FACULTAD DE INGENIERÍA Y CIENCIAS APLICADAS
UNIVERSIDAD CENTRAL DEL ECUADOR

INGENIO
2022, VOL 5(2), JULIO-DICIEMBRE, PP. 4-13

REVISTA INGENIO

Estadística Bayesiana para la Inferencia sobre el Comportamiento Electoral

Bayesian Statistics for Inference over Electoral Behavior

Atal Kumar Vivas Paspuel | iD Universidad de las Fuerzas Armadas ESPE, Quito, Ecuador
David Alfredo Vivas Paspuel | iD Universidad San Francisco de Quito, Quito, Ecuador
Alberto Benjamín Santillán Tituaña | iD Universidad de las Fuerzas Armadas ESPE, Quito, Ecuador

ARTICLE HISTORY
Received: 25/3/2022
Accepted: 27/10/2022

KEY WORDS
Markov chain, Monte Carlo, bayesian models, hierarchical models, ecological inference.

ABSTRACT
Los resultados que proporcionan las entidades electorales no permiten conocer el apoyo a los partidos por clases sociales, grupos de edad o razas. En este trabajo, se dividió a la población electoral por clases de edad y se realizó inferencias sobre las proporciones de apoyo por edad para los partidos Alianza País y CREO, los más importantes de la contienda presidencial en Ecuador en 2013. Se tomaron los resultados de la contienda política en tablas de contingencia de tamaño RxC a nivel parroquial y por medio de la inferencia ecológica se estiman las proporciones de apoyo hacia los candidatos por parte de dichas clases. Las inferencias se realizaron a través de técnicas bayesianas con un modelo jerárquico Dirichlet-Multinomial y se utilizaron métodos computacionales Markov Chain Montecarlo ejecutados por el paquete RStan.

PALABRAS CLAVE
Cadenas de Markov, Monte Carlo, modelos bayesianos, inferencia ecológica.

RESUMEN
Los resultados que proporcionan las entidades electorales no permiten conocer el apoyo a los partidos por clases sociales, grupos de edad o razas. En este trabajo, se dividió a la población electoral por clases de edad y se realizó inferencias sobre las proporciones de apoyo por edad para los partidos Alianza País y CREO, los más importantes de la contienda presidencial en Ecuador en 2013. Se tomaron los resultados de la contienda política en tablas de contingencia de tamaño RxC a nivel parroquial y por medio de la inferencia ecológica se estiman las proporciones de apoyo hacia los candidatos por parte de dichas clases. Las inferencias se realizaron a través de técnicas bayesianas con un modelo jerárquico Dirichlet-Multinomial y se utilizaron métodos computacionales Markov Chain Montecarlo ejecutados por el paquete RStan.

1. Introducción

Los datos que arrojan las instituciones oficiales sobre resultados electorales están siempre limitados a mostrar el apoyo hacia un partido político o candidato en forma agregada. Por ejemplo, para las elecciones presidenciales del Ecuador en 2013, la información de carácter general muestra que el 57,2% de los votantes apoyaron a Alianza País (AP), mientras que el 22,7% apoyó al movimiento CREO [1]. Esta información también está dada por provincias y otras subdivisiones políticas. Sin embargo, sabiendo que el voto es secreto, es lógico suponer que este conteo electoral no nos permite conocer las proporciones de apoyo de grupos o clases sociales hacia los candida-

https://doi.org/10.29166/ingenio.v5i2.3712

pISSN 2588-0829
eISSN 2697-3243
fing.revista.ingenio@uce.edu.ec

tos. Sería difícil conocer el apoyo de cierta raza, grupo de edad o cierta clase social hacia los partidos políticos. Este es el denominado problema de la inferencia ecológica y fue abordado desde el siglo anterior por varios autores que propusieron métodos para su solución como el determinístico de los intervalos, [2] la regresión ecológica, [3] el método EI [4] o más recientemente con técnicas de aprendizaje para regresiones de distribución, [5] técnicas de programación lineal, [6] técnicas de optimización, [7] entre otros. A pesar de sus deficiencias, la inferencia ecológica sigue siendo una parte necesaria de algunas áreas de inferencia cuantitativa. Algunos ejemplos en el campo de las ciencias políticas electorales son [8], [9] y [10].

Para realizar las inferencias, este trabajo aplicó un modelo jerárquico bayesiano para tablas de contingencia de tamaño RxC [11] sin el uso de covariables. Este modelo es una extensión del modelo jerárquico bayesiano para el caso de tablas 2x2 [12]. Dada la complejidad de los cálculos de las distribuciones para los parámetros de interés, se utilizaron técnicas computacionales Markov Chain Monte Carlo mediante el lenguaje RStan para la configuración del modelo. Los resultados obtenidos son las inferencias sobre los parámetros de las distribuciones *a posteriori* para las variables de interés. Esto permitió presentar los resultados en forma geográfica a nivel de parroquias[1] para los partidos de la contienda electoral y se analizaron los resultados.

Para la construcción de las tablas de contingencia se tomaron dos fuentes de datos: la información que proporciona el INEC, [13] sobre la población por edades para cada parroquia y los resultados oficiales de la contienda electoral en cada parroquia. La primera base de datos nos permite dividir al electorado en grupos de edades, más específicamente se ha dividido la población en los grupos de edad: 16-29, 30-44, 45-60 y > 60 años. Se obtuvo la cantidad de personas en cada parroquia que pertenecen a estas clases. La segunda base de datos nos permite ver el apoyo hacia los dos partidos políticos de forma agregada para la contienda electoral del 2013, de modo que se tengan tres grupos: cantidad de votantes que apoyaron al partido ganador AP, cantidad de votantes que apoyaron al partido CREO y cantidad de votantes que optaron por apoyar a otro partido político, abstenerse o anular el voto.

Estos modelos tienen algunas desventajas: 1) existe poca investigación sobre la precisión de los métodos que extienden la inferencia ecológica al caso RxC; 2) se debe tener precaución al utilizarlos, especialmente en aquellos casos en los que las estimaciones involucran tablas de contingencia grandes, las estimaciones pueden resultar sesgadas en los casos en que los partidos políticos son pequeños; [14] 3) suelen apoyarse en supuestos que son difíciles verificar, en la mayoría de casos requieren una considerable capacidad de hardware para ser implementados y aun así no lograr la convergencia para los algoritmos MCMC. Sin embargo, hay evidencias en favor de utilizar estos métodos de inferencia para obtener estimaciones válidas sobre este tipo de fenómenos [14].

2. Método

La inferencia ecológica es el proceso de aprendizaje acerca del comportamiento individual a partir de datos agrupados, es decir, hacer predicciones a nivel desagregado a partir de datos agregados [15]. Una manera de entender el problema de la inferencia ecológica es considerar una tabla de contingencia cuyas entradas dentro de ella sean desconocidas y sus marginales conocidas. Tomemos la siguiente tabla de contingencia de tamaño *RxC*, donde *R=4 y C=3* (ver Tabla 1).

Las cantidades marginales por recinto electoral T_{1i} y T_{2i}, se las puede obtener en los resultados de las contiendas electorales por cada parroquia y son las cantidades totales de votos que recibieron AP y CREO, respectivamente, la tercera columna muestra los votos anulados y abstenciones. Las cantidades marginales X_{1i}, X_{2i} y X_{3i} se las puede obtener en la información que proporcionan los censos nacionales y corresponden a los grupos de edades respectivos. Sin embargo, las cantidades al interior de la tabla no se las puede conocer directamente dado que el voto es secreto. De este modo, se trata de inferir las intersecciones de la tabla de contingencia: $\beta_{11}^{i}, \ldots, \beta_{42}^{i}$.

Tomando la nomenclatura de [11] la información consta de recintos electorales (parroquias), para cada recinto i (i=1,2,…,p), se tiene la cantidad de personas que acudieron a las urnas. Podemos observar las proporciones del electorado que apoyaron a un partido específico: T_{1i},T_{2i},…,T_{Ci}. y las fracciones del electorado en las diferentes clases de edad: X_{1i},X_{2i},…,X_{Ri}. Las variables de interés a inferir son las fracciones del electorado que pertenecen a la clase , que votaron por el partido $c: \beta_{rc}^{i}$, donde *r=1,…,R,c=1,…,C-1.*

2.1. LOS MODELOS BAYESIANOS

El paradigma bayesiano afirma que la probabilidad de un evento puede estar sujeto al grado de creencia que tengamos sobre ese evento, de hecho, en los modelos bayesianos están sujetos a esta idea, por tanto, se les asigna dicho grado de creencia. Esto implica que, de acuerdo con nuestro grado de creencia, podemos elaborar

1 En Ecuador, las parroquias son parte de la división política y en este trabajo serán tratadas como recintos electorales.

Tabla 1.
Tabla de contingencia RxC que establece las clases por edad y apoyo a AP, CREO *y el resto de los candidatos que incluye votos nulos y blancos.*

Grupo etario	Partido AP	Partido CREO	Otros	Total
16 - 29	β_{11}^{i}	β_{12}^{i}	$1-\sum_{c=1}^{2}\beta_{1c}^{i}$	X_{1i}
30 - 44	β_{21}^{i}	β_{22}^{i}	$1-\sum_{c=1}^{2}\beta_{2c}^{i}$	X_{2i}
45 - 60	β_{31}^{i}	β_{32}^{i}	$1-\sum_{c=1}^{2}\beta_{3c}^{i}$	X_{3i}
> 60	β_{41}^{i}	β_{42}^{i}	$1-\sum_{c=1}^{2}\beta_{4c}^{i}$	$1-\sum_{r=1}^{3}X_{ri}$
Total	T_{1i}	T_{2i}	$1-\sum_{c=1}^{2}T_{ci}$	

Fuente. Elaboración propia.

distribuciones de probabilidad y agregarles valores de probabilidad a estas, basados en este criterio. Por supuesto que también se requiere de una verificación de nuestras suposiciones acerca de los parámetros y esto lo podemos realizar mediante el cálculo de una verosimilitud [16].

De acuerdo con todo esto, podemos apoyarnos en el teorema de Bayes aplicado a distribuciones continuas de probabilidad. El paradigma bayesiano se puede resumir en la siguiente fórmula: [17]

$$p(\theta|x)=\frac{p(x|\theta)p(\theta)}{\int_{\Theta}p(x,\theta)d\theta} \quad (1)$$

En donde, $p(\theta)$ representa nuestro grado de creencias sobre el parámetro θ, denominada también como distribución *a priori* justamente al representar grados de creencia. Tenemos también $p(x|\theta)$ a que representa la verosimilitud para los datos. Por otro lado, la integral $\int_{\Theta}p(x,\theta)d\theta$ representa la probabilidad total de los datos si tomamos en cuenta el espacio total para el parámetro θ. Finalmente, $p(x|\theta)$ es la distribución de probabilidad para θ dados los datos, también llamada la distribución *a posteriori*. Todo esto se da con el fin de obtener $p(x|\theta)$ a partir de $p(\theta,x)$ y $p(\theta)$.

Una forma de resumir la construcción de un modelo del tipo bayesiano puede ser: [17] 1) Primero, se debe elaborar un modelo que tome en cuenta todas las variables que entran en juego para el modelo o fenómeno a estudiar. Se debe tomar en cuenta que este modelo debe mostrar coherencia con los datos históricos y el conocimiento construido con anterioridad a partir de los datos. 2) Elaborar las distribuciones de probabilidad *a posteriori* para las cantidades desconocidas, pero condicionadas a las cantidades observadas del modelo. Se deben tomar en cuenta para estos cálculos la inclusión de parámetros y posibles predicciones. 3) Finalmente, se debe realizar una evaluación del modelo que ha sido ajustado a los datos.

2.1.1. Modelos jerárquicos bayesianos

Puede ocurrir que la densidad *a priori* para nuestros datos x, $P(x|\theta)$, dependa de r parámetros, es decir, $\theta=(\theta_1,\ldots,\theta_r)$, además, si los θ_i son independientes e idénticamente distribuidos, podemos creer adicionalmente, que estos r parámetros están relacionados de algún modo por medio de una densidad con su propio parámetro ϑ, al que llamaremos hiperparámetro. Si ϑ es desconocido, ento-

ces estamos ante una hiperdensidad *a priori* que representa nuestras creencias acerca de los datos [16]. Si el modelo es construido con este criterio, estamos bajo un modelo jerárquico y puede estar conformado por varios niveles dependiendo del modelo. Este concepto puede ser aplicado cuando se intenta agrupar los datos en regiones geográficas o áreas locales como en nuestro caso y se espera que al «tomar prestada la fuerza» de las otras regiones, se mejore la eficiencia al reducir el error estándar de la estimación de cada región en particular [12].

Dado que, de acuerdo con el modelo o fenómeno a modelar, se debe escoger una distribución para el cálculo de $p(x|\theta)$, esta debería ser escogida de modo que el producto $p(x|\theta)\ p(\theta)$ resulte en una distribución que sea equivalente o del mismo tipo que $p(\theta)$, es decir, pertenezca a la misma familia de la distribución *a priori*. Desde el punto de vista matemático, la ventaja de esta condición es que cuando se dispone de nuevos datos, el modelo se actualiza de forma automática. Suele llamarse *distribución conjugada* a aquella que cumple la condición descrita anteriormente, y por lo tanto se dice que la distribución $p(\theta)$ es la distribución conjugada para $p(x|\theta)$. Para este trabajo, dado que usaremos un modelo multinomial-Dirichlet, [11] diremos que la distribución Dirichlet es la conjugada de la distribución multinomial [17].

Las suposiciones para este modelo son las siguientes: 1) Ausencia de autocorrelación espacial, [4] y es necesaria para construir la función de máxima verosimilitud y establece que condicionando en Xi, Ti y Tj son independientes en la media. Las violaciones de esta suposición de maneras empíricamente razonables (e incluso algunas no razonables) no parecen inducir mucho sesgo [15]. 2) La suposición más crítica establece que Xi deben ser independiente de los $\beta_{rc}^{i\prime}s$. Esta hipótesis es equivalente a asumir la inexistencia de sesgo en la agregación, lo que es necesario para obtener estimaciones consistentes para los parámetros de la distribución. Aunque se trata de una suposición sólida, y a menudo no se puede justificar en la práctica, sirve como un punto de partida útil para desarrollar modelos en condiciones más generales [18].

2.1.2. Modelo jerárquico multinomial Dirichlet

Para describir el modelo (ver Tabla 1) definimos: $T_i'=T_{1i},T_{2i}',\ldots,T_{Ci}'$ que definen a todos los individuos votantes de algún recinto que apoyaron con su voto a alguno de los partidos políticos. Debiendo tratar al modelo de forma jerárquica, [11] se procede a construir el primer nivel tomando en cuenta que los valores T_i' siguen una distribución multinomial con vector de parámetros $\theta_i=(\theta_{1i},\theta_{2i}\ldots,\theta_{C,i})^t$ y cantidad , donde para bajo la suposición de que $\sum_{c=1}^{C}\theta_{ci}=1$.

El segundo nivel jerárquico se puede construir tomando las proporciones de apoyo $\beta_r{}^i=(\beta_{r1},\beta_{r2},\ldots,\beta_{r,C-1})^t$ con $i=1,\ldots,p$ y $r=1,\ldots R$ y vamos a asumir que siguen distribuciones de probabilidad Dirichlet con parámetros $(\alpha_{r1},\alpha_{r2},\alpha_{r3})$. Finalmente, tenemos a los , en los que asumiremos que siguen una distribución gamma con parámetros (λ_1,λ_2). En resumen, tenemos:

Primer nivel jerárquico

$$(T_{1i}',T_{2i}',T_{3i}')\sim Multinomial\left(N_i,\sum_{r=1}^{4}\beta_{r1i}X_{ri},\sum_{r=1}^{4}\beta_{r2i}X_{ri},\sum_{r=1}^{4}\beta_{r3i}X_{ri}\right) \quad (2)$$

N_i= número de personas habilitadas a ejercer el voto en el recinto i, con

$$\sum_{c=1}^{3}\theta_{ci}=1 \quad (3)$$

Segundo nivel jerárquico

$$(\beta_{r1i},\beta_{r2i},\beta_{r3i})\sim Dirichlet(\alpha_{r1},\alpha_{r2},\alpha_{r3}) \quad (4)$$

Tercer nivel jerárquico

$$\alpha_{rc}\sim Gamma(\lambda_1,\lambda_2) \quad (5)$$

El modelo jerárquico bayesiano obtenido puede ser ahora configurado en R mediante el paquete RStan. Stan es un paquete de software que crea muestras representativas de valores de parámetros de una distribución posterior para modelos jerárquicos complejos. Podemos especificar modelos para Stan y comunicarnos con Stan desde R a través de RStan. Stan utiliza un método Hamiltonian Monte Carlo (HMC). Stan opera con C ++ compilado y permite una mayor flexibilidad de programación, especialmente útil para modelos inusuales o complejos [19].

3. Métodos Markov Chain Monte Carlo

El cálculo mediante métodos numéricos de las integrales, que suelen aparecer en los modelos bayesianos suele ser muy complejo, por lo tanto, se desarrollaron métodos más eficientes como las técnicas Monte Carlo que nos permiten obtener de forma más eficiente los diferentes parámetros de una distribución *a posteriori*. Además, las conocidas cadenas de Markov sirven para obtener muestras de la distribución *a posteriori* en lugar de trabajar con la distribución en sí. De este modo, uniendo ambas ideas, tenemos los métodos Markov Chain Monte Carlo. Los denominados modelos de Gibbs o Metropoli realizan el trabajo de construir la densidad *a posteriori* a partir de muestras [20].

Existen muchas aplicaciones actualizadas y que siguen la idea de los algoritmos de Metropoli. Este trabajo utiliza el método de muestreo Hamiltoniano Monte Carlo que se encuentra programado en el paquete Stan y que puede trabajar con R para su aplicación. Fundamentalmente, este método está basado en conceptos de las ciencias físicas en las que se aplica la teoría hamiltoniana para un sistema físico emulando las variables de la posición y sus energías. El vector de parámetro se corresponde con

la posición de un cuerpo en un espacio k-dimensional y el cálculo de la energía potencial se corresponde con la probabilidad. Las muestras son generadas por la cadena de Markov y luego se determina la energía cinética inicial de la misma, con esto se puede encontrar la trayectoria de la partícula [21].

Una de las ventajas de este tipo de software, como RStan, es que los modelos jerárquicos que en principio pueden ser difíciles de programar o configurar, se los puede configurar nivel a nivel completando la jerarquía del modelo. Para esto, se construyen las distribuciones para los datos tomando en cuenta los parámetros que los comandan. Luego, de acuerdo a la teoría de modelos jerárquicos, se crea la distribución para los parámetros, pero esta vez sujetos a los hiperparámetros. Se realiza este procedimiento el número de veces necesario hasta obtener todos los niveles de jerarquía del modelo. Este procedimiento posibilita construir un modelo completo para las diferentes cantidades y nos permite establecer en fórmulas de probabilidad lo que nosotros creíamos acerca de los datos y sus relaciones. La parte que corresponde al cálculo de la distribución *a posteriori* y la integración en el espacio de probabilidades la realiza el paquete RStan mediante las técnicas ya mencionadas.

Las métricas así como son las que el paquete RStan utiliza para medir y evaluar si la simulación de un modelo converge o no. La métrica es una medida de la precisión que se obtiene en las diferentes simulaciones efectuadas a través del cálculo del tamaño efectivo de la muestra. No todas las muestras en una cadena de Markov son efectivas, por lo tanto, la métrica es el número de muestras que efectivamente trabajaron. Los métodos Markov Chain Monte Carlo suelen producir muestras correlacionadas al efectuar las diferentes cadenas, por tanto, al realizar las diferentes estimaciones, como en el caso de las medias *a posteriori*, no resultan ser tan precisas como lo serían si se tomaran muestras independientes. Por lo tanto, el número de muestras efectivas es en realidad, una estimación del número de muestras verdaderamente independientes que conducirían a la misma precisión en el modelo [11].

Por otro lado, la métrica (factor de reducción de potencial de escala) indica el factor de escala por el cual la desviación estándar de la distribución para un parámetro determinado podría reducirse si el número de simulaciones tiende al infinito. Cuando la simulación alcanza la convergencia en la distribución *a posteriori* para el parámetro escogido, entonces el valor de debe ser 1. Debemos tomar en cuenta que estos procedimientos no equivalen a llevar a cabo una prueba de hipótesis, esto implica que no existe algún valor-p como valor de aceptación o rechazo de hipótesis y tampoco se tiene alguna significancia estadística, sino que más bien se evalúa la discrepancia de la convergencia de la distribución de forma práctica [17].

Se muestra el código de programación para el modelo de inferencia construido en lenguaje Stan:

```
data{
int N;  // número de observaciones
int R; // número de grupos edad
int C; // categorías
int fcorrea [N]; // personas que apoyan AP
int flasso [N]; // personas que apoyan CREO
int otros [N]; // personas que apoyan otros
int ni [N]; // número total de votantes
real g1 [N]; // número de personas entre 16 y 29 años
real g2 [N]; // número de personas entre 30 y 44 años
real g3 [N]; // número de personas entre 45 y 56 años
real g4 [N]; // número de personas de 60 o más años
}
transformed data{
int Ti [N ,3];
  for (n in 1:N) {
      Ti [n,1]=correa[n];
      Ti [n,2]=lasso[n];
      Ti [n,3]=Fotros[n];
     }
}
parameters {
   fsimplex [C]   fbetas [N, R];
   vector < lower=0.01>   [C]   falfas [N,R] ;
}

transformed parameters{
     fsimplex [C]  ftheta [N] ;
     for (n in 1: N) {
      theta [n,1] =fbetas [n,1,1]*g1 [n] +fbetas [n,2,1]
*g2 [n] +fbetas [n,3,1]*g3 [n] +fbetas [n,4,1]*g4 [n] ;
      ftheta [n,2] =fbetas [n,1,2]*g1 [n]+fbetas [n,2,2]
*g2 [n] +fbetas [n,3,2]*g3 [n] +fbetas [n,4,2]*g4[n] ;
      ftheta [n,3]=fbetas [n,1,3]*g1 [n] +fbetas [n,2,3]
*g2 [n]+fbetas [n,3,3]*g3 [n] +fbetas [n,4,3]*g4 [n];
    }
}
model {
  for (n in 1:N){
    for (i in 1: R) {
       for (j in 1:C){
          falfas [n,i,j]~fgamma(4,2);
        }
    }
}
for (n in 1: N) {
    for(i in 1: R) {
         fbetas [n,i ]~fdirichIet (falfas [n,i]);
    }
}
for (n in 1: N) {
     Ti[n]~fmultinomial (ftheta [n]);
   }
}
```

Tabla 2.
Parámetros simulados obtenidos para el recinto Camilo Ponce, en cada uno de los grupos etarios.

	mean	sd_mean	n_{eff}	*Rhat*
betas[1,1,1]	0.69	0.1	1311	1.00
betas[1,1,2]	0.13	0.06	1616	1.00
betas[1,1,3]	0.21	0.07	1227	1.00
betas[1,2,1]	0.51	0.13	1321	1.00
betas[1,2,2]	0.20	0.08	1641	1.00
betas[1,2,3]	0.31	0.14	1162	1.00
betas[1,3,1]	0.39	0.18	2000	1.00
betas[1,3,2]	0.22	0.10	2000	1.00
betas[1,3,3]	0.29	0.11	2000	1.00
betas[1,4,1]	0.41	0.14	2000	1.00
betas[1,4,2]	0.27	0.14	2000	1.00
betas[1,4,3]	0.31	0.15	2000	1.00

Fuente. Elaboración propia.

4. Resultados y discusión

Esta sección muestra algunos resultados de las simulaciones realizadas para inferir el apoyo que recibieron los partidos AP y CREO en la contienda electoral presidencial del 2013. Se tomaron en cuenta 1223 recintos electorales (parroquias) que se encuentran a lo largo del territorio nacional y se presentan en todas las provincias del país, por lo tanto $p=1223$. El conjunto de personas habilitadas para el sufragio fue repartido en 4 clases según su edad, por lo tanto $R=4$. El estudio se realiza sobre los dos partidos políticos que obtuvieron la mayor votación, quedando un grupo de votantes que apoyó al resto de partidos no relevantes junto con los votos nulos y abstenciones, por lo tanto $C=3$. El objetivo es conocer cuánto apoyo recibieron estos dos partidos políticos de parte de la población votante. Las proporciones de apoyo son calculadas estadísticamente e indexadas por Stan siempre que se haya configurado el modelo de forma adecuada. La indexación se da a cabo mediante tres subíndices: i,j,k. El índice i corresponde al recinto (parroquia), el índice j corresponde al grupo de edad y el índice k corresponde al partido político al que apoyan cada uno de los individuos votantes.

La ejecución se realizó mediante cuatro cadenas. Para estos procesos, las iteraciones suelen dividirse en tipo calentamiento y tipo posterior. En este caso, se asignaron 500 iteraciones de calentamiento y 500 posteriores al mismo. Con esto se obtiene una muestra de 2000 simulaciones para la distribución *a posteriori* en cada uno de los parámetros, con esto se espera alcanzar la convergencia del método.

La tabla 2 ofrece los resultados de la simulación, ejecutados específicamente para el recinto Camilo Ponce que pertenece a la provincia del Azuay. Los parámetros que han sido simulados por Stan son: media, desviación estándar de la media, y para cada una de las distribuciones *a posteriori* para esas cantidades. El método alcanzó la convergencia en las distribuciones para cada uno de los parámetros, esto lo verifica el valor que para todos los casos es 1 (ver Tabla 2).

Los denominados *traceplots* son las gráficas que Stan muestra para la observación de las cadenas generadas en las simulaciones y en las que se puede apreciar visualmente la convergencia de las simulaciones (ver Figura 1). Se muestra los *traceplot* de las cadenas generadas por el método para el caso del recinto de Cumbayá que pertenece a la provincia de Pichincha. Estos *traceplots* corresponden a las simulaciones de las proporciones de apoyo hacia el partido político AP que recibió de los grupos de edad generados a partir de los datos de población. Los diferentes colores que se aprecian en los traceplots corresponden a las cuatro cadenas simuladas para los diferentes parámetros en el recinto definido. La forma visual de comprobar si las cadenas convergen es verificando que las iteraciones estén centradas alrededor de un mismo valor en forma horizontal.

Stan puede elaborar también las distribuciones de probabilidad *a posteriori* para las variables de interés. Se presentan estas gráficas (ver Figura 2) para el recinto de Cumbayá. Nuevamente, las iteraciones se dividieron en 500 de calentamiento, 500 de poscalentamiento y 1000 iteraciones para la simulación propiamente. Los diferentes colores corresponden a cada una de las cadenas y podemos ver que cada cadena converge prácticamente a la misma distribución.

Una manera general de ver el conjunto de resultados puede ser a través de mapas con escalas de color que muestran el nivel de apoyo hacia los partidos políticos

Figura 1.
Traceplots para las variables de interés del apoyo a AP en el recinto de Cumbayá

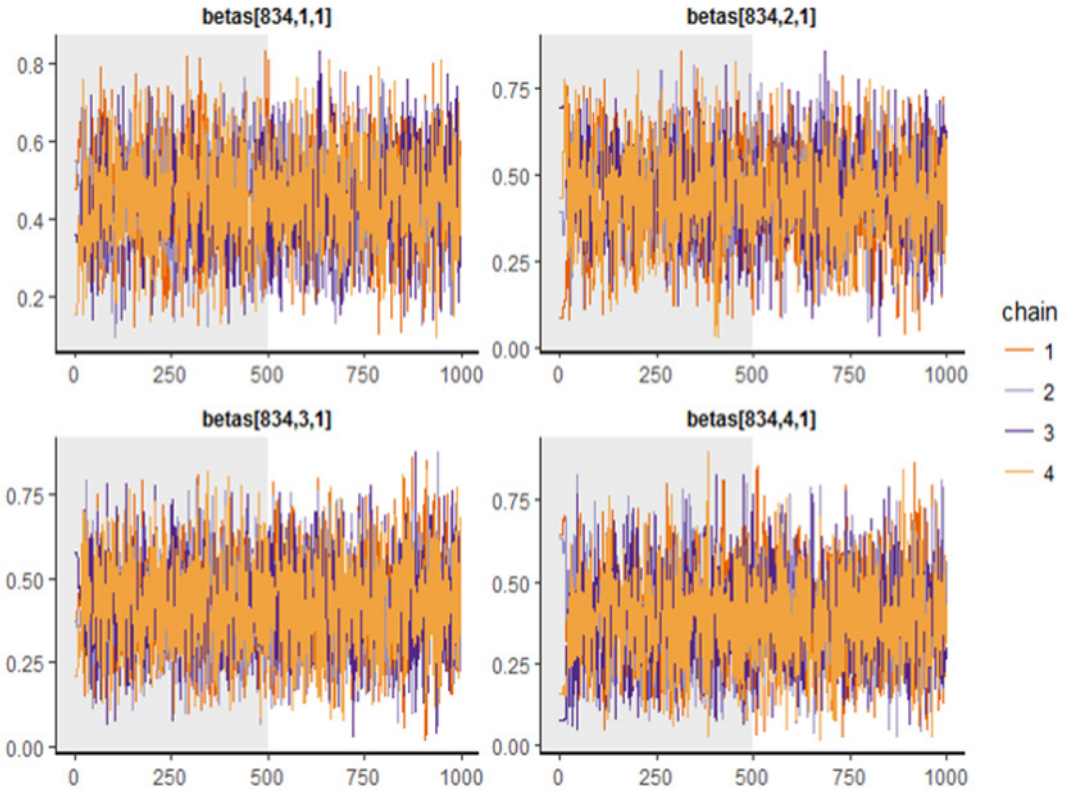
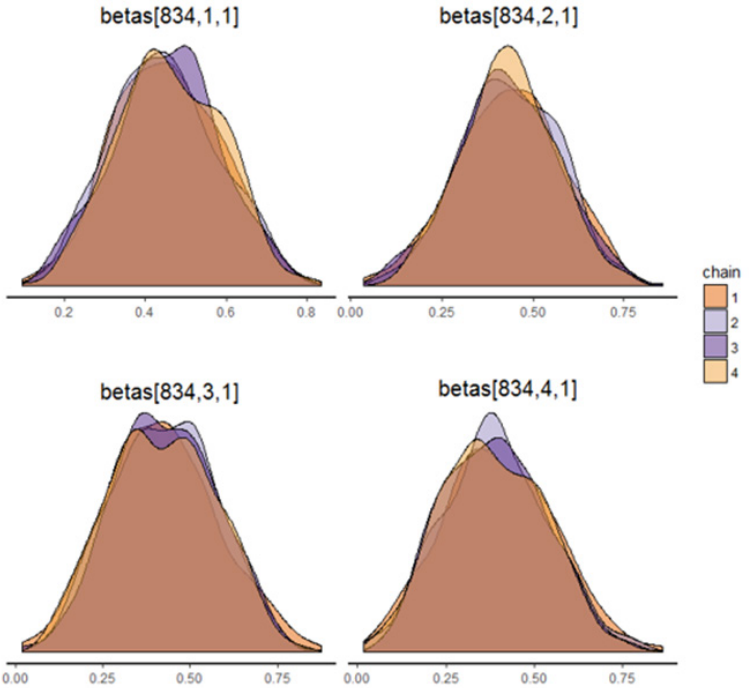
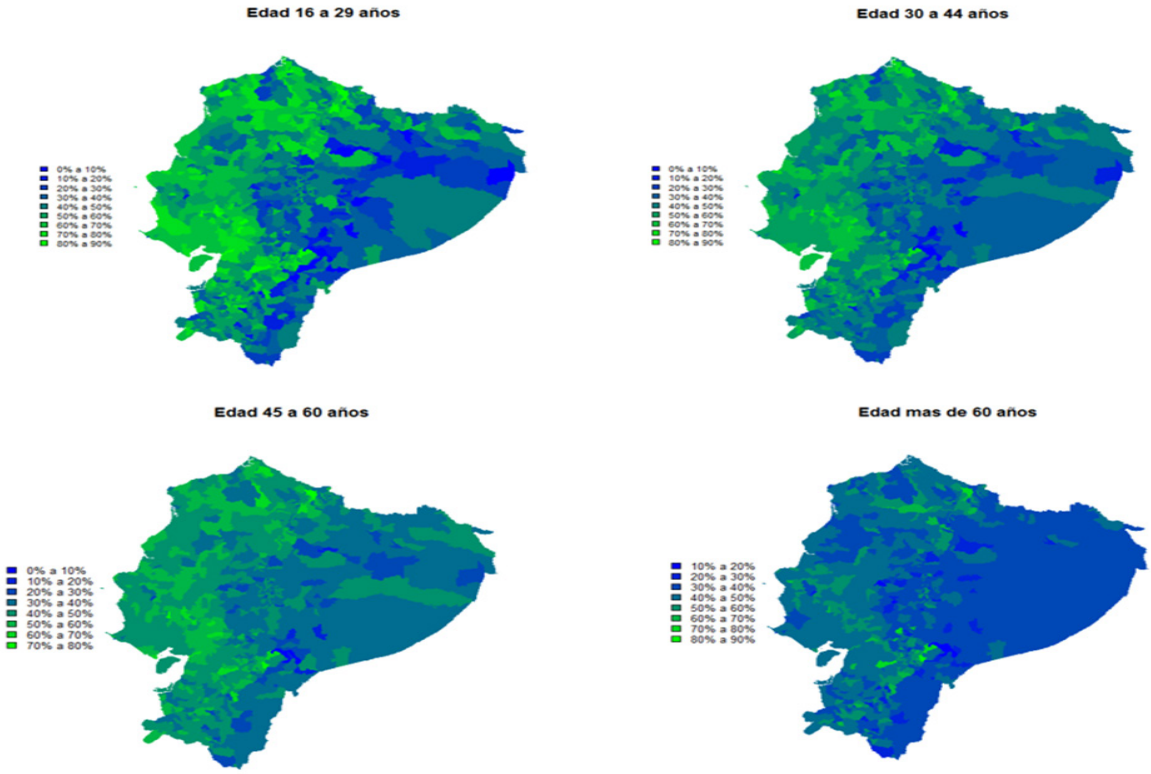


Figura 2.
Distribuciones de probabilidad a posteriori para las proporciones de apoyo de las diferentes clases etarias hacia AP en el **recinto Cumbayá**



AP y CREO dependiendo del color y su intensidad. La figura 3 muestra el mapa descrito para el apoyo que recibió el partido AP. Se puede observar de forma general que la mayor proporción de apoyo que este partido recibió corresponde al primer grupo etario, es decir, los votantes entre 16 y 29 años. Podemos apreciar que estas proporciones oscilan aproximadamente entre 50% y 70%. Además, los mapas muestran que la parte costera y una gran

Figura 3.
Apoyo que recibe AP de parte de las clases etarias en las elecciones presidenciales 2013



porción de la región andina fueron quienes apoyaron en mayor medida a este partido. La región oriental no parece presentar altos valores de apoyo. Además, parece ser que el apoyo hacia AP disminuye de forma considerable cuando aumenta la edad de los votantes (ver Figura 3).

La figura 4 muestra las proporciones de apoyo hacia el partido político CREO. La gráfica muestra que las clases etarias inferiores apoyaron en menor medida que las clases etarias más altas. Por ejemplo, para la clase etaria que está entre los 16 hasta los 29 años, se observa un apoyo que está en un rango entre el 10% y 30%. Este apoyo parece crecer a medida que la clase etaria crece. El mapa no presenta regionalismo en el apoyo de las clases etarias hacia este partido político ya que la intensidad en los colores se mantiene en forma general (ver Figura 4).

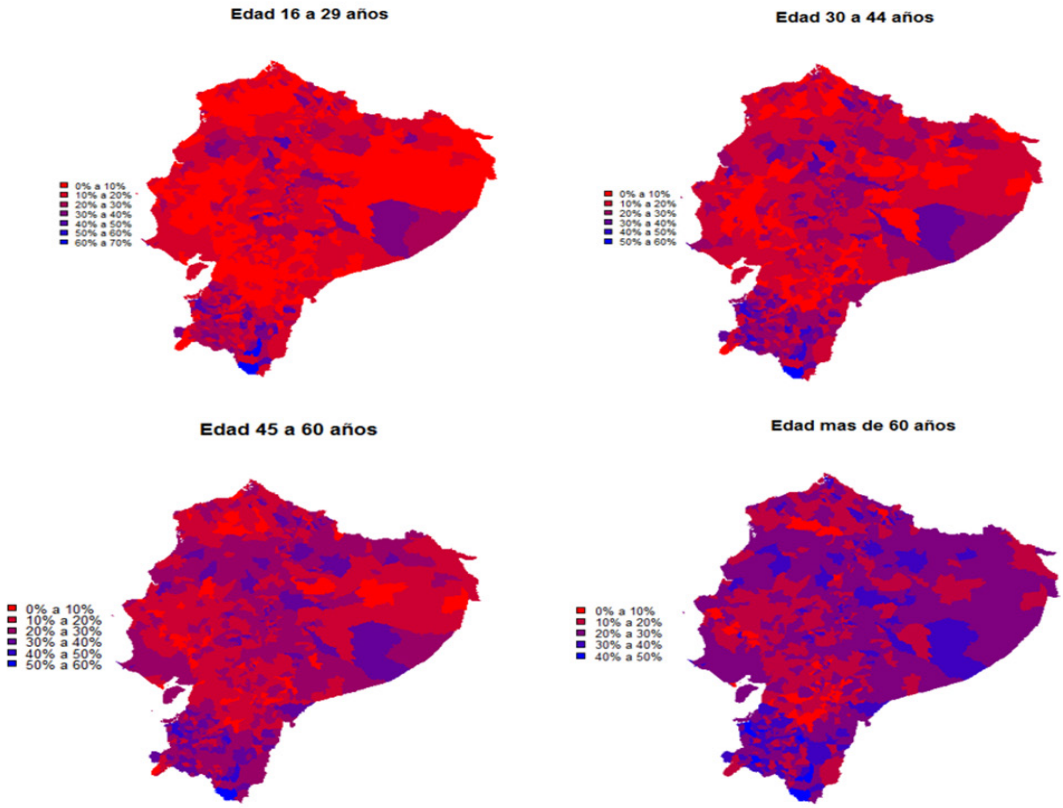
Este trabajo ha empleado métodos Markov Chain Monte Carlo para efectuar la inferencia sobre las proporciones de apoyo desde ciertas clases etarias a partidos políticos determinados en elecciones presidenciales. La inferencia a través de estos algoritmos se basa en conseguir una distribución de probabilidad muestreada *a posteriori* por lo que se hizo necesario el uso de simulaciones que utilizan cálculos computacionalmente caros. A cambio de esto se tienen ventajas: 1) Sus cálculos logran gran eficiencia, por ejemplo, en los intervalos de confianza que pueden ser más finos [11]. 2) Las probabilidades para las proporciones de apoyo *a posteriori* pudieron ser calculadas, esto es algo que no todos los paquetes calculan.

Comparando la metodología utilizada en este trabajo con [22] se concluye que, en general, el uso de metodologías bayesianas con modelos que muchas veces son jerárquicos es más eficiente que el uso de otros métodos que también pueden usar simulaciones. Se debe tomar en cuenta el cumplimiento de las suposiciones para el buen desarrollo del modelo, aunque en la práctica difícilmente se cumple con alguna de estas, por ejemplo, el supuesto de distribución del modelo.

El análisis realizado en este trabajo podría ser mejorado si se utilizan covariables en la información obtenida para los recintos electorales. Es muy conocida la utilidad que poseen las covariables en un modelo al permitir que los parámetros de interés varíen en función de ellas. Esto implica también que las distribuciones sean más flexibles, es decir, será posible trabajar con densidades más complejas. Por ejemplo, al condicionar al modelo sobre los , es posible modelar la relación que existe entre estos datos y los parámetros en lugar de asumir que ambos son independientes *a priori* como en [12].

Las desviaciones estándar para las medias en la tabla 3 tienden a ser más bajas que en los resultados del modelo MCMC Multinomial-Dirichlet, [11] esto puede deberse a que en este trabajo se utilizó el método HMC que proporciona RStan en comparación con el sampleador de Gibbs

Figura 4.
Apoyo por clases etarias hacia el partido político CREO en las elecciones 2013



utilizado en ese artículo y la cantidad de recintos que se utilizaron en este trabajo.

Stan utilizó 2000 iteraciones que fueron configuradas en principio y que fueron suficientes para llegar a la convergencia del método, en el caso de [12] se usaron 3.000.000 iteraciones mediante el paquete WinBUGS que también construye distribuciones *a posteriori*. Los *traceplots* que se mostraron en este trabajo demostraron que las cadenas efectivamente convergen mientras que en [12] no ocurre esto.

5. Conclusiones

Varios enfoques se han desarrollado para la solución del problema de la inferencia ecológica; en este trabajo se ha tomado un enfoque bayesiano mediante el modelo Multinomial-Dirichlet y el método Hamiltonian Monte Carlo para la simulación. A diferencia de los métodos estadísticos frecuentistas, los métodos bayesianos se basan en la formulación de un conjunto de distribuciones previas para los parámetros desconocidos que se basan en creencias *a priori* del investigador. Tales distribuciones previas son parte del modelo estadístico, así como la parte que expresa la distribución de probabilidad de las observaciones dadas a través del cálculo de su verosimilitud.

RStan presenta una alta efectividad al trabajar con modelos jerárquicos bayesianos. Por medio de su método HMC, los parámetros de interés fueron estimados y las cadenas llegaron a la convergencia de forma efectiva. Para mejorar la exactitud de las inferencias, es recomendable la formulación del modelo incluyendo covariables que permitan una flexibilización del modelo con respecto a las suposiciones del mismo.

Referencias

[1] «Consejo Nacional Electoral Ecuador». http://cne.gob.ec/es/ (accessed aug. 11, 2021).

[2] D. Duncan and B. Davis, «An alternative to ecological correlation», *Am. Sociol. Rev.*, vol. 18, N.º 6, dec. 1953, p. 665.

[3] L. A. Goodman, «Ecological regressions and behavior of individuals», *Am. Sociol. Rev.*, vol. 18, N.º 6, dec. 1953, p. 663.

[4] G. King, *A solution to the ecological inference problem*, 1st ed., Princeton: Princeton University Press, 1997.

[5] S. R. Flaxman, Y. X. Wang and A. J. Smola, «Who supported Obama in 2012? Ecological inference through distribution regression», in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery*

and Data Mining, 2015, pp. 289-298.

[6] J. Pavia and R. Romero, «Improving estimates accuracy of voter transitions. Two new algorithms for ecological inference based on linear programming», *Advance*, 2021, pp. 124, jun. [Online]. Available: /articles/preprint/Improving_estimates_accuracy_of_voter_transitions_Two_new_algorithms_for_ecological_inference_based_on_linear_programming/14716638/1.

[7] P. Sandoval and S. Ojeda, «Estimation of electoral volatility parameters employing ecological inference methods». https://www.researchgate.net/publication/338951636_Estimation_of_Electoral_Volatility_parameters_employing_Ecological_inference_methods/references (accessed aug. 11, 2021).

[8] S. Flaxman, D. Sutherland, Y.-X. Wang and Y. W. Teh, «Understanding the 2016 US presidential election using ecological inference and distribution regression with census microdata», 2016.

[9] G. King, O. Rosen, M. Tanner and A. Wagner, «Ordinary economic voting behavior in the extraordinary election of Adolf Hitler», *J. Econ. Hist.*, vol. 68, N.º 4, 2008.

[10] E. Castela, «Inferencia ecológica para la caracterización de abstencionistas: el caso de Portugal», *Discussion Papers - Spatial and Organizational Dynamics*, N.º 3.

[11] O. Rosen, W. Jiang, G. King and M. Tanner, «Bayesian & Frequentist inference for ecological inference: the RxC case», *Stat. Neerl.*, vol. 55, 2001.

[12] G. King, O. Rosen and M. Tanner, «Binomial beta models for ecological inference», *Sociol. Methods Res.*, vol. 28, 1999.

[13] «Instituto Nacional de Estadística y Censos». https://www.ecuadorencifras.gob.ec/institucional/home/ (accessed aug. 11, 2021).

[14] C. Plescia and L. De Sio, «An evaluation of the performance and suitability of R × C methods for ecological inference with known true values», *Qual. Quant.*, vol. 52, N.º 2, mar. 2018, pp. 669-683. DOI: 10.1007/s11135-017-0481-z.

[15] G. King, O. Rosen and M. A. Tanner, «Ecological inference: new methodological strategies», *Ecol. Inference New Methodol. Strateg*, jan. 2004, pp. 1-431. DOI: 10.1017/CBO9780511510595.

[16] P. Lee, *Bayesian statistics: an introduction*, Wiley, 2012.

[17] A. Gelman, B. Carlin, H. Stern, and B. Rubin, *Bayesian data analysis*, third edition (Statistical Science), 2014.

[18] K. Imai and Y. Lu, «Bayesian and Likelihood inference for 2 x 2 ecological tables: an incomplete-data approach», *Polit. Anal.*, vol. 16, 2007, pp. 41-69, DOI: 10.1093/pan/mpm017.

[19] J. Kruschke, *Doing bayesian data analysis: a tutorial with R, JAGS, and Stan*, second edition, Elsevier Science, 2014.

[20] M. A. Tanner, *Tools for statistical inference: methods for the exploration of posterior distributions and likelihood functions*, Springer, 1996.

[21] «Stan Reference Manual». https://mc-stan.org/docs/2_21/reference-manual/index.html (accessed dec. 10, 2019).

[22] A. Klima, T. Schlesinger, P. W. Thurner and H. Küchenhoff, «Combining aggregate data and exit polls for the estimation of voter transitions», *Sociol. Methods Res.*, vol. 48, N.º 2, may 2017, pp. 296-325.