/ISA para las elecciones presidencial y de jefe de gobierno del año 2012

T E S I S

QUE PARA OBTENER EL TÍTULO DE

LICENCIADA EN ACTUARÍA

PRESENTA:

ROCIO MARIBEL AVILA AYALA

ASESOR: DR. ARTURO ERDELY RUIZ

JUNIO 2014

II

Algo he aprendido en mi larga vida: que toda nuestra ciencia es primitiva y pueril, y, sin embargo, es lo más valioso que tenemos

Albert Einstein

Agradecimientos

A mis padres, Paty y Francisco, por su amor, esfuerzo y apoyo incondicional en cada uno de mis proyectos. Gracias por la educación y los valores que me han brindado desde pequeña y por seguir alentándome día tras día a mi superación. ¡Los amo!

A mis hermanos, Alma y Juan, por darme, cada uno a su manera, enseñanzas de vida y por compartir conmigo los momentos más importantes.

A todos los que en algún momento apoyaron en mi formación, en especial a la familia Torres Ávila, por haberme brindado su hogar durante el transcurso de mi carrera y por impulsarme a cumplir mis metas.

A mi asesor, Arturo Erdely Ruiz, por ser un pilar importante en mi formación académica. Gracias por su paciencia y su tiempo invertido en la revisión de este trabajo y por su motivación en mi desarrollo como persona, estudiante y profesionista, por su confianza y amistad.

A mis sinodales, por su valioso tiempo dedicado en la revisión de este trabajo.

A mis amigos y compañeros, en especial a Fernanda y Natalia, por todo lo que hemos pasado juntas y por su influencia en mi vida, y a Manu por sus consejos y ánimos constantes.

A Isabel Mendoza y su hermosa familia, por haberme adoptado como parte de su núcleo familiar, por su confianza, consejos y su apoyo en el desarrollo de mis proyectos.

A mi Alma Mater, la FES-Acatlán de la UNAM por haber sido el culmen de mi formación académica.

Índice general

Introducción

A lo largo del tiempo se ha buscado la mejor manera de estimar y predecir el comportamiento de diversos fenómenos en los cuales existe incertidumbre. La estadística ha sido una herramienta fundamental para lograr este fin, ya que ayuda a modelar el comportamiento de dichos fenómenos asignando a cada posible escenario una probabilidad de ocurrencia.

Las elecciones políticas son un caso donde es de particular interés para los votantes, y más aún para los candidatos, conocer el comportamiento a lo largo del tiempo de los porcentajes de preferencia por candidato. Existen varias empresas encuestadoras que se dedican a inferir dichos porcentajes usando diversas técnicas de muestreo y algunas de ellas publican los resultados obtenidos en medios de comunicación.

En el año 2012 se generó un gran revuelo debido a la magnitud del sesgo que presentaron las estimaciones de la mayoría de las encuestadoras en la elección presidencial. Se empezó a especular acerca de las causas de dicho sesgo y a desconfiar de la honestidad de las empresas encuestadoras, ya que algunas de ellas presentaban una diferencia entre los dos candidatos punteros de hasta dos o tres veces el margen de error.

Los resultados de la encuesta que realizó *Grupo Milenio* en colaboración con la casa encuestadora GEA [1]/ISA[2] fueron elegidos como objeto de análisis debido a dos razones fundamentales. La primera es que esta encuestadora, entre las diez más comunes tomadas por el sitio ADN Político (2012) fue la que más se alejó del resultado de seis puntos porcentuales de diferencia entre el primero y el segundo lugar, dado que estimaba hasta dieciocho puntos por-

[1] Grupo de Economistas y Asociados
[2] Investigaciones Sociales Aplicadas

centuales a favor del candidato ganador, una diferencia del doce por ciento. Y la segunda razón es porque es la única que realizó un seguimiento diario, lo cual es de gran importancia para un análisis estadístico, de acuerdo con Guerrero (2012):

> Otro ejemplo ilustrativo de la obtención de pronósticos surge a raíz del seguimiento diario que realizó el periódico Milenio, con la encuestadora GEA/ISA. Una de las principales ventajas que tiene el usar datos de la misma fuente es que las virtudes – y vicios – que pudiera tener la empresa encuestadora al realizar sus encuestas se mantienen a lo largo del estudio y esto hace que los datos sean comparables en el tiempo.

Paralelamente a la encuesta de GEA/ISA para la elección presidencial, en este estudio se hace análisis de los resultados de la misma encuestadora para la elección de jefe de gobierno del Distrito Federal con fines comparativos, ya que esta última no presentó un sesgo significativo. Es importante destacar que el análisis realizado en el presente trabajo puede ser utilizado para extraer conclusiones de índole político o social, sin embargo, éste sólo será abordado e interpretado desde la perspectiva de la estadística. Partiendo de esto, el objetivo de este estudio es buscar información significativa en los datos provenientes de la encuesta descrita anteriormente, así como modelar la estructura de dependencia que se presentó durante el periodo de encuesta entre los porcentajes de preferencia por cada candidato.

En el Capítulo 1 se pretende describir de manera breve y sintetizada la teoría que existe detrás de los procedimientos utilizados para el presente análisis estadístico. La Sección 1.1 aborda de manera concreta el paradigma bayesiano, al igual que conceptos como distribución a priori y a posteriori, propios de este enfoque de la estadística. En el mismo apartado se introduce el experimento aleatorio que da origen al modelo multinomial, ya que éste será útil para modelar las preferencias de la población de acuerdo a la encuesta, y por último se presenta una breve explicación de la estimación bayesiana de proporciones. Posteriormente, en la Sección 1.2 se relata una breve historia del surgimiento de las cópulas, así como algunas características y propiedades de estas funciones. Se abordan también conceptos relacionados con dependencia y la estrecha relación que ésta guarda con la cópula subyacente a los datos de estudio.

El Capítulo 2 se enfoca en hacer una descripción general de las variables que serán utilizadas para el presente análisis. Además, en este capítulo también se realiza una estimación del sesgo de los resultados de ambas encuestas involucradas, y dos métodos propuestos para la corrección del mismo.

Posteriormente, en el Capítulo 3 se propone un modelo bayesiano a partir del cual se simulan observaciones para calcular probabilidades de ganar por candidato y además estimaciones puntuales e intervalos de probabilidad para las mismas.

Por último, en el capítulo 4 se realiza un análisis de la estructura de dependencia entre cada par de candidatos contendientes tanto en la elección presidencial como en la elección de jefe de gobierno, con el fin de identificar los efectos (en signo y magnitud) que causaba la variación porcentual de uno de los candidatos en cada uno de los otros.

El análisis que se presenta en este trabajo es *a posteriori*, ya que se hace uso del cómputo final de cada una de las elecciones analizadas, por lo que en vez de predecir los porcentajes para los candidatos se pretende comparar el comportamiento de las probabilidades de ganar de cada uno de ellos en los resultados publicados de la encuesta de seguimiento diario Milenio GEA/ISA (en adelante la encuesta original) y una trayectoria estimada con un método propuesto de corrección del sesgo (en adelante encuesta corregida), para así identificar el efecto del sesgo de los porcentajes de preferencia en las probabilidades de ganar por candidato, además de modelar las dependencias bivariadas vía cópulas.

Capítulo 1

Marco teórico

En este capítulo se enunciarán los principales fundamentos teóricos necesarios para describir y explicar el objeto de estudio del presente trabajo. Como ya se mencionó anteriormente, a pesar de tratarse de un tema relacionado con la política, éste será abordado desde un punto de vista estadístico, por lo que no se incluirán conceptos ni antecedentes de las elecciones o los candidatos.

1.1. Estadística bayesiana

La *Estadística Bayesiana* debe su nombre al trabajo pionero del reverendo Thomas Bayes titulado *"An Essay towards solving a Problem in the Doctrine of Chances"* publicado en 1763. A pesar de ser un trabajo que data de varios años atrás, el desarrollo de la *Estadística Bayesiana* es relativamente nuevo, ya que existían limitaciones computacionales que se han ido superando recientemente.

La característica primordial que hace interesante al enfoque bayesiano es que éste, a diferencia de la estadística clásica, además de usar información muestral hace posible incorporar de manera consistente al modelo información subjetiva *(a priori)* derivada de las creencias previas a la realización del experimento (experiencia de expertos, información histórica, etc.) y aún con muestras pequeñas proporciona inferencias aceptables.

El primer paso a realizar en la modelación estadística paramétrica, es

elegir una familia paramétrica $\mathcal{P} = \{\wp(x \mid \theta) : \theta \in \Theta\}$ que logre describir el comportamiento probabilístico del fenómeno aleatorio de estudio, pero aún queda la incertidumbre acerca de cuál θ elegir para que el modelo esté definido de forma explícita. En contraste con el enfoque clásico o frecuentista, en estadística bayesiana se modela a θ, el parámetro del modelo probabilístico general $\wp(x \mid \theta)$, como una variable (o vector) aleatoria cuya ***distribución de probabilidad a priori o inicial*** $\wp(\theta)$ está basada en información previa. Elegir tal distribución es todo un tema, existen varios métodos para capturar la información subjetiva en una distribución de probabilidad para θ y también hay formas de expresar en dicha distribución la incertidumbre acerca del parámetro mediante una distribución a priori poco informativa[1].

Ya que se cuenta con la *distribución a priori*, se procede a obtener la muestra, y la distribución inicial se actualiza con las observaciones $\mathbf{x} = (x_1, \ldots, x_n)$ conforme a la Regla de Bayes para obtener una ***distribución Aa posteriori o final***:

$$\wp(\theta \mid \mathbf{x}) = \frac{\wp(\mathbf{x} \mid \theta)\wp(\theta)}{\int_\Theta \wp(\mathbf{x} \mid \vartheta)\wp(\vartheta)\, d\vartheta} \tag{1.1}$$

Obsérvese que la distribución $\wp(\theta \mid \mathbf{x})$ contempla tanto la información muestral como la información incorporada por $\wp(\theta)$. El denominador de (1.1) es una constante que típicamente en dimensiones paramétricas mayores es difícil de calcular de manera explícita, sin embargo existen métodos para aproximarlo numéricamente. Cuando la distribución a priori de θ sigue la misma ley de probabilidad que la distribución a posteriori se dice que es *familia conjugada*[2] para el modelo $\wp(x \mid \theta)$, este hecho hace que los cálculos de probabilidades y las estimaciones sean más sencillos por contar con un modelo ya conocido.

En la metodología bayesiana, una vez hallada la distribución a posteriori ya es posible plantear cualquier problema de inferencia: estimación puntual, estimación por intervalos, contrastes de hipótesis, etc. (Bernardo, 2003, Blasco, 2005). A continuación se presenta un breve resumen de cómo hacerlo.

[1] Para mayor detalle acerca de distribuciones a priori, véase Robert (2007).

[2] En DeGroot (2005) se profundiza en el tema deA familias conjugadas.

Estimación puntual

Estimar puntualmente es elegir **un** $\hat{\theta} \in \Theta$ que sea el representante de todos los valores posibles para θ. En el enfoque clásico esto se logra maximizando la función de verosimilitud, pero bajo el enfoque bayesiano se puede proponer un estimador puntual de θ como alguna medida de tendencia central, por ejemplo la mediana o la varianza (Robert, 2007):

$$\hat{\theta} = \mathbb{E}(\theta) = \int_{\Theta} \theta \wp(\theta \,|\, \mathbf{x}) d\theta$$

En el caso de que no se tenga una muestra observada, puede usarse $\wp(\theta)$ en vez de $\wp(\theta \,|\, \mathbf{x})$.

Estimación por intervalos

Consiste en encontrar valores a y b tales que en el intervalo (a, b) se encuentre θ con una probabilidad γ (deseablemente alta), es decir, que $\mathbb{P}(a < \theta < b) = \gamma$, esto se logra buscando el intervalo de longitud mínima donde a y b sean solución de la siguiente integral (o suma en el caso discreto):

$$\int_a^b \wp(\theta \,|\, \mathbf{x}) d\theta = \gamma.$$

Contraste de hipótesis

Sea $\{\Theta_j : j = 1, \ldots, m\}$ una partición del espacio paramétrico Θ, una hipótesis estadística acerca del parámetro θ es una proposición lógica de la forma

$$\mathcal{H}_j : \theta \in \Theta_j, \ \Theta_j \subseteq \Theta.$$

En estadística bayesiana es posible contrastar más de dos hipótesis, ya que se les asigna una probabilidad de ocurrencia en vez de *rechazar* o *aceptar*. La probabilidad de una hipótesis se calcula de la siguiente forma:

$$\mathbb{P}(\mathcal{H}_j) = \mathbb{P}(\theta \in \Theta_j) = \int_{\Theta_j} \wp(\theta \,|\, \mathbf{x}) d\theta.$$

Al igual que en los casos anteriores, si no se cuenta con una muestra observada puede sustituirse $\wp(\theta)$ por $\wp(\theta \,|\, \mathbf{x})$. Un problema de inferencia en estadística bayesiana también puede ser planteado como un problema de

decisión y para ello es necesario proponer una función de utilidad que refleje pérdidas o ganancias si se toma la decisión correcta o equivocada; para mayor detalle sobre este tópico revisar los capítulos 2 y 3 de Ghosh, Delampady y Samanta (2006).

1.1.1. Modelo multinomial

Considere una serie de n ensayos independientes de un experimento aleatorio en el cual sólo se observará uno de k eventos mutualmente excluyentes $E_1, \ldots, E_k$ y donde la probabilidad de observar el evento E_j es igual a θ_j. Es claro que $\sum_{i=1}^{k} \theta_i = 1$.

Sean $X_1, \ldots, X_k$ las variables aleatorias que denotan el número de ocurrencias de los eventos $E_1, \ldots, E_k$, respectivamente, en los n ensayos, con $\sum_{i=1}^{k} X_i = n$. Entonces la distribucion conjunta de $X_1, \ldots, X_n$ es:

$$\mathbb{P}(X_1 = x_1, \ldots, X_k = x_k \,|\, n, \boldsymbol{\theta}) = \frac{n!}{\prod_{i=1}^{k} x_i!} \prod_{i=1}^{k} \theta_i^{x_i} \mathbb{1}_{\{\sum_{i=1}^{k} x_i = n\}} \tag{1.2}$$

Este modelo es llamado *multinomial* de parametros n y $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$. A continuación se muestran resultados generales acerca de dicha distribución que se usarán en el desarrollo de este trabajo.

Algunas propiedades de la distribución multinomial

De acuerdo con Johnson, Kotz y Balakrishnan (1997), se obtiene de (1.2) que la distribución marginal de X_i es una binomial con parámetros (n, θ_i), en consecuencia:

$$\mathbb{E}[X_i] = n\,\theta_i$$

y

$$\mathbb{V}ar(X_i) = n\theta_i(1 - \theta_i).$$

Además, $X_i + X_j \sim Binom(n, \theta_i + \theta_j)$ por lo que:

$$\begin{aligned} \mathbb{V}ar(X_i + X_j) &= \mathbb{V}ar(X_i) + \mathbb{V}ar(X_j) + 2\mathbb{C}ov(X_i, X_j) \\ n(\theta_i + \theta_j)(1 - \theta_i - \theta_j) &= n\theta_i(1 - \theta_i) + n\theta_j(1 - \theta_j) + 2\mathbb{C}ov(X_i, X_j) \\ \therefore\ \mathbb{C}ov(X_i, X_j) &= -n\,\theta_i\,\theta_j. \end{aligned}$$

Otra propiedad importante es la correlación negativa entre dos entradas X_i, X_j, según Johnson y otros (1997):

$$corr(X_i, X_j) = -\sqrt{\frac{\theta_i \theta_j}{(1 - \theta_i)(1 - \theta_j)}}.$$

En el siguiente apartado se presenta la solución desde el enfoque bayesiano al problema de estimación de los parámetros de la distribución multinomial.

1.1.2. Estimación bayesiana de proporciones

En los últimos 30 años se ha prestado especial atención a la estimación de las probabilidades multinomiales tanto en literatura teórica como aplicada. La estimación (o inferencia) para la distribución multinomial parece tocar desde los aspectos básicos hasta los más profundos de la inferencia estadística moderna, especialmente la aproximación bayesiana, dentro de la cual figura el artículo de Walley (1996), que desató numerosas discusiones e investigaciones posteriores.

La estimación bayesiana se basa en suponer que los parámetros desconocidos $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ de la distribución multinomial no son fijos, sino que son variables aleatorias y en consecuencia tienen asociada una ley de probabilidad que modela la incertidumbre acerca de su valor. El método bayesiano usual para estimar las probabilidades de la distribución multinomial según Gisbert y otros (2007) consiste en suponer que la distribución a priori es una Dirichlet con parámetros $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)$, $\alpha_i \geq 0$,

$$\wp(\theta_1, \ldots, \theta_k \,|\, \alpha_1, \ldots, \alpha_k) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^{k} {\theta_i}^{\alpha_i - 1} \mathbb{1}_{\{\sum_{i=1}^{k} \theta_i = 1\}}. \tag{1.3}$$

donde el coeficiente $B(\boldsymbol{\alpha})$ es la función Beta multinomial:

$$B(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^{k} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{k} \alpha_i)}.$$

Si los datos observados $\boldsymbol{n} = (n_1, \ldots, n_k)$ se presentan en una muestra multinomial, la verosimilitud de los parámetros vendrá dada por:

$$\mathcal{L}(\theta_1, \ldots, \theta_k \,|\, X_1 = x_1, \ldots, X_k = x_k) = \frac{n!}{\prod_{i=1}^{k} x_i!} \prod_{i=1}^{k} \theta_i^{x_i}.$$

Observemos que es la misma ecuación que (1.2), sólo que ahora está en función de los parámetros $\theta_1, \ldots, \theta_k$. Bajo estas condiciones, se demuestra que la distribución a posteriori también es una Dirichlet (Agresti, 2002), cuyos parámetros son $(\alpha_1 + x_1, \ldots, \alpha_k + x_k)$:

$$\wp(\theta_1, \ldots, \theta_k \,|\, \alpha_1 + x_1, \ldots, \alpha_k + x_k) = \frac{1}{B(\boldsymbol{\alpha} + \mathbf{x})} \prod_{i=1}^{k} \theta_i^{\alpha_i + x_i - 1} \mathbb{1}_{\{\sum_{i=1}^{k} \theta_i = 1\}}. \quad (1.4)$$

Ya contando con la distribución a posteriori, es posible hacer cualquier clase de inferencia bayesiana. Para este análisis usaremos el modelo multinomial con $k = 4$, ya que en las dos elecciones en cuestión existen cuatro candidatos posibles. La interpretación de cada θ_k en éste contexto es la probabilidad de que una persona vote por el candidato k y se hará estimación acerca de dichas probabilidades.

1.2. Cópulas y dependencia

En esta sección se dará una breve descripción del concepto de cópula y su relación con las medidas de dependencia. Se abordarán los conceptos y técnicas que serán utilizados en el capítulo 4 para medir dependencias bivariadas en las variables de estudio.

Desde hace varias décadas los estadísticos han estado interesados en la relación que existe entre una función de distribución multivariada y sus marginales de menor dimensión (univariadas o no). Fréchet (1951) y Dall'Aglio (1956) publicaron trabajos interesantes acerca de este tema en los años cincuentas, estudiando funciones de distribución de dos y tres variables dadas

sus marginales univariadas. La respuesta al problema planteado para marginales univariadas fue descubierta por Sklar (1959), quien propuso una nueva clase de funciones a las cuales llamó *cópulas*. Estas nuevas funciones son restricciones al cuadrado unitario de funciones de distribución bivariadas cuyas marginales son uniformes en el intervalo $[0, 1]$(Quesada, Rodríguez y Úbeda, 2003).

La mayoría de los resultados acerca de cópulas que se publicaron entre 1959 y 1976 fueron enfocados al área de espacios métricos probabilísticos, principalmente al estudio de operaciones binarias en el espacio de las funciones de distribución de probabilidad, por ejemplo, Menger (1942) propuso una generalización de la teoría de espacios métricos reemplazando $d(p, q)$ por una función de distribución F_{pq}, donde $F_{pq}(x)$ para cualquier número real x mide la probabilidad de que la distancia entre p y q sea menor o igual que x. Una de las dificultades al proponer una medida es demostrar que cumple la desigualdad del triángulo, Menger propuso $F_{pr}(x + y) \leq T(F_{pq}(x), F_{qr}(y))$ donde T es una *norma triangular* o *t-norma*. Dentro de éste trabajo también se demostró que algunas t-normas son cópulas y de manera análoga, algunas cópulas son t-normas. Para mayor detalle del desarrollo histórico de la teoría probabilística de espacios métricos y su relación con las cópulas véase Schweizer (1991) y Schweizer y Sklar (1983a).

Más tarde se descubrió que las cópulas podían ser útiles para definir medidas de dependencia no paramétricas entre variables aleatorias. Desde entonces, el concepto de cópula juega un papel importante en la teoría de probabilidad y estadística, particularmente en problemas relacionados con dependencia dadas las marginales y funciones de variables aleatorias que son invariantes bajo transformaciones monótonas. Una revisión histórica acerca de la evolución de este campo se encuentra en Dall'Aglio (1991). Nelsen (2006), la primera y única introducción a cópulas hasta el momento escrita estudia familias de cópulas, construcción de cópulas, medidas de dependencia, entre otros tópicos importantes. En relación a la solución mediante cópulas a problemas dadas las distribuciones marginales, puede consultarse Benes y Stepan (1997), Cuadras, Fortiana y otros (2002), Dall'Aglio (1991), ó Rüschendorf, Schweizer y Taylor (1996).

1.2.1. Cópulas: definición y algunas propiedades

Comencemos con la definición de cópula para el caso bivariado, para posteriormente enunciar algunas generalidades de estas funciones.

Definición 1.2.1. Una *cópula* es una función $C : [0,1]^2 \longrightarrow [0,1]$ que satisface las siguientes propiedades:

a) Para todo u, v en $[0,1]$,

$$C(u,0) = 0 = C(0,v),$$

además

$$C(u,1) = u \text{ y } C(1,v) = v;$$

(condiciones de frontera)

b) para todo u_1, u_2, v_1, v_2 en el $[0,1]$ tal que $u_1 \leq u_2$ y $v_1 \leq v_2$,

$$C(u_2,v_2) - C(u_2,v_1) - C(u_1,v_2) + C(u_1,v_1) \geq 0.$$

(propiedad 2-creciente)

Como se mencionó anteriormente, la importancia de las cópulas dentro de la estadística se describe en el siguiente teorema:

Teorema 1.2.1 (Sklar)**.** *Sean X y Y variables aleatorias con función de distribución conjunta H y funciones de distribución marginales F y G, respectivamente. Entonces existe una cópula C tal que*

$$H(x,y) = C(F(x), G(y)) \tag{1.5}$$

para todo x, y en $\mathbb{R}$. Si F y G son funciones continuas entonces C es única, de otro modo la cópula C es única en $Ran(F) \times Ran(G)$. Más aún, si C es cualquier cópula y F y G son funciones de distribución univariadas, entonces la función H definida en (1.5) *es una función de distribución conjunta con marginales F y G.*

En el Teorema 1.2.1 se describe cómo las cópulas relacionan las funciones de distribución multivariadas con sus marginales de una dimensión; en Nelsen (2006) puede encontrarse una demostración del mismo. En la práctica, si se tiene un vector bivariado de datos y se quiere estimar la cópula subyacente, se puede realizar una aproximación de la misma; en la Sección 1.2.2 se detallará el procedimiento correspondiente.

Introduciremos algunos conceptos de acuerdo con Genest, Nešlehová y Quessy (2011).

Definición 1.2.2. Sea una muestra aleatoria $(X_1, Y_1), \ldots, (X_n, Y_n)$ de una distribución conjunta H con marginales F y G. Sea C la cópula subyacente. Cuando F y G son conocidas podemos construir una muestra aleatoria de C haciendo para todo i en $\{1, \ldots, n\}$

$$(U_i, V_i) = (F(X_i), G(Y_i)) \tag{1.6}$$

a las cuales denominaremos *observaciones de la cópula*.

Obsérvese que en la Definición 1.2.2 se requiere conocer de las distribuciones marginales F y G. Si éstas son desconocidas no es posible obtener observaciones de C, son embargo existe un concepto alternativo que se muestra en la siguiente definición:

Definición 1.2.3. Sea una muestra aleatoria $(X_1, Y_1), \ldots, (X_n, Y_n)$ de una distribución conjunta H con marginales F, G (desconocidas) y cópula C. Definimos a las *pseudo-observaciones de la cópula* para todo i en $\{1, \ldots, n\}$ como

$$\left(\hat{U}_i, \hat{V}_i\right) = (F_n(X_i), G_n(Y_i)). \tag{1.7}$$

donde F_n y G_n son las funciones de distribución empíricas de F y G dadas por

$$F_n(x) = \frac{1}{n}\sum_{k=1}^{n} \mathbb{1}_{(-\infty,x]}(X_k), \quad G_n(y) = \frac{1}{n}\sum_{k=1}^{n} \mathbb{1}_{(-\infty,y]}(Y_k)$$

A continuación se muestra un resultado que permite acotar cualquier cópula:

Teorema 1.2.2. *Sea C una cópula cualquiera, entonces para todo (u, v) en Dom C*

$$\text{máx}(u + v - 1, 0) \leq C(u, v) \leq \text{mín}(u, v). \tag{1.8}$$

Las cotas de (1.8) son también cópulas (véase Nelsen, 2006), y típicamente se denotan $M(u, v) := \text{mín}(u, v)$ y $W(u, v) := \text{máx}(u + v - 1, 0)$. Entonces, para una cópula C y $u, v \in [0, 1]^2$,

$$W(u, v) \leq C(u, v) \leq M(u, v). \tag{1.9}$$

La desigualdad (1.9) es la versión para cópulas de las *cotas de Fréchet–Hoeffding*, la cual también puede expresarse en términos de funciones de distribución. Nos referimos a M como la *cota superior de Fréchet–Hoeffding* y a W como la *cota inferior de Fréchet–Hoeffding*. Una tercera cópula muy importante que también usaremos con frecuencia es la cópula $\Pi(u, v) = uv$, que se conoce como *cópula producto* o *cópula independencia* debido a lo siguiente:

Teorema 1.2.3. *Sean X y Y variables aleatorias continuas. Entonces X y Y son independientes si y sólo si $C_{XY} = \Pi$*

Es decir, la cópula producto caracteriza a las variables aleatorias independientes cuando las funciones de distribución son continuas. La prueba es inmediata al observar que X y Y son independientes si y sólo si $H(x, y) = F(x)G(y)$.

Mucha de la utilidad de las cópulas en el estudio de la estadística no paramétrica se deriva de los resultados del siguiente teorema:

Teorema 1.2.4. *Sean X y Y variables aleatorias continuas, y C_{XY} su cópula subyacente. Sean f y g funciones estrictamente monótonas en $Ran(X)$ y $Ran(Y)$, respectivamente.*

a) Si f y g son estrictamente crecientes, entonces

$$C_{f(X),g(Y)}(u, v) = C_{XY}(u, v).$$

b) Si f es estrictamente creciente y g es estrictamente decreciente, entonces

$$C_{f(X),g(Y)}(u,v) = u - C_{XY}(u, 1-v).$$

c) Si f es estrictamente decreciente y g es estrictamente creciente, entonces

$$C_{f(X),g(Y)}(u,v) = v - C_{XY}(1-u, v).$$

d) Si f y g son ambas estrictamente decrecientes, entonces

$$C_{f(X),g(Y)}(u,v) = u + v - 1 + C_{XY}(1-u, 1-v).$$

El inciso *a)* del Teorema 1.2.4 dice que para transformaciones estrictamente crecientes de variables aleatorias, la cópula subyacente es invariante. Para ver una demostración detallada del teorema, puede consultarse Nelsen (2006).

Existen funciones $[0,1] \longrightarrow [0,1]$ derivadas de las cópulas denominadas *secciones*. La *sección horizontal* de una cópula C en a está dada por $t \longmapsto C(t,a)$, la *sección vertical* de C en a está dada por $t \longmapsto C(a,t)$; y la *sección diagonal* de C es la función $\delta_C : [0,1] \longrightarrow [0,1]$ definida por $\delta_C(t) = C(t,t)$. En Nelsen (2006) se demuestra que la sección horizontal, la sección vertical y la sección diagonal de una cópula son no decrecientes y uniformemente continuas en el intervalo $[0,1]$.

La sección diagonal de una cópula es muy importante, Frank (1996) encontró una condición suficiente, mas no necesaria para que una cópula *Arquimediana* (familia de cópulas) estuviese determinada de manera única por su sección diagonal. Dicho problema también fue estudiado en el contexto de *normas triangulares*, para mayor detalle puede consultarse por ejemplo Klement, Mesiar y Pap (2000) y Klement y Mesiar (2005).

Ejemplo. La sección diagonal de la cópula $W(u,v) := \text{máx}(u+v-1, 0)$ es

$$\begin{aligned} \delta_W(t) = W(t,t) &= \text{máx}(t+t-1, 0) \\ &= \text{máx}(2t-1, 0) \\ &= \begin{cases} 0 & \text{si } t < \frac{1}{2}, \\ 2t-1 & \text{si } t \geq \frac{1}{2}. \end{cases} \end{aligned}$$

La sección diagonal de la cópula $M(u, v) := \text{mín}(u, v)$ es

$$\begin{aligned}\delta_M(t) = M(t, t) &= \text{mín}(t, t) \\ &= t.\end{aligned}$$

Y la sección diagonal de la cópula $\Pi(u, v) := uv$ es

$$\begin{aligned}\delta_\Pi(t) = \Pi(t, t) &= t \cdot t \\ &= t^2.\end{aligned}$$

En la Figura 1.1 se muestra una gráfica de las diagonales calculadas, la cual será de gran utilidad al explorar el comportamiento de las cópulas bivariadas en el capítulo 4.

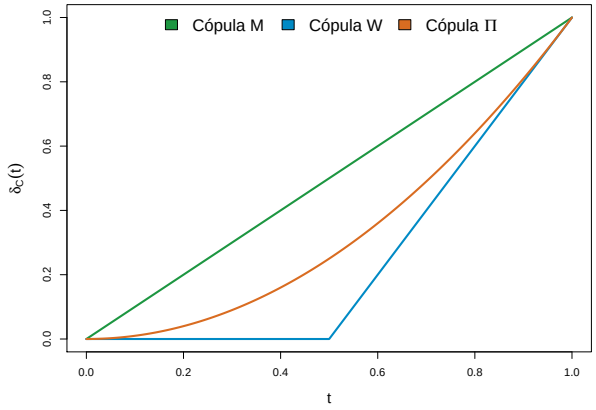

Figura 1.1: Diagonales de las cópulas M, W y Π

1.2.2. Aproximación de cópulas

En esta sección se revisará la expresión empírica de una cópula bivariada, que es una estimación de la cópula subyacente a un par de variables aleatorias; ésta puede usarse para la construcción de pruebas de hipótesis no paramétricas como pruebas de independencia, pruebas de bondad de ajuste para cópulas o para calcular versiones muestrales de algunas medidas de asociación, que será nuestro caso.

Sea $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ una muestra aleatoria observada de tamaño n de una distribución bivariada continua. Nelsen (2006) define la *cópula empírica* (bivariada) como la función

$$C_n\left(\frac{i}{n},\frac{j}{n}\right)=\frac{\text{número de pares } (x,y) \text{ en la muestra tales que } x\le x_{(i)}, y\le y_{(j)}}{n},$$

donde $x_(i)$, $y_(j)$ son los estadísticos de orden de la muestra para i y j en $\{1,\dots,n\}$ y $C_n(\frac{i}{n},0)=0=C_n(0,\frac{j}{n})$.

La *frecuencia de la cópula empírica* c_n está dada por

$$c_n\left(\frac{i}{n},\frac{j}{n}\right)=\begin{cases}1/n, & \text{si } \left(x_{(i)},y_{(j)}\right) \text{ es un elemento de la muestra,}\\ 0, & \text{en otro caso.}\end{cases}$$

Obsérvese que C_n y c_n están relacionadas mediante

$$C_n\left(\frac{i}{n},\frac{j}{n}\right)=\sum_{p=1}^{i}\sum_{q=1}^{j}c_n\left(\frac{p}{n},\frac{q}{n}\right)$$

Las cópulas empíricas fueron estudiadas en un principio por Deheuvels (1979), quien las llamó *funciones de dependencia empíricas*. Una cópula empírica no es una cópula, es una *subcópula* bidimensional, para mayor detalle acerca de subcópulas véase Nelsen (2006).

La versión muestral de la sección diagonal de una cópula es la *diagonal empírica* (bivariada) δ_n, la cual está definida por

$$\delta_n\left(\frac{j}{n}\right):=C_n\left(\frac{j}{n},\frac{j}{n}\right)\quad j=0,1,\dots,n\,. \tag{1.10}$$

Si asumimos sin pérdida de generalidad que los valores x_k en la muestra están ordenados, entonces

$$\delta_n\left(\frac{j}{n}\right)=\frac{1}{n}\sum_{k=1}^{j}\mathbb{1}_{]-\infty,y_{(j)}]}(y_k),\quad j=0,1,\dots,n-1\,,$$

y $\delta_n(0)=0$, $\delta_n(1)=1$. Dado lo anterior, es claro que δ_n es una función no decreciente en j. Más aún, por las cotas de Fréchet-Hoeffding:

$$\text{máx}\left(\frac{2j}{n}-1,0\right)\le\delta_n\left(\frac{j}{n}\right)\le\frac{j}{n}.$$

En Erdely (2007) se demuestra que

$$\delta_n\left(\frac{j+1}{n}\right) - \delta_n\left(\frac{j}{n}\right) \in \left\{0, \frac{1}{n}, \frac{2}{n}\right\},$$

lo cual significa que todos las posibles trayectorias $\left\{\delta_n(\frac{j}{n}) \,:\, j = 0, 1, \ldots, n\right\}$ se encuentran acotadas por las trayectorias $\left\{\text{máx}(\frac{2j}{n} - 1, 0) \,:\, j = 0, 1, \ldots, n\right\}$ y $\left\{\frac{j}{n} \,:\, j = 0, 1, \ldots, n\right\}$ con saltos de tamaño 0, $\frac{1}{n}$ o $\frac{2}{n}$ entre pasos consecutivos.

1.2.3. Dependencia

En este apartado se hará una breve revisión de algunas consideraciones que deben hacerse al medir asociación y dependencia entre variables aleatorias con base en el arículo de Erdely (2009).

Existen problemas en todas las áreas en los cuales es de vital interés medir de alguna forma el grado de dependencia entre variables aleatorias, de acuerdo con Mari y Kotz (2001):

> El concepto de dependencia aparece por todas partes en nuestra tierra y sus habitantes de manera profunda. Son innumerables los ejemplos de fenómenos meteorológicos interdependientes en la naturaleza, o de interdependencia en aspectos médicos, sociales, políticos y económicos de nuestra existencia. Más aún, la dependencia es obviamente no determinística, sino de naturaleza estocástica. Es por lo anterior que resulta sorprendente que conceptos y medidas de dependencia no hayan recibido suficiente atención en la literatura estadística, al menos hasta 1966 cuando el trabajo pionero de Lehmann (1966) entró en escena. El concepto de correlación (y sus modificaciones) introducido por Galton (1888) ha dominado la estadística durante unos 70 años del siglo XX, sirviendo prácticamente como la única medida de dependencia generalmente aceptada, a pesar de resultar en ocasiones claramente una medida inapropiada. La última parte del siglo XX ha sido testigo de un rápido resurgimiento en investigaciones sobre dependencia desde los puntos de vista probabilístico y estadístico. El primer libro, hasta donde sabemos, que ha sido

> dedicado a conceptos de dependencia apareció bajo la autoría de Joe (1997). Más aún, pareciera ser que no hay departamento de matemáticas o estadística en Estados Unidos o Europa que ofrezca cursos especialmente dedicados a estudiar conceptos y medidas de dependencia.

El *coeficiente de correlación lineal* (o de Pearson) entre dos variables aleatorias X y Y se denota y define

$$r(X,Y) := \frac{\mathbb{C}ov(X,Y)}{\sqrt{\mathbb{V}ar(X)\mathbb{V}ar(Y)}}. \tag{1.11}$$

Obsérvese que para calcular el coeficiente (1.11) se supone la existencia de segundos momentos no nulos, que algunas variables no cumplen. Además, según Embrechts, Lindskog y McNeil (2003):

> Las cópulas proveen una manera natural de estudiar y medir dependencia entre variables aleatorias. Como consecuencia directa del Teorema 1.2.4, las propiedades de las cópulas son invariantes bajo transformaciones estrictamente crecientes de las variables aleatorias involucradas. La correlación lineal (o de Pearson) es frecuentemente utilizada en la práctica como medida de dependencia. Sin embargo, como la correlación lineal no es una medida basada en la cópula subyacente, en ocasiones conduce a resultados aberrantes y por tanto no debiera tomarse como la medida canónica de dependencia.

En Embrechts, McNeil y Straumann (1999) se enuncian algunas consideraciones que deben tomarse en cuenta al utilizar el coeficiente de correlación lineal. Para una revisión histórica más detallada sobre medidas de asociación y conceptos de dependencia se pueden consultar Kruskal (1958) y Lehmann (1966).

Medidas de concordancia

La forma más común de medir asociación entre pares de variables es clasificarlas como concordantes o discordantes. De manera informal diremos que un par de variables aleatorias son concordantes si valores "grandes"

de una tienden a estar asociados con valores "grandes" de la otra y valores "pequeños" de una con valores "pequeños" de la otra. De forma más específica, sean (x_i, y_i) y (x_j, y_j) observaciones independientes de un vector aleatorio (X, Y) de variables aleatorias continuas. Decimos que (x_i, y_i) y (x_j, y_j) son *concordantes* si $(x_i - x_j)(y_i - y_j) > 0$ y son *discordantes* si $(x_i - x_j)(y_i - y_j) < 0$. Entonces:

$$\mathbb{P}(concordancia) = \mathbb{P}[(X_i - X_j)(Y_i - Y_j) > 0]$$

y

$$\mathbb{P}(discordancia) = \mathbb{P}[(X_i - X_j)(Y_i - Y_j) < 0]$$

Definición 1.2.4. Sean (X, Y) variables aleatorias continuas cuya cópula subyacente es C. Entonces la *tau de Kendall* (τ, $\tau_{X,Y}$ o τ_C) y la *rho de Spearman* (ρ, $\rho_{X,Y}$ o ρ_C) para (X, Y) están definidas por:

$$\tau_{X,Y} = 1 - 4 \int\int_{[0,1]^2} \frac{\partial}{\partial u} C(u, v) \frac{\partial}{\partial v} C(u, v) \, dudv \tag{1.12}$$

y

$$\rho_{X,Y} = 12 \int\int_{[0,1]^2} [C(u, v) - uv] \, dudv. \tag{1.13}$$

Las ecuaciones (1.12) y (1.13) también pueden definirse en términos de las probabilidades de concordancia y discordancia (ver Nelsen, 2006). En la siguiente definición se lista un conjunto de propiedades deseables para una medida de concordancia:

Definición 1.2.5. Una medida de asociación μ entre dos variables aleatorias continuas X, Y cuya cópula es C es una *medida de concordancia* si satisface las siguientes propiedades (la denotaremos como $\mu_{X,Y}$ o μ_C):

1. μ está definida para todo par de variables aleatorias continuas;
2. $-1 \leq \mu_{X,Y} \leq 1$, $\mu_{X,X} = 1$, y $\mu_{X,-X} = -1$;
3. $\mu_{X,Y} = \mu_{Y,X}$;
4. Si X, Y son independientes, entonces $\mu_{X,Y} = \mu_\Pi = 0$;
5. $\mu_{-X,Y} = \mu_{X,-Y} = -\mu_{X,Y}$;

6. Si C_1 y C_2 son cópulas tales que $C_1(u,v) \leq C_2(u,v)$ para todo u, v en $[0,1]$, entonces $\mu_{C_1} \leq \mu_{C_2}$;

7. Si $\{(X_n, Y_n)\}$ es una sucesión de variables aleatorias continuas con cópulas C_n, y si $\{C_n\}$ converge puntualmente a C, entonces $\lim_{n\to\infty} \mu_{C_n} = \mu_C$.

Puede demostrarse que la tau de Kendall y la rho de Spearman definidas como en (1.12) y (1.13) son medidas de concordancia, es decir, satisfacen todas las propiedades de la Definición 1.2.5. Como consecuencia de dicha definición, se enuncia el siguiente teorema:

Teorema 1.2.5. *Sea μ una medida de concordancia para variables aleatorias continuas X, Y:*

1. *Si Y es casi seguramente una función creciente de X, entonces $\mu_{X,Y} = \mu_M = 1$;*

2. *Si Y es casi seguramente una función decreciente de X, entonces $\mu_{X,Y} = \mu_W = -1$;*

3. *Si α y β son casi seguramente funciones estrictamente monótonas en $Ran(X)$ y $Ran(Y)$, respectivamente, entonces $\mu_{\alpha(X),\beta(Y)} = \mu_{X,Y}$.*

En este trabajo se usará como medida de concordancia a la rho de Spearman, en el Capítulo 4 se verá su utilidad al compararse con una medida de dependencia. Para este fin, definimos el estadístico muestral asociado a esta medida, basado en la cópula empírica:

$$\hat{\rho} = \frac{12}{n^2 - 1} \sum_{i=1}^{n} \sum_{j=1}^{n} \left[C_n\left(\frac{i}{n}, \frac{j}{n}\right) - \frac{i}{n} \cdot \frac{j}{n} \right]. \tag{1.14}$$

Medidas de dependencia

Hasta ahora no se ha llegado a un consenso sobre cuál debe ser *la* forma de medir dependencias, sin embargo existen algunas propuestas de propiedades deseables para medidas de dependencia en variables aleatorias continuas. Schweizer y Wolff (1981) proponen la siguiente definición para una medida de dependencia:

Definición 1.2.6. Una medida λ de asociación entre dos variables aleatorias continuas (X, Y) cuya cópula es C es una *medida de dependencia* (se denotará $\lambda_{X,Y}$ o λ_C) si satisface las siguientes propiedades:

1. λ debe estar definida para cada par de variables aleatorias continuas X, Y, en términos de la cópula subyacente;
2. $\lambda_{X,Y} = \lambda_{Y,X}$;
3. $0 \leq \lambda_{X,Y} \leq 1$;
4. $\lambda_{X,Y} = 0$ si y sólo si X y Y son independientes;
5. $\lambda_{X,Y} = 1$ si y sólo si alguno de X,Y es casi seguramente una función estrictamente monótona de la otra;
6. Si α y β son funciones estrictamente monótonas crecientes en $Ran(X)$ y $Ran(Y)$, respectivamente, entonces $\lambda_{\alpha(X),\beta(Y)} = \lambda_{X,Y}$;
7. Si $\{(X_n, Y_n)\}$ es una sucesión de variables aleatorias continuas con cópulas C_n, y si $\{C_n\}$ converge puntualmente a C, entonces $\lim_{n\to\infty} \lambda_{C_n} = \lambda_C$.

De acuerdo a la propiedad 4 y al Teorema 1.2.3, Schweizer y Sklar (1983b) sugieren como medida de dependencia una distancia entre la cópula correspondiente respecto de la cópula que representa la independencia.

Teorema 1.2.6. *Para cualquier p en $[1, \infty)$ la distancia L_p entre una cópula C y Π dada por:*

$$\mho_C(p) = \left(k_p \int_0^1 \int_0^1 |C(u, v) - uv|^p du dv \right)^{\frac{1}{p}} \tag{1.15}$$

en donde

$$k_p = \frac{\Gamma(2p + 3)}{2(\Gamma(p + 1))^2} \tag{1.16}$$

para cualquier $1 \leq p < \infty$ es una medida de dependencia.

Obsérvese que con k definida como (1.16), la ecuación (1.15) es igual a 1 cuando $C = M$ y cuando $C = W$. Algunos casos particulares del Teorema 1.2.6, son:

1. El índice de dependencia de Hoeffding (1940) $\mho_C^2(2)$, donde

$$\mho_C(2) = \Phi_{X,Y} = \Phi_C = \left(90 \int\int_{[0,1]^2} |C(u,v) - uv|^2 \, dudv\right)^{\frac{1}{2}} \tag{1.17}$$

2. La medida de dependencia propuesta en Schweizer y Wolff (1981), conocida como *sigma de Schweizer y Wolff*

$$\mho_C(1) = \sigma_{X,Y} = \sigma_C = 12 \int\int_{[0,1]^2} |C(u,v) - uv|^2 \, dudv \tag{1.18}$$

En Nelsen (2006) se demuestra que (1.17) y (1.18) son medidas de dependencia. A continuación se presenta la versión muestral de la sigma de Schweizer y Wolff, que es calculada a partir de la cópula empírica y será usada en el capítulo 4:

$$\hat{\sigma} = \frac{12}{n^2 - 1} \sum_{i=1}^{n} \sum_{j=1}^{n} \left| C_n\left(\frac{i}{n}, \frac{j}{n}\right) - \frac{i}{n} \cdot \frac{j}{n} \right|. \tag{1.19}$$

Dependencias estrictamente monótonas

El que un par de variables aleatorias presenten entre sí una dependencia estrictamente monótona tiene la ventaja de que es más factible ajustarles una cópula conocida, ya que la gran mayoría de ellas tienen esa propiedad. Comencemos con la definición de tal dependencia:

Definición 1.2.7. Sean X y Y variables aleatorias continuas con cópula C_{XY}, se dice que X y Y guardan una *dependencia estrictamente monótona* si $C_{XY}(u,v) - \Pi(u,v) > 0$ ó $C_{XY}(u,v) - \Pi(u,v) < 0$, $\forall\ u, v \in [0,1]$.

Obsérvese que las cópulas M y W son los extremos de la dependencia estrictamente monótona.

Recordemos las definiciones de la rho de Spearman (1.13) y la sigma de Schweizer y Wolff (1.18), es sencillo demostrar que si X y Y son variables aleatorias con dependencia estrictamente monótona, entonces:

$$|\rho_{X,Y}| = \sigma_{X,Y}. \tag{1.20}$$

En la medida que el indicador $\sigma_{X,Y} - |\rho_{X,Y}|$ se aleje de cero, diremos que existen dependencias no monótonas a modelar. Si es el caso, lo que procedería sería particionar la muestra en subconjuntos en los que se presenten dependencias monótonas e intentar ajustar a cada subconjunto una cópula conocida para después construir una única cópula a través de los segmentos; ésta podría construirse mediante la técnica de "pegado de cópulas" propuesta por Siburg y Stoimenov (2008).

En el presente estudio no se ajustarán cópulas paramétricas a los datos obtenidos, únicamente se realizará un análisis de dependencias interpretando las magnitudes de los índices presentados en esta sección.

Capítulo 2

Análisis descriptivo

Este capítulo se enfoca en tres objetivos: (1) describir los datos involucrados en este estudio, así como la notación que será utilizada a lo largo del mismo; (2) explicar cómo está conformado el grupo de la *no respuesta* y cómo éste pudo influir en el sesgo de la encuesta de seguimiento diario Milenio GEA/ISA; y (3) presentar los resultados obtenidos por la encuestadora y proponer una corrección del sesgo estimado.

2.1. Variables

Los datos analizados en el presente trabajo corresponden a los resultados de encuestas de preferencias electorales, expresados en porcentajes de votos por candidato para las elecciones presidencial y de jefe de gobierno del Distrito Federal del año 2012, publicados por el medio de comunicación Milenio y obtenidos por GEA/ISA. Dichos datos fueron extraídos de los ejemplares del periódico Milenio (2012) del 19 de marzo al 27 de junio para la elección presidencial y del 16 de abril al 22 de junio para la elección de jefe de gobierno del Distrito Federal. Todos los resultados fueron publicados en el periódico el día inmediato posterior a la encuesta correspondiente.

Los resultados de las encuestas antes mencionadas se capturaron en el software de distribución libre R (véase R Core Team, 2013) y se muestran en el Apéndice A, en los Cuadros A.1, A.2, A.3 y A.4 para la elección presidencial, y los Cuadros A.5, A.6 y A.7 para la elección de jefe de gobierno. En la sección *Datos* del Anexo B.1 pueden consultarse los detalles de los objetos

creados en R Core Team para este análisis.

A lo largo del trabajo se harán las siguientes abreviaturas de los nombres de los candidatos:

Elección Presidencial

EPN	Enrique Peña Nieto
JVM	Josefina Eugenia Vázquez Mota
AMLO	Andrés Manuel López Obrador
Quadri	Gabriel Ricardo Quadri de la Torre

Elección de Jefe de Gobierno

MAM	Miguel Ángel Mancera Espinosa
BPR	Beatriz Paredes Rangel
IMdW	Isabel Miranda de Wallace
RG	Rosario Guerra Díaz

Las variables de estudio en este trabajo son los votos por candidato para cada elección y se cuenta con información acerca de los porcentajes estimados de dichos votos, sin embargo, para convertir los porcentajes de preferencia a números absolutos de votos estimados fue necesario consultar el total de votantes de cada una de las elecciones.

El cómputo final de la elección presidencial (véase Cuadro 2.1) fue obtenido del sitio electrónico del Tribunal Electoral del Poder Judicial de la Federación (2012) y el correspondiente a la elección de jefe de gobierno (véase Cuadro 2.2) se extrajo del "Acta de cómputo total de la elección de Jefe de Gobierno del Distrito Federal 2012" publicada en el sitio electrónico del Instituto Electoral del Distrito Federal (2012). Cabe mencionar que el análisis estadístico será efectuado por candidato, por lo cual fueron sumados los votos de algunos de ellos que contendían por su partido político y además por alguna coalición.

Candidato	Total de votos
EPN	19,158,592
AMLO	15,848,827
JVM	12,732,630
Quadri	1,146,085

Cuadro 2.1: Cómputo final para la elección presidencial

De los Cuadros 2.1 y 2.2 se tiene que el total de votos válidos para la elección presidencial es de $48,886,134$ y para la elección de jefe de gobierno de $4,681,115$ votos, estos totales serán denotados por n_{pres} y n_{jg}, respectivamente. Para estos indicadores no se tomarán en cuenta los votos nulos ni los votos a candidatos no registrados.

Candidato	Total de votos
MAM	3,031,156
BPR	941,921
IMdW	649,345
RG	58,693

Cuadro 2.2: Cómputo final para la elección de jefe de gobierno

Se llamará *indecisos* a todas las personas que respondieron la encuesta, pero que declararon no tener claro por cuál de los candidatos votarían; en las publicaciones del periódico le llaman *"indefinido"* a esta misma variable. El último día de publicación de ambas encuestas analizadas, únicamente aparecen los porcentajes correspondientes a cada uno de los candidatos y no existe información acerca de los *indecisos*; para continuar con el análisis de los mismos se realizó una estimación del porcentaje de votos que les correspondía tomando en cuenta el registro inmediato anterior.

Algunos días las estimaciones publicadas en el periódico Milenio presentaban errores de redondeo que ocasionaban que los porcentajes pronosticados (para los candidatos y los indecisos) no sumaran el 100 %. Para la elección presidencial, este problema sucedió los días 20 de marzo(99.9 %), 1º de abril(99.8 %), 10 de abril(100.5 %), 2 de mayo(100.3 %), 3 de mayo(100.8 %), 6 de junio (99.74 %) y 25 de junio(101 %); y para la elección de jefe de gobierno el 17 de abril(100.2 %), 30 de abril (100.5 %) y 3 de junio(100.5 %).

Para corregir el problema anterior, en los registros correspondientes se reescalaron los porcentajes guardando la misma proporción. A continuación se muestra un ejemplo de este procedimiento aplicado al día 20 de marzo para la elección presidencial.

Ejemplo. Porcentajes publicados correspondientes al 20 de marzo.

EPN	**JVM**	**AMLO**	**Quadri**	**Indecisos**	**Suma**
32.6	22.1	14.5	0.6	30.1	**99.9**

Ajuste de los porcentajes:

EPN	**JVM**	**AMLO**	**Quadri**	**Indecisos**	**Suma**
$32.6 \cdot \frac{100}{99.9}$	$22.1 \cdot \frac{100}{99.9}$	$14.5 \cdot \frac{100}{99.9}$	$0.6 \cdot \frac{100}{99.9}$	$30.1 \cdot \frac{100}{99.9}$	$\mathbf{99.9} \cdot \frac{100}{99.9}$
32.63263	22.12212	14.51451	0.6006006	30.13013	**100**

En general, si en un registro la suma de los porcentajes (*suma*) difería del 100 %, se multiplicó toda la fila por $100/suma$ para corregirlo.

A continuación se abordará un tema que fue muy controversial en los medios de comunicación por tratarse de un asunto delicado acerca del cual muchos especialistas sociales siguen buscando explicación.

2.2. La tasa de no respuesta

En el levantamiento de una encuesta ocurre inevitablemente que es imposible obtener, de parte de algunas de las unidades interrogadas, respuestas para una o más variables que la encuesta desea medir. Se dice entonces que se está en presencia de *no respuestas*.

En la aplicación de cuestionarios de preferencias electorales, como fue el caso de la encuesta de seguimiento diario Milenio GEA/ISA, la no respuesta engloba tanto a las personas que respondieron la encuesta pero que no proporcionaron su preferencia por alguno de los candidatos (*indecisos*) como a las personas que se negaron a responder (*encuestas rechazadas*). La información que proporcionan los datos obtenidos del periódico Milenio (2012),

denotados como *indefinido* corresponden al primero de estos grupos, sin embargo, para poder obtener la no respuesta total, necesitamos conocer algún índice de rechazo de la encuesta. En Aristegui Noticias (2012) se publica una tasa de rechazo para la encuestadora GEA/ISA entre 30 y 33.8 %. Se utilizará en análisis posteriores la cota superior de dicho intervalo para las dos encuestas en estudio.

La no respuesta puede deberse a distintos factores, como lo afirma Ricardo de la Peña, presidente ejecutivo de GEA/ISA (Everdy, 2012):

> Una parte de los votantes catalogados en la 'No respuesta' son votantes indecisos, de los cuales encontramos dos tipos: unos son aquellos que poseen información sobre los procesos electorales pero no tienen una decisión tomada, y otros son personas que simplemente no tienen información ni idea del proceso electoral, por lo que son incapaces de emitir una valoración al respecto.
>
> Además, existe un segmento de personas que no va a votar por ninguno de los candidatos o que piensa anular su voto sumados a un movimiento denominado 'anulista', que busca manifestar su inconformidad con los políticos y los partidos, así como su exigencia de una reforma en la materia y otro grupo que simplemente se queda callado o que alude a que el voto es secreto. Todos ellos figuran en esta categoría.

Como no se cuenta con información acerca de los anulistas o personas que van a votar por algún candidato no registrado, se supondrá que todas las personas pertenecientes al grupo de la no respuesta pretendían votar por alguno de los cuatro candidatos registrados en la elección correspondiente. A continuación se presenta el procedimiento utilizado para estimar las preferencias electorales de la no respuesta.

Como se dijo antes, la no respuesta corresponde a la suma de los indecisos más las encuestas rechazadas, tomando como indicador el 33.8 % de rechazo. El primer paso a realizar, entonces, es crear una nueva variable que mida el porcentaje de la no respuesta ajustando el porcentaje de indecisos publicado diariamente en el periódico con la tasa de rechazo mencionada previamente. También se requería modificar los tamaños de muestra sumando a los originales las personas que rechazaron la encuesta de acuerdo a la información

obtenida de Aristegui Noticias (2012). Hecho esto se convirtieron los porcentajes ya ajustados a número de votos estimados, tomando como referencia la suma del cómputo final de la elección correspondiente (véase la función *votos* en el Anexo B.1).

En la siguiente sección se realiza una comparación entre los resultados de la encuesta y el cómputo final. También se obtiene una estimación de las preferencias de la no respuesta proponiendo un método de corrección de sesgo.

2.3. Análisis del sesgo

Recordemos que el *sesgo* ($\mathbb{S}$) de un estimador $\hat{\theta}$ del parámetro o vector de parámetros desconocido θ, se define como la diferencia entre la esperanza de dicho estimador y el verdadero valor del parámetro, es decir,

$$\mathbb{S}(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta.$$

Es deseable que los estimadores sean insesgados, ya que al ser así su esperanza es igual al parámetro que se desea estimar; y en cualquier caso, la estimación $\hat{\theta}$ difícilmente será igual al verdadero valor del parámetro θ. Un estimador insesgado de $\mathbb{S}(\hat{\theta})$ es

$$\widehat{\mathbb{S}}(\hat{\theta}) = \hat{\theta} - \theta$$

ya que $\mathbb{E}\big[\,\widehat{\mathbb{S}}(\hat{\theta})\,\big] = \mathbb{E}(\hat{\theta}) - \theta = \mathbb{S}(\hat{\theta})$.

Si atribuimos el sesgo de los resultados de GEA/ISA al hecho de que no se conocían las preferencias de la no respuesta, se tomarán en cuenta dos alternativas:

1. Un sesgo diferente cada día de la encuesta, y

2. un sesgo sistemático constante prevaleciente durante todo el periodo de encuestas.

Para modelar estadísticamente el proceso de la encuesta, usamos el modelo multinomial descrito en la Sección 1.1.1 del Capítulo 1, ya que cada

elemento del universo (total de votantes) puede elegir uno de cuatro candidatos con probabilidades desconocidas. Los porcentajes publicados diariamente son una estimación de tales probabilidades y la diferencia entre el cómputo final y los porcentajes del último día de encuesta es una estimación del sesgo del último registro. Es evidente que para medir el sesgo de los datos de estudio únicamente podemos compararlos con el cómputo final (valores "reales" de los parámetros), así que para la alternativa 1, se propone el siguiente procedimiento:

- Estimar el sesgo (en votos) de la estimación de cada día.
- Ya obtenido el número de votos estimado para la no respuesta diariamente, determinar por diferencia cuántos votos corresponden a cada candidato por día, de manera que cada uno de ellos alcance el total de votos que obtuvo en el cómputo final.
- Convertir de nuevo a porcentajes (si se desea) o presentar los resultados en votos.

La ventaja de este primer método para corregir el sesgo es que puede estimarse cómo debieron comportarse las preferencias de la no respuesta a lo largo del periodo de encuesta para obtener los resultados del cómputo final.

Del lado izquierdo en la Figura 2.1, se observa la gráfica de las estimaciones de preferencias electorales publicadas por GEA/ISA sin tomar en cuenta a los indecisos. En el día 101 aparecen marcados con un punto los resultados para cada candidato en el cómputo final; tal día se observa el gran sesgo del último resultado de la encuesta en algunos candidatos al identificar la distancia entre dicho resultado y el cómputo final.

En la misma Figura, del lado derecho aparecen las trayectorias que debieron seguir las preferencias de la no respuesta para llegar a un resultado sin sesgo. Puede verse, por ejemplo, que Enrique Peña Nieto, que estaba sobreestimado en la encuesta, debió tener menor porcentaje de preferencia en el grupo de la no respuesta para que así se llegara al resultado esperado.

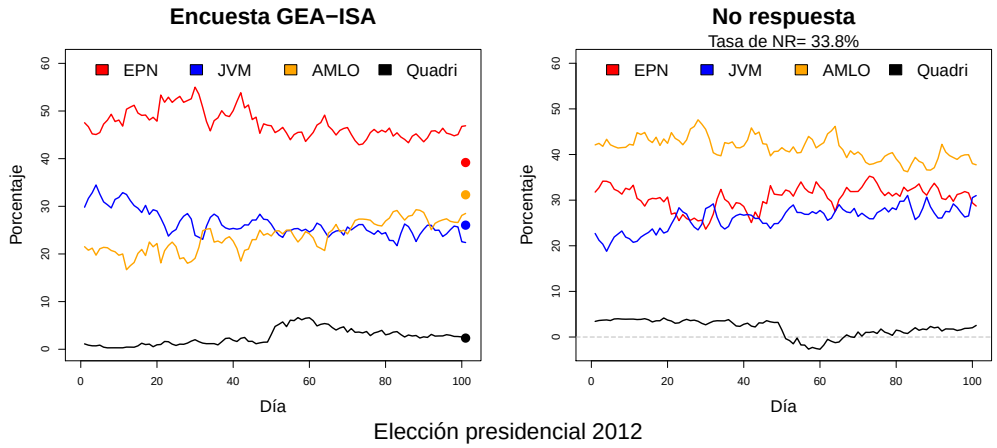

Figura 2.1: Trayectoria estimada para la no respuesta en la elección presidencial

Sucede de forma análoga con la elección de jefe de gobierno; en la Figura 2.2 del lado izquierdo se muestran las estimaciones de la encuestadora, y puede verse que el último día de encuesta (día 68) el resultado esperado en cada uno de los candidatos es muy cercano al real, es decir, el sesgo que se presentó fue mínimo. Del lado izquierdo se presentan las preferencias pronosticadas para la no respuesta por el método 1.

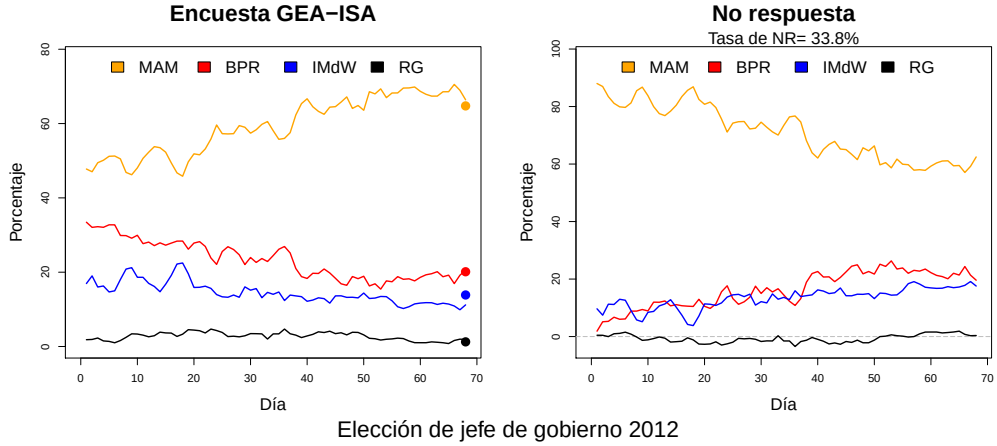

Figura 2.2: Trayectoria estimada para la no respuesta en la elección de jefe de gobierno

Sin embargo, se presenta una inconsistencia en los datos calculados, ya

que al observar la trayectoria de Gabriel Quadri en la Figura 2.1 y de la candidata Rosario Guerra en la Figura 2.2, encontramos que algunos de los porcentajes pronosticados por este método son negativos. Esto puede deberse a dos razones fundamentales, a saber: una posible explicación es que no sea realista el suponer que los porcentajes obtenidos por los candidatos en el cómputo final se conservaron a lo largo del tiempo, ya que el tiempo sí influye en el comportamiento de las preferencias de las personas (por ejemplo antes del primer debate o después de él); por otro lado, también puede deberse a que la tasa de no respuesta estimada por la encuestadora y publicada en Aristegui Noticias (2012) no era suficiente para que el sesgo fuera explicado únicamente por las preferencias de la no respuesta, es decir, se necesitaría hacer más grande al grupo de la no respuesta para que al repartirlos entre los candidatos todos los porcentajes fueran consistentes.

Otra desventaja del mismo método es que las trayectorias insesgadas de todos los candidatos son constantes en todo el periodo de encuesta (iguales al cómputo final).

A raíz de los problemas anteriores se propone un método alternativo para corregir el sesgo de la encuesta (método 2) en el cual se supondrá que existió por algun error no controlado por la encuestadora un sesgo sistemático constante a lo largo del periodo de encuesta, es decir, que el sesgo (en proporción) presentado en el último registro que se calculó anteriormente, se conservó aproximadamente constante en el tiempo.

Para aplicar este segundo método se ejecutaron los siguientes pasos:

- Estimar el sesgo (en votos) de la última estimación para cada candidato.
- Calcular la proporción de votos del grupo de la no respuesta que corresponde a cada uno de los candidatos el último día de encuesta.
- Para los días anteriores, repartir los votos calculados pertenecientes a la no respuesta entre los candidatos conservando las proporciones anteriores.
- Para la trayectoria sin sesgo, sumar los votos originales estimados por la encuesta a los obtenidos del cálculo anterior.

- Convertir de nuevo a porcentajes (si se desea) o presentar los resultados en votos.

Una ventaja de este método, además de contar con estimaciones consistentes, es que se obtiene una trayectoria insesgada más realista; no obstante, el inconveniente que presenta es el no poder pronosticar las trayectorias de la no respuesta, ya que éstas permanecen constantes.

En la Figura 2.3 se muestra para la elección presidencial una comparación entre los resultados obtenidos por GEA/ISA y los mismos corregidos mediante el método 2. Se observa que las trayectorias conservan su forma original, pero se reduce considerablemente la variabilidad de las mismas. Además, las intersecciones de trayectorias, que son de gran importancia porque determinan el momento en el que un candidato supera a otro, se presentan en momentos muy diferentes que en la encuesta original.

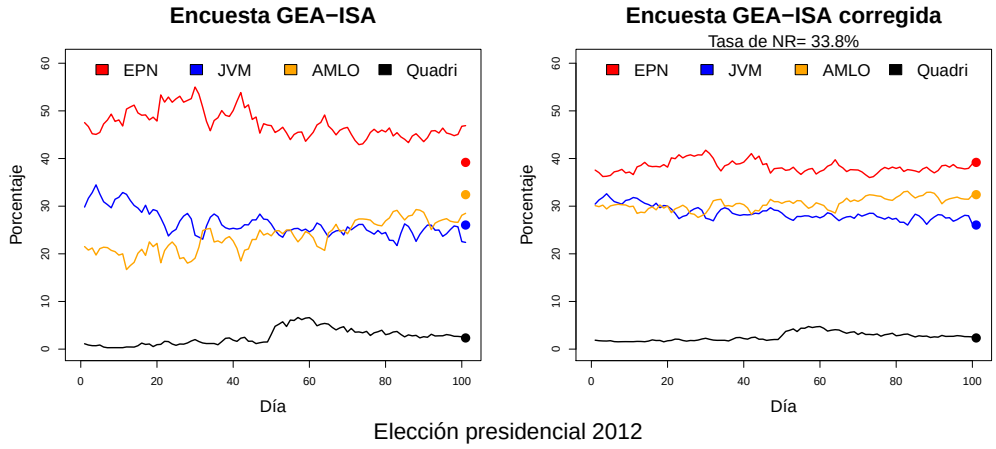

Figura 2.3: Trayectoria estimada por GEA/ISA vs. trayectoria 'sin sesgo' para la elección presidencial

Para la elección de jefe de gobierno, en la Figura 2.4 se observa igualmente una disminución de la variabilidad, pero no tan considerable como en la elección presidencial. Esto se debe a que el sesgo de la primera es considerablemente menor, y por ello el ajuste de los resultados es más inmediato. Además, todas las trayectorias estimadas conservan su tendencia, ya sea creciente o decreciente.

Considerando las ventajas y desventajas de ambos métodos propuestos, en los análisis posteriores se usará el método 2 por no presentar inconsistencias en los resultados, y para poder trabajar con las trayectorias originales y las corregidas.

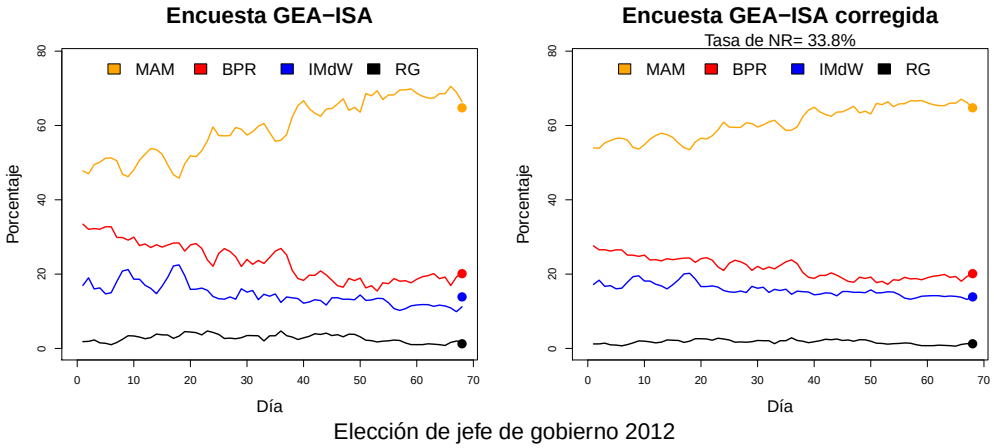

Figura 2.4: Trayectoria estimada por GEA/ISA vs. trayectoria 'sin sesgo' para la elección de jefe de gobierno

2.4. Sobre/sub estimación

Como ya se mencionó anteriormente, los resultados obtenidos por la encuesta en la elección presidencial difieren de los porcentajes reales que obtuvieron los candidatos al momento de la elección. En este apartado se realizará un análisis de dichas diferencias, tomando como el valor *real* del vector de parámetros el cómputo final de la elección correspondiente. Se llama sobreestimación cuando el valor pronosticado para una variable queda por encima del verdadero y subestimación cuando se pronostica un menor valor que el real.

La sobreestimación y la subestimación siempre existen, ya que no se puede realizar una estimación puntual exacta del parámetro o vector de parámetros debido al error muestral, sin embargo, existen niveles en los cuales es aceptable y otros en los cuales resulta catastrófico que las estimaciones se alejen demasiado del verdadero valor. En política, sobre todo, esta situación

suele generar sospechas y polémicas relacionadas con la manipulación de los datos, ya que está demostrado (véase Metodología de encuestas, 2000) que el comportamiento de las preferencias electorales en las encuestas sí influye en las decisiones de los electores.

En la Figura 2.5 se muestran los porcentajes de sobreestimación y subestimación para cada candidato de acuerdo con los resultados de GEA/ISA y tomando como porcentaje de no respuesta 33.8 %. Podemos observar que el candidato Enrique Peña Nieto estuvo sobreestimado durante todo el periodo de encuesta con niveles que alcanzaron hasta 13 % sobre el total de votos que obtuvo en el cómputo final; por otro lado, el candidato Andrés Manuel López Obrador presentó subestimación en cada uno de los resultados, llegando a estar hasta un 12 % por debajo del cómputo final. La candidata Josefina Eugenia Vázquez Mota se subestimó a partir del día 18, pero a niveles más moderados, ya que no rebasaban el 5 %. Por último el candidato Gabriel Quadri fue subestimado la primera mitad del periodo de encuesta, y de ahí en adelante se sobreestimó ligeramente.

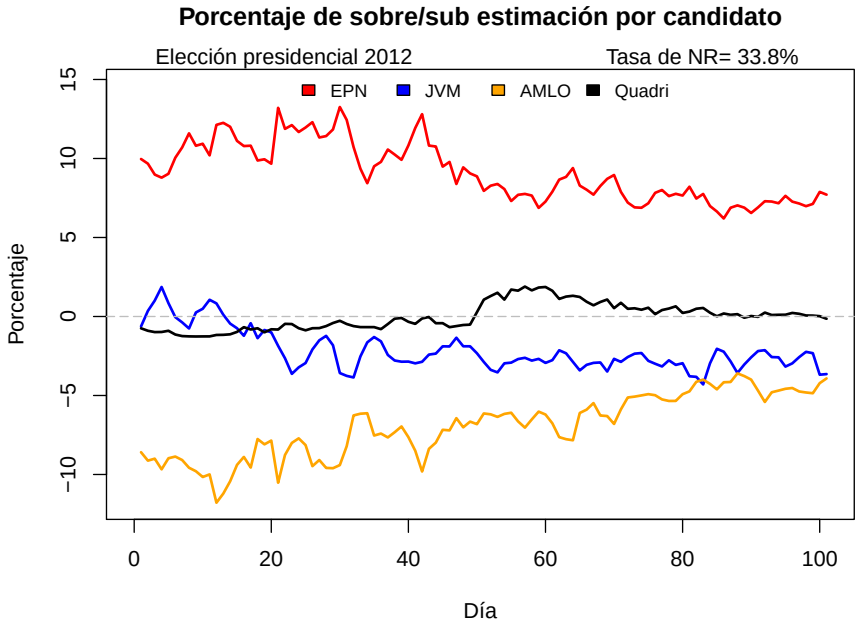

Figura 2.5: Porcentajes de sobre/sub estimación de la encuesta en la elección presidencial

Concluimos este capítulo presentando los porcentajes de sobreestimación y subestimación por candidato correspondientes a la elección de jefe de gobierno del Distrito Federal, los cuales se muestran en la Figura 2.6. Es evidente que dichos porcentajes son considerablemente menores a los obtenidos en la elección presidencial, ya que esta vez el rango de sobre/sub estimación se mueve entre -8 y 6 puntos porcentuales. En este caso todas las trayectorias (excepto la de Rosario Guerra Díaz) se intersectan con el 0, lo cual significa que en algún momento del tiempo se dieron estimaciones casi exactas. La candidata Rosario Guerra presentó una sobreestimación en todas las estimaciones pero ésta fue mínima (no mayor al 2 %). Todos los demás candidatos presentaban una tendencia en el comportamiento de su sobre/sub estimación. En el caso del candidato Miguel Ángel Mancera dicha tendencia fue creciente, por otro lado, las candidatas Beatriz Paredes Rangel e Isabel Miranda de Wallace presentan una tendencia decreciente.

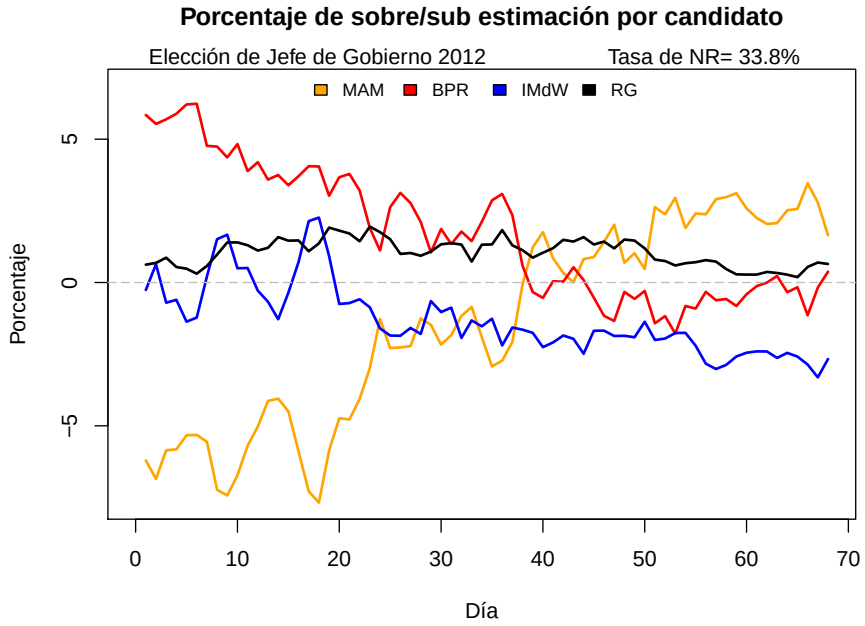

Figura 2.6: Porcentajes de sobre/sub estimación de la encuesta en la elección de jefe de gobierno

Capítulo 3

Modelo bayesiano

Este capítulo tiene como objetivo principal realizar simulación de un modelo bayesiano propuesto para nuestros datos, para posteriormente hacer cálculos relacionados con las probabilidades de ganar de los candidatos involucrados en cada elección y estimaciones de los porcentajes de preferencia de dichos candidatos.

La propuesta de monitorear las probabilidades de ganar de los candidatos en vez de los porcentajes de preferencia fue presentada por Nate Silver (véase O'Hara, 2012), este estadístico predijo al ganador por estado en la elección presidencial de Estados Unidos del 2012. En un principio su método fue muy criticado por varios analistas, sin embargo, al corroborar con los resultados del día de la elección, sus estimaciones resultaron muy acertadas.

De la misma manera, en este estudio se hará cálculo de las probabilidades de ganar, pero considerando cuatro candidatos elegibles en vez de dos. A continuación se presenta el modelo estadístico paramétrico que será utilizado con dicho propósito.

3.1. Aplicación del modelo multinomial

Recordemos el modelo multinomial con cuatro parámetros, visto en la Sección 1.1.1. Si se desean simular las proporciones diarias de votos para cada uno de los candidatos en la elección presidencial y suponemos una distribución a priori Dirichlet poco informativa *(a priori de Jeffreys)*, es de-

cir $\boldsymbol{\alpha} = (\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2})$; la distribución a posteriori del vector de proporciones $\mathbf{p} = (p_1, p_2, p_3, p_4)$ con p_1 asociado al candidato EPN, p_2 asociado a la candidata JVM, p_3 asociado a AMLO y p_4 asociado a Quadri resulta:

$$\wp\big(p_1, \ldots, p_4 \,\big|\, \frac{1}{2} + x_1, \ldots, \frac{1}{2} + x_4\big) = \frac{1}{B(\boldsymbol{\alpha} + \mathbf{x})} \prod_{i=1}^{4} p_i^{x_i - \frac{1}{2}} \, \mathbb{1}_{\{\sum_{i=1}^{4} p_i = 1\}}. \quad (3.1)$$

donde $\mathbf{x} = (x_1, x_2, x_3, x_4)$ son las observaciones de una realización del modelo multinomial asociadas con cada uno de los candidatos de la elección presidencial en el orden ya mencionado.

De manera análoga para la elección de jefe de gobierno, si asociamos cada entrada del vector de proporciones $\mathbf{q} = (q_1, q_2, q_3, q_4)$ a los candidatos MAM, BPR, IMdW y RG, respectivamente, y usando la distribución a priori de Jeffreys para la multinomial, tenemos como distribución a posteriori de $\mathbf{q}$:

$$\wp\big(q_1, \ldots, q_4 \,\big|\, \frac{1}{2} + y_1, \ldots, \frac{1}{2} + y_4\big) = \frac{1}{B(\boldsymbol{\alpha} + \mathbf{y})} \prod_{i=1}^{4} q_i^{y_i - \frac{1}{2}} \, \mathbb{1}_{\{\sum_{i=1}^{4} q_i = 1\}}. \quad (3.2)$$

con $\mathbf{y} = (y_1, y_2, y_3, y_4)$ las observaciones de una realización del modelo multinomial asociadas con cada uno de los candidatos en el orden previsto anteriormente.

Es decir, como se vio en la Sección 1.1.2 para ambas elecciones la distribución a posteriori del vector de parámetros resulta una Dirichlet con parámetros $(1/2+x_1, \ldots, 1/2+x_4)$ y $(1/2+y_1, \ldots, 1/2+y_4)$, respectivamente.

En resumen, cada registro diario en los resultados de la encuesta (convertidos a encuestas efectivas) se puede ver como la observacion de una realización del modelo multinomial; entonces, para realizar la simulación de la distribución a posteriori de $\mathbf{p}$ y $\mathbf{q}$ se tomaron como (x_1, x_2, x_3, x_4) y (y_1, y_2, y_3, y_4) las encuestas efectivas estimadas en cada día de encuesta de la elección presidencial y de la elección de jefe de gobierno, respectivamente.

Para realizar la simulación se hizo uso del paquete MCMCpack (véase Martin, Quinn y Park, 2011), del software estadístico R Core Team, ya que

éste contiene funciones precompiladas para calcular la densidad y simular números aleatorios con distribución Dirichlet, dados los parámetros. Así, se realizaron simulaciones de las proporciones por día dada la distribución a priori de Jeffreys y los datos proporcionados por los resultados de la encuesta. Si se desea mayor detalle del procedimiento, puede encontrarse en la función *simula.posteriori* del Anexo B.2.

3.2. Probabilidades de ganar

Para obtener las probabilidades empíricas de ganar por candidato día a día, se realizaron diez mil simulaciones del vector de proporciones siguiendo la distribución Dirichlet de las ecuaciones (3.1) y (3.2). Se calcularon probabilidades globales de ganar y por pares de candidatos (para mayor detalle puede consultarse la función *probab.ganar* del Anexo B.2). Los resultados se muestran a continuación.

3.2.1. Probabilidades globales

Definiremos la *probabilidad global de ganar* del candidato i como

$$\mathbb{P}(\theta_i = \text{máx}\{\theta_1, \theta_2, \theta_3, \theta_4\})$$

con $i \in \{1, 2, 3, 4\}$ y donde $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3, \theta_4)$ es el vector que contiene las proporciones de votantes por candidato en la elección correspondiente. Regresando a la notación anterior, $\boldsymbol{\theta} = \mathbf{p}$ para la elección presidencial y $\boldsymbol{\theta} = \mathbf{q}$ en la elección de jefe de gobierno.

En la Figura 3.1 podemos observar el comportamiento de las probabilidades de ganar de los candidatos correspondientes a la elección presidencial, tanto para la encuesta original, como para la encuesta corregida, usando el método antes descrito y una tasa de no respuesta de 33.8 %. Notamos que para Enrique Peña Nieto, con los datos de la encuesta original se obtiene que la probabilidad de ganar es casi constante e igual a uno; en contraste con los demás candidatos, cuyas probabilidades de ganar usando la misma encuesta son muy cercanas cero a lo largo del periodo de análisis. Por otro lado, las trayectorias con la encuesta corregida presentan algunas fluctuaciones, lo cual puede dar lugar a un análisis posterior donde se busque explicar las causas de las mismas de acuerdo al periodo en el que ocurrieron y los hechos

relevantes que hayan ocurrido en ese periodo de la campaña, sin embargo, al ser las fluctuaciones anteriores muy pequeñas, serán despreciadas en el presente análisis.

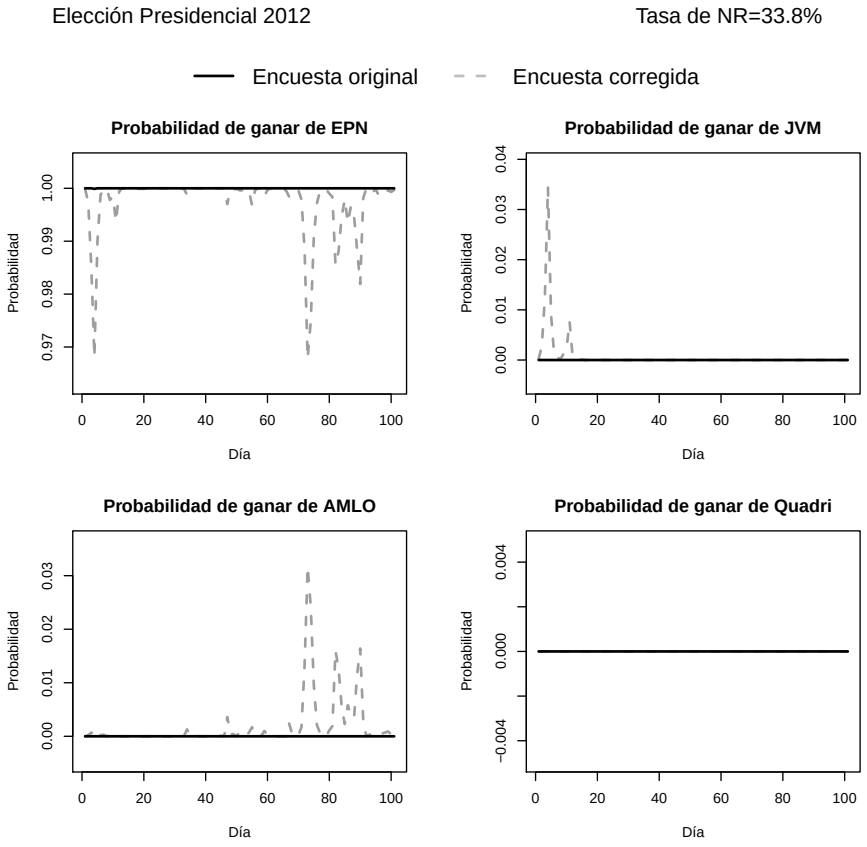

Figura 3.1: Probabilidades de ganar por candidato en la elección presidencial

Por otro lado, encontramos en la Figura 3.2 que ambas encuestas (la original y la 'corregida') arrojan que la probabilidad de ganar del candidato Miguel Ángel Mancera era casi constante e igual a uno durante todo el periodo de encuesta. En este caso, por la polarización de preferencias de la que se habló anteriormente, la elección ya estaba prácticamente definida. Es decir, con probabilidad muy cercana a uno, si se realizaba la elección en vez de la encuesta de seguimiento diario, hubiera ganado el candidato Miguel Ángel Mancera en la gran mayoría de los casos.

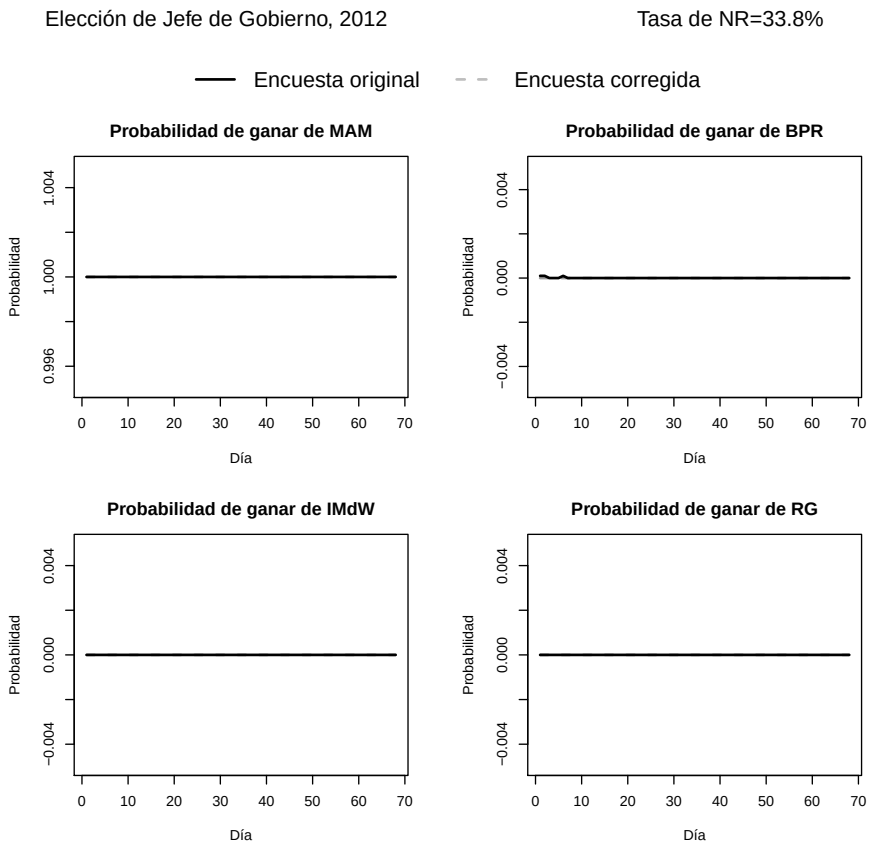

Figura 3.2: Probabilidades de ganar por candidato en la elección de jefe de gobierno

3.2.2. Probabilidades por pares de candidatos

La probabilidad de que el candidato i le gane al candidato j con $i \neq j$ es

$$\mathbb{P}(\theta_i > \theta_j)$$

con $i, j \in \{1, 2, 3, 4\}$ y donde θ_k es el parámetro de la distribución dirichlet correspondiente al candidato k, es decir, la proporción de votos que le corresponde a dicho candidato.

Haciendo el análisis de las probabilidades por pares de que un candidato le gane a otro en la elección presidencial, no se encontraron trayectorias distintas de las triviales en la mayoría de los casos. En la Figura 3.3 se muestra

un gráfico con un comportamiento muy interesante, el monitoreo de la probabilidad de que la candidata Josefina Vázquez Mota le gane a Andrés Manuel López Obrador. Puede observarse que en la encuesta original, al principio del seguimiento de la encuesta, dicha probabilidad permanecía casi constante e igual a uno hasta aproximadamente el día 20, posteriormente, ésta empezó a oscilar de una forma muy drástica hasta llegar a cero al final del periodo de encuesta. Por otro lado, al analizar la trayectoria de la encuesta corregida podemos concluir que se estaban sobreestimando tales probabilidades, ya que ésta se encuentra muy por debajo de la anterior. Además, empieza a estacionarse alrededor del cero desde el día 65, a diferencia de la encuesta original, que presenta una caída drástica hasta el día 100.

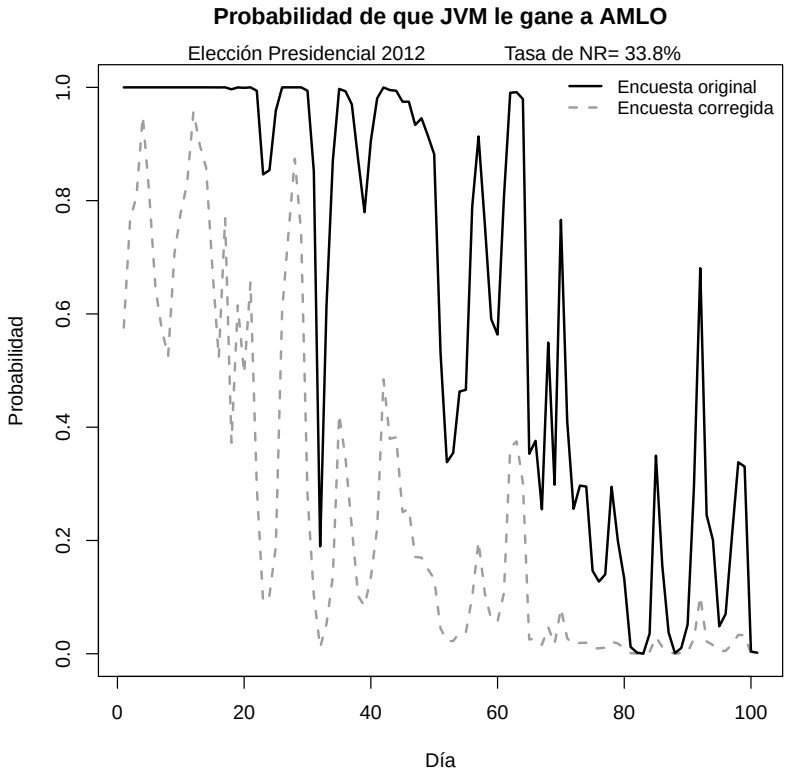

Figura 3.3: Probabilidad de que JVM le gane a AMLO

En el caso de la elección de jefe de gobierno, el único caso que ameritó nuestro estudio por ser una trayectoria distinta de la trivial fue el mo-

nitoreo de las probabilidades de que la candidata Beatriz Paredes Rangel le ganara a Isabel Miranda de Wallace. Como se observa en la Figura 3.4, la trayectoria basada en la encuesta original es muy parecida a la proveniente de la encuesta corregida. Los primeros días la candidata Beatriz Paredes Rangel tenía un triunfo casi seguro sobre la otra contendiente, sin embargo, a partir del día 15, la probabilidad de que esto sucediera comenzó a decrecer, dando un salto de hasta 20 % en el día 52.

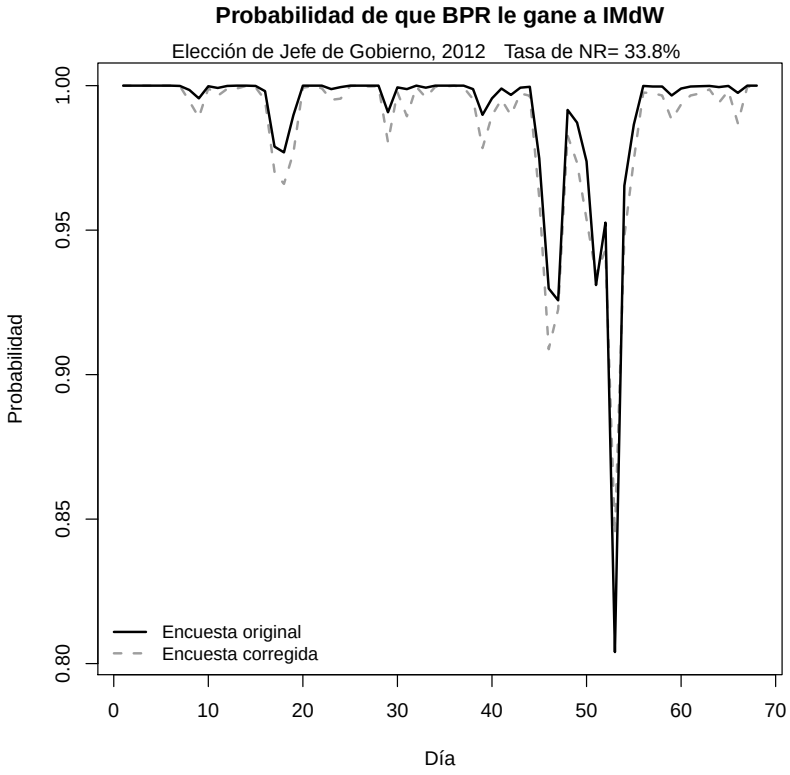

Figura 3.4: Probabilidad de que BPR le gane a IMdW

A continuación se presenta la última sección de este capítulo, donde se realizará estimación bayesiana de las proporciones correspondientes por candidato, de la manera que se describió en la Sección 1.1.2.

3.3. Estimación bayesiana

Como hemos venido mencionando, el propósito principal de las encuestas de preferencias electorales es realizar una radiografía de las mismas en un momento del tiempo que pueda servir de pronóstico para conocer al ganador de la elección y los porcentajes de votos correspondientes a cada uno de los candidatos. La estimación que se realizará en este apartado se basa en los porcentajes publicados por Milenio GEA/ISA y la encuesta corregida (sin sesgo).

En primer lugar, se realizaron 10000 simulaciones de la distribución a posteriori (ver ecuaciones (3.1), (3.2)) por cada día de encuesta, tomando como muestra observada el vector de proporciones de los resultados de la encuesta correspondiente a ese día convertido a número de encuestas efectivas.

En el Cuadro 3.1 se muestra un ejemplo de las estimaciones bayesianas correspondientes a dos días diferentes de encuesta para la elección presidencial. Se monitorean las estimaciones del día 45 y el día 55 de encuesta, ya que se desea corroborar si existe un cambio notorio en las preferencias electorales antes y después del día 49 (6 de mayo de 2012) en el cual se llevó a cabo el primer debate entre los candidatos a Presidente de los Estados Unidos Mexicanos.

	Día 45			Día 55		
	2.5 %	50 %	97.5 %	2.5 %	50 %	97.5 %
EPN	0.45	0.48	0.51	0.41	0.44	0.47
JVM	0.24	0.27	0.30	0.22	0.25	0.28
AMLO	0.21	0.23	0.25	0.23	0.25	0.28
Quadri	0.01	0.02	0.02	0.05	0.06	0.07

Cuadro 3.1: Estimaciones puntuales y por intervalo para los días 45 y 55 (elección presidencial)

Realizado lo anterior observamos que, en efecto, la estimación puntual del porcentaje de preferencia del candidato Gabriel Quadri incrementa considerablemente después del debate, al igual que el correspondiente al candidato Andrés Manuel López Obrador, que aumentó del 23 al 25 %. Por otro lado, la candidata Josefina Vázquez Mota sufrió un decremento del 2 % en su

porcentaje de preferencia; y por último el candidato Enrique Peña Nieto decayó cuatro puntos porcentuales en su porcentaje de simpatizantes.

El procedimiento anterior se realizó por cada día de encuesta, es decir, se estimaron puntualmente y por intervalos los porcentajes de preferencia diarios para los candidatos de ambas elecciones. El resultado para la elección presidencial se muestra en la Figura 3.5, donde se puede apreciar la diferencia entre la encuesta original y la encuesta corregida, que presenta menor variabilidad y un comportamiento más estable que la primera.

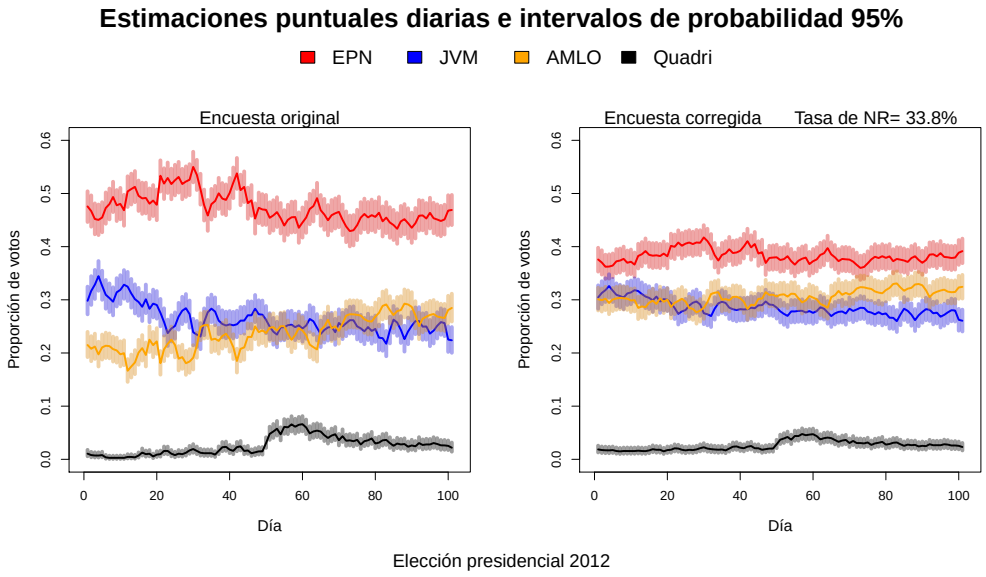

Figura 3.5: Estimaciones de las proporciones diarias por candidato para la elección presidencial

Del mismo modo, en la Figura 3.6 se presentan los resultados de la estimación bayesiana de las proporciones diarias por candidato para la elección de jefe de gobierno del Distrito Federal.

Para concluir este capítulo cabe mencionar que fue utilizada la mediana de las simulaciones como estimación puntual de las proporciones correspondientes a cada día, ya que se sabe que este estadístico es más robusto que la media por no ser sensible a datos atípicos. Además, puede hablarse de intervalos de *probabilidad* 95 %, ya que los mismos fueron calculados a partir

de la simulación de observaciones del modelo multinomial bayesiano donde se considera a los parámetros desconocidos (proporciones de votantes) como variables aleatorias.

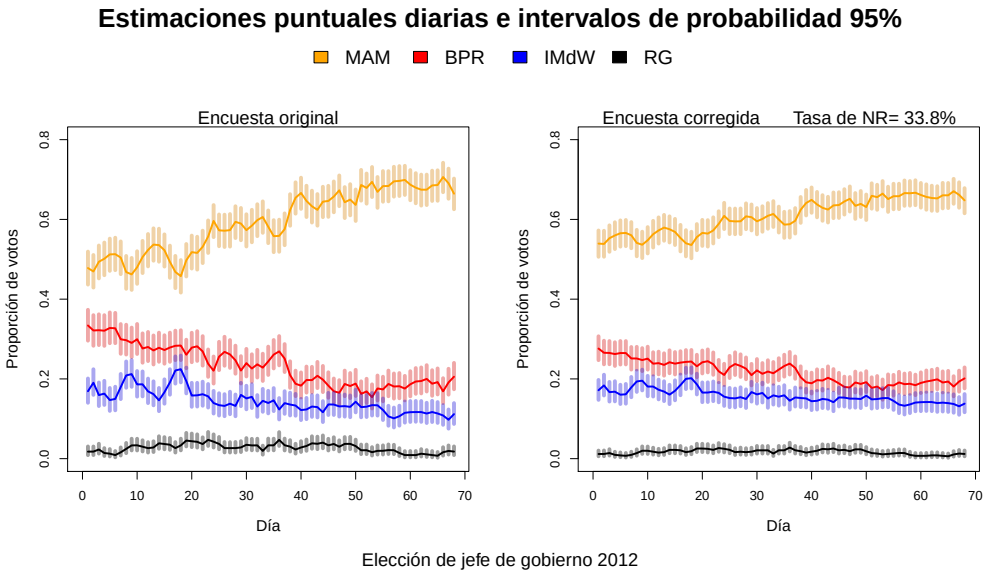

Figura 3.6: Estimaciones de las proporciones diarias por candidato para la elección de jefe de gobierno del Distrito Federal

Capítulo 4

Análisis de dependencias vía cópulas

El problema de estimar las preferencias electorales de una población generalmente es abordado desde el punto de vista marginal, es decir, se calculan estimaciones puntuales e intervalos de confianza o probabilidad para cada uno de los candidatos por separado, sin tomar en cuenta que a partir del modelo multinomial del que surgen (véase la Sección 3.1), los porcentajes o proporciones de preferencia entre candidatos guardan una relación o dependencia entre sí. Cuando no se consideran dichas dependencias pueden existir traslapes en los intervalos calculados y se escucha hablar de términos como *empate técnico*.

En este capítulo se analizarán de manera breve las dependencias bivariadas que existen tanto en la elección presidencial como en la de jefe de gobierno del Distrito Federal con el fin de identificar cómo y en qué medida un cambio en el porcentaje de preferencia de cierto candidato se refleja en cambios en los porcentajes de los candidatos restantes.

4.1. Pseudo-observaciones de las cópulas

Como se mencionó en la Sección 1.2, la teoría de cópulas se usa para modelar dependencias entre variables aleatorias continuas y los porcentajes (o proporciones) de preferencia de los candidatos caen en esa clasificación. Sin embargo, si se consideran los porcentajes publicados existen algunos valores

repetidos debido al redondeo de las cifras. Esto pudo ser corregido al hacer el ajuste del que se habló en la Sección 2.2, donde se repartía el porcentaje correspondiente a la no respuesta entre los candidatos válidos. Después de dicho ajuste se dio solución al problema de datos repetidos.

En primer lugar, podemos usar las pseudo-observaciones de la cópula subyacente a un par de variables aleatorias para darnos una idea gráfica acerca del comportamiento de la dependencia entre las mismas. La gráfica de las pseudo-observaciones es análoga a un histograma en el caso del análisis del comportamiento de una variable aleatoria.

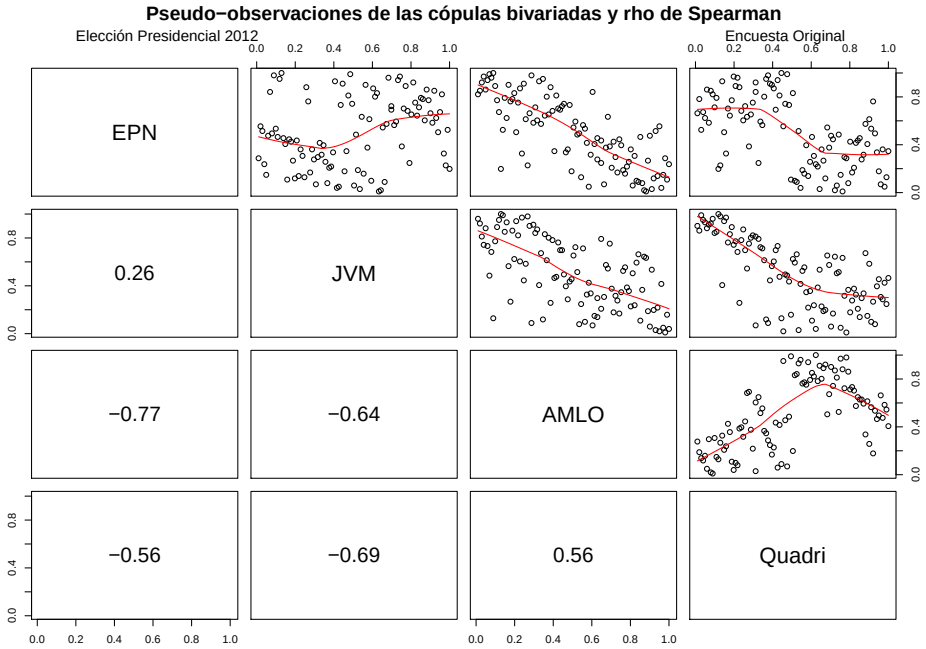

Figura 4.1: Pseudo-observaciones de las cópulas bivariadas para la elección presidencial (encuesta original)

En nuestro caso, las pseudo-observaciones de las cópulas bivariadas para cada par de candidatos en la elección presidencial tomando como datos los obtenidos en la encuesta de seguimiento diario de GEA/ISA se pueden observar en la Figura 4.1, y las correspondientes a la encuesta corregida se muestran en la Figura 4.2. En cada una de ellas, con un propósito descriptivo,

aparece en color rojo un suavizamiento de la tendencia de las dependencias obtenido mediante un método no paramétrico. Este suavizamiento será de gran ayuda para identificar dependencias no monótonas (ver Sección 4.2.2).

Recordemos que el estadístico muestral asociado a la rho de Spearman es un índice calculado a partir de la cópula empírica asociada a un par de variables aleatorias, y sirve para medir concordancia entre variables aleatorias. Los valores para dicho índice se muestran en el triángulo inferior de cada una de las gráficas anteriormente descritas (si se desea consultar el procedimiento de cálculo, véase el Anexo B.3).

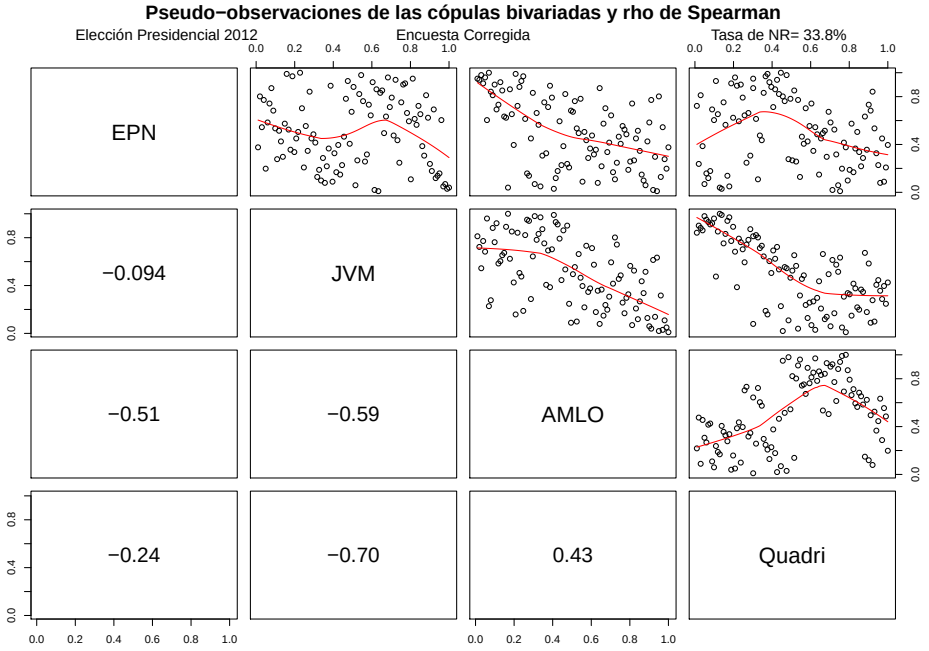

Figura 4.2: Pseudo-observaciones de las cópulas bivariadas para la elección presidencial (encuesta corregida)

Comparando las Figuras 4.1 y 4.2, podemos observar que entre la encuesta original y la encuesta corregida, para EPN-JVM, EPN-Quadri y AMLO-Quadri, existen diferencias considerables en los valores de la rho de Spearman de ambas encuestas, dichas diferencias también se observan en los suavizamientos que aparecen en rojo. Esto quiere decir que al realizar la eliminación

del sesgo, también se provocó un cambio en la estructura de las dependencias entre dichos candidatos.

Por otro lado, para la elección de jefe de gobierno, en la Figura 4.3 se muestran las pseudo-observaciones de las cópulas correspondientes a la encuesta original para cada par de candidatos, y en la Figura 4.4 para la encuesta corregida.

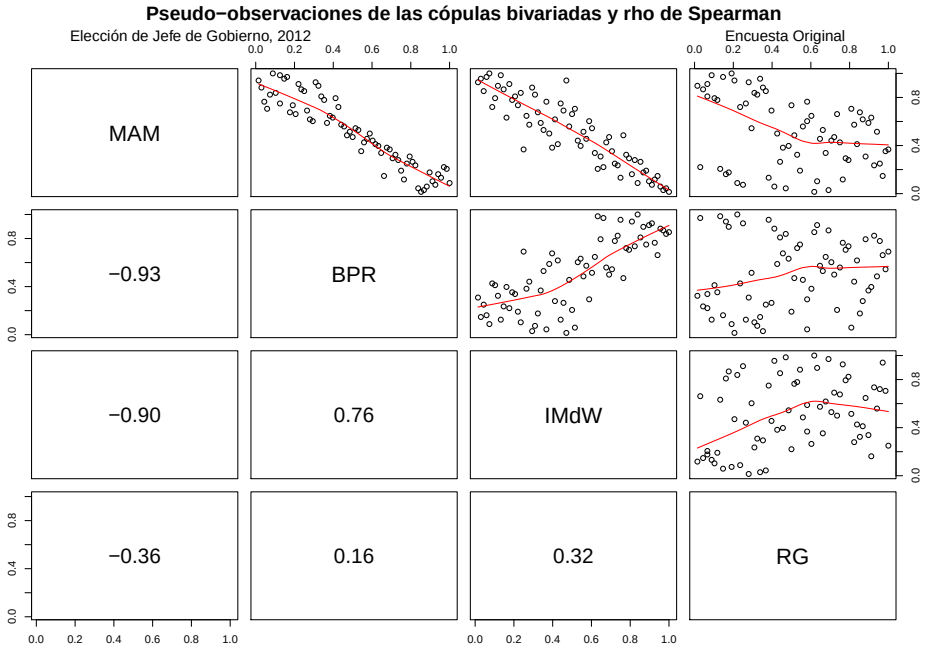

Figura 4.3: Pseudo-observaciones de las cópulas bivariadas para la elección de jefe de gobierno (encuesta original)

Al comparar ambos gráficos observamos que en este caso no existen diferencias relevantes en las rho de Spearman calculadas entre la encuesta original y la corregida. Además, analizando los suavizamientos de la tendencia de las dependencias, observamos que a excepción del par de candidatas IMdW-RG, en todos los casos parece haber dependencias monótonas (en la Sección 4.2 se profundizará el análisis de dichas tendencias). También a partir de los gráficos anteriores concluimos que las pseudo-observaciones de las cópulas bivariadas en las que aparece la candidata Rosario Guerra Díaz son las que tienen la

mayor dispersión, tanto en la encuesta original como en la encuesta corregida.

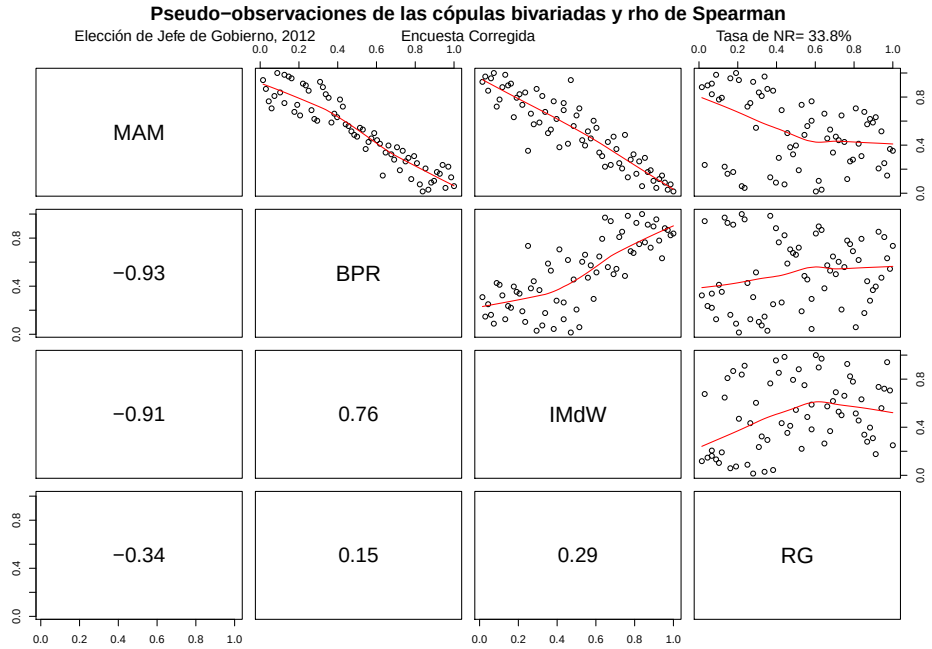

Figura 4.4: Pseudo-observaciones de las cópulas bivariadas para la elección de jefe de gobierno (encuesta corregida)

Como se dijo anteriormente, con los gráficos presentados en esta sección obtenemos un acercamiento al comportamiento conjunto del vector asociado a cada par de candidatos, sin embargo, necesitamos un índice para medir el grado de dichas dependencias y para ello también es importante saber si las dependencias son monótonas a lo largo del tiempo. Esto se analizará en la siguiente sección.

4.2. Monotonía de las dependencias

Para nuestro objeto de estudio, el que un par de candidatos presenten una dependencia estrictamente monótona entre sus porcentajes de preferencia se traduce en el signo y el grado del impacto que un aumento de un punto

porcentual de un candidato causaba en el otro.

Recordemos que Schweizer y Sklar (1983b) sugieren como medida de dependencia una distancia entre la cópula correspondiente respecto de la cópula que representa la independencia. En este trabajo usaremos como medida de dependencia la versión muestral de la sigma de Schweizer y Wolff ($\sigma_{X,Y}$), la cual será contrastada con la rho de Spearman ($\rho_{X,Y}$) calculada anteriormente, ya que, como se mencionó en la Sección 1.2.3, a medida que el indicador $\sigma_{X,Y} - |\rho_{X,Y}|$ se aleje de cero, detectaremos la existencia de dependencias no monótonas. En nuestro caso diremos que existen dependencias no monótonas a modelar si $\sigma_{X,Y} - |\rho_{X,Y}| \geq 0.05$.

En la Figura 1.1 de la Sección 1.2 se muestra la gráfica de la sección diagonal de las cotas superior e inferior de Fréchet y Hoeffding (δ_M, δ_W), así como la diagonal de la cópula producto (δ_Π), que representa a un par de variables aleatorias independientes. Como se estudió en la Sección 1.2.3, si un par de variables aleatorias presentan una dependencia estrictamente positiva, la gráfica de la diagonal empírica de la cópula subyacente a las mismas aparecerá por encima de δ_Π, de forma análoga, si existe una dependencia estrictamente negativa, la gráfica de la diagonal empírica irá por debajo de δ_Π. A continuación se clasifican las dependencias por pares de candidatos en monótonas y no monótonas usando el criterio antes mencionado.

4.2.1. Determinación de la monotonía

Las gráficas de las secciones diagonales de las cópulas empíricas correspondientes a cada par de candidatos para la elección presidencial, de acuerdo a la encuesta original, pueden apreciarse en la Figura 4.5. Por otro lado, para la encuesta corregida las diagonales empíricas se muestran en la Figura 4.6. A continuación se muestra un comparativo entre los resultados de ambas encuestas.

	Encuesta original				Encuesta corregida			
Candidatos (X,Y)	$\rho_{X,Y}$	$\sigma_{X,Y}$	$\sigma_{X,Y} - \lvert\rho_{X,Y}\rvert$	**Conclusión**	$\rho_{X,Y}$	$\sigma_{X,Y}$	$\sigma_{X,Y} - \lvert\rho_{X,Y}\rvert$	**Conclusión**
EPN, JVM	0.26	0.32	0.06	DNM	−0.09	0.22	0.13	DNM
EPN, AMLO	−0.77	0.77	0	Disc.	−0.51	0.51	0	Disc.
EPN, Quadri	−0.56	0.57	0.01	Disc.	−0.24	0.32	0.08	DNM
JVM, AMLO	−0.64	0.66	0.02	Disc.	−0.59	0.59	0	Disc.
JVM, Quadri	−0.69	0.69	0	Disc.	−0.7	0.7	0	Disc.
AMLO, Quadri	0.57	0.62	0.05	DNM	0.43	0.49	0.06	DNM

DNM: Dependencia no monótona
Disc.: Discordancia

Como se observa en el cuadro, entre los pares de candidatos EPN-JVM, EPN-Quadri en la encuesta corregida y AMLO-Quadri se encontraron dependencias no monótonas, las cuales se analizarán más a fondo en la Sección 4.2.2.

La dependencia más fuerte (basados en la sigma de Schweizer y Wolff) fue la discordancia entre EPN y AMLO, sin embargo, la discordancia entre JVM y AMLO fue de magnitud parecida. Además, se encontraron notables diferencias entre los resultados obtenidos en la encuesta original y los correspondientes a la encuesta corregida; en el caso de EPN y Quadri, por ejemplo, las magnitudes cambian a tal grado que la conclusión respecto a la dependencia es distinta para cada encuesta.

Por otro lado, en lo que respecta a la elección de jefe de gobierno del Distrito Federal, basados en el análisis de las diagonales empíricas y los cálculos de medidas de concordancia y dependencia de la encuesta original (Figura 4.7) y la encuesta corregida (Figura 4.8), se obtuvo lo siguiente:

	Encuesta original				Encuesta corregida			
Candidatos (X,Y)	$\rho_{X,Y}$	$\sigma_{X,Y}$	$\sigma_{X,Y} - \lvert\rho_{X,Y}\rvert$	**Conclusión**	$\rho_{X,Y}$	$\sigma_{X,Y}$	$\sigma_{X,Y} - \lvert\rho_{X,Y}\rvert$	**Conclusión**
MAM, BPR	−0.93	0.93	0	Disc.	−0.93	0.93	0	Disc.
MAM, IMdW	−0.9	0.9	0	Disc.	−0.91	0.91	0	Disc.
MAM, RG	−0.36	0.41	0.05	DNM	−0.34	0.4	0.06	DNM
BPR, IMdW	0.76	0.77	0	Conc.	0.76	0.77	0	Conc.
BPR, RG	0.17	0.31	0.14	DNM	0.15	0.29	0.14	DNM
IMdW, RG	0.32	0.36	0.04	Conc.	0.29	0.33	0.04	Conc.

DNM: Dependencia no monótona
Conc.: Concordancia
Disc.: Discordancia

En este caso, encontramos dependencias no monótonas entre MAM-RG, y BPR-RG, de la misma forma, dichas dependencias serán analizadas en la Sección 4.2.2.

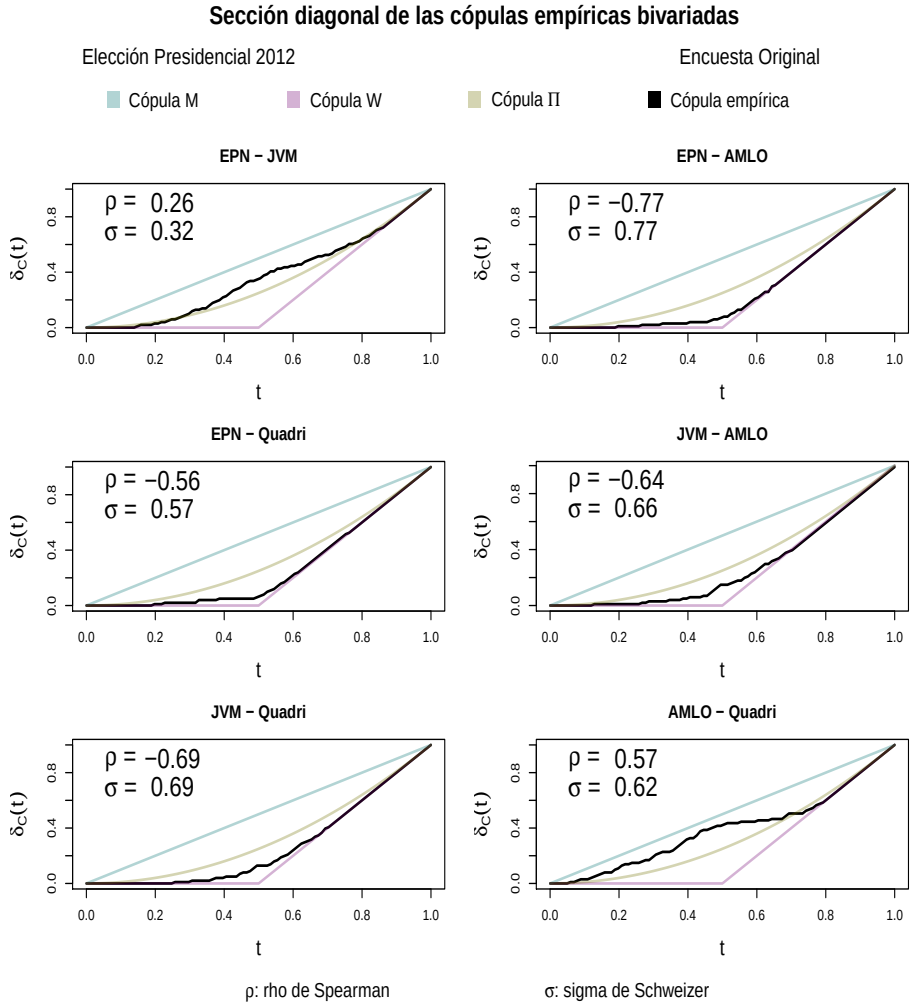

Figura 4.5: Sección diagonal de las cópulas empíricas para la elección presidencial (encuesta original)

A diferencia de los resultados obtenidos en la elección presidencial, la magnitud de las medidas de asociación y de dispersión en este caso no varía

considerablemente de la encuesta original a la corregida.

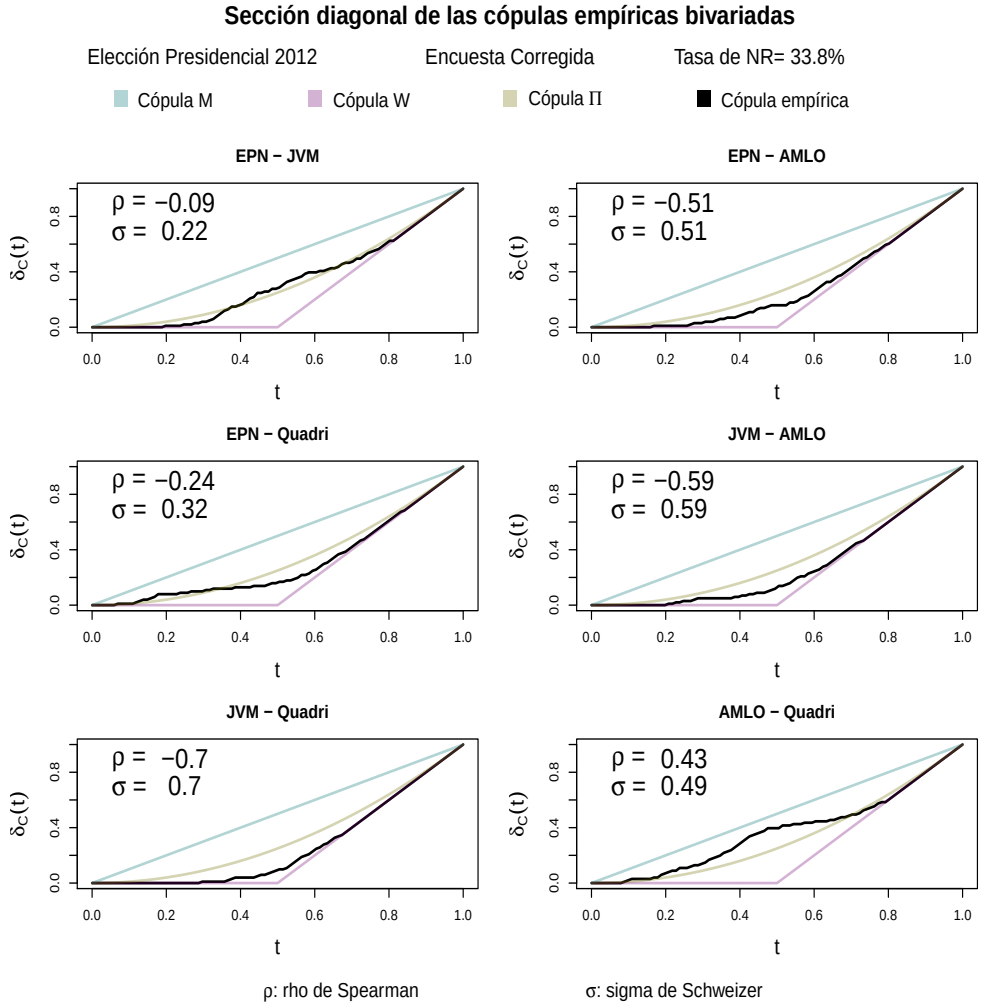

Figura 4.6: Sección diagonal de las cópulas empíricas para la elección presidencial (encuesta corregida)

El candidato MAM presentó dependencia estrictamente negativa y de magnitud muy alta (mayor a 0.9) con BPR e IMdW en ambas encuestas.

Por otro lado, BPR e IMdW reflejan una concordancia relativamente alta, y otro caso de concordancia se presenta entre IMdW y RG.

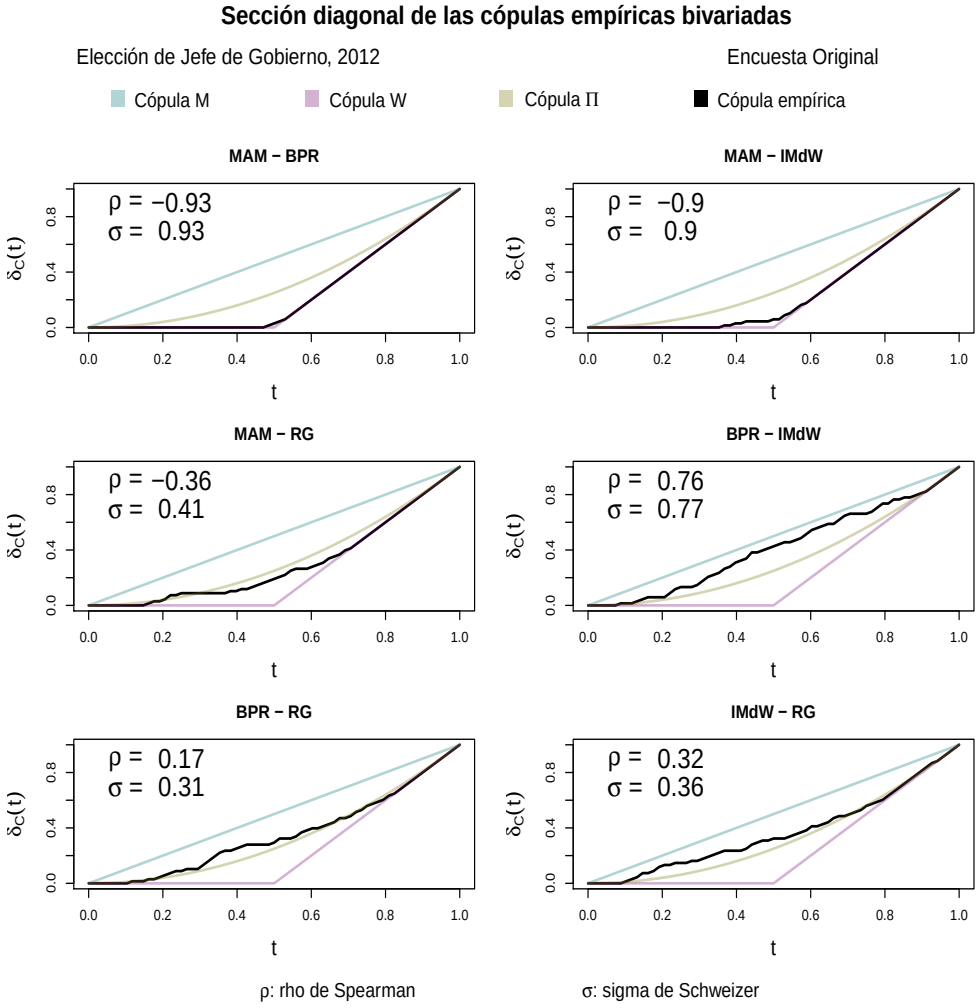

Figura 4.7: Sección diagonal de las cópulas empíricas para la elección de jefe de gobierno (encuesta original)

En los cuadros anteriores se resume la información acerca de la monotonía de las dependencias entre pares de candidatos observadas en las Figuras 4.5

a 4.8.

Obsérvese que cuando existen intersecciones de la diagonal empírica subyacente con la diagonal de la cópula producto se detectan dependencias no monótonas.

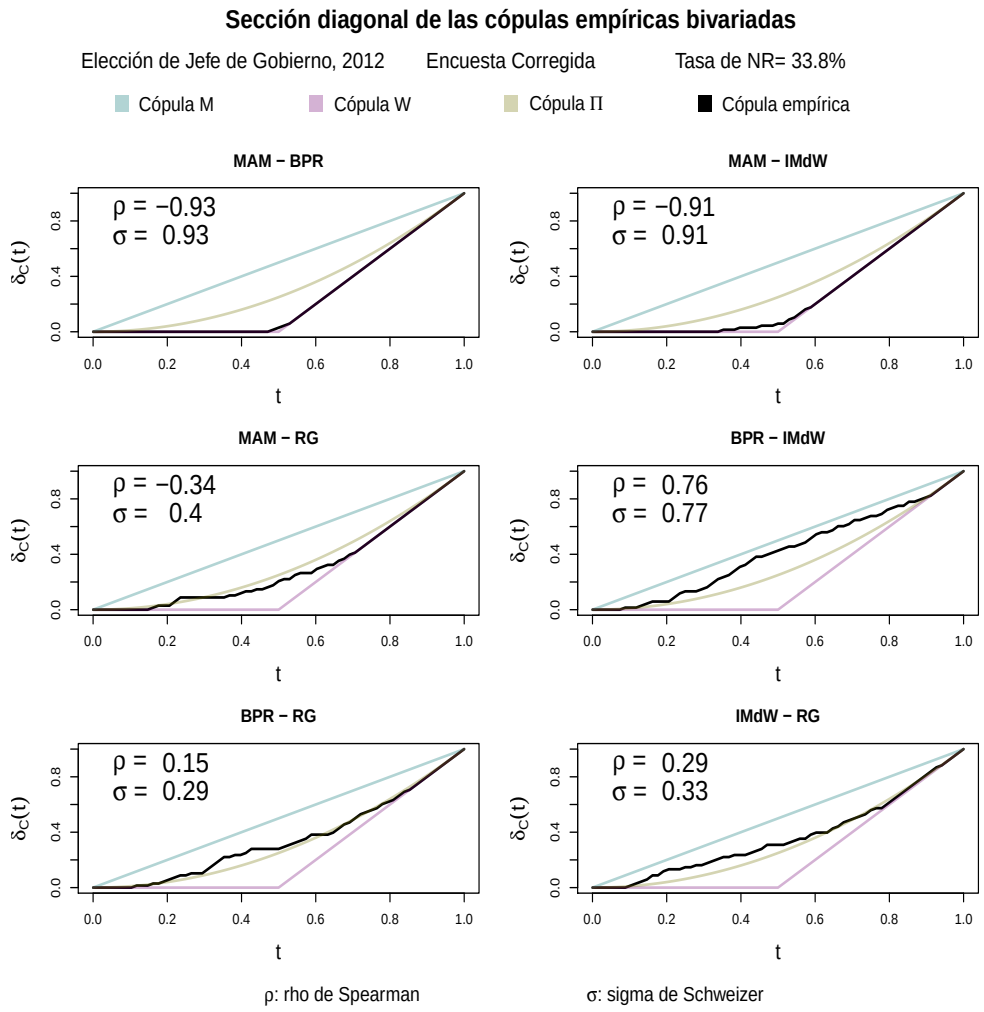

Figura 4.8: Sección diagonal de las cópulas empíricas para la elección de jefe de gobierno (encuesta corregida)

4.2.2. Tratamiento de las dependencias no monótonas

El objetivo de esta sección es obtener para los pares de candidatos en los que se identificó una dependencia no monótona (Sección 4.2), una partición de la muestra tal que la dependencia sea monótona en cada elemento de la misma, para posteriormente medir el grado de dependencia que se presentó en dichos subconjuntos. El algoritmo utilizado fue el siguiente:

1. En primer lugar, se identificaron los puntos de intersección de la diagonal empírica de la cópula subyacente con la diagonal de la cópula Π (*breakpoints*). Tales puntos de intersección serán los que determinen la partición de la muestra.

2. Posteriormente se realiza un mapeo de los puntos de la diagonal empírica para saber a qué día de encuesta corresponde cada uno de ellos.

3. Por último, se grafican los resultados de la segmentación con los registros ordenados por día para ver si existe un patrón temporal de la estructura de dependencia.

Si se desea un mayor detalle del procedimiento anterior, se pueden consultar las funciones *cambios.dependencias* y *dependencias.no.monotonas* del Anexo B.3 de este documento.

Para la dependencia entre los candidatos EPN y JVM, en la encuesta original se identificaron dos *breakpoints*, de donde se generan tres subconjuntos, dos de ellos correspondientes a una relación de discordancia (en los extremos de la diagonal empírica) y el tercero a concordancia. En la Figura 4.9 se presenta la gráfica de dicha estructura de dependencia (agrupando los subconjuntos donde se identificó discordancia), así como las versiones muestrales de la rho de Spearman y la sigma de Schweizer y Wolff calculadas para cada segmento. Al calcular las diferencias entre los indicadores antes mencionados, llegamos a la conclusión de que en uno de los subconjuntos ya existe una dependencia monótona (concordancia) del 53 %, sin embargo, en el otro subconjunto aún existe dependencia no monótona, por lo que se intentó particionar, a su vez, este subconjunto en otros menores obteniendo el mismo resultado que no es concluyente.

Debido a lo anterior, lo único que podemos afirmar respecto a la dependencia de EPN y JVM es que en un subconjunto del periodo de encuesta, existió una concordancia del 53 % entre ambos candidatos.

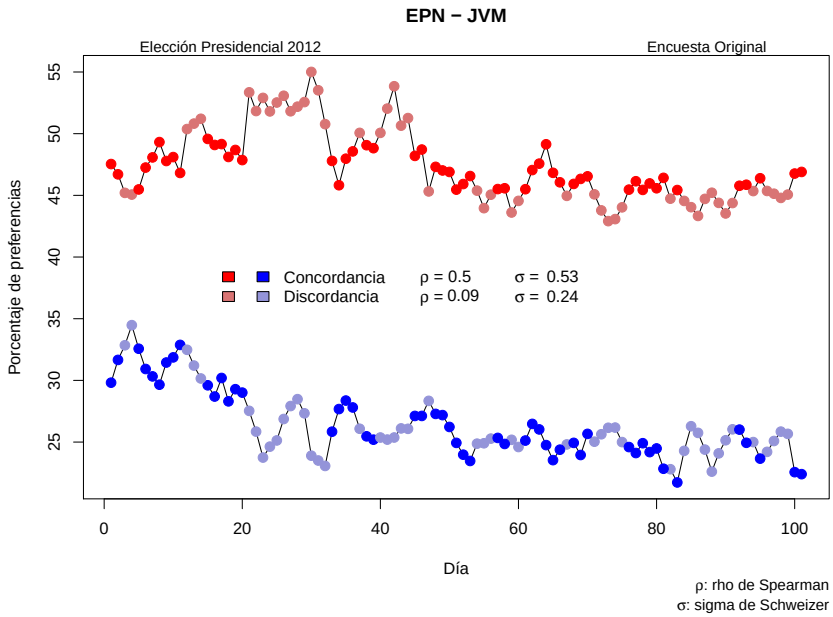

Figura 4.9: Segmentación de la estructura de dependencia entre EPN y JVM (encuesta original)

En cuanto a la encuesta corregida, para EPN y JVM se encontró el mismo problema anterior, pero esta vez para ambos subconjuntos (Figura 4.10), por lo que no se puede obtener una conclusión precisa acerca de la estructura de dependencia entre los porcentajes de preferencia de tales candidatos.

En la Sección 4.2 también se encontró que para la encuesta corregida, EPN y Quadri presentaban una dependencia no monótona. En este caso, al emplear el procedimiento descrito al inicio de este capítulo se logró obtener un grupo con dependencia estrictamente negativa (véase Figura 4.11), a partir del cual podemos decir que para un subconjunto del periodo de encuesta, si EPN subía un punto porcentual, esto afectaba a Quadri disminuyéndole un 0.45 porciento.

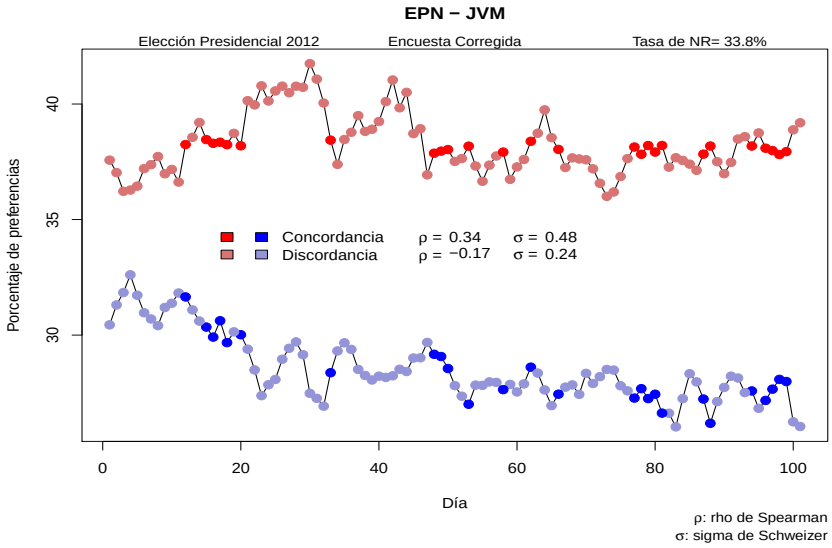

Figura 4.10: Segmentación de la estructura de dependencia entre EPN y JVM (encuesta corregida)

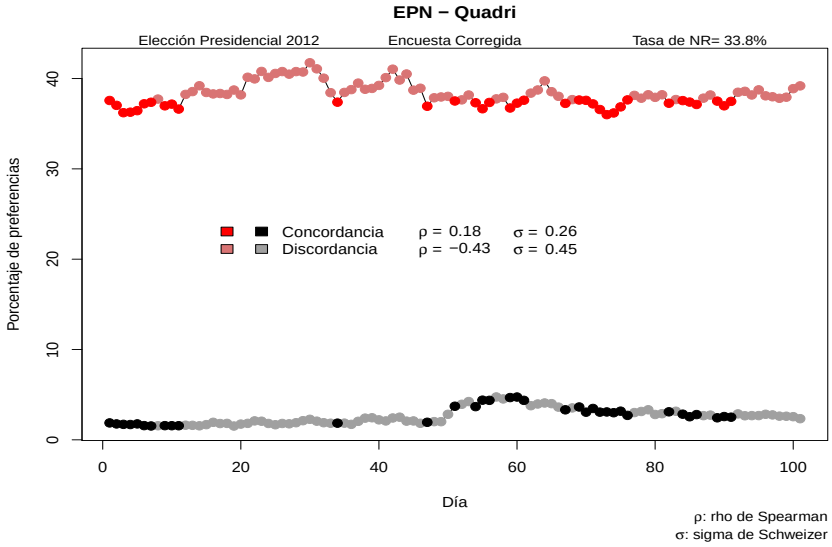

Figura 4.11: Segmentación de la estructura de dependencia entre EPN y Quadri (encuesta original)

Al particionar la muestra de la encuesta original para AMLO y Quadri, se encontró una dependencia estrictamente positiva que se presentó al principio del periodo de encuesta (días 1 al 65), como se observa en la Figura 4.12, de donde se interpreta que durante ese intervalo de tiempo una variación al porcentaje de preferencia de AMLO afectaba positivamente al porcentaje de Quadri en un 45 %.

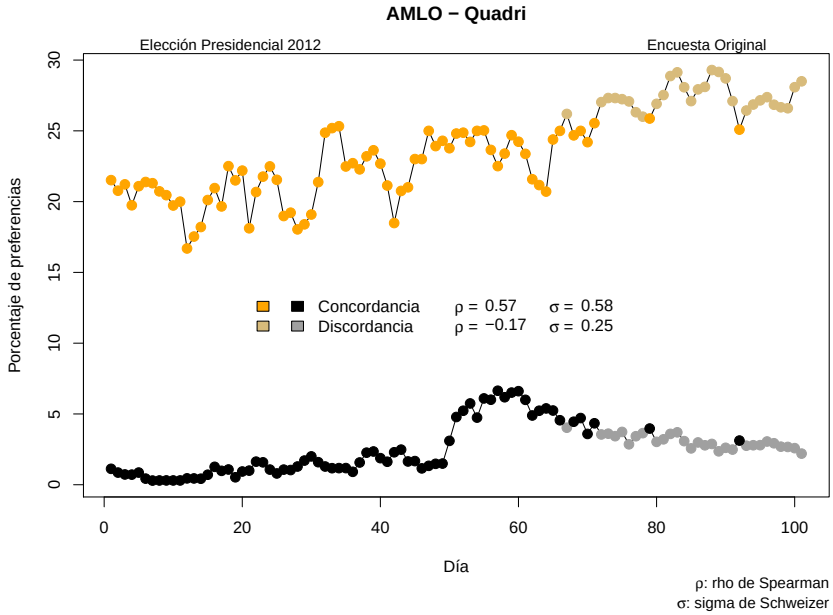

Figura 4.12: Segmentación de la estructura de dependencia entre AMLO y Quadri (encuesta original)

Para AMLO y Quadri en la encuesta corregida se encontró la misma conclusión que en la original, no obstante, el resultado de la magnitud de la dependencia positiva entre ambos candidatos en los días del 1 al 65 es del 37 % en este caso (Figura 4.13).

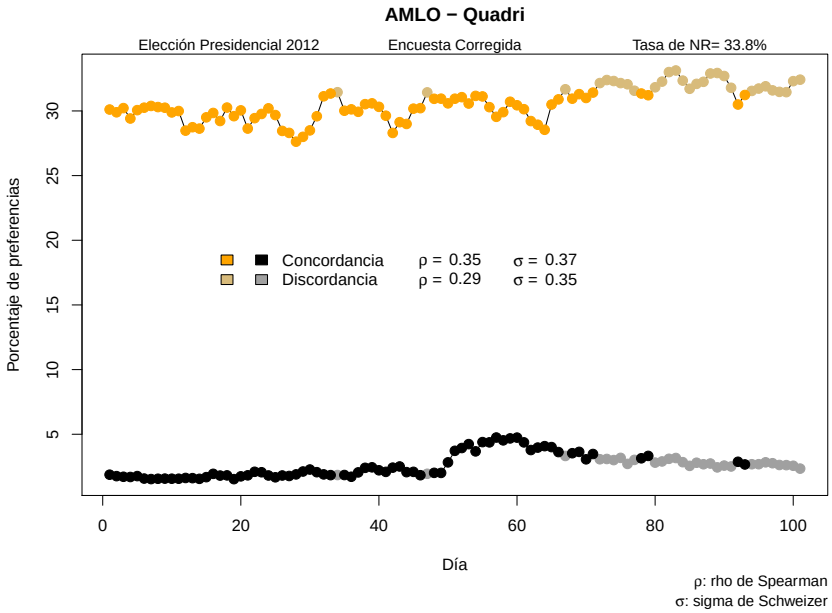

Figura 4.13: Segmentación de la estructura de dependencia entre AMLO y Quadri (encuesta corregida)

En cuanto a la elección de jefe de gobierno, el único caso para el cual se encontró dependencia no monótona fue entre las candidatas BPR e IMdW.

En la encuesta original, para BPR-IMdW se identifica un subconjunto donde la concordancia es la relación que impera en la estructura de dependencia de los porcentajes de ambas candidatas. Dicho subconjunto va del día de encuesta 1 al 44 y dentro del mismo se identifica una dependencia positiva del 49 % entre ambas candidatas (Figura 4.14). Este mismo resultado se obtuvo para la encuesta corregida, pero con una dependencia positiva del 48 % (Figura 4.15).

A modo de resumen, podemos decir que el objetivo de segmentar la muestra en subconjuntos con dependencias monótonas no se logró completamente en ninguno de los casos, sin embargo, en la mayoría de ellos pudimos encontrar un grupo donde la estructura de dependencia entre los pares de candidatos era estrictamente monótona, y así, al menos dentro de ese subconjunto resulta interpretable el indicador obtenido con la sigma de Schweizer y Wolff.

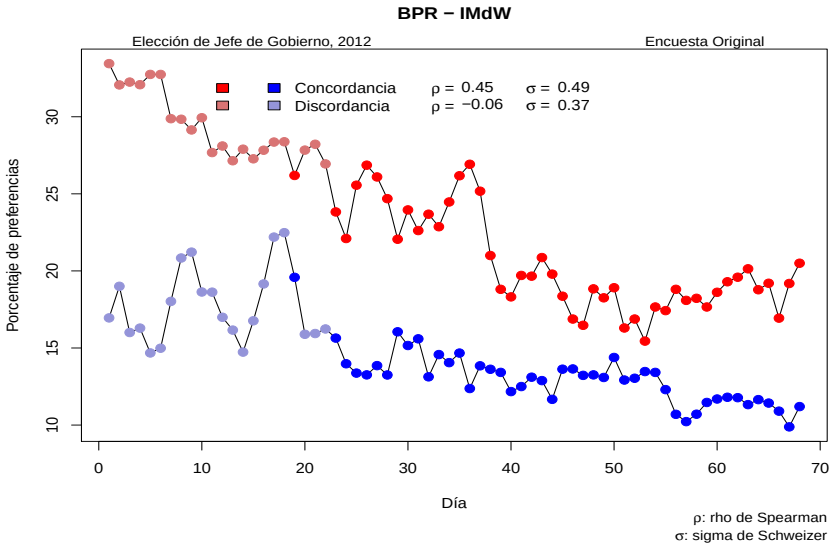

Figura 4.14: Segmentación de la estructura de dependencia entre BPR e IMdW (encuesta original)

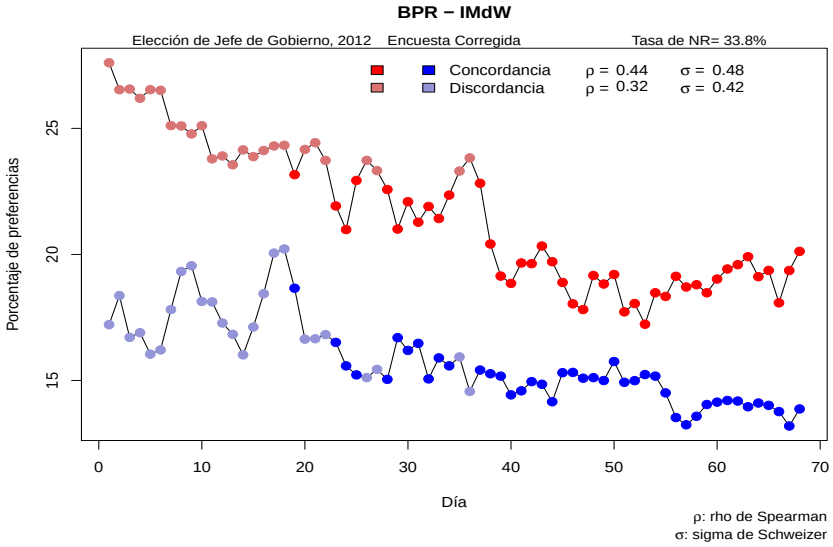

Figura 4.15: Segmentación de la estructura de dependencia entre BPR e IMdW (encuesta corregida)

Conclusiones

La motivación principal del análisis realizado en este trabajo fue el sesgo públicamente criticado de los resultados de la Encuesta de Seguimiento Diario Milenio GEA/ISA para la elección presidencial, el cual generó, además de especulaciones acerca de la confiabilidad de los datos publicados por la empresa encuestadora, la desacreditación pública del Exdirector General Adjunto del diario Milenio, Ciro Gómez Leyva. Con fines comparativos, se analizó de manera paralela a la encuesta para la elección presidencial, el comportamiento de los resultados de la encuesta para la elección de Jefe de Gobierno del Distrito Federal realizada por la misma encuestadora, en la cual sus estimaciones fueron bastante precisas.

Después de realizar un breve análisis descriptivo de los datos, basados en la hipótesis de que la no-respuesta pudo haber sido un factor determinante en el sesgo anteriormente mencionado, se procedió a corregir dicho sesgo proponiendo un algoritmo donde se estimaba el comportamiento de las preferencias electorales de este grupo, para generar una trayectoria insesgada a la que se denotó *encuesta corregida*. Tomando como porcentaje de no respuesta el 33.8 % publicado para GEA/ISA en Aristegui Noticias (2012), se llegó a la conclusión de que el sesgo sí puede ser explicado por la no-respuesta e incluso ésta debería ser mayor a la publicada para llegar a resultados congruentes.

Comparando los resultados publicados por Milenio GEA/ISA con la encuesta corregida, se estimaron los porcentajes de sobre/sub estimación por candidato y se encontró lo siguiente para la elección presidencial:

- El candidato Enrique Peña nieto estuvo sobreestimado durante todo el periodo de encuesta, llegando a alcanzar niveles de hasta el 13 % por encima del total de votos que obtuvo en el cómputo final.

- El candidato Andrés Manuel López Obrador fue subestimado durante toda la encuesta, llegando a estar hasta un 12 % por debajo de su cómputo final el día de la elección.

- Josefina Vázquez Mota estuvo subestimada a partir del día 18 de encuesta, pero a un nivel más bajo, el cual no rebasaba los cinco puntos porcentuales.

- Las estimaciones de Gabriel Quadri no estuvieron lejos del valor real que obtuvo en el cómputo final de la elección, sin embargo, presentó una ligera sobreestimación posterior al Segundo Debate Presidencial.

Inspirados por el trabajo de Nate Silver (O'Hara, 2012), con el propósito de monitorear las probabilidades de ganar de los candidatos y comparar dichas probabilidades de la encuesta original a la corregida, se procedió a modelar desde el enfoque bayesiano de la estadística la distribución a posteriori del vector de proporciones de votos de los candidatos, y se simularon diez mil observaciones de dicho vector de proporciones para cada día de encuesta.

A partir de las simulaciones de la distribución a posteriori se encontró que las probabilidades globales de ganar para los candidatos no cambiaban considerablemente de la encuesta original a la corregida. Más aún, la probabilidad de que ganara el candidato Enrique Peña Nieto era casi constante e igual a uno, con ligeras variaciones para la encuesta corregida, que no rebasaban el 3 %, por ende, las probabilidades de ganar de los demás candidatos no llegaron a estar por encima de ese 3 %.

Esto demuestra que el monitoreo de probabilidades de ganar por candidato es mucho más robusto que el de los porcentajes de preferencia, ya que los resultados no cambian considerablemente a pesar del sesgo. Además, según Silver, el interés más grande en una elección es saber quién será el candidato ganador, sin importar mucho la diferencia con la que derrote a los demás candidatos.

Adicional a las probabilidades globales de ganar, también fue posible calcular probabilidades de ganar por pares de candidatos, dentro de las cuales la única trayectoria no trivial resultó la de JVM-AMLO. La probabilidad de que la primera candidata le ganara al segundo presentó una alta variabilidad,

además de una tendencia decreciente que llevó a la probabilidad de alrededor de 1 durante los primeros dias de encuesta, a una cercana a cero durante los últimos días. También se concluye, comparando la encuesta original y la corregida, que el sesgo hacía que estas probabilidades fueran sobreestimadas.

Por último, en el capítulo 4, se realizó un análisis del comportamiento de las dependencias entre pares de candidatos tomando como referencia la encuesta original y la corregida. Para la elección presidencial se concluyó lo siguiente:

- Se encontraron diferencias considerables entre los indicadores de concordancia y dependencia de la encuesta original y la corregida, lo que indica que el sesgo de la encuesta se reflejaba en el comportamiento de las dependencias por pares de candidatos.
- Se identificaron dependencias estrictamente monótonas (y negativas) entre $EPN - AMLO$, $JVM - AMLO$ y $JVM - Quadri$; siendo todas ellas de magnitud mayor a 0.5.
- En los casos $EPN - JVM$, $EPN - Quadri$ y $AMLO - Quadri$ se concluyó la existencia de dependencias no monótonas, las cuales no fue posible descomponer en dependencias estrictamente monótonas luego de dos iteraciones.

A partir del desarrollo de este trabajo fue posible analizar a posteriori, desde el punto de vista estadístico, el monitoreo de las preferencias electorales para la elección presidencial de la encuesta de seguimiento diario Milenio GEA/ISA, sin embargo, los resultados obtenidos aquí pueden servir de base para un posterior análisis interpretativo dentro del ámbito de la política.

Apéndice A

Cuadros auxiliares

	Fecha	EPN	JVM	AMLO	Quadri	Indecisos	Tamaño de muestra
1	19 /03	33.80	21.20	15.30	0.80	28.90	1146
2	20 /03	32.60	22.10	14.50	0.60	30.10	1147
3	21 /03	31.10	22.60	14.60	0.50	31.20	1134
4	22 /03	31.50	24.10	13.80	0.50	30.10	1150
5	23 /03	31.70	22.70	14.70	0.60	30.30	1152
6	24 /03	32.70	21.40	14.80	0.30	30.80	1152
7	25 /03	32.50	20.50	14.40	0.20	32.40	1152
8	26 /03	32.60	19.60	13.70	0.20	33.90	1152
9	27 /03	31.30	20.60	13.40	0.20	34.50	1144
10	28 /03	31.70	21.00	13.00	0.20	34.10	1112
11	29 /03	30.90	21.70	13.20	0.20	34.00	1152
12	30 /03	33.50	21.60	11.10	0.30	33.50	1150
13	31 /03	34.20	21.00	11.80	0.30	32.70	1152
14	1 /04	36.00	21.20	12.80	0.30	29.50	1127
15	2 /04	35.00	20.90	14.20	0.50	29.40	1152
16	3 /04	34.90	20.40	14.90	0.90	28.90	1119
17	4 /04	35.00	21.50	14.00	0.70	28.80	1152
18	5 /04	35.70	21.00	16.70	0.80	25.80	1152
19	6 /04	36.90	22.20	16.30	0.40	24.20	1152
20	7 /04	35.80	21.70	16.60	0.70	25.20	1152
21	8 /04	37.40	19.30	12.70	0.70	29.90	1152

Cuadro A.1: Resultados para la elección presidencial, parte 1

	Fecha	EPN	JVM	AMLO	Quadri	Indecisos	Tamaño de muestra
22	9 /04	38.10	19.00	15.20	1.20	26.50	1152
23	10 /04	40.10	18.00	16.50	1.20	24.70	1152
24	11 /04	38.70	18.40	16.80	0.80	25.30	1152
25	12 /04	39.50	18.90	16.20	0.60	24.80	1152
26	13 /04	39.70	20.10	14.20	0.80	25.20	1152
27	14 /04	39.90	21.50	14.80	0.80	23.00	1152
28	15 /04	40.50	22.10	14.00	1.00	22.40	1148
29	16 /04	40.00	20.80	14.00	1.30	23.90	1148
30	17 /04	41.20	17.90	14.30	1.50	25.10	1148
31	18 /04	40.30	17.70	16.10	1.20	24.70	1152
32	19 /04	39.40	17.90	19.30	1.00	22.40	1152
33	20 /04	36.80	19.90	19.40	0.90	23.00	1152
34	21 /04	35.10	21.20	19.40	0.90	23.40	1145
35	22 /04	36.70	21.70	17.20	0.90	23.50	1152
36	23 /04	37.20	21.30	17.40	0.70	23.40	1152
37	24 /04	38.20	19.90	17.00	1.20	23.70	1152
38	25 /04	36.80	19.10	17.40	1.70	25.00	1152
39	26 /04	37.40	19.30	18.10	1.80	23.40	1152
40	27 /04	37.30	18.90	16.90	1.40	25.50	1152
41	28 /04	38.40	18.60	15.60	1.20	26.20	1147
42	29 /04	39.90	18.80	13.70	1.70	25.90	1147
43	30 /04	38.80	20.00	15.90	1.90	23.40	1147
44	1 /05	40.50	20.60	16.60	1.30	21.00	1152
45	2 /05	37.50	21.10	17.90	1.30	22.50	1152
46	3 /05	37.90	21.10	17.90	0.90	23.00	1152
47	4 /05	33.90	21.20	18.70	1.00	25.20	1152
48	5 /05	35.20	20.30	17.80	1.10	25.60	1152
49	6 /05	35.80	20.70	18.50	1.14	23.60	1152
50	7 /05	36.30	20.30	18.40	2.40	22.60	1152
51	8 /05	36.10	19.80	19.70	3.80	20.60	1152
52	9 /05	36.00	18.80	19.50	4.10	21.60	1152
53	10 /05	37.30	18.80	19.40	4.60	19.90	1152
54	11 /05	35.40	19.40	19.50	3.70	22.00	1152
55	12 /05	34.60	19.60	19.70	4.80	21.30	1152
56	13 /05	36.00	20.20	18.90	4.80	20.10	1152

Cuadro A.2: Resultados para la elección presidencial, parte 2

	Fecha	EPN	JVM	AMLO	Quadri	Indecisos	Tamaño de muestra
57	14 /05	37.00	20.60	18.30	5.40	18.70	1144
58	15 /05	37.60	20.50	19.30	5.10	17.50	1144
59	16 /05	35.50	20.50	20.10	5.30	18.60	1144
60	17 /05	36.40	20.10	19.80	5.40	18.30	1152
61	18 /05	36.40	20.10	18.70	4.80	20.00	1152
62	19 /05	37.50	21.10	17.20	3.90	20.30	1152
63	20 /05	38.20	20.90	17.00	4.20	19.70	1152
64	21 /05	40.10	20.20	16.90	4.40	18.40	1152
65	22 /05	38.40	19.30	20.00	4.30	18.00	1152
66	23 /05	37.40	19.80	20.30	3.70	18.80	1152
67	24 /05	35.70	19.70	20.80	3.20	20.60	1152
68	25 /05	36.10	19.60	19.40	3.50	21.40	1148
69	26 /05	35.40	18.30	19.10	3.60	23.60	1148
70	27 /05	35.00	19.30	18.20	2.70	24.80	1148
71	28 /05	35.30	19.60	20.00	3.40	21.70	1152
72	29 /05	34.50	20.20	21.30	2.80	21.20	1152
73	30 /05	33.30	20.30	21.20	2.80	22.40	1152
74	31 /05	33.90	20.60	21.50	2.70	21.30	1152
75	1 /06	35.40	20.10	21.90	3.00	19.60	1152
76	2 /06	36.60	19.80	21.80	2.30	19.50	1152
77	3 /06	37.70	19.70	21.50	2.80	18.30	1152
78	4 /06	37.40	20.50	21.40	3.00	17.70	1152
79	5 /06	38.20	20.10	21.50	3.30	16.90	1152
80	6 /06	37.60	20.20	22.20	2.50	17.50	1152
81	7 /06	37.60	18.50	22.30	2.60	19.00	1152
82	8 /06	36.10	18.40	23.30	2.90	19.30	1152
83	9 /06	36.80	17.60	23.60	3.00	19.00	1152
84	10 /06	37.60	20.50	23.70	2.60	15.60	1152
85	11 /06	37.70	22.50	23.20	2.20	14.40	1152
86	12 /06	37.70	22.40	24.30	2.60	13.00	1152
87	13 /06	38.50	21.00	24.20	2.40	13.90	1152
88	14 /06	39.20	19.60	25.40	2.50	13.30	1152
89	15 /06	37.60	20.40	24.70	2.00	15.30	1152
90	16 /06	36.70	21.20	24.20	2.20	15.70	1152
91	17 /06	37.50	22.00	22.90	2.10	15.50	1152

Cuadro A.3: Resultados para la elección presidencial, parte 3

	Fecha	EPN	JVM	AMLO	Quadri	Indecisos	Tamaño de muestra
92	18 /06	39.60	22.50	21.70	2.70	13.50	1152
93	19 /06	39.90	21.70	23.00	2.40	13.00	1152
94	20 /06	39.00	21.50	23.10	2.40	14.00	1150
95	21 /06	39.80	20.30	23.30	2.40	14.20	1150
96	22 /06	38.60	20.60	23.30	2.60	14.90	1150
97	23 /06	38.50	21.40	22.90	2.50	14.70	1152
98	24 /06	38.30	22.10	22.80	2.30	14.50	1152
99	25 /06	38.80	22.10	22.90	2.30	14.90	1152
100	26 /06	39.80	19.20	23.90	2.20	14.90	1152
101	27 /06	46.90	22.40	28.50	2.20		

Cuadro A.4: Resultados para la elección presidencial, parte 4

	Fecha	MAM	BPR	IMdW	RG	Indecisos	Tamaño de muestra
1	16 /04	41.70	29.20	14.80	1.60	12.70	576
2	17 /04	39.60	27.00	16.00	1.60	16.00	574
3	18 /04	41.10	26.80	13.30	1.90	16.90	572
4	19 /04	40.00	25.60	13.00	1.20	20.20	572
5	20 /04	40.80	26.10	11.70	1.10	20.30	574
6	21 /04	40.70	26.00	11.90	0.80	20.60	571
7	22 /04	40.90	24.20	14.60	1.30	19.00	576
8	23 /04	38.00	24.20	16.90	2.00	18.90	576
9	24 /04	37.90	23.90	17.40	2.80	18.00	576
10	25 /04	38.70	24.10	15.00	2.70	19.50	576
11	26 /04	39.70	21.70	14.60	2.40	21.60	576
12	27 /04	40.00	21.50	13.00	2.00	23.50	576
13	28 /04	42.60	21.50	12.80	2.30	20.80	576
14	29 /04	44.30	23.10	12.20	3.20	17.20	576
15	30 /04	44.30	23.10	14.20	3.10	15.80	576
16	1 /05	41.00	23.10	15.90	3.00	17.00	576
17	2 /05	37.90	23.00	18.00	2.20	18.90	576
18	3 /05	37.30	23.10	18.30	2.70	18.60	576
19	4 /05	40.60	21.40	16.00	3.70	18.30	576
20	5 /05	43.40	23.30	13.30	3.70	16.30	576
21	6 /05	43.70	23.90	13.50	3.60	15.30	572

Cuadro A.5: Resultados para la elección de jefe de gobierno, parte 1

	Fecha	MAM	BPR	IMdW	RG	Indecisos	Tamaño de muestra
22	7 /05	45.20	22.90	13.80	3.10	15.00	576
23	8 /05	46.40	19.80	13.00	3.90	16.90	576
24	9 /05	49.90	18.50	11.70	3.60	16.30	568
25	10 /05	48.40	21.60	11.30	3.20	15.50	576
26	11 /05	49.20	23.10	11.40	2.30	14.00	576
27	12 /05	49.60	22.60	12.00	2.40	13.40	576
28	13 /05	52.50	21.80	11.70	2.30	11.70	576
29	14 /05	51.10	19.10	13.90	2.50	13.40	576
30	15 /05	49.60	20.70	13.10	3.00	13.60	576
31	16 /05	49.00	19.00	13.10	2.90	16.00	576
32	17 /05	51.00	20.20	11.20	2.90	14.70	576
33	18 /05	51.10	19.30	12.30	1.70	15.60	576
34	19 /05	49.60	20.90	12.00	2.90	14.60	576
35	20 /05	47.50	22.30	12.50	2.90	14.80	576
36	21 /05	48.90	23.50	10.80	4.10	12.70	576
37	22 /05	50.30	22.00	12.10	3.00	12.60	571
38	23 /05	54.90	18.50	12.00	2.70	11.90	571
39	24 /05	57.00	16.40	11.70	2.10	12.80	571
40	25 /05	58.60	16.10	10.70	2.50	12.10	576
41	26 /05	57.30	17.50	11.10	2.90	11.20	576
42	27 /05	56.00	17.40	11.60	3.50	11.50	576
43	28 /05	54.80	18.30	11.30	3.30	12.30	568
44	29 /05	56.30	17.30	10.20	3.60	12.60	576
45	30 /05	55.90	15.90	11.80	3.00	13.40	576
46	31 /05	56.90	14.60	11.80	3.20	13.50	576
47	1 /06	57.90	14.20	11.40	2.70	13.80	573
48	2 /06	55.10	16.20	11.40	3.30	14.00	573
49	3 /06	56.50	15.90	11.40	3.30	13.40	573
50	4 /06	54.80	16.30	12.40	2.70	13.80	576
51	5 /06	58.90	14.00	11.10	1.90	14.10	576
52	6 /06	58.40	14.50	11.20	1.80	14.10	576
53	7 /06	59.70	13.30	11.60	1.50	13.90	568
54	8 /06	58.40	15.40	11.70	1.70	12.80	568
55	9 /06	59.90	15.30	10.80	1.80	12.20	568

Cuadro A.6: Resultados para la elección de jefe de gobierno, parte 2

	Fecha	MAM	BPR	IMdW	RG	Indecisos	Tamaño de muestra
56	10 /06	60.60	16.70	9.50	2.00	11.20	568
57	11 /06	61.90	16.10	9.10	1.90	11.00	576
58	12 /06	61.10	16.00	9.40	1.30	12.20	576
59	13 /06	60.90	15.40	10.00	0.90	12.80	576
60	14 /06	60.50	16.40	10.30	0.90	11.90	576
61	15 /06	59.80	17.00	10.40	0.90	11.90	576
62	16 /06	59.50	17.30	10.40	1.10	11.70	576
63	17 /06	58.90	17.60	9.90	1.00	12.60	576
64	18 /06	60.60	16.60	10.30	0.90	11.60	576
65	19 /06	60.00	16.80	10.00	0.70	12.50	576
66	20 /06	60.80	14.60	9.40	1.40	13.80	576
67	21 /06	59.30	16.50	8.50	1.70	14.00	576
68	22 /06	66.40	20.50	11.20	1.90		

Cuadro A.7: Resultados para la elección de jefe de gobierno, parte 3

Apéndice B

Código de programación en R

B.1. Estadística descriptiva y corrección de sesgo

Datos

```
# Encuesta.EP > Dataframe de 101x7 donde:
#
#     Columna 1: fecha: Fecha de publicación del resultado de cada registro.
#
#     Columna 2: EPN: Porcentaje de votos pronosticado para el candidato Enrique Peña Nieto
#       en la elección presidencial.
#
#     Columna 3: JVM: Porcentaje de votos pronosticado para la candidata Josefina Eugenia
#       Vázquez Mota en la elección presidencial.
#
#     Columna 4: AMLO: Porcentaje de votos pronosticado para el candidato Andrés Manuel
#       López Obrador en la  elección presidencial.
#
#     Columna 5: Quadri: Porcentaje de votos pronosticado para el candidato Gabriel Ricardo
#       Quadri de la Torre en la elección presidencial.
#
#     Columna 6: Indecisos: Porcentaje que representa a las personas que contestaron la
#       encuesta pero se declararon indecisas. (NA en el último registro)
#
#     Columna 7: Tamaño de muestra: Publicado en el periódico Milenio cada día de
#       encuesta.(NA en el último registro)
#

# Encuesta.JG > Dataframe de 68x7 donde:
#
#     Columna 1: fechaDF: Fecha de publicación del resultado de cada registro.
#
#     Columna 2: MAM: Porcentaje de votos pronosticado para  el candidato Miguel Ángel
#       Mancera en la elección para jefe de gobierno.
```

```
#
#     Columna 3: BPR: Porcentaje de votos pronosticado para la candidata Beatriz Paredes
#       Rangel en la elección para jefe de gobierno.
#
#     Columna 4: IMdW: Porcentaje de votos pronosticado para la candidata Isabel Miranda de
#       Wallace en la elección para jefe de gobierno.
#
#     Columna 5: RG: Porcentaje de votos pronosticado para la candidata Rosario Guerra Díaz
#       en la elección para jefe de gobierno.
#
#     Columna 6: Indecisos: Porcentaje que representa a las personas que contestaron la
#       encuesta pero se declararon indecisas. (NA en el último registro)
#
#     Columna 7: Tamaño de muestra: Publicado en el periódico Milenio cada día de encuesta.
#       (NA en el último registro)
#

# CompFinal.EP > Dataframe de 1x4 que contiene los resultados oficiales obtenidos del "ACTA
#     DE CALIFICACIÓN JURISDICCIONAL DEL CÓMPUTO FINAL" emitida por el Tribunal Electoral
#     del Poder Judicial de la Federación, donde:
#
#     Columna 1: EPN:  Número de votos obtenidos por el candidato Enrique Peña Nieto.
#
#     Columna 2: JVM: Número de votos obtenidos por la candidata Josefina Eugenia Vázquez
#       Mota.
#
#     Columna 3: AMLO: Número de votos obtenidos por el candidato Andrés Manuel López
#       Obrador.
#
#     Columna 4: Quadri: Número de votos obtenidos por el candidato Gabriel Ricardo Quadri
#       de la Torre.
#

# CompFinal.JG > Dataframe de 1x4 que contiene los resultados  oficiales obtenidos del
#     "ACTA DE CÓMPUTO  TOTAL DE LA ELECCIÓN DE JEFE DE GOBIERNO DEL DISTRITO FEDERAL 2012"
#     publicada por el Instituto Electoral del Distrito Federal, donde:
#
#     Columna 1: MAM:  Número de votos obtenidos por el candidato Miguel Ángel Mancera.
#
#     Columna 2: BPR: Número de votos obtenidos por la candidata Beatriz Paredes Rangel.
#
#     Columna 3: IMdW: Número de votos obtenidos por la candidata Isabel Miranda de
#       Wallace.
#
#     Columna 4: RG: Número de votos obtenidos por la candidata Rosario Guerra Díaz.
#
```

Conversión de la encuesta a número de votos

```
votos <- function(Encuesta,CompFinal,PorcNoResp){
# Input:
#
#     Encuesta > Encuesta.EP o Encuesta.JG como se describieron en Datos.
#
#     CompFinal > CompFinal.EP o CompFinal.JG como se describieron en Datos.
#
```

```
#     PorcNoResp > Número entre 0 y 100 que indique el porcentaje de no respuesta que se
#       contemplará para hacer los cálculos.
#
# Output:
#
#     Dataframe de n*9 (n=101 para la elección presidencial y n=68 para la elección de jefe
#     de gobierno) donde:
#
#     Columna 1: fecha: Fecha de publicación del resultado de cada registro.
#
#     Columna 2: EPN/MAM:  Número de votos pronosticado para el candidato Enrique Peña
#       Nieto/ Miguel Ángel Mancera en la elección correspondiente.
#
#     Columna 3: JVM/BPR: Número de votos pronosticado candidata Josefina Eugenia Vázquez
#       Mota/ Beatriz Paredes Rangel en la elección correspondiente.
#
#     Columna 4: AMLO/IMdW: Número de votos pronosticado para Andrés Manuel López Obrador/
#       Isabel Miranda de Wallace en la elección correspondiente.
#
#     Columna 5: Quadri/RG: Número de votos pronosticado para Gabriel Ricardo Quadri de la
#       Torre/ Rosario Guerra Díaz en la elección correspondiente.
#
#     Columna 6: Indecisos: Número de votos que representan las personas que contestaron la
#       encuesta pero se declararon indecisas.
#
#     Columna 7: Rechazadas: Número de votos estimados que representan las personas que se
#       negaron a responder la encuesta.
#
#     Columna 8: Indecisos+Rechazadas: Suma de las columnas 6 y 7.
#
#     Columna 9: Tamaño de muestra: Modificado considerando el (PorcNoResp)% de no
#       respuesta.
#
####
#
# Estimación del porcentaje de indecisos y tamaño de muestra del último registro tomando en
# cuenta el dato inmediato anterior
#
     n<-dim(Encuesta)[1]
     Encuesta[n,6]<-Encuesta[n-1,6]
     Encuesta[n,2:6]<-Encuesta[n,2:6]/sum(Encuesta[n,2:6])
     Encuesta[n,7]<-Encuesta[n-1,7]
#
# Corregir registros donde la suma de porcentajes difiere del 100%
#
     k<-which(apply(Encuesta[,2:6],1,sum)!=100)
     Encuesta[,2:6]<-Encuesta[,2:6]/100
     for(i in k) Encuesta[i,2:6]<-Encuesta[i,2:6]/sum(Encuesta[i,2:6])
#
# Calcular el número de encuestas efectivas por candidato
# y añadir a los que no respondieron.
#
# Num encuestas rechazadas = Tamaño de muestra/(1-Porc. de no resp.)*Porc. de no resp.
#
     Encuesta[,2:6]<-Encuesta[,2:6]*Encuesta[,7]
     Encuesta<-cbind(Encuesta,Encuesta[,7]/(1-PorcNoResp/100)*PorcNoResp/100)
     names(Encuesta)[8]<-"Rechazadas"
```

```
     Encuesta[,7]<-Encuesta[,7]/(1-PorcNoResp/100)
#
# Reordenar columnas y agregar una nueva con la suma de Indecisos + Encuestas Rechazadas
#
     Encuesta<-cbind(Encuesta,Encuesta[,"Indecisos"]+Encuesta[,"Rechazadas"])
     E<-Encuesta
     Encuesta[,7]<-E[,8] ; names(Encuesta)[7]<-names(E)[8]
     Encuesta[,8]<-E[,9] ; names(Encuesta)[8]<-"Indecisos+Rechazadas"
     Encuesta[,9]<-E[,7] ; names(Encuesta)[9]<-names(E)[7]
#
# Convertir a número de votos
#
     Encuesta[,2:7]<-Encuesta[,2:7]/apply(Encuesta[,2:7],1,sum)
     Encuesta[,"Indecisos+Rechazadas"]<-Encuesta[,"Indecisos"]+Encuesta[,"Rechazadas"]
     Encuesta[,2:8]<-Encuesta[,2:8]*sum(CompFinal)
     return(Encuesta)
}
```

Estimación de las preferencias de la no respuesta

```
desglose <- function(Encuesta, CompFinal, PorcNoResp, met, porc=F){
# Input:
#
#     Encuesta > Encuesta.EP o Encuesta.JG como se describieron en Datos.
#
#     CompFinal > CompFinal.EP o CompFinal.JG como se describieron en Datos.
#
#     PorcNoResp > Número entre 0 y 100 que indique el porcentaje de no respuesta que se
#       contemplará para hacer los cálculos.
#
#     met > Método a emplear para corregir el sesgo
#       met=1 Diferencias con el cómputo final
#       met=2 Porcentaje sistemático constante
#
#     porc > Valor lógico, FALSE por default. Si porc=T, el dataframe obtenido presentará
#       los porcentajes de votos por rubro, porc=F presenta el  número absoluto de votos
#       por rubro.
#
# Output:
#
#     Dataframe de nx11 donde:
#
#     Columna 1: fecha: Fecha de publicación del resultado de cada registro.
#
#     Columna 2: EPN/MAM:  Número de votos pronosticado para el candidato Enrique Peña
#       Nieto/ Miguel Ángel Mancera en la elección correspondiente.
#
#     Columna 3: JVM/BPR: Número de votos pronosticado candidata Josefina Eugenia Vázquez
#       Mota/ Beatriz Paredes Rangel en la elección correspondiente.
#
#     Columna 4: AMLO/IMdW: Número de votos pronosticado para Andrés Manuel López Obrador/
#       Isabel Miranda de Wallace en la elección correspondiente.
#
#     Columna 5: Quadri/RG: Número de votos pronosticado para Gabriel Ricardo Quadri de la
#       Torre/ Rosario Guerra Díaz en la elección correspondiente.
```

```
#
#     Columna 6: Indecisos+Rechazadas: Número de votos que representan las personas que no
#       revelaron sus preferencias, ya sea por no responder la encuesta o por declararse
#       indecisos.
#
#     Columna 7: IR_EPN/IR_MAM
#     Columna 8: IR_JVM/IR_BPR
#     Columna 9: IR_AMLO/IR_IMdW
#     Columna 10: IR_Quadri/IR_RG
#
#     Si met=1, las columnas 7 a 10 contienen el número de votos necesarios por candidato
#     para llegar al cómputo final, es decir, cuántos de la columna 6 tuvieron que votar
#     por cada candidato para llegar al resultado obtenido el día de la elección.
#
#     Si met=2, las columnas 7 a 10 contienen un ajuste suponiendo un sesgo constante
#     calculado a partir de la diferencia del último registro con el cómputo final.
#
####
#
     Votos<-votos(Encuesta,CompFinal,PorcNoResp)
#
# Reacomodo de columnas y adición de las nuevas que contendrán las proyecciones de votos de
# indecisos y los que no respondieron.
#
     V<-Votos
     Votos<-cbind(Votos,0,0)
     Votos[,6]<-V[,8]; names(Votos)[6]<-names(V)[8]
     Votos[,7:8]<-0; names(Votos)[7]<-paste("IR_",names(V)[2],sep="")
     names(Votos)[8]<-paste("IR_",names(V)[3],sep="")
     Votos[,11]<-V[,9]; names(Votos)[11]<-names(V)[9]
     Votos[,9]< 0; names(Votos)[9]<-paste("IR_",names(V)[4],sep="")
     names(Votos)[10]<-paste("IR_",names(V)[5],sep="")
#
# Una matriz diferente para cada uno de los métodos de proyección (diferencias y
# porcentajes)
#
# PRIMER MÉTODO: Diferencias con el cómputo final
#
     m1<-Votos
     for(j in 1:4) m1[,j+6]<-CompFinal[,j]-m1[,j+1]
#
# SEGUNDO MÉTODO: Porcentajes suponiendo un sesgo sistemático constante
#
     m2<-Votos
#
# Calcular el factor de inflación (sesgo)
#
     f<-(CompFinal-Votos[dim(Encuesta)[1],2:5])/Votos[dim(Encuesta)[1],6]
#
# Corregir sesgo en los registros anteriores
#
     for(j in 1:4) m2[,j+6]<-f[,j]*m2[,6]
#
     if(met==1) L<-m1
     else L<-m2
#
# Si se requiere, hacer el cálculo de porcentajes de votos
```

```
#
     if(porc==T){
     L[,c(2,3,4,5,7,8,9,10)]<-L[,c(2,3,4,5,7,8,9,10)]/apply(L[,c(2,3,4,5,7,8,9,10)],1,sum)
     L[,6]<-apply(L[,7:10],1,sum)
     L[,2:10]<-L[,2:10]*100
     }
     return(L)
}
```

Gráficas. Etapa 1

```
graf1 <- function(Eleccion, PorcNoResp, met, porc=T,tray="Sobre/subest", cand=1:4, plot=T){
# Input:
#
#     Eleccion > "EP" si se desea trabajar con los datos de la elección presidencial, "JG"
#       en caso de de querer usar los datos de la elección de jefe de gobierno.
#
#     PorcNoResp > Número entre 0 y 100 que indique el porcentaje de no respuesta que se
#       contemplará para hacer los cálculos.
#
#     met > Método a emplear para corregir el sesgo
#       met=1 Diferencias con el cómputo final,
#       met=2 Porcentaje sistemático constante.
#
#     porc > Valor lógico, FALSE por default. Si porc=T, el dataframe obtenido presentará
#       los porcentajes de votos por rubro, porc=F presenta el  número absoluto de votos
#       por rubro.
#
#     tray > "Enc" en caso de que se desee que en la gráfica aparezcan las trayectorias que
#       pronosticó  la encuestadora, "IR" para graficar las trayectorias pronosticadas para
#       indecisos y no respuesta, "Real" en caso de que se desee que en la gráfica
#       aparezcan las trayectorias insesgadas de todos los candidatos. "Sobre/subest"
#       grafica los porcentajes de sobre/sub estimación de todos los candidatos de la
#       elección correspondiente.
#
#     cand > Candidadato del cual se desea graficar trayectorias,
#       1: EPN/MAM
#       2: JVM/BPR
#       3: AMLO/IMdW
#       4: Quadri/RG
#
#     plot > Valor lógico, determina si se mostrarán las gráficas (TRUE) o los dataframes
#       (FALSE).
#
# Output:
#
#     plot=T > Gráfica de las trayectorias con las características deseadas.
#
#     plot=F > Dataframe con las trayectorias deseadas
#
########
#
# Datos a utilizar
#
     if(Eleccion=="EP"){
          datos <- list(Encuesta=Encuesta.EP,
```

```
          CompFinal=CompFinal.EP)
          color<-c("red","blue","orange","black")
          elecc<-"Elección presidencial 2012"
          }
        else{
          datos <- list(Encuesta=Encuesta.JG,
          CompFinal=CompFinal.JG)
          color<-c("orange","red","blue","black")
          elecc<-"Elección de Jefe de Gobierno 2012"
          }
#
# Hacer los cálculos correspondientes
#
     G<-desglose(datos$Encuesta,datos$CompFinal,PorcNoResp=PorcNoResp,met=met,porc=porc)
#
# Gráficas: t1 > Trayectoria pronosticada por la encuesta,
# t2 > Trayectoria pronosticada para los indecisos y no respuesta. t3 > Trayectoria "real"
# (suma de las dos anteriores)
#
     t1<-G[,2:5]
     t2<-G[,7:10]
     t3<-G[,2:5]+G[,7:10]
#
     t1[,1:4]<-t1[,1:4]/apply(t1[,1:4],1,sum)*(100*porc+sum(datos$CompFinal)*(1-porc))
     t2[,1:4]<-t2[,1:4]/apply(t2[,1:4],1,sum)*(100*porc+sum(datos$CompFinal)*(1-porc))
     t3[,1:4]<-t3[,1:4]/apply(t3[,1:4],1,sum)*(100*porc+sum(datos$CompFinal)*(1-porc))
#
#
#
     if(porc){
          laby<-"Porcentaje"
          datos$CompFinal<-datos$CompFinal/sum(datos$CompFinal)*100
     }
     else{
          laby<-"Votos"
     }
#
     if(tray=="Enc") titulo<-"Encuesta GEA-ISA"
     else if(tray=="Real") titulo<-"Encuesta GEA-ISA corregida"
          else titulo<-"Indecisos y no respuesta"
#
     s<-c(1,1,1)*(length(cand)==1)+c(tray=="Enc",tray=="IR",tray=="Real")
     m<-min(s[1]*min(t1),s[2]*min(t2),s[3]*min(t3))
     M<-max(max(t1),s[2]*max(t2),max(t3))
#
     if(plot){
        if(tray=="Sobre/subest")
        {
          t4<-t1-t3
          plot(0,main="Porcentaje de sobre/sub estimación por candidato", ylim=c(min(t4),
            max(t4)*1.1), xlim=c(0,dim(t1)[1]), xlab="Día", ylab="Porcentaje",type="n")
          mtext(elecc,side=3,adj=0.1)
          mtext(paste("Tasa de NR= , PorcNoResp, "%", sep=""), side=3, adj=0.9)
          for(l in 1:4) lines(t4[,l], type="l", col=color[l],lwd=2)
          abline(h=0,lty=5,col="gray")
          legend("top", names(t1)[cand], fill=color[cand], bty="n",horiz=T,cex=0.8)
#
```

```
        }
        else
        {
          plot(0, main=titulo, ylim=c(m,M*(1.1)), xlim=c(0,dim(datos$Encuesta)[1]),
            ylab=laby, xlab="Día", type="n")
          if(s[1]==0){
            mtext(elecc,side=3,adj=0.1)
            mtext(paste("Tasa de NR= ", PorcNoResp, "%", sep=""),side=3,adj=0.9)}
          else mtext(elecc,side=3)
#
       if(tray!="IR"){
         for(x in cand) points(dim(datos$Encuesta)[1], datos$CompFinal[x], col=color[x],
            lwd=6,pch=19)
         if(s[1]==1) for(i in cand)
           lines(t1[,i],type="l",col=color[i],lty=1,lwd=2)
         if(s[2]==1){ abline(h=0,lty=5,col="gray")
           for(j in cand) lines(t2[,j], type="l", col=color[j], lty=1, lwd=2) }
         if(s[3]==1) for(k in cand) lines(t3[,k], type="l", col=color[k], lty=1, lwd=2)
#
         legend("top",names(t1)[cand],fill=color[cand],bty="n", horiz=(sum(s)==1),cex=0.8)
        }
       }
# Si no se pide gráfica, regresar el dataframe
     else
     {
     S<-list(Encuesta=t1,Real=t3)
     return(S)
     }
}
```

Encuestas efectivas por candidato

```
enc.ef <- function(Eleccion, PorcNoResp, tray, mch=0){
# Input:
#
#     Eleccion > "EP" si se desea trabajar con los datos de la elección presidencial, "JG"
#       en caso de querer usar los datos de la elección de jefe de gobierno.
#
#     PorcNoResp > Número entre 0 y 100 que indique el porcentaje de no respuesta que se
#       contemplará para hacer los cálculos.
#
#     tray > "Enc.O" en caso de que se desee tomar como referencia la encuesta original,
#       "Enc.C" si se desea la encuesta corregida.
#
#     mch > Valor lógico que indica si se desea o no considerar tamaños de muestra de 100
#       encuestas válidas diarias.
#
# Output:
#
#     Dataframe de nx4 con el número de encuestas efectivas por candidato.
#
#
########
#
# Datos a utilizar
#
```

```
     if(Eleccion=="EP") datos<-list(Encuesta=Encuesta.EP, CompFinal=CompFinal.EP)
     else datos<-list(Encuesta=Encuesta.JG,CompFinal=CompFinal.JG)
#
# Hacer los cálculos correspondientes
#
#  tray=Enc.C => E=2; tray=Enc.O => E=1
#
     E<-(tray=="Enc.C")+1
#
# Obtener proporciones diarias para cada candidato
#
     G<-graf1(Eleccion, PorcNoResp, met=2, plot=F, porc=T)[[E]]/100
     E<-E-1
#
# Tamaños de muestra: para la encuesta original se usará el publicado en el periódico, para
# la encuesta corregida se aplicará el ajuste usando  el porcentaje de no respuesta.
#
     TM<-datos$Encuesta[,"Tamaño de muestra"]*(1-E)+votos(datos$Encuesta,datos$CompFinal,
             PorcNoResp)[,9]*E
     TM[dim(datos$Encuesta)[1]]<-TM[dim(datos$Encuesta)[1]-1]
     if(mch) TM<-rep(100,dim(datos$Encuesta)[1])
#
# Multiplicar las proporciones por el tamaño de muestra
# para obtener encuestas efectivas por candidato
#
     G<-G*TM
     return(G)
}
```

B.2. Ajuste del modelo Bayesiano

Simulación a partir de la distribución a posteriori

```
simula.posteriori <- function(n, Eleccion, PorcNoResp=33.8, tray, mch=0, dia, alpha=0.5){
# Input:
#
#     n > Número de observaciones a simular
#
#     Eleccion > "EP" si se desea trabajar con los datos de la elección presidencial, "JG"
#       en caso de querer usar los datos de la elección de jefe de gobierno.
#
#     PorcNoResp > Número entre 0 y 100 que indique el porcentaje de no respuesta que se
#       contemplará para hacer los cálculos.
#
#     tray > "Enc.O" en caso de que se desee tomar como referencia la encuesta original,
#       "Enc.C" si se desea la encuesta corregida.
#
#     mch > Valor lógico que indica si se desea o no considerar tamaños de muestra de 100
#       encuestas válidas diarias.
#
#     dia > Número entre 1 y 101 para la elección presidencial y entre 1 68 para la
#       elección de jefe de gobierno. Día que será tomado como muestra.
#
#     alpha > Escalar entre 0 y 1 que indica el parámetro de la distribución Dirichlet a
```

```
#       priori. (Se supone alpha_1 = alpha_2 = alpha_3 = alpha_4)
#
# Output:
#
#     mcmcobject de nx4 que contiene n simulaciones de proporciones por candidato  a partir
#       de la muestra observada en un día.
#
########
#
# Datos
#
     if(Eleccion=="EP"){
          datos <- list(Encuesta=Encuesta.EP,
          CompFinal=CompFinal.EP)
          elecc<-"Elección presidencial 2012"
          }
        else{
          datos <- list(Encuesta=Encuesta.JG,
          CompFinal=CompFinal.JG)
          elecc<-"Elección de Jefe de Gobierno 2012"
          }
#
# Ajuste a enteros de las encuestas efectivas por candidato para poder realizar la
# simulación
#
     muestra<-enc.ef(Eleccion, 33.8, tray, mch=0)[dia,]
     muestra<-as.integer(muestra)
#
# Simulación de n proporciones por candidato usando el objeto "muestra"
#
     output <- rdirichlet(n, muestra + rep(alpha,4)),
       thin=1)
     colnames(output)<-colnames(CompFinal)
     return(output)
}
```

Cálculo de las probabilidades de ganar

```
probab.ganar<-function(n, Eleccion, PorcNoResp=33.8, tray,
 alpha=0.5, cand=1:4){
# Input:
#
#     n > Número de observaciones a simular
#
#     Eleccion > "EP" si se desea trabajar con los datos de la elección presidencial, "JG"
#       en caso de querer usar los datos de la elección de jefe de gobierno.
#
#     PorcNoResp > Número entre 0 y 100 que indique el porcentaje de no respuesta que se
#       contemplará para hacer los cálculos.
#
#     tray > "Enc.O" en caso de que se desee tomar como referencia la encuesta original,
#       "Enc.C" si se desea la encuesta corregida.
#
#     alpha > Escalar entre 0 y 1 que indica el parámetro de la distribución Dirichlet a
#       priori. (Se supone alpha_1 = alpha_2 = alpha_3 = alpha_4)
#
```

```
#     cand > Parámetro con los índices de candidatos para los cuales se desean calcular las
#       probabilidades.
#
# Output:
#
#     Matriz de nx4 que contiene probabilidades empíricas de ganar por candidato y por día
#       calculadas a partir de las simulaciones a posteriori.
#
########
#
# Datos necesarios
#
     if(Eleccion=="EP"){
          datos <- list(Encuesta=Encuesta.EP,
          CompFinal=CompFinal.EP)
          }
     else{
          datos <- list(Encuesta=Encuesta.JG,
          CompFinal=CompFinal.JG)
          }
#
# Declarar objetos que se van a utilizar
#
     P<-vector()
     matriz.p<-NULL
#
     for(i in 1:dias){
#
# Simulación de n proporciones por candidato usando como muestra el día i
#
      x<-simula.posteriori(n, Eleccion, PorcNoResp, tray, dia=i, alpha=alpha, thin=1)
#
#  Probabilidad de ganar
#
      if(length(cand)==4) for(j in 1:4) P[j]<- mean(apply(x[,j]>x[,-j],1,sum)==3)
      else{ for(j in 1:2) P[j]<-mean(x[,cand[j]]>x[,cand[-j]])}
#
#  Matriz con probabilidades diarias de ganar por candidato
#
     matriz.p<-rbind(matriz.p,P)
     }
     rownames(matriz.p)<-NULL
     colnames(matriz.p)<-colnames(CompFinal[cand])
     return(matriz.p)
}
```

Estimaciones de los porcentajes de preferencia

```
estimaciones <- function(n, Eleccion, PorcNoResp=33.8,
 tray, alpha=0.5,cand=1:4, ai=0.95, dia=0){
# Input:
#
#     n > Número de observaciones a simular
#
#     Eleccion > "EP" si se desea trabajar con los datos de la elección presidencial, "JG"
#       en caso de querer usar los datos de la elección de jefe de gobierno.
```

```
#
#     PorcNoResp > Número entre 0 y 100 que indique el porcentaje de no respuesta que se
#       contemplará para hacer los cálculos.
#
#     tray > "Enc.O" en caso de que se desee tomar como referencia la encuesta original,
#       "Enc.C" si se desea la encuesta corregida.
#
#     alpha > Escalar entre 0 y 1 que indica el parámetro de la distribución Dirichlet a
#       priori. (Se supone alpha_1 = alpha_2 = alpha_3 = alpha_4)
#
#     ai> Escalar entre 0 y 1 que indica la probabilidad del intervalo a calcular, por
#       default es al 95%.
#
#     cand > Parámetro con el índice del candidato para el cual se desean calcular las
#       estimaciones, por default, c=1:4 y se obtiene un concentrado de todos los
#       candidatos en el dia establecido en el parámetro "dia".
#
#     dia > Día para el cual se desean estimar los porcentajes para todos los candidatos.
#
# Output:
#
#     Dataframe con las estimaciones puntuales y por intervalo.
#
########
#
# Datos necesarios
#
     if(Eleccion=="EP"){
          datos <- list(Encuesta=Encuesta.EP,
          CompFinal=CompFinal.EP)
          }
     else{
          datos <- list(Encuesta=Encuesta.JG,
          CompFinal=CompFinal.JG)
          }
#
# Límites del intervalo
#
     limInt<-c((1-ai),(1+ai))/2
#
# Cuantiles a calcular
#
     q<-c(limInt[1], .5, limInt[2])
#
# Si se desean todas las estimaciones diarias para un
# candidato...
#
     if(length(cand)==1){
#
# Declarar objetos que se van a utilizar
#
       est<-matrix(NA,nrow=dias,ncol=3)
       for(i in 1:3) colnames(est)<-paste(names(CompFinal[cand]),"_",q*100,"%",sep="")
#
       for(i in 1:dias){
#
# n simulaciones diarias de la distribución a posteriori
```

```
# del vector de parámetros usando como muestra el día i
#
       x<-simula.posteriori(n, Eleccion, PorcNoResp, tray, dia=i, alpha=alpha, thin=1)
#
# Cálculo de los cuantiles de interés
#
       est[i,]<-quantile(x[,cand], prob=q)
       }
     }
#
# Si se desean tablas de todos los candidatos para el día
# especificados en el parámetro dia...
#
     else{
       est <- matrix(NA, ncol=3, nrow=4)
       colnames(est) <- paste(q*100,"%",sep="")
       rownames(est) <- names(CompFinal)
#
# n simulaciones de la distribución a posteriori del vector
# de parámetros para el día=dia
#
       x <- simula.posteriori(n, Eleccion, PorcNoResp, tray, dia=dia, alpha=alpha, thin=1)
#
# Cálculo de los cuantiles de cada candidato
#
       for(i in 1:4) est[i,] <- quantile(x[,i],prob=q)
     }
     return(est)
}
```

Gráficas. Etapa 2.

```
graf2 <- function(n, Eleccion, PorcNoResp=33.8, tray=0, alpha=0.5, ai=0.95, graf, cand=0){
# Input:
#     n > Número de observaciones a simular por día
#
#     Eleccion > "EP" si se desea trabajar con los datos de la elección presidencial, "JG"
#       en caso de querer usar los datos de la elección de jefe de gobierno.
#
#     PorcNoResp > Número entre 0 y 100 que indique el porcentaje de no respuesta que se
#       contemplará para hacer los cálculos.
#
#     tray > "Enc.O" en caso de que se desee tomar como referencia la encuesta original,
#       "Enc.C" si se desea la encuesta corregida.
#
#     alpha > Escalar entre 0 y 1 que indica el parámetro de la distribución Dirichlet a
#       priori. (Se supone alpha_1 = alpha_2 = alpha_3 = alpha_4)
#
#     ai> Escalar entre 0 y 1 que indica la probabilidad del intervalo a calcular, por
#       default es al 95%.
#
#     graf > graf="estimaciones" para graficar estimaciones puntuales e intervalos de
#       probabilidad graf="probabilidades" para graficar probabilidades de ganar con las
#       especificaciones del parámetro "cand".
#
#     cand > Candidato o candidatos para los cuáles se desea graficar las probabilidades de
```

```
#       ganar la elección. Si es un candidato se trata de probabilidades globales, y si son
#       2 candidatos son probabilidades de que el primero le gane al segundo.
#
# Output:
#
#     graf="estimaciones" > Gráfica con estimaciones
#       puntuales y por intervalo de los porcentajes de
#       preferencia.
#
#     graf="probabilidades" > Gráfica con probabilidades
#       de ganar.
#
########
#
# Datos
#
     if(Eleccion=="EP"){
       CompFinal<-CompFinal.EP
       color1<-c(rgb(252,31,31,maxColorValue=300,alpha=115),rgb(39,28,255,maxColorValue=300)
        rgb(254,164,29,maxColorValue=300,alpha=115),rgb(0,0,0,maxColorValue=300,alpha=115))
       color2<-c("red","blue","orange","black")
       elecc<-"Elección Presidencial 2012"
       E<-1
     }
     else{
       CompFinal<-CompFinal.JG
       color1<-c(rgb(254,164,29,maxColorValue=300,alpha=115),rgb(252,3,31,maxColorValue=300)
        rgb(39,28,255,maxColorValue=300,alpha=115),rgb(0,0,0,maxColorValue=300,alpha=115))
       color2<-c("orange","red","blue","black")
       elecc<-"Elección de Jefe de Gobierno, 2012"
       E<-0
     }
     if(tray=="Enc.O") enc<-"Encuesta original"
     else if(tray=="Enc.C") enc<-"Encuesta corregida"
          else enc<-c("Encuesta original","Encuesta corregida")
#
# Si se desea gráfica de las estimaciones
#
     if(graf=="estimaciones"){
#
#         Declarar objetos necesarios
#
       X<-list(u=NA,d=NA,t=NA,c=NA)
#
#         Estimaciones para todos los candidatos(lista)
#
       for(i in 1:4) X[[i]]<-estimaciones(n,Eleccion=Eleccion,tray=tray,ai=ai,cand=i,
          PorcNoResp=PorcNoResp)
#
#         Gráfica en blanco
#
       plot(c(0,0), ylim=c(0,0.6*E+0.8*(1-E)),xlim=c(0,102*E+68*(1-E)),type="n",
         main=paste("Estimaciones puntuales diarias e intervalos de probabilidad
         ",ai*100,"%",sep=""), xlab="Día", ylab="Proporción de votos")
       mtext(elecc,side=3,adj=0.1)
       mtext(enc,side=3)
       mtext(paste(" Tasa de NR= ", PorcNoResp, "%", sep=""), side=3, adj=0.9)
```

```
#
# Graficar estimaciones puntuales y por intervalo para
# todos los candidatos
#
       for(i in 1:4){
         for (j in 1:dim(X[[i]])[1])
           lines(c(j,j), c(X[[i]][j,1],X[[i]][j,3]), col=color1[i], lwd=4)
           lines(X[[i]][,2],col=color2[i],type="l",lwd=2)
     }
#
# Leyendas con los nombres de los candidatos
#
       legend("top",names(CompFinal), fill=color2, horiz=T, bty="n", cex=0.8)
     }
#
# Si se desean graficar las probabilidades de ganar
#
     else{
#
# Si son probabilidades globales
#
       if(length(cand)==1){
         X <- probab.ganar(n=n, Eleccion=Eleccion, tray='Enc.O')[,cand]
         Y<-probab.ganar(n=n, Eleccion=Eleccion, tray='Enc.C', PorcNoResp=PorcNoResp)[,cand]
         plot(X,type='l', col=color2[4], main=paste("Probabilidad de ganar de",
           names(CompFinal)[cand]),xlab="Día",ylab="Probabilidad",ylim=c(min(X,Y),max(X,Y)))
         lines(Y,col=color1[4],lty=2,lwd=2)
         mtext(elecc,side=3,adj=0.2)
         mtext(paste("Tasa de NR= ", PorcNoResp, "%", sep=""), adj=0.8)
         legend("top",enc,lwd=2,lty=1:2,horiz=T, col=c(color2[4],color1[4]),bty='n')
       }

#
# Si son probabilidades por pares
#
       else{
         X<-probab.ganar(n=n, Eleccion=Eleccion, tray='Enc.O', cand=cand)[,1]
         Y<-probab.ganar(n=n, Eleccion=Eleccion, tray='Enc.C', cand=cand,
           PorcNoResp=PorcNoResp)[,1]
         plot(X,type='l',main=paste("Probabilidad de que",names(CompFinal)[cand[1]],
           "le gane a", names(CompFinal)[cand[2]]), xlab="Día", ylab="Probabilidad", lwd=2)
         lines(Y, lty=2, lwd=2)
         mtext(elecc, side=3, adj=0.2)
         mtext(paste("Tasa de NR= ", PorcNoResp, "%", sep=""), adj=0.8)
         legend("topright", enc, lwd=2, lty=1:2, bty='n', cex=0.9)
       }
     }
}
```

B.3. Cópulas

Cópula empírica

```
cop.emp <- function(muestra){
# Programa que genera la cópula empírica a partir de un vector bivariado.
#
# Input:
#
#      muestra > Matriz de dos columnas (nx2) provenientes de un vector bivariado
#
# Output:
#
#     Matriz de nxn con la cópula empírica calculada para cada (i/n,j/n): i,j=1,2,...n
#
########
#
# Datos
     n<-dim(muestra)[1]
     muestra.ord<-apply(muestra,2,rank)
#
# Matriz de frecuencias de la cópula empírica
#
     cn<-matrix(0,ncol=n,nrow=n)
     cn[muestra.ord]<-1/n
#
# Cópula empírica
#
     Cn<-apply(cn,2,cumsum)
     for (i in 2:n) Cn[,i]<-Cn[,i-1]+Cn[,i]
     return(Cn)
}
```

Sigma de Schweizer y Wolff

```
sigma.sw<-function(x,y,digits=2,prefix=""){
# Input:
#     muestra > Matriz de dos columnas (nx2) provenientes de un vector bivariado.
#
# Output:
#     Valor del cálculo la sigma de Schweizer y Wolff.
#
########
#
     muestra<-cbind(x,y)
     n<-dim(muestra)[1]
     Cn<-cop.emp(muestra)
     x<-(1:n)/n
     A<-Cn-outer(x,x,FUN="*")
     sigma<-(12/(n^2-1))*sum(abs(A))
     return(sigma)
}
```

Pseudo observaciones de la cópula bivariada

```
pseudo.obs<-function(Eleccion,tray,method){
# Input:
#
#     Eleccion > "EP" si se desea trabajar con los datos de la elección presidencial, "JG"
#       en el caso de la elección de jefe de gobierno
#
#     tray > "Enc.Orig" en caso de que se desee que en la gráfica aparezcan las
#       pseudo-observaciones calculadas a partir de lo que pronosticó la encuestadora,
#       "Enc.Corr" en caso de que se desee que en la gráfica aparezcan las trayectorias
#       insesgadas de todos los candidatos.
#
#     method > 'spearman' para regresar en el triángulo inferior de la gráfica la rho de
#       Spearman, 'schweizer' para regresar en su lugar la sigma de Schweizer y Wolff.
#
# Output:
#
#     Gráfica con las pseudo-observaciones de la cópula bivariada que modela la dependencia
#     entre los diferentes pares de candidatos.
#
########
#
# Datos a utilizar
#
#  Trayectoria:
#
     if(tray=="Enc.Orig"){ L<-1; enc<-"Encuesta Original"}
     else{ L<-2; enc<-"Encuesta Corregida"}
     votos<-graf1(Eleccion, PorcNoResp=33.8, met=2, plot=F, porc=F)[[L]]
#
#  Elección
#
     if(Eleccion=="EP"){
       CompFinal<-CompFinal.EP
       elecc<-"Elección Presidencial 2012"
     }
     else{
       CompFinal<-CompFinal.JG
       elecc<-"Elección de Jefe de Gobierno, 2012"
     }
#
#  Medida de dependencia
#
     if(method=='spearman') dep<-"rho de Spearman"
     else dep<-"sigma de Schweizer"
#
# Función para escribir la rho de spearman o la sigma de Schweizer de un par de variables
#
     panel.cor <- function(x, y, digits=2, prefix="",
      cex.cor, ...)
     {
       usr <- par("usr"); on.exit(par(usr))
       par(usr = c(0, 1, 0, 1))
       if(method=='spearman')
         r <- cor(x, y, method='spearman')
       else r<-sigma.sw(x,y)
```

```
      txt <- format(c(r, 0.123456789), digits=digits)[1]
      text(0.5,0.5,txt,cex=1.9)
    }
#
# Pseudo observaciones de la cópula
#
    D<-apply(votos,2,rank)/dim(votos)[1]
    par(mar=c(5.1,4.1,10.1,2.1))
    pairs(D, lower.panel=panel.cor, upper.panel=panel.smooth, main=paste("Pseudo-
      observaciones de las cópulas bivariadas y ", dep, sep=""))
    mtext(elecc, side=3, adj=0.1, line=7, cex=0.9)
    if(tray=="Enc.Corr") { mtext(enc, side=3, line=7, cex=0.9)
      mtext(paste(" Tasa de NR= ", 33.8, "%", sep=""), side=3, adj=0.9, line=7, cex=0.9)
    }
    else mtext(enc,side=3, adj=0.9, line=7, cex=0.9)
}
```

Correlaciones

```
razon<-function(x){
    y<-NULL
    for(i in 2:length(x)) y<-c(y,(x[i]/x[i-1])-1)
    return(y)
    }

votos.encuesta<-graf1("EP", PorcNoResp=33.8, met=2, plot=F,
  porc=F)$Encuesta
votos.correcc<-graf1("EP", PorcNoResp=33.8, met=2, plot=F,
  porc=F)$Real

#Diferencias absolutas
    X<-apply(votos.encuesta,2,diff)
    Y<-apply(votos.correcc,2,diff)
    cor(X)
    cor(Y)

#Diferencias porcentuales
    X<-apply(votos.encuesta,2,razon)
    Y<-apply(votos.correcc,2,razon)
    cor(X)
    cor(Y)
```

Secciones diagonales de las cópulas empíricas

```
diag.copula<-function(Eleccion,tray){
# Input:
#
#     Eleccion > "EP" si se desea trabajar con los datos de la elección presidencial, "JG"
#       en el caso de la elección de jefe de gobierno.
#
#     tray > "Enc.Orig" en caso de que se desee que la gráfica se base en lo que pronosticó
#       la encuestadora, "Enc.Corr" en caso de que se desee usar las trayectorias insesgadas
#       detodos los candidatos.
#
```

```
# Output:
#     Scatterplot para todos los candidatos donde en el panel superior aparecen las gráficas
#       de las diagonales de lacópula empírica y en el inferior los valores de la rho de
#       Spearman y la sigma de Schweizer y Wolff.
#
###################
#
# Datos:
#
#  Trayectoria
#
     if(tray=="Enc.Orig"){ L<-1; enc<-"Encuesta Original"}
     else{ L<-2; enc<-"Encuesta Corregida"}
     votosT<-graf1(Eleccion, PorcNoResp=33.8, met=2, plot=F, porc=T)[[L]]
#
#  Elección
#
     if(Eleccion=="EP"){
       CompFinal<-CompFinal.EP
       elecc<-"Elección Presidencial 2012"
     }
     else{
       CompFinal<-CompFinal.JG
       elecc<-"Elección de Jefe de Gobierno, 2012"
     }
     n<-dim(votosT)[1]
#
#  Colores
     colores<-c(rgb(4,134,134,maxColorValue=300,alpha=90),
       rgb(134,4,134,maxColorValue=300,alpha=90),
       rgb(134,134,4,maxColorValue=300,alpha=90), 'black')
#
# Gráficas de las diagonales de M, pi y W y de la cópula
# empírica para un par de variables
#
     grafica.diagonales<-function(muestra,puntos=1000){
       n<-dim(muestra)[1]
       valores<-(1:puntos)/puntos
       diag.M<-valores
       diag.W<-((2*valores-1)>0)*(2*valores-1)
       diag.Pi<-valores^2
       diag.Cop<-diag(cop.emp(muestra))
       par(mar=c(4.2,4.4,3.2,2.1))
       plot(c(0,(1:n)/n), c(0,diag.Cop), lwd=2, col=colores[4],type='l', xlab="t",
         ylab=expression(delta[C](t)), cex.lab=1.5, main=paste(names(muestra)[1], "-",
         names(muestra)[2]))
       lines(c(0,valores),c(0, diag.M), type='l', lwd=2, col=colores[1])
       lines(c(0,valores), c(0,diag.W), col=colores[2], lwd=2)
       lines(c(0,valores), c(0,diag.Pi), lwd=2, col=colores[3])
       rho<-cor(muestra,method='spearman')[2]
       sigma<-sigma.sw(muestra[,1],muestra[,2])
       txt<-c(round(rho,2),round(sigma,2))
       text(0.1,0.88,expression(paste(rho," =")),cex=1.9)
       text(0.25,0.9,txt[1],cex=1.9)
       text(0.1,0.68,expression(paste(sigma," =")),cex=1.9)
       text(0.25,0.7,txt[2],cex=1.9)
     }
```

```
#
# Gráfica total
#
     par(oma=c(5,2,8,2))
     par(mfrow=c(3,2))
     for(i in 2:4) grafica.diagonales(votosT[,c(1,i)])
     for(j in 3:4) grafica.diagonales(votosT[,c(2,j)])
     grafica.diagonales(votosT[,c(3,4)])
#
# Especificaciones
#
     title(main="Sección diagonal de las cópulas empíricas bivariadas", outer=T,
       cex.main=1.8, line=6)
     mtext(elecc,side=3,adj=0.1,outer=T,line=3.5)
     if(tray=="Enc.Corr") {
       mtext(enc, side=3, outer=T, line=3.5)
       mtext(paste(" Tasa de NR= ", 33.8, "%", sep=""), side=3, adj=0.8, line=3.5, outer=T)}
     else mtext(enc,side=3,adj=0.85,line=3.5,outer=T)
#
# Legend a mano
#
     for (i in 1:4)
       mtext(".",side=3, adj=0.1+0.2*(i-1), line=1, col=colores[i], font=2, outer=T, cex=5)
     mtext("Cópula M", side=3, outer=T, line=1, adj=0.138, cex=0.9)
     mtext("Cópula W", side=3, outer=T, line=1, adj=0.35 cex=0.9)
     mtext(expression(paste("Cópula ",Pi)), side=3, outer=T, line=0.84, adj=0.56, cex=0.9)
     mtext("Cópula empírica", side=3, outer=T, line=1, adj=0.816, cex=0.9)
#
# Rho y sigma
#
     mtext(expression(paste(rho,": rho de Spearman")), side=1, line=1.5, adj=0.3, outer=T)
     mtext(expression(paste(sigma,": sigma de Schweizer")), side=1, line=1.5, adj=0.7)
}
```

Dependencias no monótonas

Encontrar breakpoints donde cambia el signo de las dependencias

```
cambios.dependencias <-function(Eleccion,tray){
# Input:
#
#     Eleccion > "EP" si se desea trabajar con los datos de la elección presidencial, "JG"
#       en el caso de la elección de jefe de gobierno.
#
#     tray > "Enc.Orig" en caso de que se desee usar los pronósticos de la
#       encuestadora, "Enc.Corr" en caso de que se desee usar las trayectorias
#       insesgadas de todos los candidatos.
#
# Output:
#
#     Diferencias entre la cópula empírica y la cópula Pi.
#
###################
#
# Datos
```

```
#
#   Trayectoria:
     if(tray=="Enc.Orig"){ L<-1; enc<-"Encuesta Original"}
     else{ L<-2; enc<-"Encuesta Corregida"}
     votosT<-graf1(Eleccion, PorcNoResp=33.8, met=2, plot=F, porc=T)[[L]]
#
#   Elección
#
     if(Eleccion=="EP"){
       CompFinal<-CompFinal.EP
       elecc<-"Elección Presidencial 2012"
     }
     else{
       CompFinal<-CompFinal.JG
       elecc<-"Elección de Jefe de Gobierno, 2012"
     }
     n<-dim(votosT)[1]
#
     dif_diagonales<-function(muestra){
       n<-dim(muestra)[1]
       valores<-(1:n)/n
       diag.Pi<-valores^2
       diag.Cop<-diag(cop.emp(muestra))
       diferencia<-(diag.Cop-diag.Pi)
       X<-cbind(c(0,valores),c(0,diferencia))
       colnames(X)<-c("Valores", paste(names(muestra)[1], "-", names(muestra)[2], sep=""))
       return(X)
     }
     matriz.dif<-cbind(dif_diagonales(votosT[,c(1,2)]),
       dif_diagonales(votosT[ c(1,3)]),
       dif_diagonales(votosT[,c(1,4)]),
       dif_diagonales(votosT[,c(2,3)]),
       dif_diagonales(votosT[,c(2,4)]),
       dif_diagonales(votosT[,c(3,4)])        )
     return(matriz.dif[-c(3,5,7,9,11)])
}
```

Particionar las dependencias no monótonas en dependencias monótonas y calcular para cada subconjunto la rho y sigma

```
dependencias.no.monotonas<-function(Eleccion, tray, cand, breaks, coords, pp=0){
# Input:
#
# Input:
#
#     Eleccion > "EP" si se desea trabajar con los datos de la elección presidencial, "JG"
#       en el caso de la elección de jefe de gobierno.
#
#     tray > "Enc.Orig" en caso de que se desee usar los pronósticos de la
#       encuestadora, "Enc.Corr" en caso de que se desee usar las trayectorias
#       insesgadas de todos los candidatos.
#
#     cand > Par de candidatos para los cuales se graficará la trayectoria.
#
```

```
#     breaks > Punto(s) en los que se dan los cambios de concordancia a discordancia o
#       viceversa.
#
# Output:
#
#  Gráfica de las trayectorias con dependencias monótonas segmentadas por colores.
#
##################
#
# Datos
#
#  Trayectoria:
#
     if(tray=="Enc.Orig"){ L<-1; enc<-"Encuesta Original"}
     else{ L<-2; enc<-"Encuesta Corregida"}
     votosT<-graf1(Eleccion, PorcNoResp=33.8, met=2, plot=F, porc=T)[[L]]
#
#   Elección
#
     if(Eleccion=="EP"){
       CompFinal<-CompFinal.EP
       color1<-c(rgb(255,136,136,maxColorValue=300),
         rgb(174, 174, 255, maxColorValue=300),
         rgb(254, 220, 145, maxColorValue=300),
         rgb(188, 188, 188, maxColorValue=300))
       color2<-c("red", "blue", "orange", "black")
       elecc<-"Elección Presidencial 2012"
     }
     else{
       CompFinal<-CompFinal.JG
       color1<-c(rgb(254, 220, 145, maxColorValue=300),
       rgb(255, 136, 136, maxColorValue=300),
       rgb(174, 174, 255, maxColorValue=300),
       rgb(188, 188, 188, maxColorValue=300))
       color2<-c("orange", "red", "blue", "black")
       elecc<-"Elección de Jefe de Gobierno, 2012"
     }
     n<-dim(votosT)[1]
#
# Programa para agrupar dependencias y mapear a los datos originales
#
     mapeo<-function(muestra,bp,color1,color2,coords){
       limy<-c(min(muestra),max(muestra))
       plot(1:length(muestra[,1]), muestra[,1], type='l', ylim=limy, xlab='Día', ylab=
         'Porcentaje de preferencias',main=paste(names(muestra)[1],"-",names(muestra)[2]))
       lines(1:length(muestra[,2]), muestra[,2])
       if(length(bp)==1){
         dep.pos<-cbind(order(muestra[,1])[1:bp], muestra[order(muestra[,1])[1:bp],])
         dep.neg<-cbind(order(muestra[,1])[(bp+1): dim(muestra)[1]],
         muestra[order(muestra[,1])[(bp+1): dim(muestra)[1]],])
         points(dep.neg[,1], dep.neg[,2],col=color1[1], pch=19, cex=1.3)
         points(dep.pos[,1], dep.pos[,2],col=color2[1], pch=19, cex=1.3)
         points(dep.neg[,1], dep.neg[,3], col=color1[2], pch=19, cex=1.3)
         points(dep.pos[,1], dep.pos[,3], col=color2[2], pch=19, cex=1.3)
         dep.neg<-dep.neg[,2:3]
     }
     else{
```

```
      dep.neg1<-cbind(order(muestra[,1])[1:bp[1]], muestra[order(muestra[,1])[1:bp[1]],])
      dep.pos<-cbind(order(muestra[,1])[(bp[1]+1):bp[2]],
        muestra[order(muestra[,1])[(bp[1]+1):bp[2]],])
      dep.neg2<-cbind(order(muestra[,1])[(bp[2]+1): dim(muestra)[1]],
        muestra[order(muestra[,1])[(bp [2]+1):dim(muestra)[1]],])
      points(dep.neg1[,1],dep.neg1[,2],col=color1[1], pch=19, cex=1.3)
      points(dep.pos[,1], dep.pos[,2], col=color2[1], pch=19, cex=1.3)
      points(dep.neg2[,1], dep.neg2[,2], col=color1[1], pch=19, cex=1.3)
#
      points(dep.neg1[,1], dep.neg1[,3], col=color1[2], pch=19, cex=1.3)
      points(dep.pos[,1], dep.pos[,3], col=color2[2], pch=19, cex=1.3)
      points(dep.neg2[,1], dep.neg2[,3], col=color1[2], pch=19, cex=1.3)
      dep.neg<-rbind(dep.neg1[,2:3],dep.neg2[,2:3])
    }
#
#
#
    rho.pos<-round(cor(dep.pos[,2:3], method='spearman') [2],2)
    rho.neg<-round(cor(dep.neg, method='spearman')[2],2)
    sigma.pos<-round(sigma.sw(dep.pos[,2],dep.pos[,3]),2)
    sigma.neg<-round(sigma.sw(dep.neg[,1],dep.neg[,2]),2)
    legend(coords[1], coords[2], c("Concordancia", "Discordancia"), fill=c(color2[2],
      color1[2]), bty='n')
    legend(coords[1]-5,coords[2], c(" "," "), fill=c(color2[1], color1[1]), bty='n',
      adj=c(1,0.3))
    if(pp==1){
      text(coords[1],coords[2],expression(paste(rho,' = ')),adj=c(-6.4,2.4))
      text(coords[1],coords[2],rho.pos,adj=c(-9,2.1))
      text(coords[1],coords[2],expression(paste(rho,' = ')),adj=c(-6.4,4.1))
      text(coords[1],coords[2], rho.neg, adj=c(-6.4,3.7))
    }
    else if(pp==2){
      text(coords[1],coords[2],expression(paste(rho,' = ')),adj=c(-6.4,2.4))
      text(coords[1], coords[2], rho.pos, adj=c(-6.5,2.1))
      text(coords[1],coords[2],expression(paste(rho,' = ')),adj=c(-6.4,4.1))
      text(coords[1],coords[2],rho.neg,adj=c(-6.5,3.7))
    }
    else{
      text(coords[1],coords[2],expression(paste(rho,' = ')),adj=c(-6.4,2.4))
      text(coords[1],coords[2],rho.pos,adj=c(-6.5,2.1))
      text(coords[1],coords[2],expression(paste(rho,' = ')),adj=c(-6.4,4.1))
      text(coords[1],coords[2], rho.neg, adj=c(-5.5,3.7))
    }
# sigma
      text(coords[1],coords[2],expression(paste(sigma,' = ')),adj=c(-9.5,3))
      text(coords[1],coords[2],sigma.pos,adj=c(-9.3,2.1))
      text(coords[1],coords[2],expression(paste(sigma,' = ')),adj=c(-9.5,5.3))
      text(coords[1],coords[2],sigma.neg,adj=c(-9.3,3.8))
    }
      mapeo(votosT[,cand], bp=breaks, color1=color1[cand], color2=color2[cand],          coords=coords)
      mtext(elecc, side=3, adj=0.1, line=7, cex=0.9)
      if(tray=="Enc.Corr") { mtext(enc, side=3, line=7, cex=0.9)
        mtext(paste(" Tasa de NR= ",33.8, "%", sep=""), side=3, adj=0.9, line=7, cex=0.9)}
      else mtext(enc, side=3, adj=0.9, line=7, cex=0.9)
}
```

Referencias

ADN Político (2012). «Intención de voto por candidato, según agencia encuestadora (%)». Disponible en línea en `http://www.adnpolitico.com/encuestas`. Último acceso en 09-10-2013.

Agresti, Alan (2002). *Categorical data analysis*. volumen 359. John Wiley & Sons.

Aristegui Noticias (2012). «Hasta 12 % de diferencia de encuestas y resultados del PREP». Disponible en línea en `http://aristeguinoticias.com/0207/post-elecciones/2da-parte-nota-encuestas-final-de-la-elecion-draft/`.

Benes, Viktor y Stepan, Josef (1997). *Distributions with given marginals and moment problems*. Kluwer Academic Publishers.

Bernardo, José M (2003). *Bayesian Statistics 7: Proceedings of the Seventh Valencia International Meeting*. Oxford University Press.

Blasco, Agustín (2005). «Bayesian Statistic Course».

Cuadras, Carles M; Fortiana, Josep y otros (2002). *Distributions with given marginals and statistical modeling*. Springer.

Dall'Aglio, G (1991). «Frechet classes: the beginnings». En: *Advances in Probability Distributions with Given Marginals*, pp. 1–12. Springer.

Dall'Aglio, Giorgio (1956). «Sugli estremi dei momenti delle funzioni di ripartizione doppia». *Annali della Scuola Normale Superiore di Pisa-Classe di Scienze*, **10(1-2)**, pp. 35–74.

DeGroot, Morris H (2005). *Optimal statistical decisions*. volumen 82. Wiley-Interscience.

Deheuvels, Paul (1979). «La fonction de dépendance empirique et ses propriétés. Un test non paramétrique d'indépendance». *Acad. Roy. Belg. Bull. Cl. Sci.(5)*, **65(6)**, pp. 274–292.

Embrechts, Paul; Lindskog, Filip y McNeil, Alexander (2003). «Modelling dependence with copulas and applications to risk management». *Handbook of heavy tailed distributions in finance*, **8(1)**, pp. 329–384.

Embrechts, Paul; McNeil, Alexander y Straumann, Daniel (1999). «Correlation: pitfalls and alternatives», **12**, pp. 69–71.

Erdely, Arturo (2007). *Diagonal properties of the empirical copula and applications. Construction of families of copulas with given restrictions.* Tesis doctoral, UNAM.

Erdely, Arturo (2009). «Cópulas y dependencia de variables aleatorias: Una introducción». *Miscelánea Matemática*, **48**, pp. 7–28.

Everdy, Luis (2012). «Los indecisos, el nuevo puntero según la encuesta de GEA-ISA». *ADN Político*. Disponible en línea en `http://www.adnpolitico.com/encuestas/2012/03/20/la-no-respuesta-se-come-a-josefina-en-encuesta-de-gea-isa`.

Frank, MJ (1996). «Diagonals of copulas and Schröder's equation». *Aequationes Math*, **51(150)**.

Fréchet, Maurice (1951). «Sur les tableaux de correlation dont les marges sont données». *Ann. Univ. Lyon Sci. Sect. A*, **9**.

Galton, Francis (1888). «Co-relations and their measurement, chiefly from anthropometric data». *Proceedings of the Royal Society of London*, **45(273-279)**, pp. 135–145.

Genest, C.; Nešlehová, J. y Quessy, J.F. (2011). «Tests of symmetry for bivariate copulas». *Annals of the Institute of Statistical Mathematics*, pp. 1–24.

Ghosh, Jayanta K; Delampady, Mohan y Samanta, Tapas (2006). *An introduction to bayesian analysis: theory and methods.* Springer.

Gisbert, Xavier y otros (2007). «Estimación Bayesiana de proporciones: Aplicación a la detección de cambios de clima».

Guerrero, Víctor Manuel (2012). «Pronósticos electorales». *Nexos en línea*. Disponible en línea en `http://www.nexos.com.mx/?P=leerarticulo&Article=2102976`.

Hoeffding, W. (1940). *Masstabinvariante Korrelationstheorie*. Schriften des Matematischen Instituts und des Instituts für Angewandte Mathematik der Universität Berlin 5 Heft 3:179-233.

Instituto Electoral del Distrito Federal (2012). «Acta de cómputo total». Disponible en línea en `http://www.iedf.org.mx/secciones/inicio/botones/ActasPEO2011-2012/totalGDF.pdf`. Último acceso en 08-10-2013.

Joe, Harry (1997). *Multivariate models and dependence concepts*. volumen 73. CRC Press.

Johnson, Norman Lloyd; Kotz, Samuel y Balakrishnan, Narayanaswamy (1997). *Discrete multivariate distributions*. volumen 165. Wiley New York.

Klement, Erich Peter y Mesiar, Radko (2005). *Logical, algebraic, analytic and probabilistic aspects of triangular norms*. Elsevier.

Klement, Erich Peter; Mesiar, Radko y Pap, Endre (2000). *Triangular norms*. Kluwer Academic Publishers Dordrecht.

Kruskal, William H (1958). «Ordinal measures of association». *Journal of the American Statistical Association*, **53(284)**, pp. 814–861.

Lehmann, Erich Leo (1966). «Some concepts of dependence». *The Annals of Mathematical Statistics*, pp. 1137–1153.

Mari, Dominique Drouet y Kotz, Samuel (2001). *Correlation and dependence*. volumen 2. World Scientific.

Martin, Andrew D.; Quinn, Kevin M. y Park, Jong Hee (2011). «MCMCpack: Markov Chain Monte Carlo in R». *Journal of Statistical Software*, **42(9)**, p. 22.
`http://www.jstatsoft.org/v42/i09/`

Menger, Karl (1942). «Statistical metrics». *Proceedings of the National Academy of Sciences of the United States of America*, **28(12)**, pp. 535–537.

Metodología de encuestas (2000). «Encuestas electorales. Artículos de opinión». *Metodología de Encuestas*, **2(1)**, pp. 135–165.

Milenio (2012). «Encuesta de seguimiento diario Milenio – GEA/ISA». *Sección: Política*. Marzo 19 – Junio 27.

Nelsen, Roger B (2006). *An introduction to copulas*. Springer, 2ª edición.

O'Hara, Bob (2012). «How did Nate Silver predict the US election?» *GRLLScientist*. Disponible en línea en `http://www.theguardian.com/science/grrlscientist/2012/nov/08/nate-sliver-predict-us-election`.

Quesada, José Juan; Rodríguez, José Antonio y Úbeda, Manuel (2003). «What are Copulas?» En: *Monografías del Semin. Matem. García de Galdeano*, 27, pp. 499–506.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. `http://www.R-project.org/`

Robert, Christian (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer, 2ª edición.

Rüschendorf, Ludger; Schweizer, Berthold y Taylor, Michael Dee (1996). «Distributions with fixed marginals and related topics». IMS.

Schweizer, Berthold (1991). «Thirty years of copulas». En: *Advances in probability distributions with given marginals*, pp. 13–50. Springer.

Schweizer, Berthold y Sklar, Abe (1983a). *Probabilistic metric spaces*. Dover.

Schweizer, Berthold y Sklar, Abe (1983b). *Probabilistic metric spaces*. Dover.

Schweizer, Berthold y Wolff, Edward F (1981). «On nonparametric measures of dependence for random variables». *The annals of statistics*, pp. 879–885.

Siburg, K.F. y Stoimenov, P.A. (2008). «Gluing copulas». *Communications in Statistics Theory and Methods*, **37(19)**, pp. 3124–3134.

Sklar, M (1959). *Fonctions de répartition à n dimensions et leurs marges*. Université Paris 8.

Tribunal Electoral del Poder Judicial de la Federación (2012). «Dictamen del cómputo final». Disponible en línea en `http://portal.te.gob.mx/contenido/dictamen-del-computo-final`. Último acceso en 08-10-2013.

Walley, Peter (1996). «Inferences from multinomial data: learning about a bag of marbles». *Journal of the Royal Statistical Society. Series B (Methodological)*.