

Memorias del
**XXXI Foro Internacional
de Estadística**

y del

**XXXII Foro Nacional
de Estadística**



**Memorias del XXXI Foro Internacional
de Estadística y del XXXII Foro Nacional
de Estadística**



Obras complementarias publicadas por el INEGI sobre el tema:
Memoria del XXIII Foro Nacional de Estadística.

Catalogación en la fuente INEGI:

310.4 Memorias del XXXI Foro Internacional de Estadística y del XXXII Foro Nacional de Estadística / Instituto Nacional de Estadística y Geografía, Asociación Mexicana de Estadística.-- México : INEGI, c2018.

v, 227 p.

“Foro Internacional de Estadística. Universidad Autónoma de Chapingo. Texcoco, Estado de México, del 26-30 de septiembre de 2016”; “Foro Nacional de Estadística. Universidad Nacional Autónoma de México. Ciudad de México, del 25-29 de septiembre de 2017”

1. Estadística - Alocuciones, Ensayos, Conferencias. I. Instituto Nacional de Estadística y Geografía (México); II. Asociación Mexicana de Estadística

Conociendo México

01 800 111 4634
www.inegi.org.mx
atencion.usuarios@inegi.org.mx



Presentación

Bajo la coordinación de la Asociación Mexicana de Estadística, anualmente se realiza un foro de Estadística en donde académicos y estudiantes de todo el país presentan trabajos de divulgación e investigación en el área. En su XXXI edición, el Foro Internacional de Estadística, organizado por la Universidad Autónoma Chapingo, se llevó a cabo en la ciudad de Texcoco, México, del 26 al 30 de septiembre de 2016, teniendo como anfitrión a la División de Ciencias Forestales de esta Universidad. Para el siguiente año, del 25 al 29 de septiembre del 2017, la Universidad Nacional Autónoma de México fue la encargada de organizar el XXXII Foro Nacional de Estadística en la Unidad de Posgrado de Ciudad Universitaria.

Los resúmenes *in extenso* recibidos se sometieron a un proceso de revisión por pares y de estilo, con el fin de asegurar que el abordaje estadístico y la redacción fueran correctos. Durante este proceso, se buscó que existiera un mínimo de originalidad o aspectos novedosos en la metodología, resultados y/o aplicaciones presentadas. De esta forma, los resúmenes contenidos en este volumen corresponden a aquellos trabajos que recibieron un dictamen favorable por parte de los árbitros.

Agradecemos a todos los autores por su participación y por la calidad de los trabajos presentados. De igual manera, expresamos nuestro agradecimiento a todos los árbitros por su invaluable colaboración. A nombre de la Asociación Mexicana de Estadística, agradecemos también a la Universidad Autónoma de Chapingo y a la Universidad Nacional Autónoma de México por el apoyo, entusiasmo y dedicación en la organización de estos Foros, así como al Instituto Nacional de Estadística y Geografía por el apoyo en la publicación del presente volumen.

El Comité Editorial:
Asael Fabian Martínez Martínez,
Lizbeth Naranjo Albarrán,
Paulino Pérez Rodríguez,
Luz Judith Rodríguez Esparza,
Carlos Erwin Rodriguez Hernández Vela.

Contenido

XXXI Foro Internacional de Estadística

Modelación Geoespacial de la Contaminación Atmosférica en la Ciudad de México	3
<i>Alejandro Ivan Aguirre Salado, Silvia Reyes Mora, Ana Delia Olvera Cervantes, Humberto Vaquera Huerta, Carlos Arturo Aguirre Salado</i>	
Uso de Componentes Principales y Correlación Canónica para la Caracterización de Poblaciones de Maíz Nativo	15
<i>Juan Elías Solís Alonso, María Gúzman Martínez, Dolores Briones Reyes, Flaviano Godínez Jaimes, Elisa Martínez Azoños</i>	
Análisis Bayesiano de un Modelo Gamma Bivariado Bajo Proyección	29
<i>Gabriel Nuñez Antonio, Juan de Dios Aguilar Gámez, Emiliano Geneyro Squarzon, Gabriel Escarela Pérez</i>	
¿Este Ítem Funciona Igual para Todos? ¿Quién lo Dice? Análisis DIF con Distintos Métodos. Coincidencias y Discrepancias	41
<i>Alma Yadhira López García</i>	
Una Propuesta Bayesiana para Medir el Grado de Traslape Entre Dos Especies de Animales	53
<i>Gabriel Núñez Antonio, Alberto Contreras Cristán, Eduardo Gutiérrez Peña, Manuel Mendoza Ramírez, Eduardo Mendoza Ramírez</i>	
Estudio Morfométrico de la Plaga <i>B. Cockerelli (Sulc)</i> en Dos Variedades de Jitomate Mediante Análisis Factorial y Componentes Principales	61
<i>Eduardo Pérez Castro, María Guzmán Martínez, Ramón Reyes Carreto, David Alejandro</i>	

Ozuna Santiago, Haidel Vargas Madriz

- Un Método para Construcción de Pruebas de No Inferioridad con Regiones Críticas Convexas.** 75
José Juan Castro Alva, Hortensia Josefina Reyes Cervantes, Félix Almendra Arao

- Averaged Shifted Histograms (ASH) or Weighted Averaging of Rounded Points (WARP), Efficient Methods to Calculate Kernel Density Estimators for Circular Data** 89
Isaías Hazarmabeth Salgado-Ugarte, Verónica Mitsui Saito-Quezada, Marco Aurelio Pérez-Hernández

- Pruebas de No Inferioridad Comparando Dos Distribuciones Poisson** 97
María de Lourdes Morales Sánchez, Hortensia Reyes Cervantes, Félix Almendra Arao

- Pruebas de Correlación Máxima, Correlación de Distancia y Covarianza para Optimización en el Problema de Selección de Variable. Avance de Investigación** 111
Yamil Burguete Fourzali, Gustavo Ramírez Valverde, David Sotres Ramos, Benito Ramírez Valverde

- La Teoría Estable Acotada. Una Alternativa para Predecir el Estado Estable del Saldo Neto Migratorio en México** 123
Javier González Rosas, Iliana Zárate Gutiérrez

- Inferencia sobre Modelos Epidemiológicos en Redes de Contactos** 135
Rocío M. Ávila Ayala, J. Andrés Christen Gracia, L. Leticia Ramírez Ramírez

- La Trata de Personas en México: Un Modelo para Identificar Patrones de Conducta de Posibles Tratantes** 149
Paulina Martínez Rosas, Blanca Rosa Pérez Salvador

- Proyecciones Aleatorias Tipo Random Fourier Features Basadas en Información Distribucional para Kernel PCA** 157
Flor de María Martínez Sermeño, Johan Van Horebeek

Pronóstico del ITAEE del Estado de Veracruz a través de la Metodología Box-Jenkins	169
<i>Ángel Luis López Morales, Juan Ruiz Ramírez, Edson Valdés Iglesias</i>	
Visualización de Datos Espaciales en R: Elecciones Gubernamentales 2016 en Zacatecas	181
<i>Iván Pacheco Soto</i>	

XXXII Foro Nacional de Estadística

Análisis Estadístico de Trayectorias sobre la Esfera: un Caso de Estadística sobre Variedades	193
<i>Lilia Karen Rivera Escovar</i>	
Desempeño de Intervalos de Confianza para una Proporción y Criterios para su Aplicación	205
<i>Marcos Morales Cortés, Hortensia J. Reyes Cervantes, Félix Almendra Arao</i>	
JRStat: Una Plataforma de Código Abierto para Implementación de Análisis Estadístico Usando Interfaces Gráficas.	215
<i>Nallely Izel Bautista Pérez, Paulino Pérez Rodríguez</i>	

**Trabajos presentados en el
XXXI Foro Internacional de
Estadística**

Modelación Geoespacial de la Contaminación Atmosférica en la Ciudad de México

Alejandro Ivan Aguirre Salado^a, Silvia Reyes Mora, Ana Delia Olvera Cervantes

Universidad Tecnológica de la Mixteca

Humberto Vaquera Huerta
Colegio de Postgraduados

Carlos Arturo Aguirre Salado
Universidad Autónoma de San Luis Potosí

Se realizó un análisis de eventos extremos de contaminación por partículas menores a 10 micrómetros (PM10) en la zona metropolitana de la Ciudad de México mediante modelos de valores extremos no estacionarios para datos con censura aleatoria. La estimación de parámetros se realizó empleando un enfoque jerárquico bayesiano. En este trabajo, utilizamos funciones de base radial gaussianas para modelar el efecto de las covariables en la distribución generalizada de valores extremos (GEV). Los resultados mostraron una tendencia espacial en el mapa de los períodos de retorno de contaminación por PM10 a 25 años.

Área-MSC: Estadística Aplicada, Estadística Bayesiana.

Subárea-MSC: Contaminación del aire, Teoría de valores extremos, No estacionariedad, MCMC.

1. Introducción

El análisis de los valores extremos ha sido usado en muchas áreas tales como las ciencias ambientales y financieras, entre otras. En particular, el análisis de valores extremos en datos espaciales, tomados en diferentes momentos del tiempo, ha venido ganando interés porque estas condiciones se presentan en una gran cantidad de situaciones reales, por ejemplo, datos diarios de temperatura, cantidad de lluvia y contaminación, entre otros. Este tipo de datos

^aaleaguirre@mixteco.utm.mx

son registrados en estaciones dentro de una región a lo largo del tiempo (Sang and Gelfand, 2010). En algunos casos, en el proceso de la recolección de datos se llegan a presentar problemas en la medición, lo que da lugar a observaciones censuradas. En tales casos, sería un error grave analizar la información sin considerar esta situación. De esta manera, en este trabajo proponemos una modificación en la teoría estándar de los valores extremos para poder analizar el caso cuando se tienen datos censurados y de esta manera modelar satisfactoriamente la contaminación espacial de PM10 en la Ciudad de México.

Los valores máximos pueden ser modelados mediante la distribución de valores extremos que emplea tres parámetros correspondientes a la localización, la escala y la forma (Jenkinson, 1955). La estimación de los parámetros se realiza generalmente por el método de máxima verosimilitud (Smith, 1985), sin embargo, este método no suele ser robusto cuando el tamaño de muestra es pequeño, es por eso que se han propuesto otros estimadores, tales como el método de momentos, el de L momentos y el método de momentos con probabilidades ponderadas (Hosking et al., 1985; Hosking, 1990; Madsen et al., 1997).

Recientemente se han propuesto nuevas metodologías para estudiar los valores extremos, principalmente aplicadas a diversos datos climatológicos, todas ellas teniendo como punto central la distribución de valores extremos, así es como Gaetan and Grigoletto (2007) propusieron usar campos aleatorios de Markov aproximados por suavizamiento kernel para modelación de los parámetros de la distribución GEV, Reich et al. (2014) estudió las ondas de calor mediante un modelo jerárquico bayesiano con distribución generalizada de Pareto y asignó a los parámetros un modelo de markov dependiente del tiempo. (Cooley and Sain, 2010) estudiaron eventos de precipitaciones máximas asignando un modelo normal con covariables temporales a los parámetros de la distribución GPD. Sang and Gelfand (2010) estudiaron procesos estocásticos espaciales para valores extremos y modelaron la tendencia como función de covariables.

En un escenario real, como el caso de máximos climatológicos, es común que las condiciones futuras cambien y que los supuestos de estacionariedad requeridos en el análisis tradicional de valores extremos no se cumplan, esto debido a que generalmente existen tendencias en valores extremos (Leadbetter et al., 1983; Wang et al., 2004; Kharin and Zwiers, 2005). Por ello, varios investigadores han introducido funciones de covariables dentro de la distribución de valores extremos, para modelar ya sea el parámetro de localización o el

parámetro de escala. Por ejemplo, para el parámetro de escala, Weissman (1978) empleó una función senoidal, Scarf (1992) y Tawn (1988) propusieron una función lineal, Rosen and Cohen (1996) y Pauli and Coles (2001) usaron splines; para modelar el parámetro de escala, Yee and Stephenson (2007) usaron modelos aditivos para modelar el logaritmo del parámetro de escala, Rodriguez et al. (2012) y El Adlouni et al. (2007) emplearon funciones lineales, Cannon (2010) propuso el uso de redes neuronales.

2. Metodología

2.1. Máximos por Bloques

Uno de los métodos para obtener los valores extremos de una serie, es el llamado máximo por bloques, en donde se dividen los datos en secciones de igual tamaño y se escoge el valor más grande dentro de cada bloque. La ventaja de este método es que se escogen valores sobre todo el conjunto de datos, sin embargo, se pueden omitir los siguientes valores extremos dentro del mismo bloque que posiblemente sean mayores que el máximo dentro de otro bloque. Entonces se asume que la distribución que siguen los máximos por bloques es la distribución de valores extremos generalizada (GEV del inglés Generalized Extreme Value distribution).

2.2. Distribución GEV

Dada una muestra de variables aleatorias independientes e idénticamente distribuidas, $X_1, \dots, X_n \sim F_X(x)$. El valor extremo es definido como:

$$M_n = \max \{X_i | i = 1, \dots, n\}$$

Puesto que las X 's son independientes,

$$F_{M_n}(x) = P(M_n \leq x) = P(X_1 \leq x, \dots, X_n \leq x) = (F_X(x))^n$$

Dado que $0 \leq F_X(x) \leq 1$ entonces $(F_X(x))^n \rightarrow \{0, 1\}$. Para resolver esto $F_X(x)$ es reescalada por constantes μ y σ tal que:

$$M_n^* = \frac{M_n - \mu}{\sigma}$$

La distribución de valores extremos, fue propuesta originalmente por Fisher and Tippett (1928) e incluía tres familias: Gumbel, Fréchet y Weibull. Posteriormente Jenkinson (1955) combinó las tres familias en la distribución de valores extremos (GEV):

$$F(x, \mu, \alpha, \kappa) = \begin{cases} \exp \left\{ - \left(1 + \kappa \frac{(y-\mu)}{\sigma} \right)^{-\frac{1}{\kappa}} \right\}, & \text{si } \kappa \neq 0, 1 - \kappa \frac{(y-\mu)}{\sigma} > 0; \\ \exp \left\{ - \exp \left(- \frac{(y-\mu)}{\sigma} \right) \right\}, & \text{si } \kappa = 0. \end{cases}$$

Con función de densidad de probabilidades dada por:

$$f(y, \mu, \alpha, \kappa) = \begin{cases} \frac{1}{\sigma} \left\{ \left(1 + \kappa \frac{(y-\mu)}{\sigma} \right)^{-\frac{(1+\kappa)}{\kappa}} \right\} \exp \left\{ - \left(1 + \kappa \frac{(y-\mu)}{\sigma} \right)^{-\frac{1}{\kappa}} \right\}, & \text{si } \kappa \neq 0, 1 - \kappa \frac{(y-\mu)}{\sigma} > 0; \\ \exp \left\{ - \frac{(y-\mu)}{\sigma} \right\} \exp \left\{ - \exp \left(- \frac{(y-\mu)}{\sigma} \right) \right\}, & \text{si } \kappa = 0. \end{cases}$$

Donde $\mu + \sigma/\kappa \leq y \leq +\infty$ cuando $\kappa < 0$ (Fréchet), $-\infty \leq y \leq +\infty$ cuando $\kappa = 0$ (Gumbel)

y $-\infty \leq y \leq \mu + \sigma/\kappa$ cuando $\kappa > 0$ (Weibull). Aquí $\mu \in \mathbb{R}$, $\sigma > 0$ y $\kappa \in \mathbb{R}$ son los parámetros de localidad, escala y forma respectivamente.

2.3. Censura Aleatoria

En la medición de fenómenos ambientales, es frecuente que se tengan problemas en las mediciones debido a la precisión de los equipos, lo que nos genera de datos censurados. Así, nosotros obtenemos los valores máximos M_i , en bloques temporales de información y asumimos que en tal bloque puede existir un valor censurado C_i . Asumimos que las variables M_i y C_i son independientes. Sea $Y_i = \min(M_i, C_i)$ y sea δ_i la variable que indica si el valor de Y_i es censurado $\delta_i = 0$ o no $\delta_i = 1$. Entonces la distribución de M_i , esta dada por:

$$M_i \sim GEV(\mu_i, \alpha_i, \kappa)$$

Sea G la función de distribución de M_i , y g su función de densidad de probabilidades, asuma que $\theta_i = (\mu_i, \alpha_i, \kappa)$. Similarmente sea F y f la función de distribución y la función de densidad de C_i , respectivamente.

Sea $S_i = M_i \wedge C_i$ y $\delta_i = 1$ ($Y_i = M_i$). Así, Y_i es el valor observado y posiblemente censurado valor extremo. Para datos con censura aleatoria tenemos:

$$\begin{aligned} P[Y_i = y, \delta_i = 1; s_i, \theta_i] &= P[M_i = y, C_i > y; s_i, \theta_i] \\ &= F(y) g(y; \theta, s_i) \end{aligned}$$

y

$$\begin{aligned} P[Y_i = y, \delta_i = 0; s_i, \theta] &= P[C_i = y, M_i > y; s_i, \theta] \\ &= f(y) G^*(y; \theta, s_i) \end{aligned}$$

Asuma que las s_i, \dots, s_n , las parejas (Y_i, δ_i) , son independientes. La función de verosimilitud sobre los datos $(Y_i = y_i, \delta_i, s_i)$, $i = 1, \dots, n$, condicional sobre las s_i, \dots, s_n es

$$L(\mu, \alpha, \kappa | y) = \prod_{i=1}^n [G(y_i; \mu, \alpha, \kappa) f(y_i)]^{1-\delta_i} [F(y_i) g(y_i; \mu, \alpha, \kappa)]^{\delta_i}$$

Arreglando los términos obtenemos finalmente:

$$L(\mu, \alpha, \kappa | y) = \left\{ \prod_{i=1}^n [F(y_i)]^{\delta_i} [f(y_i)]^{1-\delta_i} \right\} \left\{ \prod_{i=1}^n [G(y_i; \mu, \alpha, \kappa)]^{1-\delta_i} [g(y_i; \mu, \alpha, \kappa)]^{\delta_i} \right\}$$

Así, si la censura es no informativa, i.e., $F_i(y)$ no contiene a los parámetros en θ , obtenemos:

$$L(\mu_i, \sigma_i, \kappa | y_i) \propto \prod_{i=1}^n [G^*(y_i; \mu_i, \sigma_i, \kappa)]^{1-\delta_i} [g(y_i; \mu_i, \sigma_i, \kappa)]^{\delta_i}$$

2.4. Implementación Bayesiana

Asumiendo que $\pi(y_t | \theta_t)$ es la distribución GEV con parámetros $\theta_t = (\mu_t, \sigma_t, \kappa_t)$, y que la relación que liga a los parámetros con las covariables es:

$$\mu_t = X\beta + Z_x u_x + Z_s u_s,$$

$$\sigma_t = \sigma$$

$$\kappa_t = \kappa$$

Donde $Z_x = [C(x_i - k_j)]_{1 \leq i \leq n; 1 \leq j \leq K_x} \cdot [C(k_h - k_j)]_{1 \leq h, j \leq K_x}^{-1/2}$; $Z_s = [C(s_i - k_j)]_{1 \leq i \leq n; 1 \leq j \leq K_s} \cdot [C(k_h - k_j)]_{1 \leq h, j \leq K_s}^{-1/2}$ y $C(v) = \exp(\|v\|^2)$.

Una formulación bayesiana para el modelo de valores extremos es la siguiente:

$$\pi(\omega^* | y_t) \propto \pi(y_t | \omega^*) \pi(\omega^*) \quad (1)$$

Donde $\pi(y_t | \omega)$ es la densidad GEV y $\omega^* = (\beta_1, \beta_2, u_{x1}, u_{x2}, u_{s1}, u_{s2}, \kappa)$. La función apriori $\pi(\omega^*)$, es tal que: $\beta \sim N(0, 10^4 I)$, $u_{(x)} | \sigma_x \sim N(0, \sigma_x I_{K_x})$, $u_{(s)} | \sigma_s \sim N(0, \sigma_s I_{K_s})$, $\sigma \sim Half-Cauchy(25)$ y $\kappa \sim Uniform(-5, 5)$. También tenemos que $\omega^{**} = \{\sigma_x, \sigma_s\}$ los cuales tienen las densidades apriori $\sigma_x \sim Half-Cauchy(25)$ y $\sigma_s \sim Half-Cauchy(25)$. Note que este modelo puede considerarse como la extensión del modelo propuesto por Bocci et al. (2013) para el caso de censura aleatoria.

2.5. Niveles Máximos de Contaminante PM10

Las partículas suspendidas (PM por sus siglas en inglés) forman una mezcla compleja de materiales sólidos y líquidos suspendidos en el aire, que pueden variar en tamaño, forma y composición, dependiendo fundamentalmente de su origen. El tamaño de las partículas varía desde 0.005 hasta $100\mu m$ de diámetro y su unidad de medida son los microgramos por metro cúbico ($\mu g/m^3$).

Realizamos un ajuste al modelo GEV no estacionario para datos censurados para el conjunto de datos de máximos de contaminación atmosférica de partículas menores a 10 micrómetros (PM10). Estos datos corresponden a 1479 observaciones de niveles máximos de PM10, tomados de la base de datos de la Red Automática de Monitoreo Atmosférico (RAMA), entre los años de 1995 al 2014 en 11 estaciones ubicadas en la Zona Metropolitana del Valle de México, específicamente, Tlalnepanta (TLA), Xalostoc (XAL), Merced (MER), Pedregal (PED), Tultitlán (TLI), Villa de Flores (VIF), Tlahuac (TLA), Santa Úrsula (SUR), FES Acatlán (FAC), San Agustín (SAG) e Iztacalco (IZT). Estos datos contienen un 13.25 por ciento de observaciones censuradas. Las covariables fueron la longitud, latitud y tiempo.

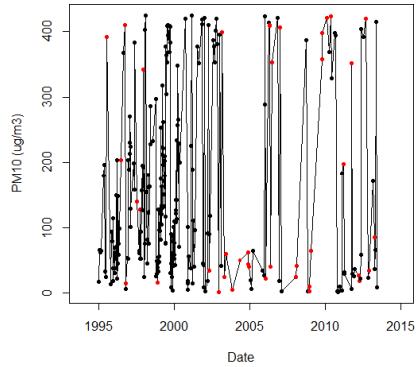


Fig. 1. Valores extremos con datos censurados (puntos rojos) en la estación pedregal, Mex.

3. Resultados

Para ajustar el modelo 1, corrimos 10,000 muestras con el método MCMC en el software estadístico R 3.0.1. Los resultados se muestran en la tabla 1, donde podemos ver los estimadores de los efectos principales y su intervalo de credibilidad al 95 %, (calculados mediante los percenciles 2.5 y 97.5 de la cadena del respectivo parámetro). Para que el algoritmo encontrara la distribución estacionaria de manera más eficiente, la semilla usada como valor inicial de los estimadores fue la moda de la distribución a posteriori indicada en la ecuación 1.

El umbral a partir del cual un valor extremo es excedido con probabilidad p es conocido como el nivel de retorno Z_p , el cual se espera que sea excedido una vez cada $1/p$ años Coles (2001). De acuerdo a la figura 2, en los años 2015 permanece la tendencia de incrementarse en la región noroeste del área de estudio, alcanzando los más grandes niveles de riesgo en áreas cercanas a Villa Flores y San Agustín, y niveles más pequeños de riesgo en áreas alrededor de la estación Pedregal.

No. Parámetros = 59			
Log Verosimilitud	-7837.742	-7857.514	
Parámetro	moda aposteriori	media aposteriori	Intervalo de credibilidad 95 %
$\beta_{0,\mu}$	-4.539918	-4.250326	(-15.85850, 7.547524)
$\beta_{1,\mu}$	-0.0024388	-0.002456182	(-0.00258, -0.00234)
$\beta_{s1,\mu}$	-0.8777787	-0.9028299	(-1.26754, -0.51131)
$\beta_{s2,\mu}$	3.753880	3.616109	(1.923589, 5.487732)
σ	4.370657	4.370988	(4.368459, 4.37352)
κ	0.3180246	0.3187162	(0.3156759, 0.3220128)
σ_x	$8.896771e - 12$	$3.565706e - 08$	($4.037948e - 12$, 3.5219)
σ_s	$2.101541e - 04$	$4.975610e - 04$	($6.055454e - 05$, 0.0108)

Tabla 1. Resultados del modelo GEV bayesiano con censura aleatoria.

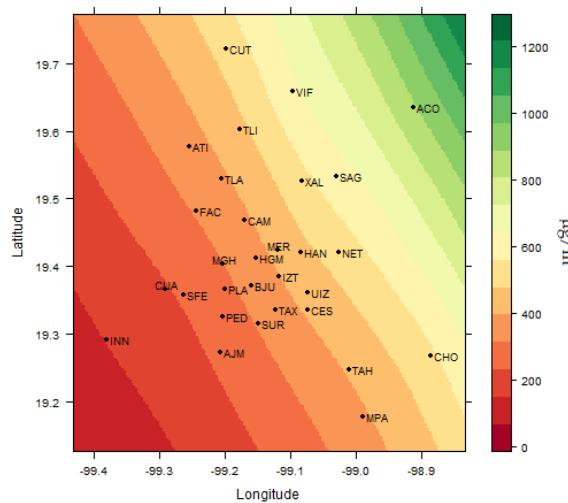


Fig. 2. Niveles de retorno con periodo de 25 años para valores de contaminación por PM10 en la región de estudio.

Para verificar la validez del modelo 1, ajustamos otros dos modelos usados de manera regular en el análisis de valores extremos, el modelo GEV para el caso estacionario y el modelo VGAM, ver Yee and Stephenson (2007), con 5 nodos por cada efecto principal.

Para el caso GEV estacionario los resultados son mostrados en la tabla 2, la verosimilitud para este modelo fue de -8888.409 .

% Parametro	Estimador	Desviación Estándar
μ	140.5459183	2.12555811
σ	70.6888574	1.80065885
κ	0.3029932	0.02483777

Tabla 2. Estimadores de máxima verosimilitud para el caso estacionario a los datos de PM10.

Para el caso VGAM, la verosimilitud resultante fue de -8701.792 . En este modelo se ajustaron los efectos principales con funciones b-splines, con cinco nodos para cada variable: tiempo, longitud y latitud, resultando en un modelo con más parámetros que el modelo en 1. Con base en estos resultados concluimos que el modelo 1 fue notablemente mejor.

4. Conclusiones

En este trabajo presentamos una análisis de los valores extremos no estacionarios con datos censurados de datos de contaminación por partículas menores a 10 micrómetros (PM10) en el área metropolitana de la Ciudad de México. La estimación de los parámetros de la distribución de los máximos de PM10 se realizó mediante un enfoque bayesiano semiparamétrico similar al usado en Bocci et al. (2013). Una vez conocidos los parámetros, calculamos los niveles de retorno de las concentraciones de PM10 para períodos de retorno de 25 años. Los resultados muestran una clara tendencia espacial de incremento en los niveles máximos de PM10 en la dirección noroeste de la zona de estudio, así como también se observa una tendencia de cambio en los niveles máximos en el tiempo. Concluimos que podemos usar esta metodología para generar mapas de riesgo de eventos extremos de lluvia, vientos, ondas de calor, etc. En particular cuando se tienen datos censurados que presentan información adicional y posiblemente multivariada, con el objetivo de medir sus efectos e implícitamente el de sus interacciones, para así poder encontrar relaciones entre estos y los valores extremos.

Bibliografía

- Bocci, C., Caporali, E., and Petrucci, A. (2013). Geoadditive modeling for extreme rainfall data. *ASTA Adv Stat Anal*, 97(2):181–193.
- Cannon, A. J. (2010). A flexible nonlinear modelling framework for nonstationary generalized extreme value analysis in hydroclimatology. *Hydrol Process*, 24(6):673–685.
- Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*, volume 208. Springer.
- Cooley, D. and Sain, S. R. (2010). Spatial hierarchical modeling of precipitation extremes from a regional climate model. *J Agric Biol Environ Stat*, 15(3):381–402.
- El Adlouni, S., Ouarda, T., Zhang, X., Roy, R., and Bobée, B. (2007). Generalized maximum likelihood estimators for the nonstationary generalized extreme value model. *Water Resour Res*, 43:W03410.
- Fisher, R. A. and Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Proc Camb Philos Soc*, 24:180–190.
- Gaetan, C. and Grigoletto, M. (2007). A hierarchical model for the analysis of spatial rainfall extremes. *J Agric Biol Environ Stat*, 12(4):434–449.
- Hosking, J. R. (1990). L-moments: analysis and estimation of distributions using linear combinations of order statistics. *J R Stat Soc Series B*, pages 105–124.
- Hosking, J. R. M., Wallis, J. R., and Wood, E. F. (1985). Estimation of the generalized extreme-value distribution by the method of probability-weighted moments. *Technometrics*, 27(3):251–261.
- Jenkinson, A. F. (1955). The frequency distribution of the annual maximum (or minimum) values of meteorological elements. *Quart J Roy Meteor Soc*, 81(348):158–171.
- Kharin, V. V. and Zwiers, F. W. (2005). Estimating extremes in transient climate change simulations. *J Clim*, 18:1156–1173.

-
- Leadbetter, M. R., Lindgren, G., and Rootzen, H. (1983). *Extremes and related properties of random sequences and processes*. 336pp, Springer, New York.
- Madsen, H., Rasmussen, P. F., and Rosbjerg, D. (1997). Comparison of annual maximum series and partial duration series methods for modeling extreme hydrologic events: 1. at-site modeling. *Water Resour Res*, 33(4):747–757.
- Pauli, F. and Coles, S. (2001). Penalized likelihood inference in extreme value analyses. *J Appl Stat*, 28:547–560.
- Reich, B., Shaby, B., and Cooley, D. (2014). A hierarchical model for serially-dependent extremes: A study of heat waves in the western us. *J Agric Biol Environ Stat*, 19(1):119–135.
- Rodriguez, S., Reyes, H., Perez, P., and Vaquera, H. (2012). Selection of a subset of meteorological variables for ozone analysis: Case study of pedregal station in mexico city. *Environ Sci Eng A*, 1:11–20.
- Rosen, O. and Cohen, A. (1996). Extreme percentile regression. In *Statistical Theory and Computational Aspects of Smoothing*, pages 200–214. Springer.
- Sang, H. and Gelfand, A. E. (2010). Continuous spatial process models for spatial extreme values. *J Agric Biol Environ Stat*, 15(1):49–65.
- Scarf, P. A. (1992). Estimation for a four parameter generalized extreme value distribution. *Commun Stat Theor M*, 21:2185– 2201.
- Smith, R. L. (1985). Maximum likelihood estimation in a class of non-regular cases. *Biometrika*, 72:67 –92.
- Tawn, J. (1988). Bivariate extreme value theory: models and estimation. *Biometrika*, 75:397–415.
- Wang, X. L., Zwiers, F. W., and Swail, V. (2004). North atlantic ocean wave climate scenarios for the 21st century. *J Clim*, 17:2368– 2383.
- Weissman, I. (1978). Estimation of parameters and large quantiles based on the k largest observations. *J Am Stat Assoc*, 73:812–815.

- Yee, T. W. and Stephenson, A. G. (2007). Vector generalized linear and additive extreme value models. *Extremes*, 10:1–19.

Uso de Componentes Principales y Correlación Canónica para la Caracterización de Poblaciones de Maíz Nativo*

Juan Elías Solís Alonso, María Gúzman Martínez

Universidad Autónoma de Guerrero

Dolores Briones Reyes

INIFAP Aguascalientes

Flaviano Godínez Jaimes, Elisa Martínez Azoños

Universidad Autónoma de Guerrero

Los estudios descriptivos, genéticos y estadísticos, sobre la caracterización morfológica y agronómica de razas, accesiones y/o poblaciones de maíz nativo, promueven y justifican la conservación de este recurso fitogenético, para su posterior aprovechamiento en programas de mejoramiento genético. La amplia diversidad de maíz en México, representa una fuente importante de genes para el mejoramiento del maíz, a nivel nacional e internacional. El objetivo de este trabajo fue identificar similitudes y diferencias entre 54 poblaciones nativas de maíz, provenientes de seis estados: Oaxaca, México, Tlaxcala, Puebla, Hidalgo y Guerrero a partir de 8 características morfométricas, utilizando Componentes Principales (CP) y Correlación Canónica (CC).

El análisis de CP explicó un 64.85 % de la variabilidad total presente en los datos, con un componente principal morfométrico de mazorca y un componente principal morfométrico de grano. De acuerdo con los resultados, en general las poblaciones del estado de México presentan características morfométricas mayores en mazorca que el resto de las poblaciones; mientras que las poblaciones de Guerrero presenten características morfométricas mayores en grano que el resto de las poblaciones.

El análisis de CC da como resultado un par de variables con una correlación canónica de 0.91, lo cual dio evidencia de la pertinencia del análisis. Es decir, de realizar la caracterización

*Facultad de Matemáticas, Universidad Autónoma de Guerrero

de las 54 poblaciones por grupos homogéneos de variables. Los resultados muestran que las poblaciones de Guerrero se identifican mejor por sus características morfométricas de grano mientras que las poblaciones del estado de México por sus características morfométricas de mazorca.

Clasificación: Trabajo de investigación.

Área-MSC: Estadística Multivariada.

Subárea-MSC: Análisis de Componentes Principales y Correlación Canónica.

1. Introducción

De acuerdo con cifras de la División de Estadísticas de la Organización de las Naciones Unidas para la Alimentación y la Agricultura (FAOSTAT, por sus siglas en inglés), en cuanto a maíz se refiere, México ocupa el quinto lugar como productor de grano; son Estados Unidos y China quienes ocupan los primeros lugares. En cuanto a la demanda de alimentos, México es el mayor consumidor de maíz en el mundo, este grano representa alrededor del 30 % del consumo diario calórico de los mexicanos, lo cual se refleja en el consumo per cápita anual de 120 kg, una cifra muy por encima del consumo promedio mundial de 17 kg (FAOSTAT, 2013). Dependencias como el Instituto Nacional de Estadística y Geografía (INEGI, 2012) y el Instituto Nacional de Salud Pública (INSP) también reportan que el maíz es uno de los principales alimentos en el país.

De acuerdo con la FAO (2009), en México el maíz es uno de los cultivos de mayor importancia a nivel nacional e internacional tanto por la superficie sembrada como por el volumen de producción y su diversidad de usos (CONABIO, 2008). La interacción de las diferentes condiciones orográficas con los factores climáticos que existen en México, ha dado como resultado una amplia diversidad ambiental y nichos ecológicos que han permitido la existencia de una gran variedad de poblaciones nativas de maíz. México es considerado como uno de los centros más importantes en diversidad de maíz, ya que en él existen una gran variedad de razas de maíz nativo (*Zea mays L.*) (Matsuoka *et al.*, 2002; Kato *et al.*, 2009). Esto le da la oportunidad a México de disponer de un amplio recurso fitogenético. La gran variedad genética de maíz constituye una riqueza no solo para México, sino para la población mundial (Ballesteros-Martínez, 2013).

Las poblaciones de maíz se agrupan con base en la categoría racial, para entender mejor su amplitud de adaptación ambiental y características morfológicas apropiadas para diversos usos. Por décadas varios investigadores han señalado a las variables morfológicas como una herramienta útil para la clasificación racial del maíz (Anderson and Cutler, 1942; Anderson, 1945; Wellhausen *et al.*, 1951; Goodman and Paterniani, 1969; Hernández and Flores, 1970; Doebley *et al.*, 1985; Sánchez *et al.*, 2000). Ballesteros-Martínez (2013); González-Castro *et al.* (2013) mencionan que el estudio de las poblaciones nativas de maíz a partir de su descripción y clasificación permite conocerlas y conservarlas para su posible uso en el mejoramiento genético del maíz y preservar la diversidad del mismo. Por su parte, González-Castro *et al.* (2013) mencionan que la evaluación de la diversidad de maíces nativos es importante en el planteamiento de estrategias de conservación, caracterización y uso del germoplasma en el mejoramiento genético, además de la caracterización de nuevas poblaciones. Las variedades nativas de maíz (*Zea may L.*) preservadas por los agricultores a través de generaciones, constituyen más del 80 % del área sembrada con maíz en México.

Existen varios estudios sobre caracterización morfológica de poblaciones de maíz en México; Rocandio-Rodríguez *et al.* (2014), realizaron un estudio para valorar la diversidad morfológica y agronómica de siete razas de maíz cultivadas en los Valles Altos Centrales de México; Ramírez-Jaspeado *et al.* (2013) realizaron una caracterización morfológica de 108 accesiones de maíz obtenidas en 17 localidades de los distritos de Zimatlán, Ocotlán y Ejutla, en Valles Centrales de Oaxaca; Ballesteros-Martínez (2013) realizó una caracterización morfológica de maíces nativos del estado de Jalisco; González-Castro *et al.* (2013) evaluaron la diversidad de 196 poblaciones de maíz nativo provenientes de 21 estados de México. Briones-Reyes (2013) realizó un estudio sobre poblaciones de maíz nativo que pertenecen a seis estados de la República Mexicana: Oaxaca, México, Tlaxcala, Puebla, Hidalgo y Guerrero. En dichos trabajos se realizan estudios descriptivos, genéticos o estadísticos. Los análisis estadísticos multivariados que generalmente se llevan a cabo son: Componentes Principales y Análisis de Conglomerados.

El objetivo de este trabajo fue utilizar las técnicas de CP y CC para identificar similitudes y diferencias entre 54 poblaciones de maíz nativo que provienen de los estados: Guerrero, México, Hidalgo, Oaxaca, Puebla y Tlaxcala; a partir de ocho características morfométricas: una de olate, tres de mazorca y cuatro de grano. Estas dos técnicas nos permitirán explicar

la variabilidad de las 54 poblaciones en función de las características morfométricas más importantes.

El trabajo tiene la estructura siguiente. En la Sección 2 se dan los fundamentos teóricos de CP y CC, en la Sección 3 se da la descripción del conjunto de datos y se presentan los resultados mas relevantes de la aplicación, en la Sección 4 se dan las conclusiones del trabajo.

2. Marco Teórico

El análisis estadístico multivariado, con técnicas como Componentes Principales y Correlación Canónica permiten al investigador estudiar la variabilidad de un conjunto de datos en términos de parejas de combinaciones lineales de las variables del estudio, investigar la existencia de estructuras latentes y/o modelar la estructura de covarianzas de los datos. Esto es de gran ayuda cuando se quiere reducir la dimensionalidad de los datos y tener una mejor interpretación de la variabilidad de éstos, ademas de conocer su comportamiento. La elección de CP y CC para el análisis de las 54 poblaciones de maíz nativo fue de acuerdo con el objetivo del estudio; CP nos permitirá reducir la dimensionalidad de los datos y así determinar diferencias y similitudes entre las 54 poblaciones de maíz nativo, en términos de las características morfométricas más relevantes. Por otra parte, dado que se tienen mediciones de mazorca y grano, parece razonable formar dos grupos de variables y utilizar CC para explicar la variabilidad de las poblaciones de maíz nativo.

A continuación se describe cada una de las técnicas antes mencionadas.

2.1. Análisis de Componentes Principales

El análisis de CP tiene como objetivo explicar la variabilidad presente en p variables aleatorias $\mathbf{X}' = [X_1, X_2, \dots, X_p]$ a través de p combinaciones lineales de éstas, es decir,

$$Y_1 = \mathbf{a}_1' \mathbf{X} = a_{11}X_1 + a_{12}X_2 + \dots + a_{1p}X_p$$

$$Y_2 = \mathbf{a}_2' \mathbf{X} = a_{21}X_1 + a_{22}X_2 + \dots + a_{2p}X_p$$

⋮

$$Y_p = \mathbf{a}_p' \mathbf{X} = a_{p1}X_1 + a_{p2}X_2 + \dots + a_{pp}X_p$$

Aunque las p variables X_1, X_2, \dots, X_p , pueden estar correlacionadas, la metodología de CP garantiza que las p combinaciones lineales, Y_i , deberán ser incorrelacionadas.

Para realizar CP se puede utilizar la matriz de varianzas y covarianzas de \mathbf{X}' o su matriz de correlaciones, denotadas con Σ y ρ respectivamente, donde:

$$\Sigma = Cov(\mathbf{X}) = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix}$$

$$\rho = Corr(\mathbf{X}) = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{12} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1p} & \rho_{2p} & \cdots & 1 \end{bmatrix}$$

con $\rho_{ik} = \frac{\sigma_{ik}}{\sqrt{\sigma_{ii}\sigma_{kk}}}$. Generalmente se utiliza ρ para realizar el análisis de CP, ya que se tiene mejor apreciación de los pesos de las variables en las combinaciones lineales establecidas.

A partir de la descomposición espectral de Σ , CP genera p parejas de eigenvalores-eigenvectores, dadas por:

$$(\lambda_1, \mathbf{e}_1), (\lambda_2, \mathbf{e}_2), \dots, (\lambda_p, \mathbf{e}_p)$$

donde $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$, tal que la i -ésima combinación lineal está dada por:

$$Y_i = e_{i1}X_1 + e_{i2}X_2 + \cdots + e_{ip}X_p, \quad i = 1, \dots, p$$

Donde:

$$Var(Y_i) = \mathbf{e}'_i \Sigma \mathbf{e}_i = \lambda_i, \quad i = 1, \dots, p$$

luego la proporción de varianza explicada por la i -ésima componente principal está dada por:

$$\frac{\lambda_i}{\lambda_1 + \lambda_2 + \cdots + \lambda_p}, \quad i = 1, \dots, p$$

y la varianza total de los datos está dada por:

$$\sum_{i=1}^p Var(Y_i) = \lambda_1 + \lambda_2 + \cdots + \lambda_p = \sigma_{11} + \sigma_{12} + \cdots + \sigma_{pp} = \sum_{i=1}^p Var(x_i)$$

De esta manera CP simplifica el análisis de la estructura de covarianzas de un conjunto de datos dado. A las p combinaciones lineales se les conoce como componentes principales.

2.2. Análisis de Correlación Canónica

El análisis de Correlación Canónica se utiliza para buscar relaciones entre dos grupos homogéneos de variables de un conjunto de datos dado, a partir de parejas de combinaciones lineales de estos grupos se busca estudiar la variabilidad presente en los datos. Dado $\mathbf{X}_{p \times 1}^{(1)} = [X_1^{(1)}, X_2^{(1)}, \dots, X_p^{(1)}]'$ y $\mathbf{X}_{q \times 1}^{(2)} = [X_1^{(2)}, X_2^{(2)}, \dots, X_p^{(2)}]'$ dos grupos de variables con $p \leq q$, se generan combinaciones lineales incorrelacionadas en cada grupo, a partir de éstas se crean p parejas de combinaciones lineales, de tal manera que la primera pareja sea la que tenga la correlación más grande, después que la segunda pareja tenga menor correlación con respecto a la primera, la tercera menor con respecto a la segunda y así sucesivamente. Si

$$\mathbf{X}_{(p+q) \times 1} = \begin{bmatrix} \mathbf{X}^{(1)} \\ \vdots \\ \mathbf{X}^{(2)} \end{bmatrix}, \quad \boldsymbol{\mu}_{(p+q) \times 1} = \begin{bmatrix} E(\mathbf{X}^{(1)}) \\ \vdots \\ E(\mathbf{X}^{(2)}) \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu}^{(1)} \\ \vdots \\ \boldsymbol{\mu}^{(2)} \end{bmatrix}$$

$$\boldsymbol{\Sigma}_{(p+q)(p+q)} = E[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})'] = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \vdots & \boldsymbol{\Sigma}_{12} \\ \vdots & \ddots & \vdots \\ \boldsymbol{\Sigma}_{21} & \vdots & \boldsymbol{\Sigma}_{22} \end{bmatrix}$$

con $\boldsymbol{\Sigma}_{11} = Cov(\mathbf{X}^{(1)})$ de dimensión $p \times p$, $\boldsymbol{\Sigma}_{22} = Cov(\mathbf{X}^{(2)})$ de dimensión $q \times q$ y $\boldsymbol{\Sigma}'_{21} = \boldsymbol{\Sigma}_{12} = Cov(\mathbf{X}^{(1)}, \mathbf{X}^{(2)})$ de dimensión $p \times q$. $\boldsymbol{\Sigma}_{12}$ mide la asociación entre los dos conjuntos de variables. CC tiene como objetivo explicar la asociación existente en $\mathbf{X}^{(1)}$ y $\mathbf{X}^{(2)}$ en términos de pocas covarianzas, elegidas cuidadosamente, en lugar de las pq covarianzas de $\boldsymbol{\Sigma}_{12}$.

Si $U = \mathbf{a}' \mathbf{X}^{(1)}$ y $V = \mathbf{b}' \mathbf{X}^{(2)}$ son dos combinaciones lineales, entonces:

$$Var(U) = Var(\mathbf{a}' \mathbf{X}^{(1)}) = \mathbf{a}' Var(\mathbf{X}^{(1)}) \mathbf{a} = \mathbf{a}' \boldsymbol{\Sigma}_{11} \mathbf{a}$$

$$Var(V) = Var(\mathbf{b}' \mathbf{X}^{(2)}) = \mathbf{b}' Var(\mathbf{X}^{(2)}) \mathbf{b} = \mathbf{b}' \boldsymbol{\Sigma}_{22} \mathbf{b}$$

$$Cov(U, V) = Cov(\mathbf{a}' \mathbf{X}^{(1)}, \mathbf{b}' \mathbf{X}^{(2)}) = \mathbf{a}' Cov(\mathbf{X}^{(1)}, \mathbf{X}^{(2)}) \mathbf{b} = \mathbf{a}' \boldsymbol{\Sigma}_{12} \mathbf{b}$$

luego el coeficiente de correlación de U y V está dado por:

$$Corr(U, V) = \frac{\mathbf{a}' \boldsymbol{\Sigma}_{12} \mathbf{b}}{\sqrt{\mathbf{a}' \boldsymbol{\Sigma}_{11} \mathbf{a}} \sqrt{\mathbf{b}' \boldsymbol{\Sigma}_{22} \mathbf{b}}}$$

donde \mathbf{a} y \mathbf{b} se calculan tal que $Corr(U, V)$ sea lo más grande posible (Johnson and Wichern, 2007).

3. Aplicación

La base de datos está compuesta por 54 poblaciones de maíz nativo, cada observación morfométrica de mazorca y olate fue obtenida de una muestra de 5 mazorcas. Las observaciones morfométricas de grano se obtuvieron del promedio de 50 granos de la misma muestra. Cada una de las 54 poblaciones provino de un lote de diferente productor, la metodología se describe en el trabajo de Briones-Reyes (2013). El estudio se llevó a cabo en un campo experimental del Colegio de Postgraduados, ubicado en Montecillo, Texcoco, estado de México, durante el ciclo primavera-verano del año 2012. Las características morfométricas (una de olate, tres de mazorca y cuatro de grano) son las siguientes:

- POM : Peso del olate de la mazorca (g).
- PSM : Peso en seco de la mazorca (g).
- DPM : Diámetro de la punta de la mazorca (cm).
- DBM : Diámetro de la base de la mazorca (cm).
- PG : Peso del grano (g).
- LG : Largo del grano (mm).
- AG : Ancho del grano (mm).
- EG : Espesor del grano (mm).

La Tabla 1 muestra la distribución de las 54 poblaciones en cada estado.

Estados	Total de poblaciones
Oaxaca	13
México	12
Tlaxcala	11
Guerrero	11
Puebla	6
Hidalgo	1

Tabla 1: Distribución por estado de las 54 poblaciones de maíz nativo.

Los dos grupos que se proponen para realizar el análisis de CC son: **Mazorca:** $\{POM, PSM, DPM, DBM\}$ y **Grano:** $\{PG, LG, AG, EG\}$. Cabe mencionar que Briones-Reyes (2013) sugiere poner a la variable peso del grano (PG) dentro de las características de mazorca.

3.1. Resultados del Análisis de Componentes Principales

Para el análisis de CP se utilizó la matriz de correlaciones de las 8 variables y la función *PCA* del paquete *FactoMiner* que se encuentra disponible en la versión 3.3.2 del software estadístico R (Team, 2016).

De los 8 componentes principales (CP) obtenidos, *CP1* y *CP2* explicaron un 64.85 % de la variabilidad total presente en las poblaciones de maíz nativo (Figura 1). La contribución de las variables morfométricas de mazorca, olate y grano en *CP1* y *CP2* se observan en la Figura 2. Se observa que en *CP1* las variables de diámetro son las que más aportan y en *CP2* aportan más las de grano. *CP1* puede plantearse como un componente morfométrico de marzoca y *CP2* como un componente morfométrico de grano.

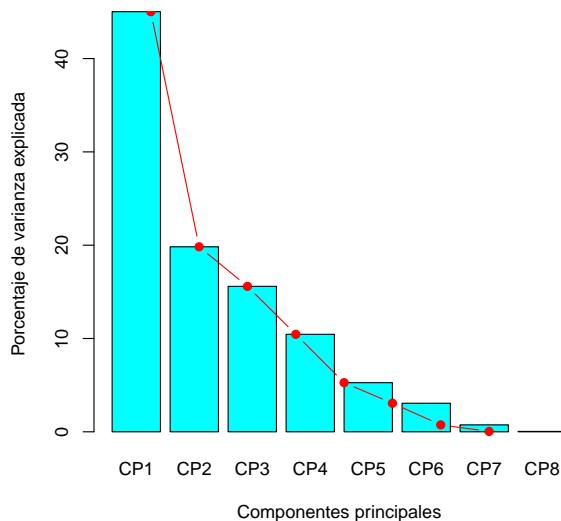


Figura 1: Proporción de varianza explicada por cada componente principal.

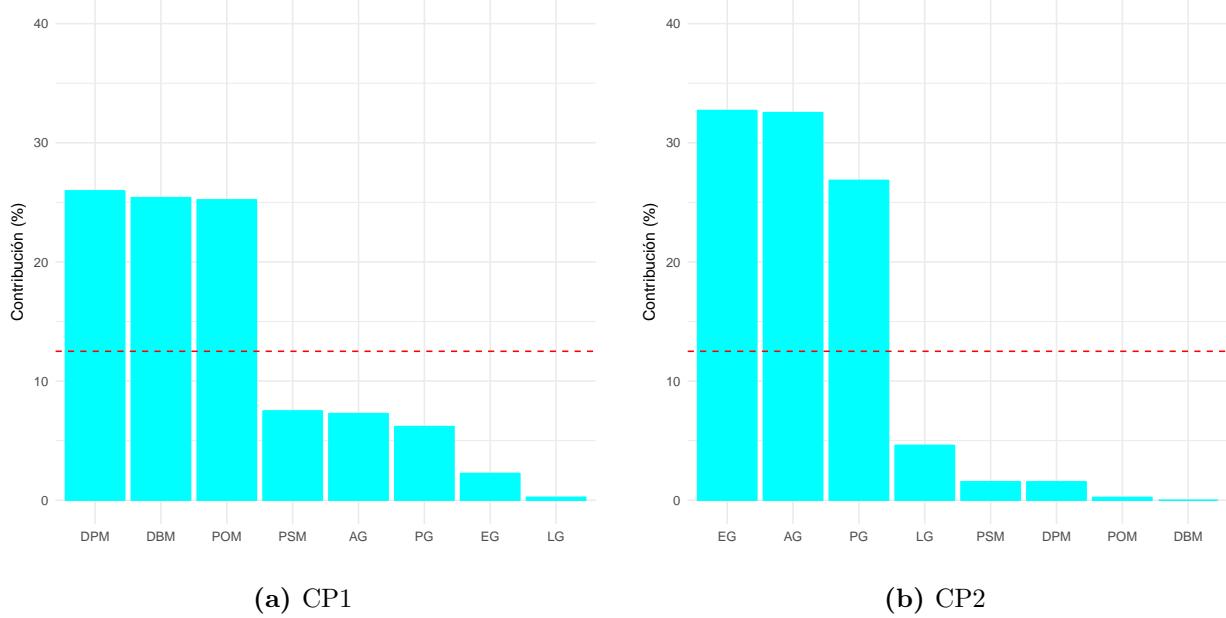


Figura 2: Contribución de las características morfométricas en *CP1* y *CP2*.

De acuerdo con los resultados, los primeros dos componentes principales están dados por:

$$CP1 = 0.520PSM + 0.472PG + 0.956DBM + 0.967DPM - 0.284 + 0.094LG - 0.512AG$$

$$CP2 = 0.157PSM + 0.653PG + 0.061DBM - 0.001DPM + 0.720 - 0.270LG + 0.718AG$$

En la Figura 3 se muestra la dispersión de las 54 poblaciones de maíz nativo. En el gráfico se forman basicamente dos grupos, el primero contiene las poblaciones con características morfométricas similares en diámetro que en su mayoría son del estado de Tlaxcala y el estado de México (primer y cuarto cuadrante); en el segundo grupo están las poblaciones con características morfométricas similares en espesor del grano y ancho del grano que son principalmente del estado de Guerrero (segundo cuadrante), hay poblaciones con características morfométricas del largo del grano que corresponden a los estados de Oaxaca y Tlaxcala. Las poblaciones del estado de Puebla se identifican por tener características morfométricas promedio respecto al *CP1* de mazorca y *CP2* de grano.

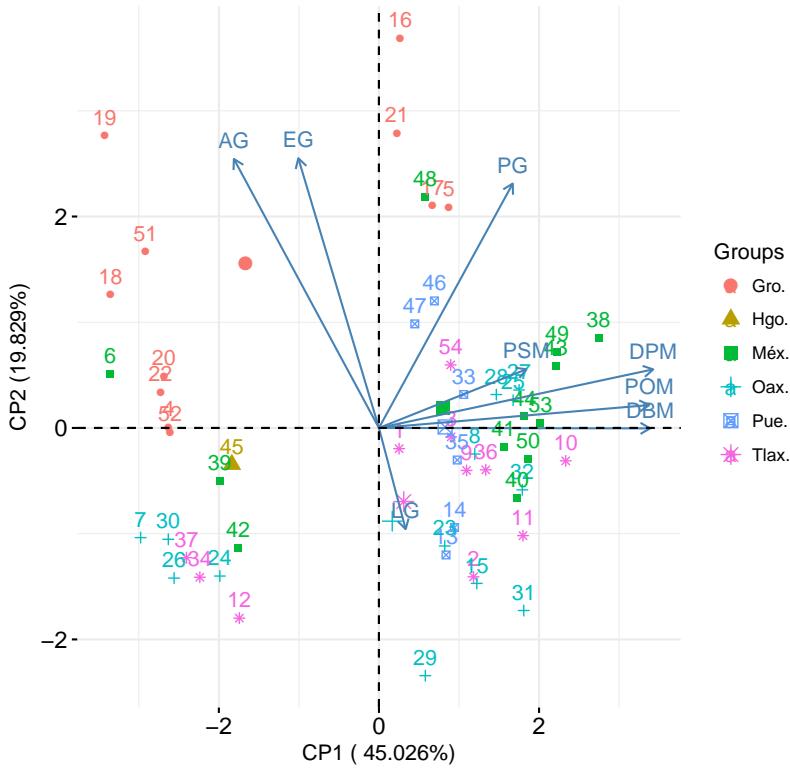


Figura 3: Agrupamiento de las 54 poblaciones de maíz nativo por características morfológicas.

3.2. Resultados del Análisis de Correlación Canónica

Para el análisis de CC, las funciones que se utilizaron fueron *cc* del paquete *CCA* y *CCorA* del paquete *vegan* que se encuentran disponibles en la versión 3.3.2 del software estadístico R (Team, 2016).

En la Figura 4 se muestran los resultados para cada uno de los grupos establecidos. El gráfico (a) muestra la dispersión de las 54 poblaciones de maíz nativo en función de las primeras dos combinaciones lineales de las variables que corresponden a la caracterización morfométrica de mazorca. Claramente se observan dos grupos; el grupo grande esta compuesto por las poblaciones con características morfométricas en mazorca, las cuales son mayores a las del grupo pequeño, gráficamente se observa que las variables de este grupo están muy correlacionadas. En el gráfico (b) se observa una mayor dispersión de las poblaciones, esto se debe a que el peso del grano (PG) y el ancho del grano (AG) presentan baja correlación.

Las poblaciones con mayor longitud en grano (LG) son del estado de México y Tlaxcala, y las poblaciones de grano más ancho son las que pertenecen al estado de Guerrero.

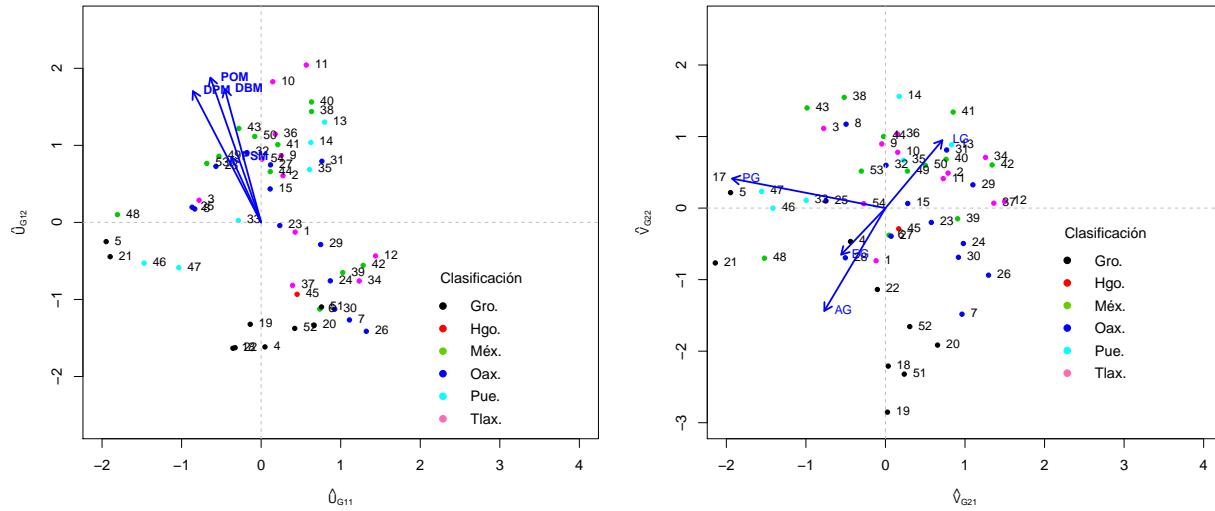


Figura 4: Agrupamiento de las 54 poblaciones por características morfológicas con respecto a Mazorca y Grano.

De acuerdo con los resultados la pareja dada por:

$$\hat{U}_{G11} = -0.100PSM + 2.038POM + 3.321DBM - 5.574DPM$$

$$\hat{V}_{G21} = -0.912PG - 0.134EG + 0.057LG - 0.167AG$$

es la que mayor correlación canónica presenta con un valor de 0.91, la cual es una correlación muy alta y corresponde a un 70.33 % de la variabilidad total explicada por los datos.

4. Conclusiones

El análisis de CP reportó como variables más importantes para la clasificación de poblaciones de maíz nativo, de las consideradas, el espesor del grano y diámetro de la punta de la mazorca. El análisis mostró que la mayoría de las poblaciones de maíz nativo del estado de México, presentan características morfométricas en mazorca más grandes que en grano; mientras que las poblaciones con características morfométricas más grandes en grano son

las que pertenecen al estado de Guerrero. En el análisis de CP que realizó Briones-Reyes (2013) con 19 características morfológicas de planta, mazorca, espiga y grano, explicó, con los dos primeros componentes, un 53.5 % de la variabilidad total presente en sus poblaciones; mientras que aquí con sólo tres características morfométricas de mazorca, una de olate y cuatro de grano, se explicó un 64.85 % de la variabilidad total presente en las poblaciones con los dos primeros componentes principales. Por lo tanto, no es necesaria la inclusión de muchas características morfológicas para realizar caracterización de poblaciones.

En cuanto al análisis de CC, se observó que la propuesta de agrupar las variables que reportan características morfométricas de mazorca en un grupo y las de grano en otro grupo, dio buenos resultados, ya que se obtuvó una correlación canónica del 0.91 para el primer par de combinaciones lineales, la cual es una correlación muy alta y corresponde a un 70.33 % de la variabilidad total explicada por los datos. Esta técnica permitió realizar un análisis más detallado de las diferencias y similitudes que existen entre las poblaciones en cuanto a sus características morfométricas de grano y mazorca. Por lo tanto, este trabajo muestra la pertinencia de la implementación de un análisis de Correlación Canónica para la caracterización morfométrica de poblaciones de maíz nativo.

Dado que la diversidad genética y clasificación racial del maíz puede ser estudiada a partir de las características morfológicas (Anderson and Cutler, 1942; Anderson, 1945; Wellhausen *et al.*, 1951; Goodman and Paterniani, 1969; Hernández and Flores, 1970; Doebley *et al.*, 1985; Sánchez *et al.*, 2000), éste trabajo puede ser de utilidad para los investigadores que buscan realizar caracterizaciones morfológicas a partir de variables morfométricos de las poblaciones de maíz nativo.

Bibliografía

- Anderson, E. (1945). Maize in the new world. *New Crops in the New World*. CM Wilson (ed). McMillan Co. New York, pages 27–42.
- Anderson, E. and Cutler, H. C. (1942). Races of zea mays: I. their recognition and classification. *Annals of the Missouri Botanical Garden*, 29(2):69–88.
- Ballesteros-Martínez, G. (2013). Caracterización morfológica de las razas de maíz Elotes Occidentales y Ancho en el estado de Jalisco. Te-

sis de maestría, Universidad de Guadalajara, CUCBA. Recuperado de: <http://repositorio.cucba.udg.mx:8080/xmlui/handle/123456789/5642>.

Briones-Reyes, D. (2013). Diversidad genética en el patosistema maíz-fusarium en el altiplano de México. Tesis doctoral, COLPOS. Recuperado de: <http://hdl.handle.net/10521/2163>.

CONABIO (2008). *Comisión Nacional para el Conocimiento y Uso de la Biodiversidad. Capital Natural de México, Volumen I: Conocimiento Actual de la Biodiversidad.* www.conabio.gob.mx. Accesado en Noviembre 2013.

Doebley, J. F., Goodman, M., and Stuber, C. W. (1985). Isozyme variation in the races of maize from Mexico. *American Journal of Botany*, 72:629–639.

FAO (2009). *Food and Agriculture Organization. Base de Datos de Estadísticas Agropecuarias.* [www.faostat.fao.org](http://www faostat fao org). Accesado en Noviembre 2013.

FAOSTAT (2013). *Food and Agriculture Organization of United Nations. Statistics Division.* <http://faostat.fao.org/>. Accesado en Mayo 2013.

González-Castro, M. E., Palacios-Rojas, N., Espinoza-Banda, A., and Bedoya-Salazar, C. A. (2013). Diversidad genética en maíces nativos mexicanos tropicales. *Revista fitotecnia mexicana*, 36:239–338.

Goodman, M. M. and Paterniani, E. (1969). The races of maize. III. Choices of appropriate characters for racial classification. *Economic Botany*, 23(3):265–273.

Hernández, X. E. and Flores, G. A. (1970). Estudio morfológico de cinco nuevas razas de maíz de la sierra madre occidental de México. implicaciones filogenéticas y fitogeográficas. *Agrociencia Chapingo*, 5:3–30.

INEGI (2012). *Encuesta Nacional de Ingresos y Gastos en los Hogares. Tabulados básicos.* www.inegi.org.mx/Sistemas/TabuladosBasicos/tabdirecto.aspx?s=est&c=33501. Accesado en Febrero 2013.

Johnson, R. A. and Wichern, D. W. (2014). *Applied multivariate statistical analysis*, volume 4. Prentice-hall New Jersey.

- Kato, T. A., Mapes, C., Mera, L., Serratos, J., and Bye, R. (2009). Origen y diversificación del maíz: Una revisión analítica. *Universidad Nacional Autónoma de México, Comisión Nacional para el Conocimiento y Uso de la Biodiversidad. México, DF*, 116.
- Matsuoka, Y., Vigoroux, Y., Goodman, M. M., Sánchez, J., Buckler, E., and Doebley, J. (2002). A single domestication for maize shown by multilocus microsatellite genotyping. *Proceedings of the National Academy of Sciences*, 99(9):6080–6084.
- Ramírez-Jaspeado, A., García de los Santos, G., Carballo-Carballo, A., Castillo-González, F., Serratos, J. A., and Cadena-Iñiguez, J. (2013). Caracterización morfológica de una muestra etnográfica de maíz (*zea mays l.*) raza bolita de Oaxaca. *Revista mexicana de ciencias agrícolas*, 4(6):895–907.
- Rocandio-Rodríguez, M., Santacruz-Varela, A., Córdova-Téllez, L., López-Sánchez, H., Castillo-González, F., Lobato-Ortiz, R., García-Zavala, J. J., and Ortega-Paczka, R. (2014). Caracterización morfológica y agronómica de siete razas de maíz de los valles altos de méxico. *Revista fitotecnia mexicana*, 37(4):351–361.
- Sánchez, G. J., Goodman, M. M., and Stuber, C. W. (2000). Isozymatic and morphological diversity in the races of maize of mexico. *Economic Botany*, 54(1):43–59.
- Team, R. C. (2016). R: A Language and Environment for Statistical Computing. Vienna, Austria. R Foundation for Statistical Computing. Disponible en: <https://www.R-project.org/>.
- Wellhausen, E. J., Roberts, L. M., and Hernández, X. E. (1951). Razas de maíz en méxico. su origen, características y distribución. in xocolotzi. *Revista de Geografía Agrícola*, 2.

Análisis Bayesiano de un Modelo Gamma Bivariado Bajo Proyección*

Gabriel Nuñez Antonio^a, Juan de Dios Aguilar Gámez^b, Emiliano Geneyro Squarzon^c, Gabriel Escarela Pérez^d

Universidad Autónoma Metropolitana, unidad Iztapalapa, México

En el análisis de fenómenos reales existen variables direccionales que por su naturaleza se ven definidas solo en ciertos subconjuntos de la esfera unitaria p –dimensional, \mathbb{S}^p . Por ejemplo, cuando se analizan datos *composicionales* usando variables circulares (variables definidas sobre el círculo unitario), el espacio muestral asociado resulta ser el intervalo $(0, \pi/2)$. Por otro lado, cuando se trabajan con datos *axiales*, el rango de posibles valores de las variables circulares asociadas, resulta ser el intervalo $(0, \pi]$. Así, desde el punto de vista metodológico es importante contar con distribuciones de probabilidad definidas en subconjuntos acotados sobre \mathbb{S}^p . Este trabajo pretende contribuir a la propuesta de modelos para describir variables circulares que no están definidas en todo el intervalo $(0, 2\pi)$. Específicamente, para modelar variables circulares restringidas al intervalo $(0, \pi/2]$, en este trabajo se introduce un modelo denominado Gamma proyectado, y se muestra la manera de llevar a cabo inferencias desde un punto de vista Bayesiano. La metodología propuesta se ejemplifica usando datos simulados.

Área-**MSC:** Datos Circulares

Subárea-**MSC:** Estadística Bayesiana

*Este trabajo fue apoyado parcialmente por el SNI, México. El apoyo del Departamento de Matemáticas de la UAM-I también es reconocido ampliamente

^agab.nuneza@gmail.com (autor responsable)

^bjuanfisica86@gmail.com

^csquarzon@gmail.com

^dge@xanum.uam.mx

1. Introducción

En diversas áreas de las ciencias el investigador se puede encontrar con variables que representan direcciones, es decir, variables direccionales. Este tipo de datos son especialmente comunes en las ciencias biológicas, geofísicas, meteorológicas, ecológicas y del medio ambiente. Algunas aplicaciones se encuentran en el análisis de la dirección de traslape de varias especies en alguna reserva ecológica, direcciones de viento, dirección de migración de aves, direcciones de propagación de fisuras en concreto y otros materiales, orientación de yacimientos geológicos, análisis de datos composicionales, análisis de datos axiales, etc. Los datos direccionales en el plano 2-dimensional se denominan datos *circulares*. Cuando los datos circulares se restringen a algún rango de la circunferencia unitaria, como es el caso de datos axiales o datos composicionales, distribuciones circulares como la von Mises, la Normal proyectada, la Normal envuelta, o cualquier distribución circular definidas sobre todo el círculo unitario, pueden no ser adecuadas para describir estos conjuntos de datos. Ver, por ejemplo, Mardia y Jupp (1999).

Para modelar datos circulares definidos en el rango $(0, \pi/2)$, en este trabajo se propone un análisis Bayesiano de la distribución circular que se genera al proyectar radialmente una distribución bivariada cuyas densidades marginales son densidades Gammas univariadas. A este modelo se le denominará modelo Gamma proyectado.

2. El Modelo Gamma Proyectado

Una manera simple de generar distribuciones sobre el círculo unitario es proyectando radialmente distribuciones originalmente definidas en el plano 2-dimensional. Así, sea \mathbf{Y} un vector bivariado aleatorio tal que $Pr(\mathbf{y} = \mathbf{0}) = 0$, entonces $||\mathbf{Y}||^{-1}\mathbf{Y}$ es un punto aleatorio sobre el círculo unitario. Un caso importante es aquel en el que \mathbf{Y} tiene una distribución Normal bivariada con vector de medias $\boldsymbol{\mu}$ y matriz de varianzas–covarianzas $\boldsymbol{\Sigma}$. En este caso, se dice que $||\mathbf{Y}||^{-1}\mathbf{Y}$ tiene una distribución Normal proyectada.

Aunque el modelo Normal proyectado ha recibido mucha atención en los últimos años, modelos adecuados para variables circulares cuando el soporte no está definido sobre todo el círculo unitario parece que han sido poco analizados hasta ahora. Para modelar datos circulares definidos en el rango $(0, \pi/2)$, en este trabajo se introduce una distribución que se obtiene al proyectar radialmente, sobre el círculo unitario, una distribución bivariada que

tiene como densidades marginales distribuciones Gamma univariadas. Como se menciona en Kotz *et al.* (2000) existen varias formas de definir una distribución Gamma en el caso multivariado. Sin embargo, en este trabajo al igual que en Kotz *et al.* (2000) nos referiremos a una distribución Gamma multivariada, simplemente, si sus marginales son distribuciones Gamma.

Los datos direcciones pueden ser especificadas usando varias representaciones. Por ejemplo, direcciones en el plano p -dimensional pueden ser reconocidos como vectores unitarios o como ángulos. En el caso general p -dimensional, es conveniente representar una dirección como un vector \mathbf{Y} sobre la esfera unitaria de dimensión p . Así, se tiene la siguiente definición.

Un vector p -dimensional $\mathbf{U} = \mathbf{Y}/R$, donde $R = \|\mathbf{Y}\|$, tiene una distribución Gamma proyectada p -variada si \mathbf{Y} tiene una distribución Gamma p -variada. Se debe notar que desde que \mathbf{U} es un vector unitario, éste también se puede especificar a través de $p - 1$ ángulos.

En el caso en el que \mathbf{Y} sea un vector bivariado con densidad conjunta, defina por:

$$\begin{aligned} f(\mathbf{y}|\boldsymbol{\alpha}, \boldsymbol{\beta}) &= \text{Gamma}(y_1|\alpha_1, \beta_1) \cdot \text{Gamma}(y_2|\alpha_2, \beta_2) \\ &= \frac{\beta_1^{\alpha_1}}{\Gamma(\alpha_1)} y_1^{\alpha_1-1} e^{-y_1\beta_1} \cdot \frac{\beta_2^{\alpha_2}}{\Gamma(\alpha_2)} y_2^{\alpha_2-1} e^{-y_2\beta_2}, \end{aligned}$$

el correspondiente modelo Gamma proyectado se obtiene al considerar el cambio de variable definido por:

$$\mathbf{y} = r(\cos \theta, \sin \theta)', \quad \text{con } \theta \in (0, \pi/2) \text{ y } r \in \mathbb{R}^+.$$

Así, la densidad conjunta del vector (R, Θ) , está dada por:

$$f(r, \theta|\boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{\beta_1^{\alpha_1} \beta_2^{\alpha_2}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \cdot (\cos \theta)^{\alpha_1-1} (\sin \theta)^{\alpha_2-1} r^{(\alpha_1+\alpha_2)-1} e^{-r(\beta_1 \cos \theta + \beta_2 \sin \theta)}. \quad (1)$$

Finalmente, la función de densidad del ángulo aleatorio Θ , es decir, la densidad de la correspondiente Gamma proyectada, $\text{PGa}(\theta|\boldsymbol{\alpha}, \boldsymbol{\beta})$, se obtiene al marginalizar la expresión (1) con respecto a R . Es decir,

$$\text{PGa}(\theta|\boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{\beta_1^{\alpha_1} \beta_2^{\alpha_2} (\cos \theta)^{\alpha_1-1} (\sin \theta)^{\alpha_2-1} (\beta_1 \cos \theta + \beta_2 \sin \theta)^{-(\alpha_1+\alpha_2)}}{B(\alpha_1, \alpha_2)} I_{(0, \pi/2)}(\theta), \quad (2)$$

donde $B(\alpha_1, \alpha_2)$ es la función matemática Beta.

El modelo Gamma proyectado, $PGa(\theta|\boldsymbol{\alpha}, \boldsymbol{\beta})$ es muy flexible ya que puede modelar comportamientos simétricos, asimétricos, unimodales y/o multimodales. Las Figuras 1, 2 ,3 y 4, generadas en Mathematica v10.2, muestran los diferentes comportamientos que se pueden obtener bajo la familia de distribuciones Gammas proyectadas. También se muestran las correspondientes distribuciones conjuntas (Gammas bivariadas) que las generan y sus respectivas curvas de nivel.

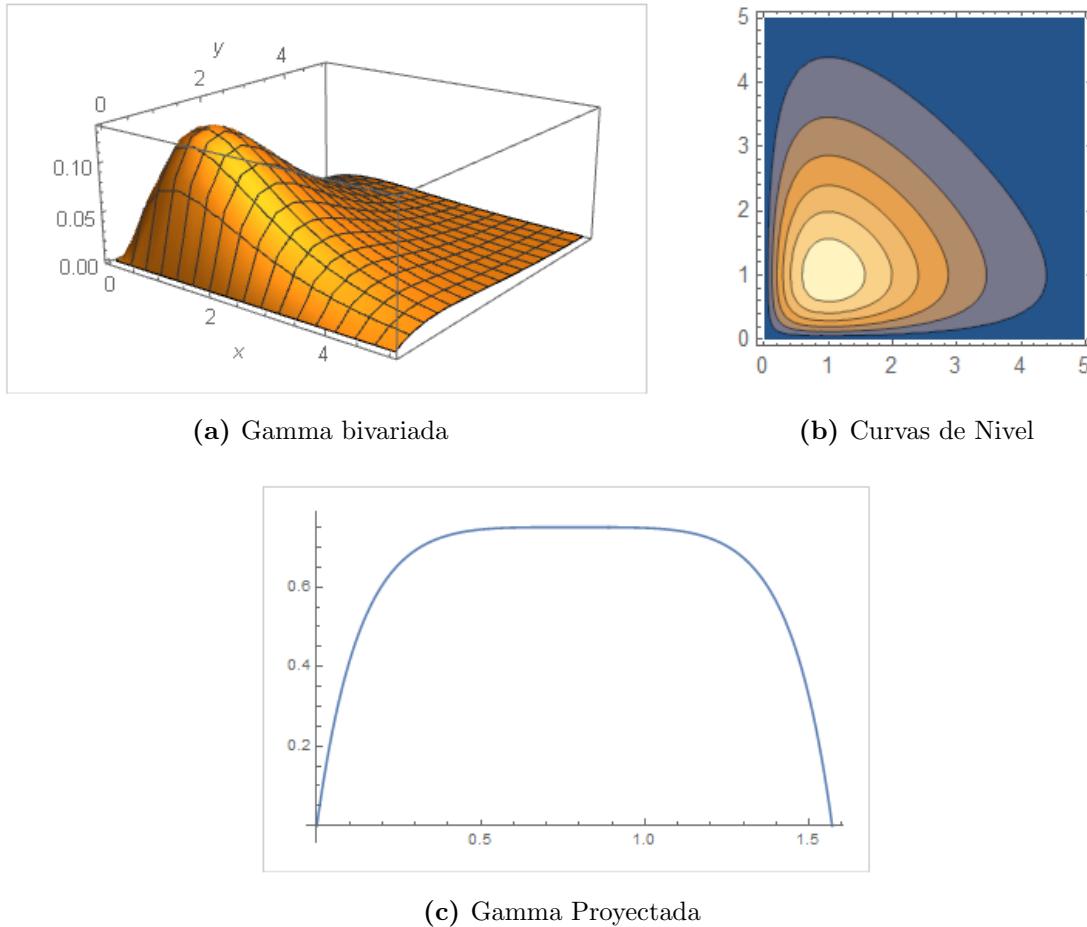


Figura 1: Distribución Gamma proyectada simétrica. Obtenida con $\alpha_1 = \alpha_2 = 2$, $\beta_1 = \beta_2 = 1$. La simetría se presenta respecto a $\theta = \pi/4$.

De la ecuación (2) se debe notar que si $\boldsymbol{\beta}^* = k\boldsymbol{\beta} \forall k > 0$, entonces $PGa(\theta|\boldsymbol{\alpha}, \boldsymbol{\beta}^*) = PGa(\theta|\boldsymbol{\alpha}, \boldsymbol{\beta})$. Es decir, los parámetros $\boldsymbol{\alpha}$ y $\boldsymbol{\beta}$ son estimables pero no identificables si no se

imponen algunas restricciones. Para dirigir este problema, y sin pérdida de generalidad, en este trabajo se considerará $\beta = (1, \beta_2)'$. Hay que señalar que bajo este modelo, aún se pueden obtener comportamientos simétricos o asimétricos y/o unimodales o multimodales.

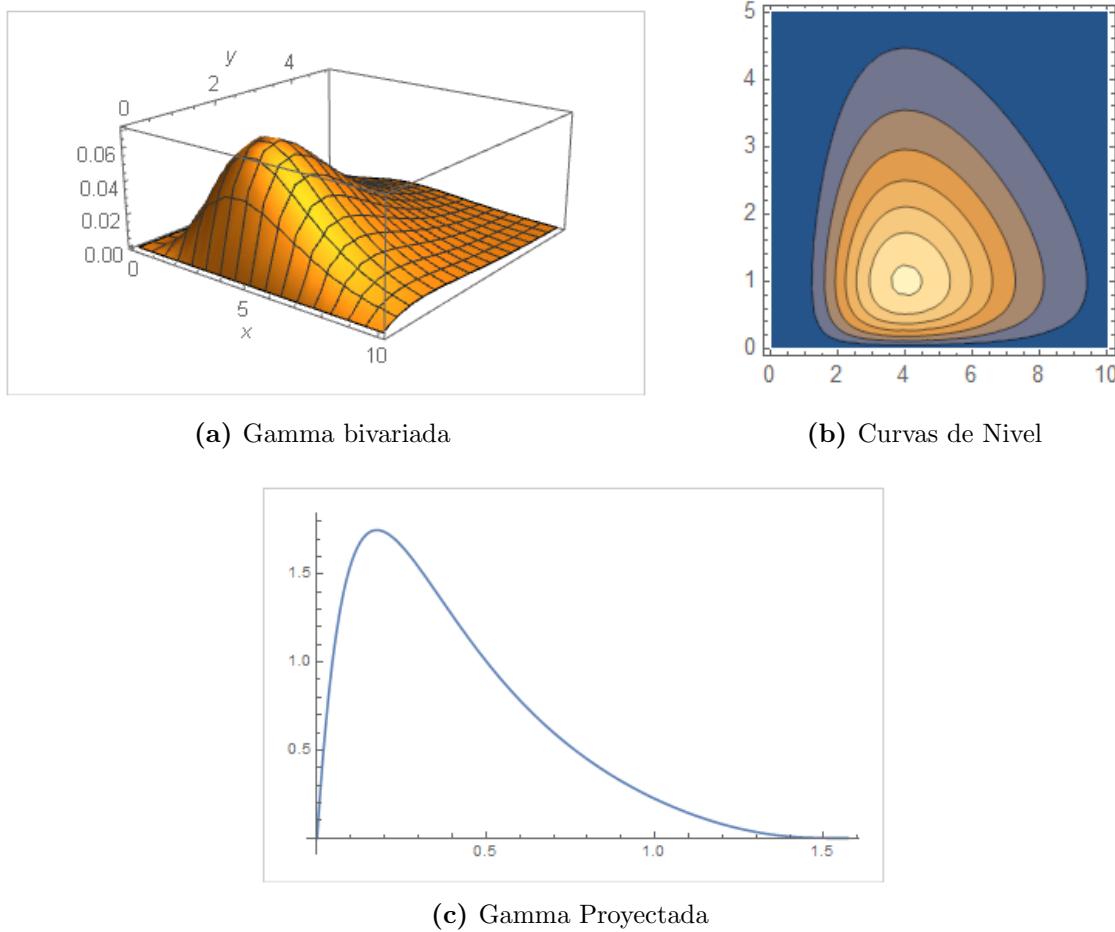


Figura 2: Distribución Gamma proyectada asimétrica. Obtenida con $\alpha_1 = 5$, $\alpha_2 = 2$, $\beta_1 = \beta_2 = 1$.

Los ejemplos aquí señalados muestran que el modelo propuesto Gamma proyectado obtenido a partir de una distribución conjunta definida por distribuciones Gammas independientes, puede ser adecuado para describir una gran variedad de comportamientos.

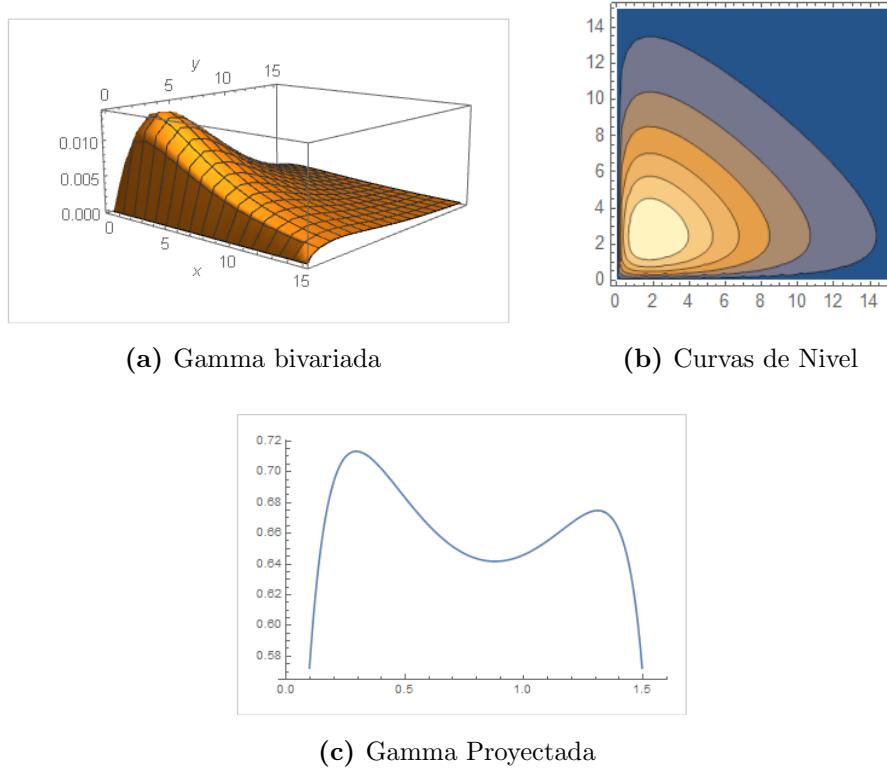


Figura 3: Distribución Gamma proyectada bimodal asimétrica. Obtenida con $\alpha_1 = 1.4$, $\alpha_2 = 1.7$, $\beta_1 = 1/4.5$, $\beta_2 = 1/3.5$.

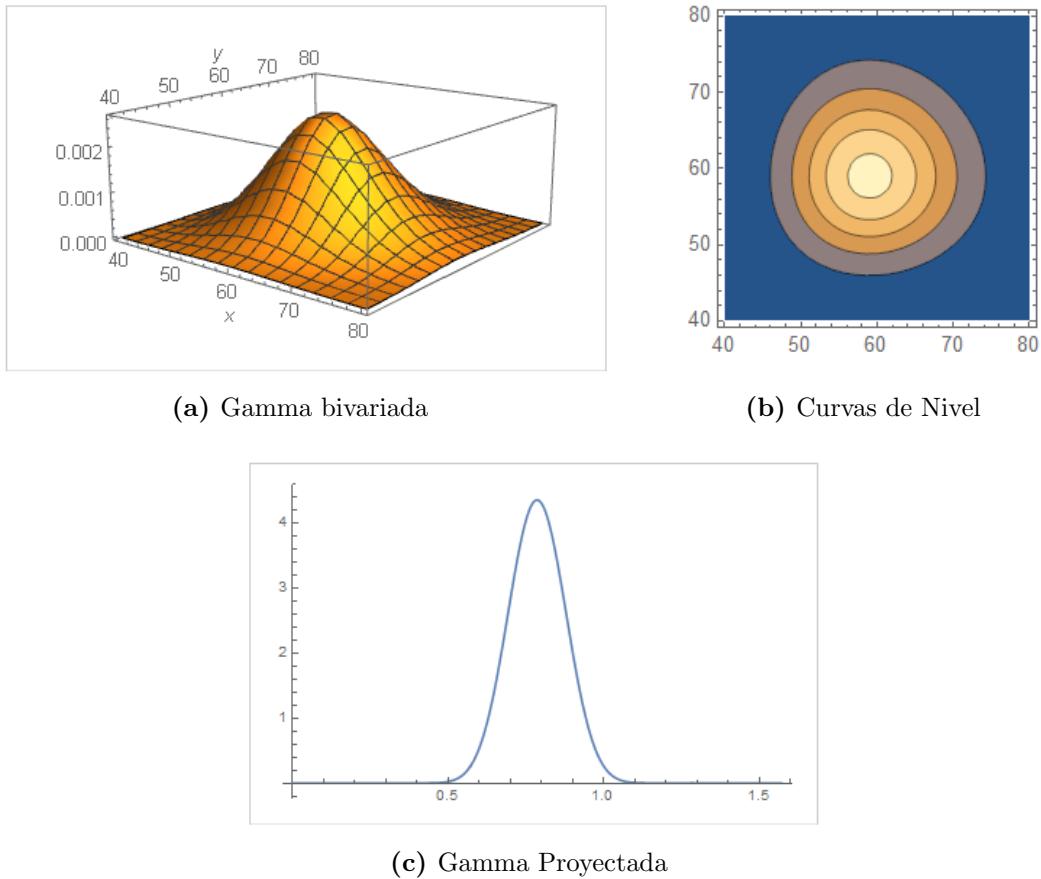


Figura 4: Distribución Gamma proyectada simétrica unimodal. Obtenida con $\alpha_1 = \alpha_2 = 60$, $\beta_1 = \beta_2 = 1$. La simetría se presenta respecto a $\theta = \pi/4$.

2.1. Inferencias Bayesianas

El objetivo es realizar inferencia sobre $\boldsymbol{\alpha}$ y β_2 con base en una muestra aleatoria de ángulos $\{\theta_1, \dots, \theta_n\}$. Se debe notar que si se tuviera una muestra aleatoria $(r, \theta) = \{(r_1, \theta_1), \dots, (r_n, \theta_n)\}$ de la densidad $f(r, \theta | \boldsymbol{\alpha}, \boldsymbol{\beta})$, entonces se estaría en condiciones de realizar inferencias sobre $\boldsymbol{\alpha}$ y β_2 . El problema es que sólo se cuenta con una m.a. de ángulos $\{\theta_1, \dots, \theta_n\}$. La estructura del modelo sugiere tratar los R_i , $i = 1, \dots, n$, como variables latentes. Así, el modelo para los datos-completos sería un modelo Gamma bivariado, con componentes independientes. Esta es la misma idea seguida por Núñez-Antonio y Gutiérrez-Peña (2005) para el caso del modelo Normal proyectado.

Recordando que $\mathbf{y} = (y_1, y_2)' = r(\cos \theta, \sin \theta)'$, entonces la *verosimilitud* queda definida por:

$$f(\mathbf{y}_1, \dots, \mathbf{y}_n | \boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{i=1}^n \text{Gamma}(y_1 | \alpha_1, 1) \cdot \text{Gamma}(y_2 | \alpha_2, \beta_2) \quad (3)$$

Por otro lado si se considera una inicial para el vector $(\boldsymbol{\alpha}, \beta_2)$, definida por el producto de las siguientes densidades.

$$\begin{aligned} \alpha_1 &\sim \text{Gamma}(a_1, b_1) \\ \alpha_2 &\sim \text{Gamma}(a_2, b_2) \\ \beta_2 &\sim \text{Gamma}(c, d), \end{aligned}$$

entonces,

$$\begin{aligned} \pi(\alpha_1 | \boldsymbol{\theta}, \mathbf{r}) &\propto \frac{\prod_{i=1}^n (r_i \cos \theta_i)^{\alpha_1 - 1}}{\Gamma^n(\alpha_1)} \text{Gamma}(\alpha_1 | a_1, b_1) \\ \pi(\alpha_2 | \boldsymbol{\theta}, \mathbf{r}) &\propto \frac{\Gamma(n\alpha_2 + c)}{\Gamma^n(\alpha_2)} \frac{\prod_{i=1}^n (r_i \operatorname{sen} \theta_i)^{\alpha_2 - 1}}{(d + \sum_{i=1}^n (r_i \operatorname{sen} \theta_i))^{(n\alpha_2 + a_2)}} \text{Gamma}(\alpha_2 | a_2, b_2) \\ \pi(\beta_2 | \boldsymbol{\theta}, \mathbf{r}, \boldsymbol{\alpha}) &= \text{Gamma}(n\alpha_2 + c, d + \sum_i^n r_i \operatorname{sen} \theta_i) \\ \pi(r_i | \boldsymbol{\theta}, \mathbf{r}, \boldsymbol{\alpha}, \beta_2) &= \text{Gamma}(\alpha_1 + \alpha_2, \cos \theta_i + \beta_2 \operatorname{sen} \theta_i) \quad \forall i = 1, \dots, n, \end{aligned}$$

donde $\boldsymbol{\theta} = (\theta_1, \dots, \theta_n)$ y $\mathbf{r} = ((r_1, \dots, r_n))$.

Las densidades anteriores se pueden usar en un Gibbs sampler para obtener muestras de la distribución final de $(\boldsymbol{\alpha}, \beta_2)$. Particularmente, para simular de las densidades $\pi(\alpha_1|\boldsymbol{\theta}, \mathbf{r})$ y $\pi(\alpha_2|\boldsymbol{\theta}, \mathbf{r})$ en este trabajo se empleó un muestreo por rechazo adaptativo, comúnmente denominado ARS por sus siglas en inglés. Ver Gilks y Wild, 1992. Sin embargo, se puede usar algún otro algoritmo como los asociados a los Métodos de Monte Carlo vía Cadenas de Markov.

3. Ejemplo

A continuación se presenta un ejemplo de la metodología expuesta en este trabajo. Aunque se analizaron varios casos, a continuación solo se presenta uno de ellos. Para este ejemplo, se simuló una muestra de 500 ángulos de la distribución PGa($\theta|\boldsymbol{\alpha}, \boldsymbol{\beta}$), con $\boldsymbol{\alpha} = (0.7, 5)$ y $\beta_2 = 3$. Se debe recordar que en el modelo PGa($\theta|\boldsymbol{\alpha}, \boldsymbol{\beta}$) propuesto, se tomo $\beta_1 = 1$.

Empleando la metodología descrita en este trabajo se obtuvo una muestra de la distribución final de $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)'$ y de β_2 . La Figura 5, muestra las distribuciones finales marginales de $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)'$ y de β_2 . Por otra parte, los intervalos de máxima probabilidad a posteriori al 95 %, para α_1 , α_2 y β_2 resultaron ser $(0.6045569, 0.7986958)$, $(2.42397, 15.20689)$ y $(1.145645, 11.018773)$ respectivamente. Se puede apreciar que los verdaderos valores de todos los parámetros del modelo se encuentran bien contenidos en los respectivos intervalos de probabilidad.

4. Conclusiones

En este trabajo se ha proporcionado un análisis Bayesiano completo de un nuevo modelo para datos circulares. El modelo propuesto es generado al proyectar radialmente una densidad conjunta bivariada cuyas marginales son densidades Gammas, por lo cual a este modelo se le ha denominado modelo Gamma proyectado. El modelo es bastante flexible para describir datos circulares acotados en el intervalo $(0, \pi/2)$, y su análisis es simple a través de un esquema de muestreo de Gibbs. Consideramos que la metodología discutida en este trabajo ofrece las bases de un análisis bayesiano completo y versátil para el estudio de datos direccionales intercambiables definidos en subconjuntos acotados de \mathbb{S}^p . Adicionalmente, a diferencia de otros modelos para datos direccionales, nuestro modelo se puede extender de manera natural

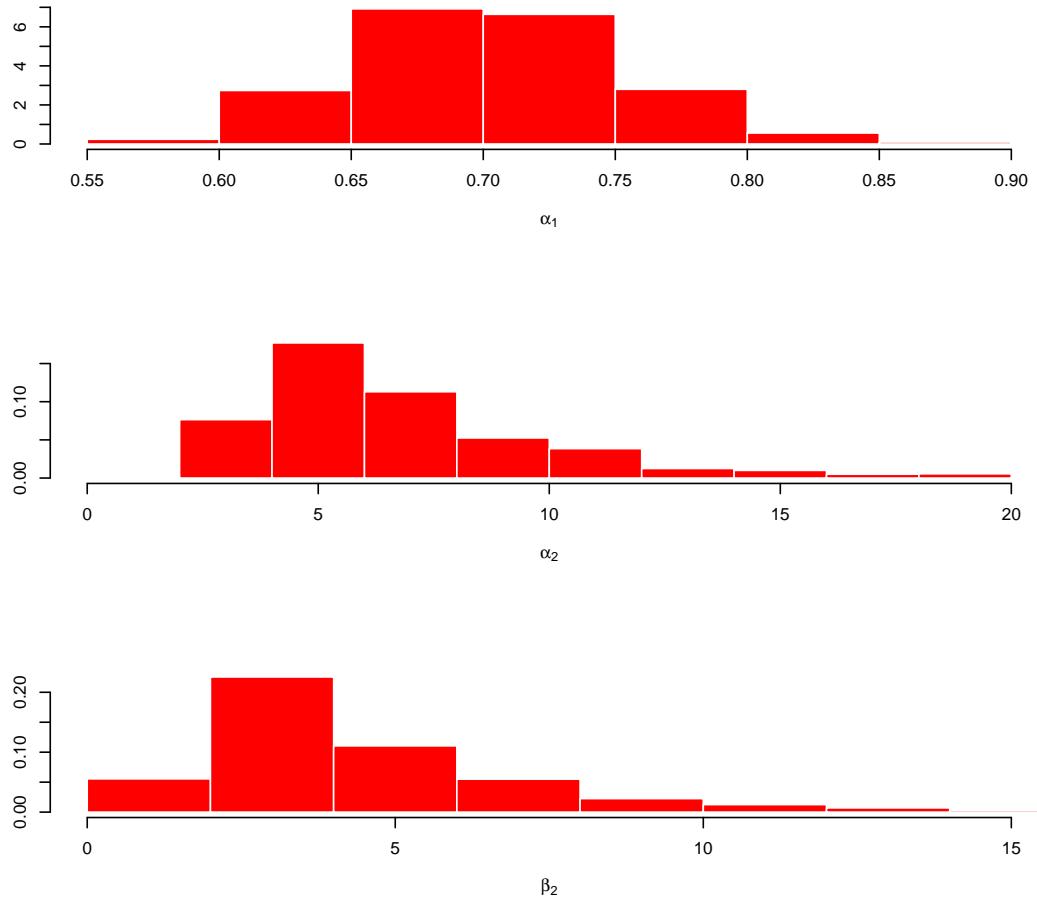


Figura 5: Distribuciones finales de α_1 , α_2 y β_2

a mayores dimensiones. Actualmente, se está trabajando en esta extensión, así como, en la generalización de la densidad conjunta de la que se deriva el modelo Gamma proyectado.

Agradecimientos

Los autores desean agradecer los comentarios de un árbitro anónimo, los cuales mejoraron sustancialmente la presentación de este trabajo.

Bibliografía

Gilks, W. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, 41:337–348.

Mardia, K. V. and Jupp, P. E. (1999). *Directional Statistics*. John Wiley and Sons, New York.

Nuñez Antonio, G. and Gutiérrez-Peña, E. (2005). A Bayesian analysis of directional data using the projected normal distribution. *Journal of Applied Statistics*, 32(10):995–1001.

Samuel Kotz, N. B. and Johnson, N. L. (2000). *Continuous Multivariate Distributions*, volume 1: Models and Applications. John Wiley and Sons, New York, 2nd edition edition.

¿Este Ítem Funciona Igual para Todos? ¿Quién lo Dice? Análisis DIF con Distintos Métodos. Coincidencias y Discrepancias*

Alma Yadhira López García^a

University of New Brunswick

Las evaluaciones estandarizadas se basan en el principio de que los ítems de las pruebas observan propiedades psicométricas similares entre distintas subpoblaciones, por ejemplo, entre hombres y mujeres o entre grupos con nivel socioeconómico alto y bajo. Para asegurar lo anterior se utiliza el análisis conocido como DIF, por sus siglas en inglés (Differential Item functioning). El análisis DIF puede llevarse a cabo a través de distintas metodologías. El objetivo de este trabajo es comparar los resultados que arrojan dos metodologías utilizadas para realizar el análisis DIF de los ítems de la Evaluación Infantil Temprana. El primer método utilizado se basa en la comparación de las probabilidades de respuesta correcta calculadas aplicando la Teoría de Respuesta al Ítem (TRI). El segundo método consiste en calcular el DIF a través de una regresión logística que considera la habilidad de los sustentantes, el idioma y la interacción de ambas variables. Los resultados obtenidos abonan a la discusión sobre la inconsistencia que existe entre diversos métodos utilizados para calcular el DIF. Las técnicas analizadas coinciden en la mayoría de los ítems que presentan un DIF significativo, pero la regresión logística identifica algunos que no son reconocidos por la TRI. Además, en los ítems donde hay coincidencia, las magnitudes del DIF son mayores con la regresión logística que las calculadas a través de la TRI. En conclusión, es necesario seguir afinando las técnicas y probando soluciones para conseguir un consenso en la forma de medir el DIF.

Área-**MSC:** Estadística Aplicada

Sub área-**MSC:** Aplicación en las Ciencias Sociales

* ¿Este Ítem Funciona Igual para Todos? ¿Quién lo Dice? Análisis DIF con Distintos Métodos...

^aalma.lopez@unb.ca

1. Introducción

En el ámbito de la educación, la evaluación ha sido y es parte inherente de los procesos de enseñanza y aprendizaje. Los resultados son, con frecuencia, el fundamento de decisiones de alto impacto en los distintos niveles organizativos del sistema educativo; al nivel federal o estatal podrían ser el detonante de nuevas reformas educativas, de la implementación de programas compensatorios o del fortalecimiento de la capacitación y actualización docente; al nivel escuela y aula, los resultados pueden conducir, por ejemplo, a la implementación de nuevas estrategias de enseñanza, a una organización diferente del tiempo de enseñanza o a establecer un criterio diferente para la conformación de los grupos. Por lo tanto, es crucial garantizar que los instrumentos de evaluación cumplan con los más altos estándares técnicos y que las interpretaciones y el uso que se hace de los resultados sean válidos y confiables APA/AERA/NCME (2014).

El análisis DIF (*Differential Item Functioning*) es una prueba que permite aportar evidencia de que los ítems de un instrumento presentan (o no) sesgo entre diversas subpoblaciones, por ejemplo entre alumnos de nivel socioeconómico alto y bajo, entre hombres y mujeres, o entre alumnos pertenecientes y no pertenecientes a una etnia. Existen diversos métodos para estimar el DIF, y por lo general se espera que estos distintos métodos coincidan en sus resultados y permitan confirmar los hallazgos.

El objetivo de este estudio es comparar las estimaciones del DIF a través de dos técnicas: utilizando las probabilidades de respuesta correcta obtenidas con la Teoría de Respuesta al Ítem (TRI) y aplicando la Regresión Logística. Los resultados muestran que si bien existe coincidencia entre ambas técnicas en más de la mitad de los ítems, las discrepancias en el resto son notorias.

2. La Evaluación Infantil Temprana (EIT)

Los datos utilizados en este estudio provienen de la aplicación de la Evaluación Infantil Temprana (EIT) en una muestra de alumnos de Canadá y de Uruguay.

La Evaluación Infantil Temprana (EIT) es una herramienta aplicada por los docentes, cuya intención es proporcionar a los educadores y a los tomadores de decisiones información sobre el desarrollo de las habilidades de los niños de 4 a 6 años de edad. Esta evaluación consiste en la observación de ciertos indicadores que los docentes pueden fácilmente identificar

durante las actividades normales de la jornada escolar. La EIT no está alineada a un currículum específico, en vez de esto, éste instrumento evalúa cinco dominios del desarrollo de los niños: *Conciencia de sí mismo y del entorno, Habilidades sociales y enfoques para el aprendizaje, Habilidades cognitivas, Lenguaje y Comunicación, y Desarrollo físico*. The Learning Bar (2016)

Este instrumento fue desarrollado por un grupo de investigadores canadienses y ha sido ampliamente utilizado en Canadá y en Australia; recientemente fue traducido al español y aplicado en Uruguay. En 2015, 151 escuelas de los departamentos de Canelones y Colonia, Uruguay, aplicaron la EIT al inicio del ciclo escolar, y 126 de esas escuelas también aplicaron la evaluación al final. Después de esta primera aplicación de la EIT en español se consideró oportuno y necesario analizar sus propiedades psicométricas, a fin de garantizar que la pertinencia de la EIT en diversos contextos The Learning Bar (2016).

Para los fines de este estudio se seleccionó una muestra estratificada por edad de los alumnos evaluados en 2015 a inicio del ciclo escolar. Se seleccionaron 150 niños de cada una de las edades (4, 5 y 6 años). Después, la muestra de Uruguay fue comparada con una muestra similar de los alumnos evaluados en Canadá durante la primavera del 2015 (150 niños por cada grupo de edad).

3. El Análisis DIF y su Estimación

El DIF, como ya se mencionó previamente, es una metodología para analizar el comportamiento estadístico de los ítems. De acuerdo con Zumbo, el DIF ocurre cuando las personas con la misma habilidad para responder un ítem específico muestran distintas probabilidades de responderlo correctamente Zumbo (1999). Algunos autores señalan que el DIF permite identificar inconsistencias del instrumento de medición con factores independientes del constructo a medir, como la pertenencia a grupos minoritarios o a grupos con distintas características demográficas Osterlind and Everson (2009).

La estimación del DIF se compone generalmente de varios indicadores: la magnitud, el tipo (uniforme o no uniforme), y el grupo al que favorece Zumbo (1999). En algunas aproximaciones gráficas también es posible identificar el conjunto de sustentantes que está directamente afectado por el DIF Willms et al. (2007).

En los últimos 30 años se ha desarrollado un importante conjunto de herramientas analíticas para la estimación del DIF y se ha avanzado en su conceptualización. Andriola hace un interesante recuento sobre las distintas técnicas estadísticas existentes y sus diversas clasificaciones Andriola (2001), y Zumbo, por su parte, propone que el DIF ha atravesado por tres generaciones Zumbo (2007). En la primera generación el término comúnmente utilizado era sesgo del ítem, el cual evolucionó al actualmente conocido Funcionamiento Diferencial del Ítem (DIF, por sus siglas en inglés). En la segunda generación se aceptó el cambio de terminología y se buscó desarrollar nuevas metodologías estadísticas para la detección del DIF. Finalmente, en la tercera, existe un énfasis en identificar las causas del DIF. Al parecer la segunda y tercera generación se encuentran vigentes, pues siguen surgiendo nuevas propuestas sobre cómo detectar el DIF Soares et al. (2009) Woods (2009), y se ha avanzado en la identificación de los factores que podrían estar relacionados con el origen del DIF Roussos and Stout (2004).

En este estudio se realiza un análisis de dos de las técnicas más comúnmente utilizadas para la estimación del DIF: el modelo de la Teoría de Respuesta al Ítem (TRI) (conocido también como Chi-cuadrado de Lord) Andriola (2001) Langer et al. (2008) y la Regresión Logística (RL) Zumbo (1999). La primera técnica se basa en la comparación de los parámetros obtenidos a través de la TRI (dificultad y discriminación), utilizando la prueba chi-cuadrado. Su representación matemática se muestra en la ecuación (1) Andriola and Soto (2002).

$$x^2 = V \sum V' \quad (1)$$

Donde:

x^2 tiene dos grados de libertad;

V es el vector de dimensión (1 x 2) de las diferencias entre los parámetros a y b de los grupos de referencia y focal;

V' es el vector traspuesto de V ;

\sum es la inversa de la matriz suma de varianzas-covarianzas de V para los grupos de referencia y focal, cuya dimensión es 2 x 2.

En relación a la segunda técnica, y siguiendo a Benito y colegas, la ecuación general para el análisis de regresión logística adopta la forma descrita en la ecuación (2) Benito et al. (2005).

$$P(y = 1/X) = \frac{e^z}{1 + e^z} \quad (2)$$

Donde:

$P(Y = 1/X)$: es la probabilidad de obtener una respuesta correcta condicionado a X (puntuación observada del sujeto en el test)

Z : representa la combinación lineal de las variables predictoras.

En el análisis del DIF, el modelo de regresión logística se parametriza en los términos descritos en la ecuación (3).

$$z = \beta_0 + \beta_1 X + \beta_2 G + \beta_3 XG \quad (3)$$

Donde:

X : es la puntuación observada de un sujeto en un test

G : la variable de grupo de pertenencia de los sujetos (cultura, idioma, étnia, género);

β_0 : representa el intercepto,

β_1 : es el coeficiente de la habilidad,

β_2 : es el coeficiente para la variable que indica el grupo de pertenencia, y

β_3 : es la interacción entre la habilidad y el grupo.

De esta forma, un ítem presenta DIF uniforme si el efecto del grupo (G) resulta estadísticamente significativo, mientras que la interacción habilidad por grupo (XG) no ejerce ningún efecto sobre el ítem. Por el contrario, si la interacción XG resulta estadísticamente significativa, el ítem presentaría DIF no-uniforme.

4. El DIF basado en la Teoría de Respuesta al Ítem (TRI) vs Regresión Logística (RL)

El cálculo del DIF a través de la TRI y de la técnica de RL arroja resultados que podrían clasificarse en tres grupos: coincidentes, cuando ambas técnicas arrojan el mismo resultado; similares, cuando los dos métodos reconocen que el ítem presenta un DIF significativo pero algunas de sus características son distintas; y discordantes, son los casos en donde no hay coincidencia entre los resultados obtenidos. Un ejemplo de cada uno de estos tres tipos de

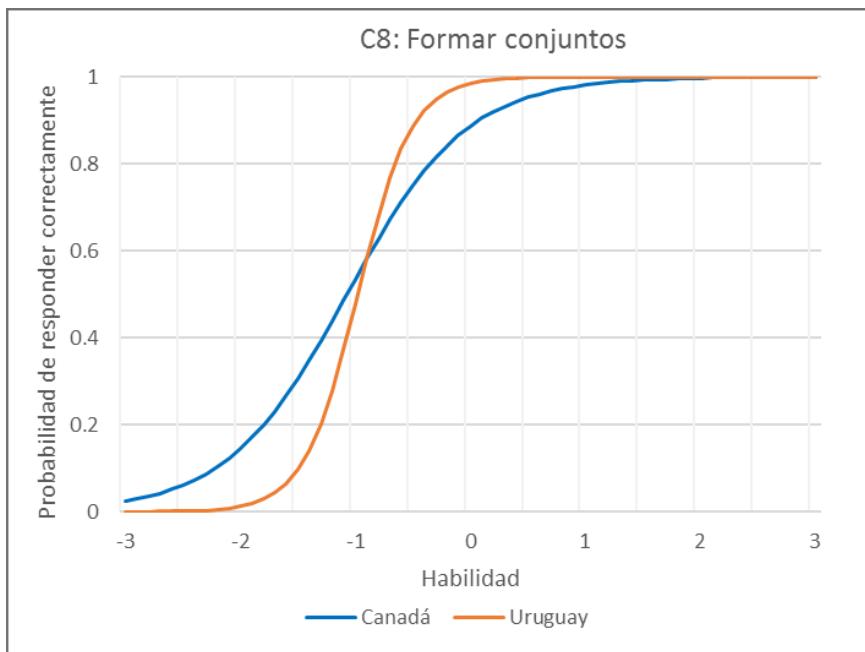


Figura 1: Comparativo de la probabilidad de respuesta correcta del ítem C8.

resultados se puede observar en los ítems C2, C5 y C8 del dominio Habilidades Cognitivas que evalúa la EIT. A continuación se describe cada uno.

Coincidente. La TRI y la RL coinciden en señalar que el ítem C2 (entender cómo se manejan los libros) tiene las mismas propiedades psicométricas en las muestras de Uruguay y Canadá. En otras palabras, el ítem no presenta DIF, los alumnos evaluados de dichos países encuentran este reactivo igualmente difícil, y ambas técnicas lo confirman.

Discordante. Con la TRI se identificó que el ítem C8 (formar conjuntos) tiene un DIF significativo, se localiza principalmente en los valores negativos de la habilidad, y es un DIF No uniforme. La Figura 1 presenta más detalles al respecto. El eje de las X representa la habilidad en un continuo que va de -3 a 3, y el eje de las Y ilustra la probabilidad de responder correctamente éste ítem. Las curvas representan la relación que existe entre estas dos variables en cada uno de los países. Se observa que el DIF (distancia entre las dos curvas) se ubica en el rango de -3 a 1.5 de la habilidad, y que en los niveles más bajos de la misma se favorece a Canadá, no obstante este comportamiento cambia a partir del -1.2 aproximadamente. De este punto de la habilidad en adelante la probabilidad de responder correctamente este ítem es más alta en Uruguay que en Canadá.

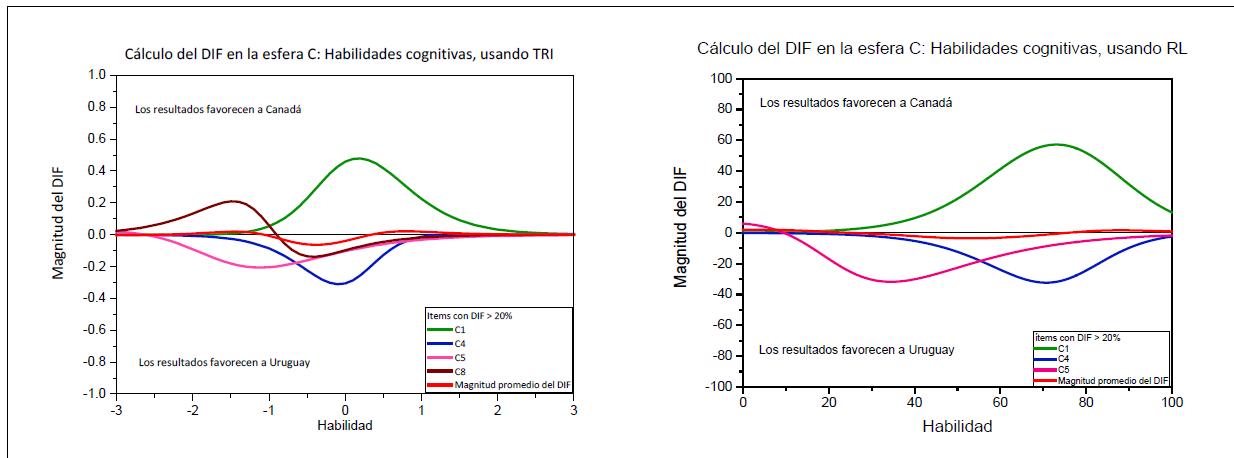


Figura 2: Comparativo del cálculo del DIF en el dominio C, usando TRI y RL.

Por otra parte, la RL indica que el ítem C8 no presenta DIF. Este resultado sugiere que el ítem tiene el mismo grado de dificultad para los alumnos de las dos poblaciones evaluadas. Sin embargo, ante los discordantes resultados obtenidos por las dos técnicas se presenta el dilema de cómo reconocer cual es el resultado correcto.

Similar. El ítem C5 (agrupar objetos) es un ejemplo de resultados similares entre la TRI y la RL. Sin embargo, la similitud tiene que ver sólo con la identificación de que este ítem presenta un DIF significativo; la magnitud (21 vs. 32) y el tipo de DIF son distintos (Uniforme vs. No uniforme).

La representación gráfica del DIF permite identificar fácilmente las diferencias entre ambas técnicas. Siguiendo con el ejemplo de la esfera C, si comparamos los resultados obtenidos se verían de la forma como se ilustra en la Figura 2. En este figura cada línea representa un ítem con DIF (la línea se obtiene de la diferencia entre las probabilidades de respuesta correcta de los dos países), el punto más alto que alcanza la línea representa la magnitud del DIF, y su ubicación indica el país que se ve favorecido.

La primera diferencia que destaca es la cantidad de ítems identificados, la TRI identifica 4 ítems y la RL identifica sólo 3. La segunda diferencia es la magnitud del DIF, por ejemplo, en el ítem C1 es de 48 con la TRI y de 57 con la RL.

Finalmente, un análisis de todos los reactivos del instrumento con ambas técnicas muestra que en 64 % de los casos hay coincidencia, en 19 % los resultados son similares y en 17 % son discordantes (ver Tabla 1). Las implicaciones de estos resultados son variadas. Por una

parte, los resultados similares, aún y cuando no muestran un total acuerdo en todos los indicadores, permiten confirmar que el ítem en cuestión presenta cierto sesgo en uno de los grupos evaluados, por lo tanto es necesario revisarlo, corregirlo y probarlo nuevamente con el objetivo de que el reactivo tenga un comportamiento estadístico equiparable en las distintas poblaciones que se desean evaluar.

En el caso de los resultados discordantes es posible que se requiera una postura exigente y asumir que el DIF identificado es correcto. Esto probablemente ocasione revisiones innecesarias cuando el DIF no es real, pero también evitara que pasemos por alto casos en donde efectivamente existe un sesgo en el ítem. Vale la pena considerar el ítem no solo de manera individual, sino también en el contexto de la esfera a la que pertenece, esto ayudará a hacer una mejor valoración de la importancia de corregir el ítem. Por ejemplo, corregir un ítem discordante como en la esfera A, Conciencia de sí mismo y del entorno, puede no ser tan urgente como corregir 3 ítems discordantes, de los 8 que conforman la esfera de Lenguaje y Comunicación.

5. Conclusiones

El objetivo de este estudio fue analizar las estimaciones del DIF a través de dos técnicas: la Teoría de Respuesta al Ítem (TRI) y la Regresión Logística, con el fin de identificar coincidencias y discrepancias. Los resultados muestran que en la mayoría de los casos las estimaciones de ambas técnicas coinciden, sin embargo, los ítems identificados presentan discrepancias en la magnitud y tipo de DIF.

Estos hallazgos confirman lo que otros hallazgos ya habían encontrado Finch (2005) Kristjansson et al. (2005) Woods (2009), que un ítem puede o no presentar sesgo ante una población, dependiendo de la técnica que se utilice para su estimación. La inversión de recursos involucrados en la corrección de un reactivo no es vana, por lo que seleccionar el método que se va a utilizar se convierte en una decisión delicada. Pero más importante que los recursos es la posible interpretación sesgada que se podría derivar de una serie de reactivos no identificados con DIF, pero que en realidad sí tienen un comportamiento estadístico diferente entre las subpoblaciones evaluadas.

En suma, es necesario analizar la calidad de los instrumentos de evaluación desde diversas perspectivas analíticas, con la finalidad de tomar decisiones suficientemente informadas en

Esfera	Ítem	Resultado	Teoría de Respuesta al Ítem			Regresión Logística		
			Magnitud DIF (%)	Tipo de DIF	A favor de	Magnitud DIF (%)	Tipo de DIF	A favor de
Conciencia de sí mismo y del entorno	a1	Similar	22	U	Canadá	31	U	Canadá
	a7		31	U	Uruguay	37	U	Uruguay
	a5	Discordante	No presenta DIF			27	U	Canadá
	a2, a3, a4, a6 y a8	Coincidente	No presentan DIF					
Habilidades sociales y enfoques para el aprendizaje	b1	Similar	35	NU	Uruguay	36	U	Uruguay
	b3		25	U	Canadá	28	U	Canadá
	b7		23	U	Canadá	26	U	Canadá
	b2, b4, b5, b6 y b8	Coincidente	No presentan DIF					
Habilidades cognitivas	c1	Similar	48	U	Canadá	57	U	Canadá
	c4		31	U	Uruguay	32	U	Uruguay
	c5		21	U	Uruguay	32	NU	Uruguay
	c8	Discordante	21	NU	Canadá			
	c2, c3, c6 y c7	Coincidente	No presentan DIF					
Lenguaje y Comunicación	d1	Discordante	No presenta DIF		24	U	Uruguay	
	d6		No presenta DIF		28	U	Canadá	
	d8		No presenta DIF		27	U	Canadá	
	d2, d3, d4, d5 y d7	Coincidente	No presentan DIF					
Desarrollo físico	e1	Discordante	No presenta DIF		24	U	Canadá	
	e2		No presenta DIF		22	U	Uruguay	
	e3, e4, e5, e6, e7, e8, e9 y e10	Coincidente	No presentan DIF					

Tabla 1: Comparativo del análisis DIF aplicado a todos los ítems de la EIT usando TRI y RL.

relación a su validez en diversos contextos y con sujetos de distintas características.

Por otra parte, también se concluye que es necesario seguir avanzando en el desarrollo de técnicas que aporten información confiable sobre las propiedades psicométricas de los instrumentos de medición, y permitan identificar con certeza las áreas de mejora.

Bibliografía

Andriola, W. B. (2001). Descrição dos principais métodos para detectar o funcionamento diferencial dos itens (dif). *Psicologia: Reflexão e Crítica*, 14(3):643–652.

Andriola, W. B. and Soto, J. L. G. (2002). *Detección del funcionamiento diferencial del ítem (DIF) en tests de rendimiento: aportaciones teóricas y metodológicas*. Universidad Complutense de Madrid.

APA/AERA/NCME (2014). *Standards for educational and psychological testing*. APA AERA NCME.

Benito, J. G., García, J. L. P., et al. (2005). Regresión logística: alternativas de análisis en la detección del funcionamiento diferencial del ítem. *Psicothema*, 17(3):509–515.

Finch, H. (2005). The mimic model as a method for detecting dif: Comparison with mantel-haenszel, sibtest, and the irt likelihood ratio. *Applied Psychological Measurement*, 29(4):278–295.

Kristjansson, E., Aylesworth, R., McDowell, I., and Zumbo, B. D. (2005). A comparison of four methods for detecting differential item functioning in ordered response items. *Educational and Psychological Measurement*, 65(6):935–953.

Langer, M. M., Hill, C. D., Thissen, D., Burwinkle, T. M., Varni, J. W., and DeWalt, D. A. (2008). Item response theory detected differential item functioning between healthy and ill children in quality-of-life measures. *Journal of clinical epidemiology*, 61(3):268–276.

Osterlind, S. J. and Everson, H. T. (2009). *Differential item functioning*, volume 161. Sage Publications.

- Roussos, L. A. and Stout, W. (2004). Differential item functioning analysis. *The Sage handbook of quantitative methodology for the social sciences*, pages 107–116.
- Soares, T. M., Gonçalves, F. B., and Gamerman, D. (2009). An integrated bayesian model for dif analysis. *Journal of Educational and Behavioral Statistics*, 34(3):348–377.
- The Learning Bar (2016). *Early Years Evaluation Annual Report. Administración Nacional de Educación Pública (ANEP) de Uruguay. 2015 Summary*. The Learning Bar.
- Willms, J., Tramonte, L., and Chin, N. (2007). The use of item response theory for scaling behavioural outcomes in the nlscy for the successful transitions project. *Unpublished report. Ottawa, ON: Human Resources and Skills Development Canada*.
- Woods, C. M. (2009). Evaluation of mimic-model methods for dif testing with comparison to two-group analysis. *Multivariate Behavioral Research*, 44(1):1–27.
- Zumbo, B. D. (1999). A handbook on the theory and methods of differential item functioning (dif). *Ottawa: National Defense Headquarters*.
- Zumbo, B. D. (2007). Three generations of dif analyses: Considering where it has been, where it is now, and where it is going. *Language assessment quarterly*, 4(2):223–233.

Una Propuesta Bayesiana para Medir el Grado de Traslape Entre Dos Especies de Animales

Gabriel Núñez Antonio^a
UAM-I

Alberto Contreras Cristán, Eduardo Gutiérrez Peña^b
IIMAS-UNAM

Manuel Mendoza Ramírez^c
ITAM

Eduardo Mendoza Ramírez
Instituto de Investigaciones sobre los Recursos Naturales (INIRENA)
Universidad Michoacana de San Nicolás de Hidalgo

Actualmente las técnicas de foto-trampeo, el desarrollo de modelos estadísticos más flexibles, el avance en la teoría sobre datos direccionales y la capacidad de procesamiento computacional de la información, han permitido un mayor desarrollo en ciertas áreas de las ciencias ecológicas. En particular, un mejor análisis de modelos para estudiar el traslape entre especies ofrece la oportunidad de entender de mejor manera los mecanismos de coexistencia de ciertos animales en su habitat natural.

Con el objetivo de contribuir al estudio del traslape entre especies de mamíferos a través de datos de foto-trampeo, en este trabajo se analiza un modelo bayesiano no paramétrico para variables circulares. Se propone una estimación de las densidades predictivas para los registros de avistamientos de animales y se ofrecen inferencias sobre una medida de traslape propuesta originalmente por Weitzman (1970).

Área-MSC: Estadística Bayesiana Aplicada, Ecología.

Subárea-MSC: Modelos no paramétricos, Índices de traslape.

^agab.nuneza@gmail.com (autor responsable)

^balberto@sigma.iimas.unam.mx, eduardo@sigma.iimas.unam.mx

^cmendoza@itam.mx

1. Introducción

En ecología es importante entender los mecanismos que permiten la coexistencia de dos o más especies de animales en alguna zona geográfica. El surgimiento de la técnica de foto-trampeo ha permitido contar con información relevante para evaluar patrones espacio-temporales en especies de mamíferos difíciles de detectar en su hábitat natural. En este caso, los datos consisten de registros durante el día de la hora en que las especies son fotografiadas desde cámaras-trampa. Así, los registros se pueden considerar como datos circulares a lo largo del día y, por lo tanto, estos patrones de actividad se pueden describir a través de algún modelo semiparamétrico o no paramétrico de distribuciones de probabilidad circulares. Con el objetivo de describir el traslape de especies de animales, en este trabajo se propone un modelo bayesiano no paramétrico para variables circulares (variables definidas sobre el círculo unitario). La metodología se ejemplifica con datos simulados. Se propone una estimación de las densidades predictivas para las mediciones de avistamiento de cada especie y se ofrecen inferencias sobre un índice de traslape definido por el área bajo la curva que se forma al tomar el mínimo de las dos densidades en cada punto del tiempo.

2. Cálculo de Traslape

En la literatura se han propuesto varias medidas de traslape; ver, por ejemplo, Ridout y Linkie (2009) para una revisión de este tema. Este trabajo se enfoca en el *coeficiente de traslape* propuesto por Weitzman (1970) para variables reales, y se adapta al caso de variables circulares, aquí denotadas por Θ . Así, un índice de traslape entre dos modelos (densidades) f y g , se puede definir como:

$$\Delta(f, g) = \int \min\{f(\theta), g(\theta)\}d\theta. \quad (1)$$

Una propuesta natural para estimar $\Delta(f, g)$ es estimar las densidades f y g y calcular $\Delta(f, g)$ de manera numérica. En el contexto de datos de foto-trampeo, éstos presentan cierta periodicidad (avistamientos por día, mes, año, etc.) y por tanto se pueden considerar como datos circulares (ver, por ejemplo, Mardia y Jupp, 2000). En general, este tipo de datos presentan características tales como multimodalidad y asimetría (ver Godínez, 2014, y las referencias allí incluidas). En estos casos, puede ser preferible considerar modelos no paramétricos para estimar las densidades f y g . En estadística Bayesiana el enfoque usual

es emplear mezclas basadas en procesos Dirichlet (MPD) considerando como distribuciones base algún modelo paramétrico definido para variables con soporte el círculo unitario. En este trabajo se analiza el uso de modelos MPD con distribuciones normales proyectadas (Nuñez-Antonio *et al.*, 2015).

2.1. El Modelo MPD Normal Proyectado

Un modelo de MPD se puede definir de manera jerárquica y es equivalente a un modelo de mezclas infinitas numerables de densidades paramétricas (Sethuraman, 1994). En nuestro caso, para construir un modelo MPD normal proyectado se puede proseguir de la siguiente manera. Sea:

$$\begin{aligned}\mathbf{X}|\boldsymbol{\mu} &\sim N(\cdot|\boldsymbol{\mu}, \mathbf{I}) \\ \boldsymbol{\mu}|H &\sim H \\ H &\sim DP(\alpha, H_0)\end{aligned}$$

donde H_0 es la distribución normal bivariada $N(\boldsymbol{\mu}_0, \Lambda_0^{-1})$. Se considera que $\boldsymbol{\mu}_0$ y Λ_0 son conocidos y se supone una distribución inicial $\alpha \sim Ga(a_0, b_0)$ para el parámetro de concentración del proceso Dirichlet. Así, la densidad de \mathbf{X} se puede expresar como una mezcla infinita de distribuciones normales bivariadas,

$$f(\mathbf{x} | \boldsymbol{\rho}, \boldsymbol{\mu}) = \sum_{s=1}^{\infty} \rho_s \phi(\mathbf{x} | \boldsymbol{\mu}_s),$$

donde $\phi(\cdot | \boldsymbol{\mu})$ denota la función de densidad de una distribución normal bivariada con matriz de covarianzas la matriz identidad.

Ahora, si se construye una variable circular Θ , proyectando radialmente \mathbf{X} sobre el círculo unitario, entonces la variable Θ seguirá una mezcla infinita de distribuciones normales proyectadas. Es decir,

$$f(\theta | \boldsymbol{\rho}, \boldsymbol{\mu}) = \sum_{s=1}^{\infty} \rho_s \phi^{PN}(\theta | \boldsymbol{\mu}_s),$$

donde $\phi^{PN}(\cdot | \boldsymbol{\mu})$ es la densidad de una distribución normal proyectada (ver Nuñez-Antonio *et al.* 2015.)

2.2. Estimación del Traslape Entre Dos Densidades

Usando el algoritmo propuesto por Kalli *et al.*, (2011) y los resultados de Nuñez-Antonio y Gutiérrez-Peña (2005), se pueden llevar a cabo inferencias para el modelo MPD normal proyectado y en particular ofrecer estimadores de cada uno de los modelos f y g en (1). En este trabajo se proponen como estimadores de $f(\theta)$ y de $g(\theta)$ a las correspondientes densidades predictivas finales $f(\theta_{n+1} | \theta_1, \dots, \theta_n)$ y $g(\theta_{n+1} | \theta_1, \dots, \theta_n)$. Así, un estimador del indice de traslape (1), bajo este procedimiento, resulta ser

$$\hat{\Delta}(f, g) = \int \min\{f(\theta_{n+1} | \theta_1, \dots, \theta_n), g(\theta_{n+1} | \theta_1, \dots, \theta_n)\} d\theta.$$

3. Ilustración (Datos Simulados)

Para este ejemplo se consideraron los siguientes dos modelos

$$\begin{aligned} f(\theta) &= \phi^{PN}(\theta | \boldsymbol{\mu}_1) \\ g(\theta) &= 0.3 \phi^{PN}(\theta | \boldsymbol{\mu}_1) + 0.7 \phi^{PN}(\theta | \boldsymbol{\mu}_2), \end{aligned}$$

donde $\boldsymbol{\mu}_1 = (-1, 0)^t$ para el modelo $f(\cdot)$, y $\boldsymbol{\mu}_1 = (0, 1)^t$ y $\boldsymbol{\mu}_2 = (0, -1)^t$ para el modelo $g(\cdot)$. Estos modelos pretenden representar la distribución de los registros de foto-trampeo de dos especies de animales. En la Figura 1 se puede apreciar una representación (lineal) de los modelos f y g . Para estos dos modelos, el verdadero valor de coeficiente de traslape, definido en (1), resulta ser de 0.5653 y está representado por el área en color gris en la Figura 1.

Para ejemplificar la metodología propuesta para la estimación del coeficiente de traslape $\Delta(f, g)$, se simuló una muestra aleatoria, $\theta_1, \dots, \theta_n$, de tamaño $n = 100$ de cada uno de los modelos (angulares) $f(\theta)$ y $g(\theta)$. En la Figura 2 se exhibe una representación (lineal) de cada uno de los dos conjuntos de datos.

Para este ejemplo se consideró la siguiente especificación inicial para el modelo MPD normal proyectado:

$$\begin{aligned} \boldsymbol{\mu}_0 &= (0, 0)^t \\ \boldsymbol{\Lambda}_0 &= \text{Diag}(1, 1) \\ a_0 &= 2 \\ b_0 &= 4. \end{aligned} \tag{2}$$

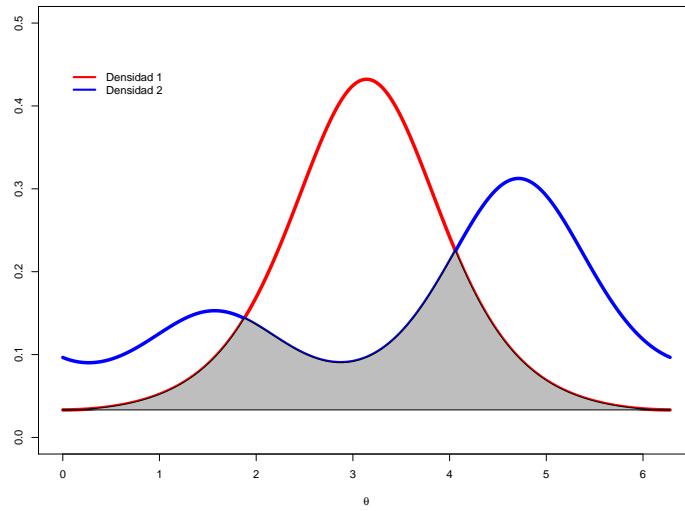


Figura 1: Representación lineal de los dos modelos de mezcla para variables angulares que representan el momento de avistamiento por foto-trampeo de dos especies de animales, y del área que representa el traslape entre los dos modelos.

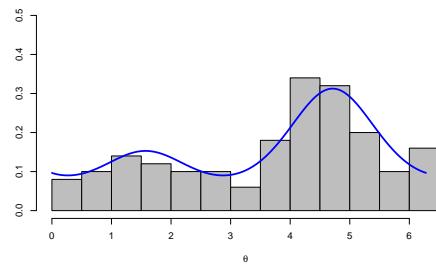
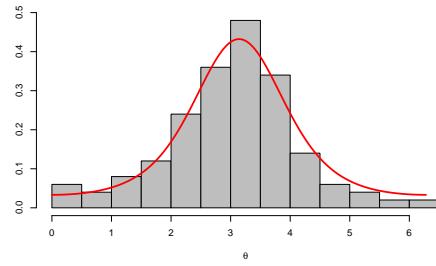


Figura 2: Muestras de tamaño 100 bajo cada uno de los modelos f (arriba) y g (abajo).

En la Figura 3 se muestra la distribución final de $\Delta(f, g)$ basada en muestras Monte Carlo de tamaño 1000 obtenidas por medio del algoritmo de *slice sampling* propuesto por Kalli *et al.* (2011) para modelos de mezclas. Cada muestra representa una realización de la distribución final de las densidades f y g bajo el modelo de mezclas no paramétrico descrito en la Sección 2.1, y para cada una de estas realizaciones se calculó $\Delta(f, g)$ de manera numérica. El intervalo al 95 % para $\Delta(f, g)$, bajo este procedimiento, resultó ser $(0.4267, 0.622)$. Se puede observar que el verdadero valor del índice de traslape queda contenido dentro del intervalo.

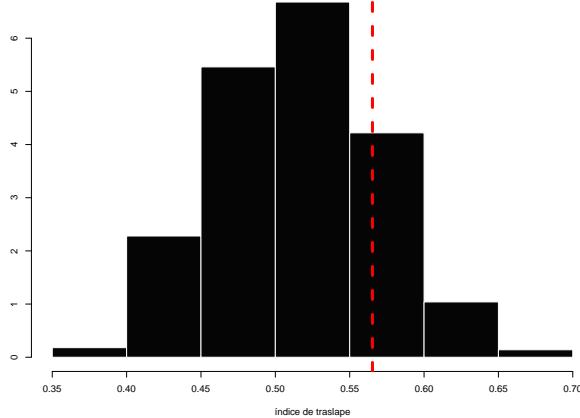


Figura 3: Distribución final del índice de traslape $\Delta(f, g)$. La línea roja punteada representa el verdadero valor del índice de traslape.

4. Comentarios Finales

La metodología propuesta en este trabajo representa una alternativa para la modelación del índice de traslape $\Delta(f, g)$. Aunque la Figura 3 y el correspondiente intervalo de probabilidad para $\Delta(f, g)$ son el resultado de una sola corrida, sirven para ilustrar el tipo de análisis que se puede realizar. Específicamente, el análisis propuesto no sólo proporciona una estimación puntual de $\Delta(f, g)$, sino que produce una distribución final a partir de la cual se puede llevar a cabo cualquier tipo de inferencia sobre este parámetro de interés.

Cabe señalar que este trabajo de investigación se encuentra en una etapa preliminar, pues hay varios aspectos que se deben estudiar con mayor atención. Por un lado, se debe analizar

el desempeño del algoritmo con conjuntos de datos pequeños, digamos de tamaño $n = 10$. Lo anterior, debido a que en muchas situaciones reales los registros de foto-trampeo sobre mamíferos pueden ser escasos (Godínez, 2014). Por otra parte, se debe analizar el papel de la especificación inicial (2) en el modelo MPD normal proyectado, debido a que el peso relativo de la distribución inicial es mayor cuando se tienen muy pocos datos, respecto a la situación en la que se tienen grandes conjuntos de datos. Actualmente se está trabajando en estos aspectos, así como en la aplicación del procedimiento propuesto al análisis de datos reales.

Bibliografía

Geange, Pledger, Burns, and Shima (2011). A unified analysis of niche overlap incorporating data of different types. *Methods in Ecology & Evolution*, 2:175–184.

Godínez, G. O. (2014). *Patrones de Actividad Espacio Temporal de los Ungalos de la Reserva de la Biosfera El Triunfo, Chiapas, México*. Universidad Michoacana de Sn. Nicolás de Hidalgo. Tesis de licenciatura.

Horn, H. (1966). Measurement of overlap in comparative ecological studies. *The American Naturalist*, 100(914):419–424.

Kalli, M., Griffin, J. E., and Walker, S. G. (2011). Slice sampling mixture models. *Statistics and Computing*, 21(1):93–105.

Lu, R., Smith, E. P., and Good, I. (1989). Multivariate measures of similarity and niche overlap. *Theoretical Population Biology*, 35:1–21.

Mardia, K. V. and Jupp, P. E. (1999). *Directional Statistics*. John Wiley and Sons Ltd, London, UK.

May, R. and Mac Arthur, R. (1972). Niche overlap as a function of environmental variability. *Proc. Nat. Acad. Sci. USA*, 69(5):1109–1113.

Nuñez Antonio, G., Ausín, C., and Wiper, M. (2015). Bayesian nonparametric models of circular variables based on Dirichlet process mixture of normal distributions. *Journal of Agricultural, Biological and Environmental Statistics*, 20(1):47–64.

- Nuñez Antonio, G. and Gutiérrez-Peña, E. (2005). A Bayesian analysis of directional data using the projected normal distribution. *Journal of Applied Statistics*, 32:995–1001.
- Ridout, M. S. and Linkie, M. (2009). Estimating overlap of daily activity patterns from camera trap data. *Journal of Agricultural, Biological and Environmental Statistics*, 14(3):322–337.
- Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650.
- Weitzman, M. S. (1970). Measure of the overlap of income distribution of white and negro families in the United States. Technical Report 22 U.S., Department of Commerce, Bureau of the Census, Washington, DC.

Estudio Morfométrico de la Plaga *B. Cockerelli (Sulc)* en Dos Variedades de Jitomate Mediante Análisis Factorial y Componentes Principales

Eduardo Pérez Castro^a, María Guzmán Martínez, Ramón Reyes Carreto,
David Alejandro Ozuna Santiago

Universidad Autónoma de Guerrero

Haidel Vargas Madríguez

Centro Universitario de la Costa Sur, Universidad de Guadalajara

El objetivo de este trabajo fue evaluar el efecto de dos variedades de jitomate, Charanda F1 y Rafaello, en la morfometría de la plaga *B. cockerelli (Sulc)* en su etapa adulta en condiciones de invernadero, mediante el uso de componentes principales y análisis factorial. Para el estudio se consideraron 5 variables morfométricas: LARGO DEL CUERPO, ANCHO DEL CUERPO EN TORAX, LONGITUD DE LAS ANTENAS, LONGITUD DEL ALA y ANCHO DEL ALA.

El análisis de componentes principales se realizó para las dos variedades de jitomate, Charanda F1 y Rafaello. Del análisis para Charanda F1, muestra que las Hembras tienen la LONGITUD DEL ALA más grande que los Machos. Pero son los Machos quienes tienen la LONGITUD DEL CUERPO y las antenas más grandes que las Hembras. En el caso de Rafaello, las Hembras tienen la LONGITUD DEL ALA más grande, pero la LONGITUD DE LAS ANTENAS más pequeñas al igual que el cuerpo.

El análisis factorial también se realizó por variedad del jitomate. El análisis para Charanda F1, muestra que las variables más importantes para explicar la morfometría del insecto son LONGITUD DEL ALA, ANCHO DEL ALA y la LONGITUD DEL CUERPO. Mientras que para Rafaello son LONGITUD DE LAS ANTENAS y la LONGITUD DEL CUERPO.

^alaloperezcastro@gmail.com

Clasificación: Trabajo de investigación.

Área-*MSC*: Estadística multivariada.

Subárea-*MSC*: Análisis factorial y componentes principales.

1. Introducción

La morfometría ha sido usada para estimar variaciones interpoblacionales en diversas especies (Martínez-Ibarra *et al.* 2006). Existe en la literatura una gran cantidad de estudios, que reportan el efecto que tienen las plantas hospederas en la morfometría de los insectos, entre ellos se encuentra el trabajo de investigación de Vargas-Madríz (2010). Este investigador realizó un estudio morfométrico de la plaga *B. cockerelli (Sulc)* en dos variedades de jitomate: Charanda F1 y Rafaello. Realizando un análisis de varianza, encontró efectos de las variedades de jitomate Rafaello y Charranda F1 sobre las variables LARGO DEL CUERPO, LARGO DE ANTENAS Y ANCHO DE ALAS. Reportó que el factor sexo del insecto influye en la diferenciación del LARGO DE ANTENAS, LARGO DE ALAS Y ANCHO DE ALAS.

El objetivo de este trabajo fue investigar el comportamiento morfométrico de *B. cockerelli (Sulc)* en las variedades de jitomate Charanda F1 y Rafaello a través de 5 variables morfométricas cuantitativas, LARGO DEL CUERPO, ANCHO DEL CUERPO EN TORAX, LONGITUD DE LAS ANTENAS, LONGITUD DEL ALA y ANCHO DEL ALA, mediante el uso de componentes principales y análisis factorial.

En la Sección 2, se da una introducción al análisis factorial y componentes principales. En la Sección 3, se muestran los resultados del estudio y finalmente en la Sección 4 se dan las conclusiones.

2. Metodología

El análisis multivariado juega un papel importante en las distintas ramas de la ciencias, una de ellas, si no es que la más importante, las ciencias agrícolas.

Dentro del análisis multivariado están el análisis factorial (AF) y el análisis de componentes principales (ACP), estas técnicas ayudan a reducir la dimensionalidad de los datos. Algunos autores consideran el ACP como una etapa del AF y otros las consideran simplemente como técnicas diferentes.

2.1. Fundamentación Teórica

El análisis de componentes principales tiene como objetivo transformar un conjunto de variables aleatorias dado, en un nuevo conjunto de combinación lineales con estas variables, denominadas componentes principales. Estas combinaciones lineales son generadas a partir de la descomposición espectral de la matriz de covarianza o de correlaciones de la muestra. Estos componentes principales se caracterizan por estar incorrelacionados entre sí. Esta técnica genera la primera combinación lineal de tal manera que explique la mayor proporción de varianza muestral, después el segundo componente explicará la mayor proporción de la varianza total restante, es decir, de la que no explicó el primer componente; así sucesivamente. El ACP genera tantas combinaciones lineales como variables existentes en el conjunto de datos dados (Jhonson & Wichern, 2007).

Mientras que el análisis de componentes principales genera combinaciones lineales de las variables originales de tal manera que expliquen la mayor proporción de la variación muestral; el análisis factorial pretende hallar un nuevo conjunto de variables, menor en número que las variables de la muestra, que exprese lo que es común a esas variables. Existen principalmente dos diferencias entre el AF y ACP, en primer lugar los componentes principales se generan para explicar un porcentaje de la variabilidad existente en los datos dados, mientras que los factores se construyen para explicar las covarianzas o correlaciones entre las variables dadas. En segundo lugar, componentes principales es una herramienta meramente descriptiva, mientras que el análisis factorial presupone la existencia de un modelo estadístico, el cual viene dado por (Peña, 2002):

$$\mathbf{X} = \mu + \mathbf{LF} + \mathbf{e} \quad (1)$$

donde:

- \mathbf{F} es un vector de $m \times 1$ variables aleatorias no observadas, llamadas *factores comunes*. Se asume que los factores son variables de media cero e independientes entre sí y con distribución normal.
- \mathbf{L} es una matriz de $p \times m$, $m < p$, de valores desconocidos llamados pesos. Ésta contiene los coeficientes que describen a los factores \mathbf{F} , que afectan a las variables observadas \mathbf{X} .

- e es un vector de $p \times 1$ de perturbaciones no observadas. Recoge el efecto de todas las variables distintas de los factores que influyen sobre \mathbf{X} .

2.2. Descripción de la Base de Datos

La base de datos está compuesta por 103 observaciones de la plaga *B. cockerelli (Sulc)* en etapa adulta en condiciones de invernadero. La Tabla 1 muestra la distribución de los datos.

Refaello	Charanda F1
Hembras: 16	Hembras: 29
Macho: 30	Machos: 28

Tabla 1: Distribución de la muestra.

La Tabla 2 muestra las variables que se consideraron para el estudio.

Variables	Acrónimo
SEXO	Macho: M, Hembra: H
VARIEDAD	Rafaello, Charanda F1
LARGO DEL CUERPO (μm)	LC
ANCHO DEL CUERPO EN TORAX (μm)	AC_TORAX
LONGITUD DE LAS ANTENAS (μm)	ANT
LONGITUD DEL ALA (μm)	ALA_LARGO
ANCHO DEL ALA (μm)	ALA_ANCHO

Tabla 2: Variables del estudio.

En la siguiente sección se muestra los resultados del estudio morfométrico para la plaga *B. cockerelli (Sulc)*.

El ACP, se realizó utilizando la función *PCA* del paquete *FactoMineR*, para el AF se utilizó la función *factanal* del paquete *stats*. Ambos paquetes se encuentran disponibles en la versión 3.3.0 del software estadístico R, (R Core Team , 2016). Los parámetros del modelo (1), fueron estimados por máxima verosimilitud (ml) y la matriz de pesos \mathbf{L} , fue rotada utilizando el método de rotación *varimax* (Jhonson & Wichern, 2007).

3. Resultados

La Figura 1, muestra la distribución de las cinco variables morfométricas. Gráficamente se observa que las variables ALA_LARGO y ANT tienen una distribución simétrica.

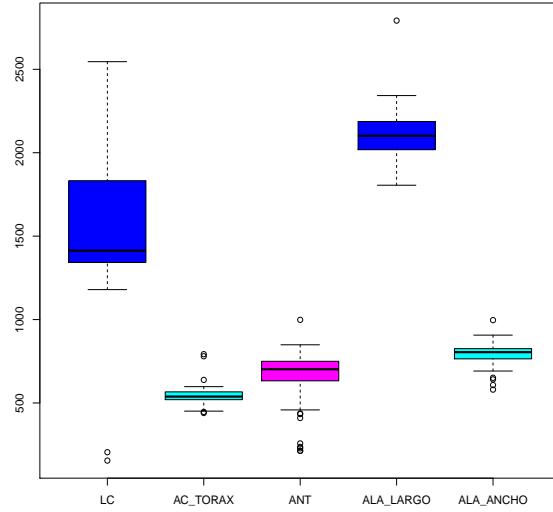


Figura 1: Distribución de los datos por variable.

3.1. Análisis de Componentes Principales

El análisis de componentes principales se realizó por variedad de jitomate, Charanda F1 y Rafaello.

Variedad Charanda F1

Los primeros dos componentes, CP1 y CP2, explican el 66.294 % de la varianza muestral total (Tabla 3). Esta proporción de varianza no es nada despreciable. Las variables con más peso y por consiguiente las más relevantes para el estudio morfométrico de la plaga son ALA_LARGO y ALA_ANCHO para CP1, y LC para CP2 (Figura 2). De esa manera CP1 es un componente de *tamaño del ala* y CP2 un componente de *tamaño de cuerpo*.

Este análisis muestra que si el insecto es criado bajo la variedad Charanda F1, entonces los Machos desarrollaron, en general, un cuerpo más largo que las Hembras. Mientras que las Hembras desarrollan, en general, más alas (Figura 3).

Variable	CP1	CP2
LC	0.459	0.781
AC-TORAX	0.567	-0.416
ANT	0.693	0.340
ALA_LARGO	0.875	-0.139
ALA_ANCHO	0.713	-0.332
Proporción de varianza explicada	45.717	20.577
Proporción de varianza acumulada	45.717	66.294

Tabla 3: Pesos de los componentes en Charanda F1.

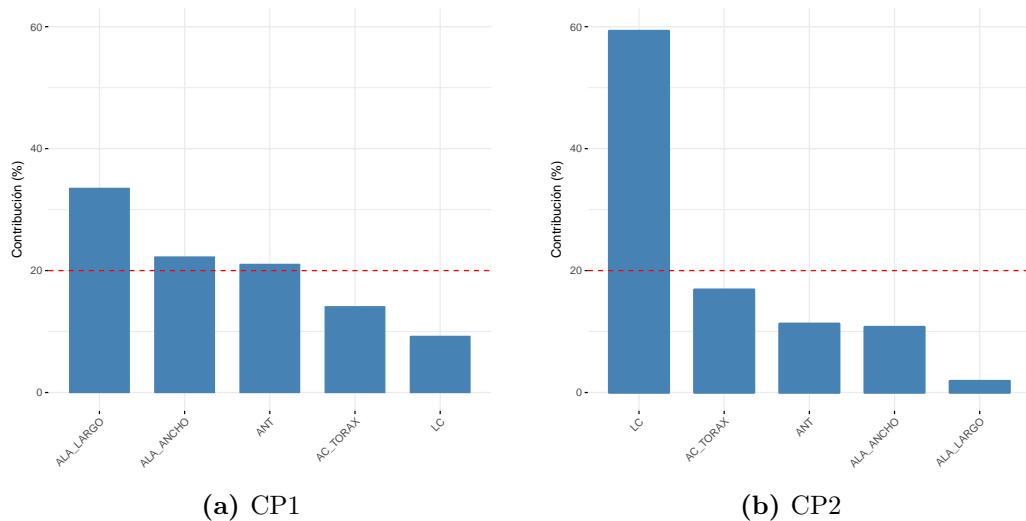


Figura 2: Contribución de las variables por componente en Charanda F1.

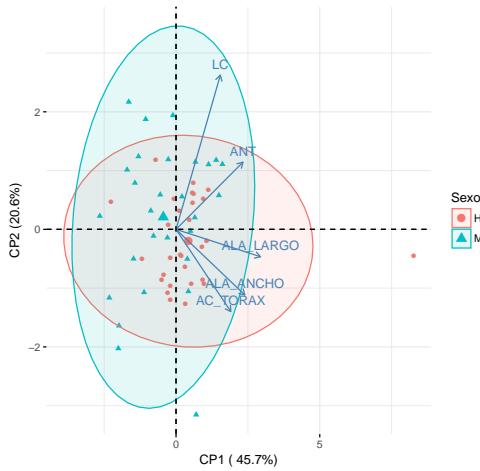


Figura 3: Distribución por sexo para Charanda F1.

Variedad Rafaello

Para esta variedad los primeros dos componentes, CP1 y CP2, explican el 53.725 % de la varianza total presente en los datos. Las variables más importantes son ALA_LARGO y ANT para CP1, y LC para CP2 (Tabla 4).

Variable	CP1	CP2
LC	0.194	0.831
AC_TORAX	0.530	0.146
ANT	-0.756	0.402
ALA_LARGO	0.756	-0.133
ALA_ANCHO	0.393	0.422
Proporción de varianza explicada	32.331	21.394
Proporción de varianza acumulada	32.331	53.725

Tabla 4: Pesos de los componentes en Rafaello.

En la Figura 4 se muestra la contribución que cada variable morfométrica tiene en los componentes. De acuerdo a estos resultados CP1 es un componente de *tamaño de ala y de antena* y CP2 es un componente de *tamaño de cuerpo*.

Este análisis reporta que, en general, los Machos tienen el cuerpo y las antenas más largas y las alas más cortas, en cambio las Hembras tienen el cuerpo más corto, pero las alas más largas (Figura 5).

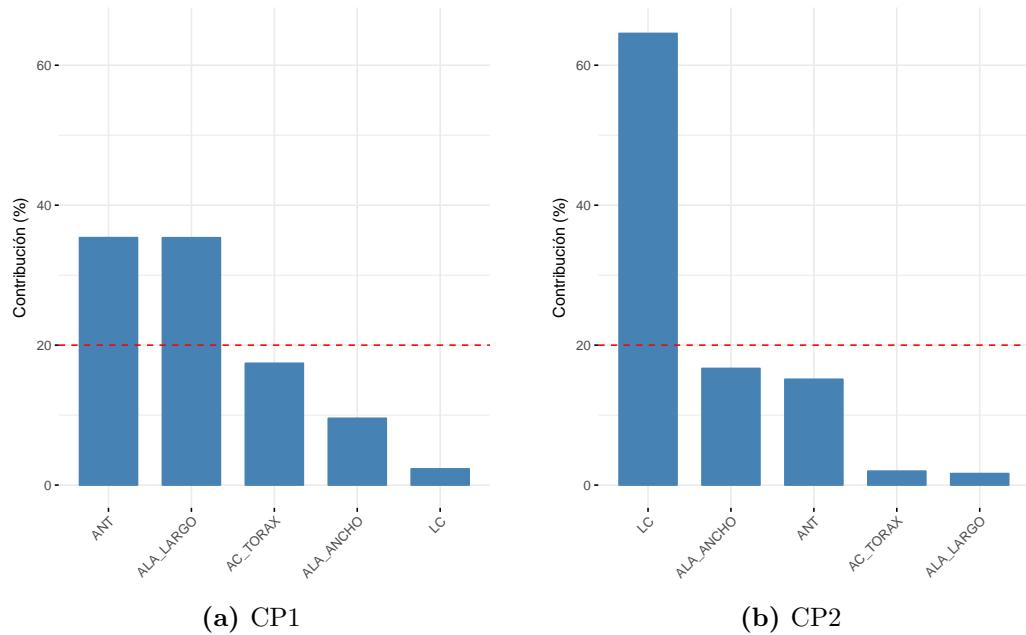


Figura 4: Contribución de las variables por componente en Rafaello.

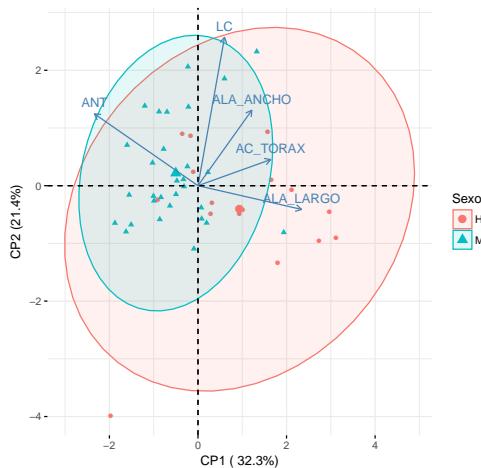


Figura 5: Distribución por sexo para Rafaello.

3.2. Análisis Factorial

El análisis factorial se realizó también por variedad de jitomate.

Variedad Charanda F1

La Figura 6 (a), muestra tres parejas de variables con alta correlación: ALA_LARGO y AC_TORAX, ALA_ANCHO y ALA_LARGO y finalmente AC_TORAX y ANT. La Figura 6 (b), muestra la existencia de poca correlación lineal entre las variables. Entonces las asociaciones lineales entre las variables están en función del sexo del insecto.

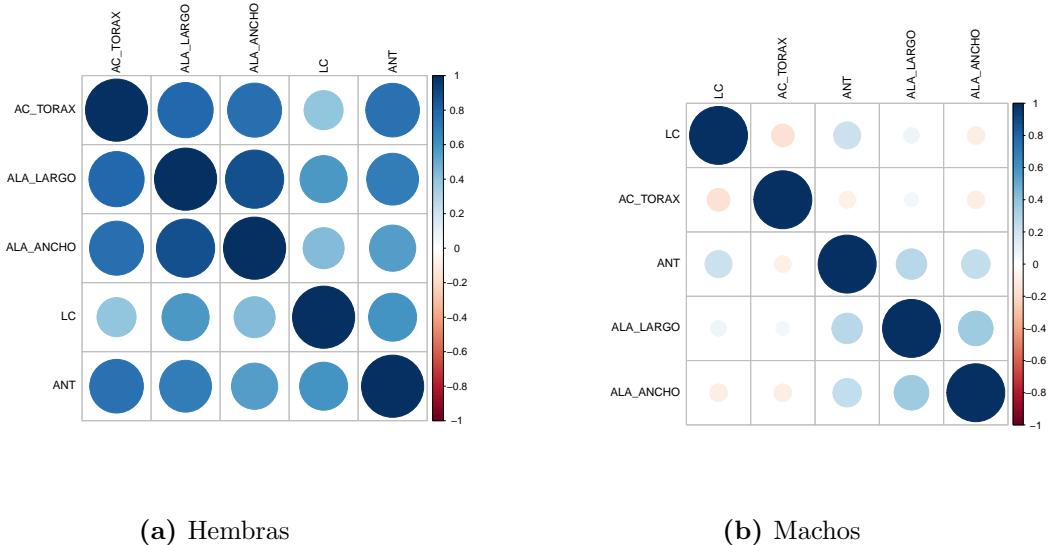


Figura 6: Correlaciones de las variables por sexo para Charanda F1.

La Tabla 5 muestra la proporción de varianza explicada por los dos primeros factores que es de 56.2 %. En el Factor1 las variables con mayor puntuación son ALA_LARGO y ALA_ANCHO, lo cual indica que es un factor de tamaño de ala; para el Factor2 la mayor puntuación es para la variable LC, indicando que es un factor de tamaño de cuerpo. De acuerdo a estos resultados los Machos tienen el cuerpo más grande que las Hembras (Figura 7).

Variable	Factor1	Factor2
LC	0.061	0.996
AC_TORAX	0.450	0.032
ANT	0.390	0.320
ALA_LARGO	0.946	0.238
ALA_ANCHO	0.635	0.046
Proporción de varianza explicada	0.331	0.231
Proporción de varianza acumulada	0.331	0.562

Tabla 5: Factores estimados por ml para Charanda F1.

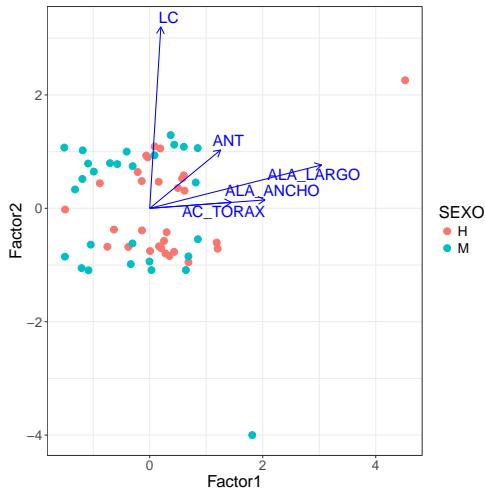


Figura 7: Dispersión de las observaciones para Charanda F1.

Variedad Rafaello

La Figura 8 muestra las correlaciones lineales de las variables morfométricas por sexo. En el caso de las Hembras ALA_LARGO y ANT, ALA_LARGO y LC, presentan alta asociación lineal (Figura 8 (a)). Por otra parte, para los Machos se observa poca correlación entre las variables (Figura 8 (b)). Comparando estos resultados con los de la variedad Charanda F1 (Figura 6), se puede inferir que las correlaciones lineales entre las variables morfométricas, además de depender del sexo del insecto también dependen de la variedad del jitomate.

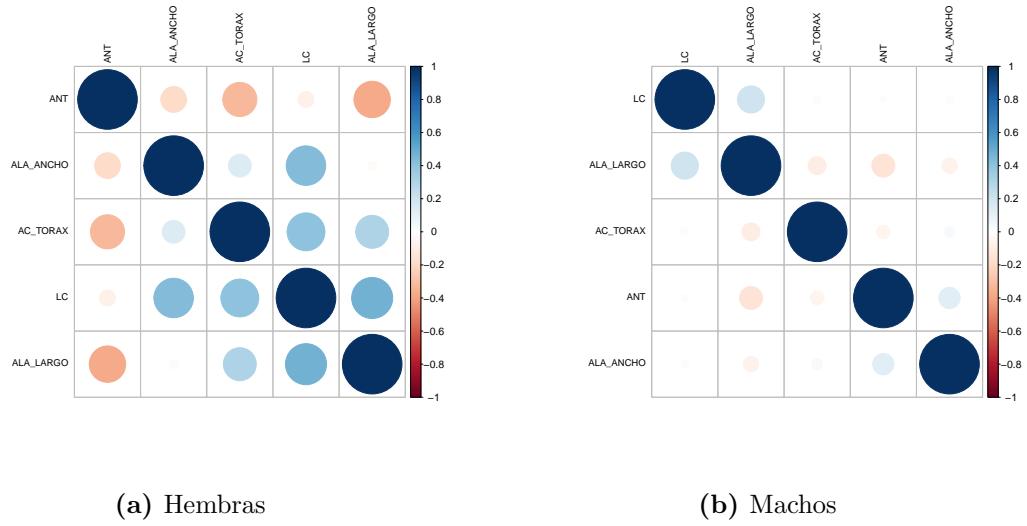


Figura 8: Correlaciones de las variables por sexo para Rafaello.

La Tabla 6 muestra los factores estimados. Como se puede observar, para el primer factor la variable con mayor puntuación es ANT; mientras que para el segundo factor es LC.

Variable	Factor1	Factor2
LC	0.026	0.444
AC_TORAX	-0.207	0.228
ANT	0.997	0.074
ALA_LARGO	-0.429	0.298
ALA_ANCHO	-0.116	0.229
Proporción de varianza explicada	0.248	0.078
Proporción de varianza acumulada	0.248	0.326

Tabla 6: Factores estimados por ml para Rafaello.

Con una proporción de varianza explicada de apenas 32.6 % por los factores, no se puede inferir mucho sobre el comportamiento de la muestra, sin embargo se puede inferir que los Machos tienen las antenas y el cuerpo más grande que las Hembras (Figura 9).

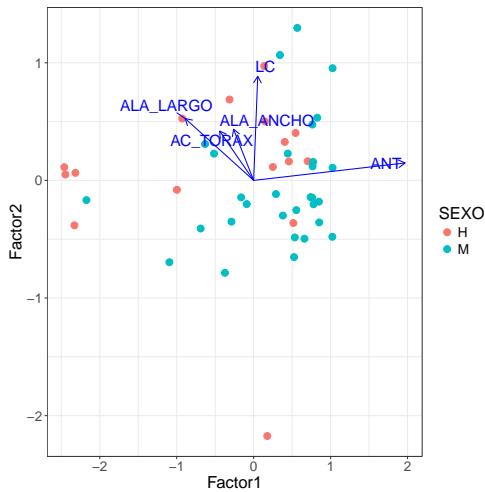


Figura 9: Dispersión de las observaciones para Rafaello.

4. Conclusiones

De acuerdo con los resultados del estudio, el AF y ACP pueden ser de ayuda para el investigador que busca cuantificar la variabilidad morfométrica de la plaga *B. cockerelli* (*Sulc*) en dos variedades de jitomate. Además estos dos métodos multivariados, permitieron identificar las variables morfométricas más importantes para caracterizar la morfometría del insecto, de acuerdo a su sexo y en este caso también tomando en cuenta la variedad del jitomate. En términos comparativos el ACP, puede ser de mayor utilidad ya que explica una proporción de varianza muestral más alta. Pero no por esto el AF, deja de ser útil, pues muestra el análisis de los datos, desde otro enfoque.

Bibliografía

Johnson, R. A. and Wichern, D. W. (2007). *Applied multivariate statistical analysis*. Prentice Hall, Upper Saddle River.

Martínez-Ibarra, J. A., Bárcenas-Ortega, N. M., Romero-Nápoles, J., Nogueda-Torres, B., and H., R.-L. M. (2006). Diferencias métricas entre poblaciones de *Meccus longipennis* (Usinger) (Hemiptera: Reduviidae) en el occidente de México. *Folia Entomol*, 45(2):83–90.

Peña, D. (2002). *Análisis de datos multivariantes*. Editorial desconocida, Barcelona, España.

R Core Team (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.

Vargas-Madríz, H. (2010). *Morfometría y tabla de vida de Bactericera cockerelli (Sulc) en dos variedades de jitomate en invernadero*. PhD thesis, Colegio de Postgraduados, Texcoco, Edo. de México, México. Tesis doctoral.

Un Método para Construcción de Pruebas de No Inferioridad con Regiones Críticas Convexas

José Juan Castro Alva, Hortensia Josefina Reyes Cervantes^a

Facultad de Ciencias Físico Matemáticas de la Benemérita Universidad Autónoma de Puebla

Félix Almendra Arao

Unidad Profesional Interdisciplinaria en Ingeniería y Tecnologías Avanzadas del Instituto Politécnico Nacional

Las pruebas de no-inferioridad (NI) son procedimientos estadísticos desarrollados con el objetivo de verificar si existe evidencia muestral de que un tratamiento nuevo es igual, superior, o inferior, a un tratamiento estándar o control activo. Tanto en las pruebas estadísticas de no-inferioridad como en las de superioridad (S), el que la región crítica sea un conjunto convexo de Barnard (CCB), juega un papel central, debido a dos razones principales. La primera es de tipo computacional, y consiste en la reducción del tiempo de cómputo al calcular los tamaños de la prueba y la segunda es de sentido clínico, que corresponde a la naturaleza del ensayo clínico. Sin embargo, algunas de las pruebas existentes en la literatura no cumplen esta condición, por lo cual es importante poder construir, a partir de una prueba de NI/S dada, una prueba que garantice que la región crítica sea un CCB. En este trabajo se presenta un procedimiento, mediante el cual, a partir de una prueba de NI/S dada, se construye otra prueba de NI/S cuya región crítica sea un CCB.

Área-MSC: Estadística Inferencial.

Subárea-MSC: Pruebas de hipótesis

^a jjcasatrhoa@gmail.com

1. Introducción

En algunas áreas científicas es común que se requiera realizar comparación entre grupos. Un tipo de comparación de grupos que resulta ser útil se basa en las llamadas pruebas de no inferioridad y también en las pruebas de superioridad, éstos son procedimientos utilizados especialmente en ensayos clínicos. Las pruebas de no-inferioridad tienen como objetivo verificar si existe evidencia muestral de que un tratamiento nuevo es igual, superior, o ligeramente inferior, que un tratamiento estándar o control activo. A su vez las pruebas de superioridad buscan evidencia en la muestra de que el tratamiento nuevo es superior al estándar. Bajo el marco de NI, para que sea razonable la aceptación de un nuevo tratamiento que puede ser incluso “ligeramente peor” que un tratamiento bien conocido comúnmente llamado estándar, naturalmente, es necesario que el nuevo tratamiento ofrezca algunas ventajas sobre el tratamiento estándar; en el caso de investigación de nuevos productos farmacéuticos estas ventajas pudieran ser por ejemplo menor costo, ventajas de seguridad, menores efectos secundarios, facilidad de aplicación, etc. Tanto en las pruebas estadísticas de no-inferioridad como en las de superioridad, el que la región crítica sea un CCB, juega un papel central, debido a dos razones principales. Una es de naturaleza numérica, ya que el cálculo de los tamaños de prueba es un problema computacionalmente intensivo, debido a la presencia de un parámetro perturbador, pero dicho cálculo se reduce considerablemente cuando la región crítica es un CCB vía el teorema de Röhmel y Mansmann Röhmel and Mansmann (1999). La segunda razón es que las regiones críticas deben ser CCB pues en caso contrario las correspondientes pruebas estadísticas carecen de sentido en el contexto de los ensayos clínicos. Por las razones expuestas es deseable que las regiones críticas de pruebas de NI/S sean conjuntos convexos de Barnard, sin embargo, algunas de las pruebas existentes en la literatura no cumplen esta condición, debido a ello, es importante poder construir, a partir de una prueba de NI/S dada, una prueba que garantice que la región crítica sea un CCB.

Adicionalmente en Almendra-Arao (2012), bajo el supuesto de que la región crítica de una prueba de no-inferioridad es un CCB, obtiene representaciones analíticas para la primera y segunda derivada de la función potencia haciendo viable la implementación del método de Newton en el cálculo de los tamaños de prueba para pruebas de no-inferioridad, ello permite reducir el tiempo de cómputo considerablemente.

2. Marco Teórico

Supóngase que se tienen n_1 y n_2 unidades experimentales para recibir los tratamientos estándar (TE) y nuevo (TN), respectivamente. Sean X_1 y X_2 las respuestas positivas de (TE) y (TN), respectivamente. Por lo tanto, se tienen dos variables aleatorias independientes $X_i \sim Bin(n_i, p_i)$, $i = 1, 2$, donde p_1 y p_2 son las probabilidades de éxito de (TE) y (TN), respectivamente. Considérese el juego de hipótesis

$$H_0 : p_2 \leq g(p_1) \quad vs \quad H_a : p_2 > g(p_1), \quad (1)$$

donde I es un intervalo cerrado y $g : I \subseteq [0, 1] \rightarrow [0, 1]$ es una función no decreciente tal que $g(p_1) \leq p_1$. Si en (1) se tiene que $g(p_1) \geq p_1$, estamos en presencia de una prueba de superioridad, por lo contrario si $g(p_1) \leq p_1$ en todo su dominio y existe un valor de p_1 en el dominio para el cual $g(p_1) < p_1$, entonces estamos en presencia de una prueba de no-inferioridad.

Considerando los supuestos mencionados anteriormente, se tiene lo siguiente: la función de masa de probabilidad de X_i está dada por $f(x_i) = \binom{n_i}{x_i} p_i^{x_i} (1 - p_i)^{n_i - x_i}$, para $i = 1, 2$, el espacio muestral es $\chi = \{(x_1, x_2) \in \{0, 1, \dots, n_1\} \times \{0, 1, \dots, n_2\}\}$, además el espacio paramétrico se puede representar convenientemente por $\Theta = [0, 1]^2$ y la función verosimilitud conjunta para X_1 y X_2 es $L(p_1, p_2, x_1, x_2) = \prod_{i=1}^2 \binom{n_i}{x_i} p_i^{x_i} (1 - p_i)^{n_i - x_i}$, así para una prueba estadística T , su función potencia está dada por $\beta_T(p_1, p_2) = \sum_{(x_1, x_2) \in R_T} L(p_1, p_2, x_1, x_2)$, donde $R_T(c) = \{(x_1, x_2) \in \chi : T(x_1, x_2) \leq c\}$ denota la región crítica de una prueba estadística T .

El espacio nulo correspondiente a la prueba de hipótesis (1) está determinado por $\Theta_0 = \{(p_1, p_2) \in \Theta : p_2 \leq g(p_1)\}$. Así, el tamaño de la prueba está dado por $\sup_{\theta \in \Theta_0} \beta_T(p_1, p_2)$.

Es importante resaltar que una gran variedad de funciones margen para NI queda incluida en (1), entre las cuales destacan las funciones margen correspondientes a la diferencia de proporciones, la razón de proporciones, la razón de momios, el margen lineal de Phillip's Phillips (2003) y las propuestas de Röhmel Röhmel and Mansmann (1999).

Definición 2.1. *Un conjunto $C \subseteq \chi$ se llamará **conjunto convexo de Barnard (CCB)** si satisface las condiciones siguientes:*

1. $(x_1, x_2) \in R_T \Rightarrow (x_1 - 1, x_2) \in R_T, \forall 1 \leq x_1 \leq n_1, 0 \leq x_2 \leq n_2$.
2. $(x_1, x_2) \in R_T \Rightarrow (x_1, x_2 + 1) \in R_T, \forall 0 \leq x_1 \leq n_1, 0 \leq x_2 \leq n_2 - 1$.

3. Determinación de Propiedades para Redefiniciones con Regiones Críticas Convexas

En este apartado se pretende identificar estrategias y propiedades convenientes para redefinir una estadística que garantice que su región crítica sea un CCB. A continuación revisaremos un par de posibles formas de redefinir una estadística.

Definición 3.1. *Dado un conjunto $A \subseteq \chi$, la cápsula convexa de Barnard de A , denotada por $[A]$, es definida como el conjunto más pequeño entre los conjuntos convexos de Barnard contenidos en χ y que contienen a A . Decimos que $[A]$ es generado por A .*

En Almendra-Arao (2011) se realiza un estudio detallado acerca de los conjuntos convexos de Barnard, cápsulas convexas de Barnard, así como algunas propiedades de importancia para el desarrollo teórico.

Observación 3.1. Dado $(a, b) \in \chi$, simbolizamos a $\{(x_1, x_2) \in \chi : x_1 \geq a, x_2 \leq b\}$ por $](a, b)[$.

Ejemplo 3.1. En la Figura 1 se muestra un ejemplo de un CCB.

Ejemplo 3.2. Sea $\chi = \{0, \dots, 6\} \times \{0, \dots, 5\}$, sea $(3, 2) \in \chi$, entonces la cápsula convexa de Barnard para el punto $(3, 2)$ es el conjunto $[(3, 2)] = \{0, \dots, 3\} \times \{2, \dots, 5\}$ y $](3, 2)[= \{3, \dots, 6\} \times \{0, 1, 2\}$ ver Figura 2.

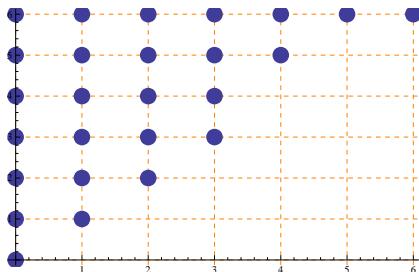


Figura 1: Conjunto Convexo de Barnard.

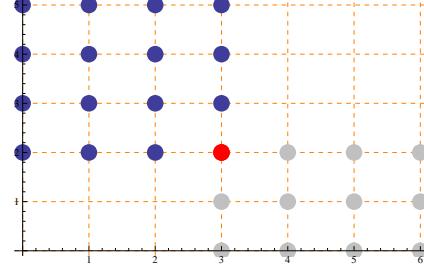


Figura 2: $[(3, 2)]$ Círculos en azul y círculo en rojo, $](3, 2)[$: círculos en gris y círculo en rojo.

3.1. Análisis de una Primera Redefinición

Para modificar una estadística, una estrategia que podría parecer natural con el objetivo de obtener regiones críticas que sean CCB , consiste en redefinir una estadística T de la

siguiente forma:

$$T_{\max}(x_1, x_2) = \max\{T(x_1, x_2), T(x_1 - 1, x_2), T(x_1, x_2 + 1)\}.$$

Esta modificación T_{\max} de la estadística general T puede aplicarse a cualquier estadística. A continuación se analiza un ejemplo de esta primera redefinición, se considera una estadística específica T construida ex profeso para analizar el comportamiento de las regiones críticas obtenidas. Así por simplicidad, para este ejemplo se considera $n_1 = n_2 = 2$ y T la estadística dada por la siguiente tabla y se obtiene la redefinición T_{\max} .

(x_1, x_2)	(0,2)	(0,1)	(0,0)	(1,2)	(1,1)	(1,0)	(2,2)	(2,1)	(2,0)
$T(x_1, x_2)$	2	1	4	3	2	1	1	3	2
$T_{\max}(x_1, x_2)$	2	2	4	3	3	4	3	3	3

2	2 3 1	2	2 3 3
x_2	1	x_2	1
1	1 2 3	0	2 3 3
0	4 1 2	0	4 4 3
	<hr/>		<hr/>
	0 1 2		0 1 2
	x_1		x_1

Figura 3: Valores de la estadística T y T_{\max} , respectivamente.

A continuación se presenta una tabla con la región crítica para cada valor de c :

	$R_{T(c)}$	¿Es $R_{T(c)}$ un CCB?	$R_{T_{\max}(c)}$	¿Es $R_{T_{\max}(c)}$ un CCB?
$c = 1$	$\{(0, 1), (1, 0), (2, 2)\}$	No	\emptyset	Sí
$c = 2$	$\{(0, 1), (1, 0), (2, 2), (0, 2), (1, 1), (2, 0)\}$	No	$\{(0, 2), (0, 1)\}$	Sí
$c = 3$	$\chi - \{(0, 0)\}$	No	$\chi - \{(0, 0), (1, 0)\}$	No
$c = 4$	χ	Sí	χ	Sí

De la tabla anterior se puede concluir que la redefinición propuesta falla en la construcción de regiones críticas que siempre sean conjuntos convexos de Barnard y por esta razón se descarta.

3.2. Análisis de una Segunda Redefinición

De la propuesta anterior parece natural establecer la siguiente redefinición

$$T_{new}(x_1, x_2) = \max\{T(x_1, x_2), T_{new}(x_1 - 1, x_2), T_{new}(x_1, x_2 + 1)\},$$

considerando el proceso en forma secuencial comenzando en la esquina superior izquierda.

Aplicando esta nueva redefinición a la estadística T anterior, se parte del punto $(0, 2)$ y se obtiene

(x_1, x_2)	(0,2)	(0,1)	(0,0)	(1,2)	(1,1)	(1,0)	(2,2)	(2,1)	(2,0)
$T_{new}(x_1, x_2)$	2	2	4	3	3	4	3	3	4

x_2	2	2	3		2	2	3	3	
x_2	1	1	2	3		2	2	3	3
x_2	0	4	1	2		4	4	4	
x_1		0	1	2			0	1	2
x_1									

Figura 4: Valores de la estadística T y T_{new} , respectivamente.

La región crítica para cada valor de c se muestra a continuación.

	$R_{T(c)}$	¿Es $R_{T(c)}$ un CCB?	$R_{T_{new}(c)}$	¿Es $R_{T_{new}(c)}$ un CCB?
$c = 1$	$\{(0, 1), (1, 0), (2, 2)\}$	No	\emptyset	Sí
$c = 2$	$\{(0, 1), (1, 0), (2, 2), (0, 2), (1, 1), (2, 0)\}$	No	$\{(0, 2), (0, 1)\}$	Sí
$c = 3$	$\chi - \{(0, 0)\}$	No	$\chi - \{(0, 0), (1, 0), (2, 0)\}$	Sí
$c = 4$	χ	Sí	χ	Sí

Con esta redefinición, las regiones críticas en este ejemplo siempre son conjuntos convexos de Barnard. Ahora comparemos las regiones críticas R_T y $R_{T_{new}}$.

Podemos apreciar que para los valores de $c = 1, 2, 3$, $R_{T_{new}}$ se reduce, mientras que para $c = 4$ se mantiene igual que la región crítica de la estadística original, el hecho de que la nueva estadística remueve puntos de la región crítica de la estadística original no es conveniente puesto que existen puntos para los cuales la estadística original rechaza la hipótesis nula, mientras que la nueva estadística para esos mismos puntos no rechaza la hipótesis nula. En otras palabras la nueva estadística pierde potencia.

3.3. Determinación de Propiedades Convenientes

Con base en los ejemplos presentados anteriormente es conveniente que la nueva estadística preserve los puntos de la región crítica original, en caso de que incremente puntos a la región crítica, admitimos que se incremente la menor cantidad de puntos posibles, es decir, admitir sólo los puntos necesarios para que la región crítica modificada corresponda a un CCB. Esta idea se expresa en la siguiente definición.

Definición 3.2. Una \bar{T} de una estadística T será llamada:

1. *B-convexificación de T si $R_{\bar{T}(c)}$ es un conjunto convexo de Barnard para toda c .*
2. *Envolvente de T , si $R_{T(c)} \subseteq R_{\bar{T}(c)}$ para todo c .*

Así, en términos de la definición previa y con base en el análisis presentado en la sección anterior, dada una estadística T , nuestro objetivo es construir una redefinición, que denotaremos por \bar{T} , de tal forma que \bar{T} sea una B-convexificación envolvente de T .

4. B-Convexificación de Estadísticas

Como una aproximación para construir una B-convexificación envolvente de la estadística T proponemos la siguiente estrategia. Rechazar $(x_1, x_2) \in \chi$ si existe $(x'_1, x'_2) \in](x_1, x_2)[$ tal que $T(x'_1, x'_2) \leq c$, esto es, $(x'_1, x'_2) \in R_T$ con $x'_1 \geq x_1, x'_2 \leq x_2$; en otras palabras, si existe $(x'_1, x'_2) \in R_T$ tal que $(x_1, x_2) \in [(x'_1, x'_2)]$.

Sea $\langle R_{T(c)} \rangle = \{(x_1, x_2) \in \chi : \text{existe } (x'_1, x'_2) \in](x_1, x_2)[\text{ con } T(x'_1, x'_2) \leq c\}$.

La siguiente proposición asegura que el conjunto $\langle R_{T(c)} \rangle$ cumple las condiciones de B-convexificación envolvente. Más aún, 2) afirma que $\langle R_{T(c)} \rangle$ no solo es un conjunto convexo de Barnard, sino que es justamente la cápsula convexa de Barnard de la región crítica $R_{T(c)}$.

Proposición 4.1. Para todo número real c se cumplen las siguientes afirmaciones.

1. $R_{T(c)} \subseteq \langle R_{T(c)} \rangle$.
2. $[R_{T(c)}] = \langle R_{T(c)} \rangle$.

Demostración. 1. Probaremos que $R_{T(c)} \subseteq \langle R_{T(c)} \rangle$.

Sea $(x_1, x_2) \in R_{T(c)}$, entonces $T(x_1, x_2) \leq c$, más aún como $(x_1, x_2) \in](x_1, x_2)[$ entonces $(x_1, x_2) \in \langle R_{T(c)} \rangle$, así hemos probado que $R_{T(c)} \subseteq \langle R_{T(c)} \rangle$.

2. Si $R_{T(c)} = \emptyset$, entonces $\forall(x_1, x_2) \in \chi, T(x_1, x_2) \leq c$, por lo tanto $\langle R_{T(c)} \rangle = \emptyset$, y de la proposición 2.1(1) de Almendra-Arao (2011), $[\emptyset] = \emptyset$ tenemos que para ese caso $[R_{T(c)}] = \langle R_{T(c)} \rangle$. Ahora consideremos el caso $R_{T(c)} \neq \emptyset$.

De 1) $R_{T(c)} \subseteq \langle R_{T(c)} \rangle$, probaremos que $\langle R_{T(c)} \rangle$ es un CCB.

Sea $(x_1, x_2) \in \langle R_{T(c)} \rangle$, entonces existe $(x'_1, x'_2) \in](x_1, x_2)[$ con $T(x'_1, x'_2) \leq c$ entonces para $(x_1 - 1, x_2)$ tenemos que $(x'_1, x'_2) \in](x_1 - 1, x_2)[$ y $T(x'_1, x'_2) \leq c$ por lo tanto $(x_1 - 1, x_2) \in \langle R_{T(c)} \rangle$.

Similarmente para $(x_1, x_2 + 1)$ tenemos que $(x'_1, x'_2) \in](x_1, x_2 + 1)[$ y $T(x'_1, x'_2) \leq c$ puesto que $(x_1, x_2 + 1) \in \langle R_{T(c)} \rangle$. Así hemos probado que $\langle R_{T(c)} \rangle$ es un CCB que contiene a $R_{T(c)}$.

Finalmente, probaremos que $\langle R_{T(c)} \rangle$ es el mínimo conjunto convexo de Barnard que contiene a $R_{T(c)}$.

Sea B un CCB con $R_{T(c)} \subseteq B \subseteq \langle R_{T(c)} \rangle$.

Sea $(x_1, x_2) \in \langle R_{T(c)} \rangle$ entonces existe $(x'_1, x'_2) \in](x_1, x_2)[$ con $T(x'_1, x'_2) \leq c$, por lo tanto $(x'_1, x'_2) \in R_{T(c)} \subseteq B$ más aún, $(x'_1, x'_2) \in](x_1, x_2)[$ implica $x'_1 \geq x_1, x'_2 \leq x_2$ y como $(x'_1, x'_2) \in B$ y B es un CCB tenemos que $(x_1, x_2) \in B$. Por lo tanto $\langle R_{T(c)} \rangle \subseteq B$ y consecuentemente $\langle R_{T(c)} \rangle = B$.

Así, hemos probado que $\langle R_{T(c)} \rangle$ es el mínimo conjunto convexo de Barnard que contiene a $R_{T(c)}$, por lo tanto $\langle R_{T(c)} \rangle$ debe ser la cápsula convexa de Barnard de $R_{T(c)}$, es decir, $[R_{T(c)}] = \langle R_{T(c)} \rangle$.

□

Basados en la proposición anterior, tenemos que los conjuntos del tipo $\langle R_{T(c)} \rangle$ satisfacen las condiciones de redefinición envolvente y B-convexificación. La pregunta que surge de manera natural es que si existe una redefinición de T de tal forma que la región crítica de esta redefinición corresponda a conjuntos de la forma $\langle R_{T(c)} \rangle$.

El hecho de que la región crítica $\langle R_{T(c)} \rangle$ pueda ser descrita por una redefinición, digamos \bar{T} ; implicaría que podría construirse vía una estadística, en este caso tendríamos $\langle R_{T(c)} \rangle = R_{\bar{T}(c)} = \{(x_1, x_2) \in \chi : \bar{T}(x_1, x_2) \leq c\}$.

En lo siguiente damos una redefinición de T la cual denotaremos por $[T]$ y en la siguiente proposición veremos que $[T]$ satisface las condiciones deseables.

Puesto que $T : \chi \rightarrow \mathbb{R}$ y χ es finito, entonces la imagen I_T de la función T , es un conjunto finito, sea $I_T = \{t_1, t_2, \dots, t_k\}$ con $t_1 < t_2 < \dots < t_k$ y $\chi_i = T^{-1}(t_i) = \{(x, y) \in \chi : T(x, y) = t_i\}$ con $i = 1, \dots, k$, claramente χ_1, \dots, χ_k es una partición del espacio muestral χ .

Con base en T , definimos $[T] : \chi \rightarrow \mathbb{R}$ como:

$$[T](x, y) = t_i \forall (x, y) \in [\chi_i] - [\chi_{i-1}] \quad \text{para } i = 1, 2, \dots, k, \quad \text{donde } [\chi_0] = \emptyset.$$

Proposición 4.2. $R_{[T](c)} = [R_{T(c)}]$ para todo $c \in I_T$.

Demostración. Si $c < t_1$, entonces $[R_{T(c)}] = [\emptyset] = \emptyset = R_{[T](c)}$.

Si $t_1 \leq c < t_k$, entonces existe $c^* \in \{1, 2, \dots, k-1\}$ tal que $t \in [t_{c^*}, t_{c^*+1})$

$$\begin{aligned} R_{[T](c)} &= \{(x, y) \in \chi : [T](x, y) \leq c\} = \{(x, y) \in \chi : [T](x, y) \leq t_{c^*}\} \\ &= \cup_{i=1}^{c^*} \{(x, y) \in \chi : [T](x, y) = t_i\} = \cup_{i=1}^{c^*} ([\chi_i] - [\chi_{i-1}]) = \cup_{i=1}^{c^*} [\chi_i] \end{aligned}$$

y por la propiedad (1) de la proposición 2.4 de Almendra-Arao (2011) tenemos $R_{[T](c)} = [\cup_{i=1}^{c^*} \chi_i]$; por otra parte $R_{T(c)} = \{(x, y) \in \chi : T(x, y) \leq c\} = \{(x, y) \in \chi : T(x, y) \leq t_{c^*}\} = \cup_{i=1}^{c^*} \chi_i$ por lo tanto $[R_{T(c)}] = [\cup_{i=1}^{c^*} \chi_i] = R_{[T](c)}$.

Si $c \geq t_k$, entonces $R_{[T](c)} = \cup_{i=1}^k [\chi_i] = [\cup_{i=1}^k \chi_i] = [R_{T(c)}]$.

□

La proposición previa dice que para construir la región crítica de la estadística $[T]$ es necesario obtener la cápsula convexa de Barnard de la región crítica T .

Ejemplo 4.1. Considerando nuevamente la estadística T como en los ejemplos anteriores, veremos que el comportamiento de la redefinición $[T]$ de T cumple con las condiciones deseables establecidas.

(x_1, x_2)	(0,2)	(0,1)	(0,0)	(1,2)	(1,1)	(1,0)	(2,2)	(2,1)	(2,0)
$T(x_1, x_2)$	2	1	4	3	2	1	1	3	2
$[T](x_1, x_2)$	1	1	1	1	1	1	1	2	2

La región crítica $R_{[T]}(c)$ se muestra a continuación

	$R_{T(c)}$	¿Es $R_{T(c)}$ un CCB?	$R_{[T](c)}$	¿Es $R_{[T](c)}$ un CCB?
$c = 1$	$\{(0, 1), (1, 0), (2, 2)\}$	No	$\chi - \{(2, 0), (2, 1)\}$	Sí
$c = 2$	$\{(0, 1), (1, 0), (2, 2), (0, 2), (1, 1), (2, 0)\}$	No	χ	Sí
$c = 3$	$\chi - \{(0, 0)\}$	No	χ	Sí
$c = 4$	χ	Sí	χ	Sí

5. Ejemplo de Aplicación

Considere la prueba clásica asintótica de Blackwelder $T(x_1, x_2) = \frac{\hat{p}_1 - \hat{p}_2 - d_0}{\hat{\sigma}}$, donde d_0 es una constante positiva y $\hat{p}_i = X_i/n_i$ son los estimadores de máxima verosimilitud de p_i para $i = 1, 2$ y $\hat{\sigma}$ es el estimador de la desviación estándar de $\hat{d} = \hat{p}_1 - \hat{p}_2$, dado por $\hat{\sigma} = \left(\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2} \right)^{1/2}$. Esta estadística se usa para contrastar (1) cuando $g(p_1) = p_1 - d_0$, y debido a que tiene una distribución asintótica normal estándar, su región de rechazo es $R_T(-z_\alpha) = \{(x_1, x_2) \in \{0, \dots, n_1\} \times \{0, \dots, n_2\} \mid T < -z_\alpha\}$, donde α es tal que $P(Z > z_\alpha) = \alpha$. En la figura 5 se muestran las regiones de rechazo de las estadísticas T y la estadística redefinida $[T]$, para $n_1 = 43, n_2 = 10, \alpha = 0.05$ y $d_0 = 0.1$. Note que R_T no es un CCB, mientras que $R_{[T]}$ si lo es. En la tabla 1 se muestran los valores que toma la estadística T y en la tabla 2 se muestran los valores que toma la redefinición $[T]$ de T .

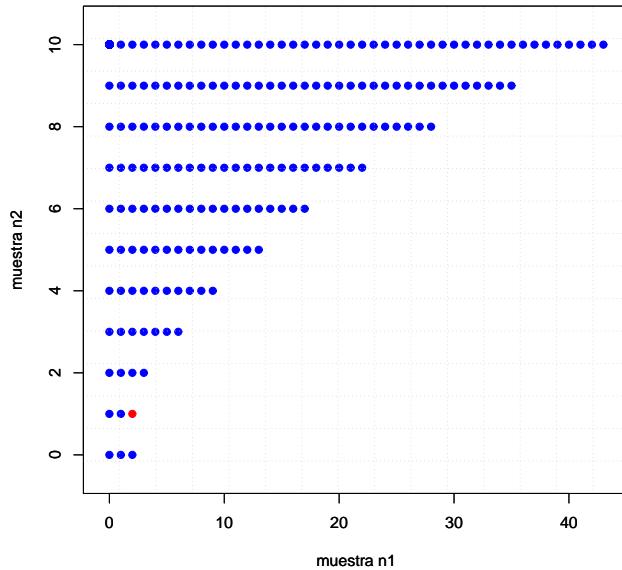


Figura 5: R_T : puntos en azul y $R_{[T]}$: puntos en azul unión punto en rojo.

10	-107.07	-46.85	-32.80	-26.52	-22.73	-20.12	-18.18	-16.65	-15.40	-14.36	-13.46	-12.69	-12.00	-11.39	-10.84	-10.34	-9.87	-9.45	-9.06	-8.69	-8.35	-8.02	
9	-10.54	-10.01	-9.52	-9.07	-8.66	-8.28	-7.92	-7.59	-7.27	-6.98	-6.69	-6.42	-6.16	-5.92	-5.68	-5.45	-5.23	-5.01	-4.80	-4.60	-4.40	-4.20	
8	-7.12	-6.82	-6.54	-6.27	-6.02	-5.78	-5.55	-5.32	-5.11	-4.90	-4.70	-4.51	-4.32	-4.13	-3.95	-3.78	-3.61	-3.44	-3.27	-3.11	-2.95	-2.79	
7	-5.52	-5.29	-5.08	-4.87	-4.67	-4.47	-4.28	-4.10	-3.92	-3.75	-3.58	-3.41	-3.25	-3.09	-2.94	-2.79	-2.63	-2.49	-2.34	-2.19	-2.05	-1.91	-1.77
6	-4.52	-4.32	-4.13	-3.95	-3.77	-3.59	-3.42	-3.26	-3.10	-2.94	-2.79	-2.63	-2.49	-2.34	-2.19	-2.05	-1.91	-1.77	-1.63	-1.5	-1.36	-1.23	
5	-3.79	-3.61	-3.43	-3.26	-3.09	-2.92	-2.76	-2.60	-2.45	-2.30	-2.15	-2.01	-1.86	-1.72	-1.58	-1.44	-1.31	-1.17	-1.04	-0.9	-0.77	-0.64	
4	-3.23	-3.04	-2.87	-2.69	-2.53	-2.36	-2.20	-2.05	-1.89	-1.74	-1.59	-1.45	-1.3	-1.16	-1.02	-0.88	-0.75	-0.61	-0.47	-0.34	-0.2	-0.07	
3	-2.76	-2.57	-2.38	-2.20	-2.03	-1.86	-1.69	-1.53	-1.37	-1.21	-1.06	-0.9	-0.75	-0.61	-0.46	-0.32	-0.17	-0.03	0.11	0.26	0.4	0.54	
2	-2.37	-2.15	-1.94	-1.74	-1.54	-1.35	-1.17	-0.99	-0.82	-0.64	-0.48	-0.31	-0.15	0.02	0.18	0.33	0.49	0.65	0.81	0.96	1.12	1.28	
1	-2.11	-1.81	-1.53	-1.27	-1.02	-0.78	-0.56	-0.34	-0.12	0.08	0.28	0.48	0.68	0.87	1.06	1.25	1.43	1.62	1.81	1.99	2.18	2.37	
0	-9.82	-3.34	-1.67	-0.78	-0.16	0.33	0.75	1.12	1.45	1.76	2.06	2.34	2.62	2.89	3.16	3.42	3.69	3.96	4.23	4.51	4.8	5.09	
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	
10	-7.72	-7.43	-7.15	-6.89	-6.64	-6.40	-6.18	-5.96	-5.74	-5.54	-5.35	-5.16	-4.99	-4.82	-4.67	-4.53	-4.42	-4.36	-4.37	-4.56	-5.36	-9.67	
9	-4.01	-3.83	-3.64	-3.46	-3.28	-3.10	-2.92	-2.74	-2.56	-2.39	-2.21	-2.03	-1.85	-1.66	-1.48	-1.28	-1.09	-0.89	-0.68	-0.46	-0.24	0	
8	-2.63	-2.47	-2.32	-2.16	-2.01	-1.86	-1.71	-1.55	-1.4	-1.25	-1.09	-0.93	-0.78	-0.62	-0.45	-0.29	-0.12	0.05	0.23	0.41	0.6	0.79	
7	-1.76	-1.62	-1.48	-1.34	-1.2	-1.06	-0.92	-0.78	-0.64	-0.49	-0.35	-0.21	-0.06	0.09	0.24	0.39	0.55	0.71	0.87	1.03	1.2	1.38	
6	-1.09	-0.96	-0.82	-0.69	-0.55	-0.42	-0.29	-0.15	-0.01	0.12	0.26	0.4	0.54	0.69	0.83	0.98	1.13	1.28	1.44	1.6	1.77	1.94	
5	-0.5	-0.37	-0.24	-0.11	0.03	0.16	0.29	0.43	0.56	0.7	0.84	0.98	1.12	1.27	1.41	1.56	1.71	1.87	2.03	2.19	2.36	2.53	
4	0.07	0.2	0.34	0.47	0.61	0.75	0.88	1.02	1.16	1.3	1.45	1.59	1.74	1.89	2.05	2.2	2.36	2.53	2.69	2.87	3.04	3.23	
3	0.68	0.82	0.97	1.11	1.26	1.4	1.55	1.7	1.85	2	2.16	2.32	2.48	2.64	2.81	2.99	3.16	3.35	3.53	3.73	3.93	4.14	
2	1.43	1.59	1.75	1.91	2.07	2.24	2.41	2.58	2.75	2.93	3.11	3.29	3.48	3.68	3.88	4.09	4.3	4.53	4.76	5.01	5.26	5.53	
0	5.4	5.72	6.05	6.4	6.77	7.16	7.58	8.04	8.53	9.08	9.68	10.36	11.13	12.03	13.09	14.39	16.03	18.22	21.37	26.58	38.15	87.58	
	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	

Tabla 1: Valores de T , para $n_1 = 43, n_2 = 10, \alpha = 0.05$ y $d_0 = 0.1$.

10	-107.07	-46.85	-32.80	-26.52	-22.73	-20.12	-18.18	-16.65	-15.40	-14.36	-13.46	-12.69	-12.00	-11.39	-10.84	-10.34	-9.87	-9.67	-9.67	-9.67	-9.67	-9.67
9	-10.54	-10.01	-9.52	-9.07	-8.66	-8.28	-7.92	-7.59	-7.27	-6.98	-6.69	-6.42	-6.16	-5.92	-5.68	-5.45	-5.23	-5.01	-4.80	-4.60	-4.40	-4.20
8	-9.82	-6.82	-6.54	-6.27	-6.02	-5.78	-5.55	-5.32	-5.11	-4.90	-4.70	-4.51	-4.32	-4.13	-3.95	-3.78	-3.61	-3.44	-3.27	-3.11	-2.95	-2.79
7	-9.82	-5.29	-5.08	-4.87	-4.67	-4.47	-4.28	-4.10	-3.92	-3.75	-3.58	-3.41	-3.25	-3.09	-2.94	-2.78	-2.63	-2.48	-2.34	-2.19	-2.05	-1.90
6	-9.82	-4.32	-4.13	-3.95	-3.77	-3.59	-3.42	-3.26	-3.10	-2.94	-2.79	-2.63	-2.49	-2.34	-2.19	-2.05	-1.91	-1.77	-1.63	-1.5	-1.36	-1.23
5	-9.82	-3.61	-3.43	-3.26	-3.09	-2.92	-2.76	-2.60	-2.45	-2.30	-2.15	-2.01	-1.86	-1.72	-1.58	-1.44	-1.31	-1.17	-1.04	-0.9	-0.77	-0.64
4	-9.82	-3.34	-2.87	-2.69	-2.53	-2.36	-2.20	-2.05	-1.89	-1.74	-1.59	-1.45	-1.3	-1.16	-1.02	-0.88	-0.75	-0.61	-0.47	-0.34	-0.2	-0.07
3	-9.82	-3.34	-2.38	-2.20	-2.03	-1.86	-1.69	-1.53	-1.37	-1.21	-1.06	-0.9	-0.75	-0.61	-0.46	-0.32	-0.17	-0.03	0.11	0.26	0.4	0.54
2	-9.82	-3.34	-1.94	-1.74	-1.54	-1.35	-1.17	-0.99	-0.82	-0.64	-0.48	-0.31	-0.15	0.02	0.18	0.33	0.49	0.65	0.81	0.96	1.12	1.28
1	-9.82	-3.34	-1.67	-1.27	-1.02	-0.78	-0.56	-0.34	-0.12	0.08	0.28	0.48	0.68	0.87	1.06	1.25	1.43	1.62	1.81	1.99	2.18	2.37
0	-9.82	-3.34	-1.67	-0.78	-0.16	0.33	0.75	1.12	1.45	1.76	2.06	2.34	2.62	2.89	3.16	3.42	3.69	3.96	4.23	4.51	4.8	5.09
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
10	-9.67	-9.67	-9.67	-9.67	-9.67	-9.67	-9.67	-9.67	-9.67	-9.67	-9.67	-9.67	-9.67	-9.67	-9.67	-9.67	-9.67	-9.67	-9.67	-9.67	-9.67	-9.67
9	-4.01	-3.83	-3.64	-3.46	-3.28	-3.10	-2.92	-2.74	-2.56	-2.39	-2.21	-2.03	-1.85	-1.66	-1.48	-1.28	-1.09	-0.89	-0.68	-0.46	-0.24	0
8	-2.63	-2.47	-2.32	-2.16	-2.01	-1.86	-1.71	-1.55	-1.4	-1.25	-1.09	-0.93	-0.78	-0.62	-0.45	-0.29	-0.12	0.05	0.23	0.41	0.6	0.79
7	-1.76	-1.62	-1.48	-1.34	-1.2	-1.06	-0.92	-0.78	-0.64	-0.49	-0.35	-0.21	-0.06	0.09	0.24	0.39	0.55	0.71	0.87	1.03	1.2	1.38
6	-1.09	-0.96	-0.82	-0.69	-0.55	-0.42	-0.29	-0.15	-0.01	0.12	0.26	0.4	0.54	0.69	0.83	0.98	1.13	1.28	1.44	1.6	1.77	1.94
5	-0.5	-0.37	-0.24	-0.11	0.03	0.16	0.29	0.43	0.56	0.7	0.84	0.98	1.12	1.27	1.41	1.56	1.71	1.87	2.03	2.19	2.36	2.53
4	0.07	0.2	0.34	0.47	0.61	0.75	0.88	1.02	1.16	1.3	1.45	1.59	1.74	1.89	2.05	2.2	2.36	2.53	2.69	2.87	3.04	3.23
3	0.68	0.82	0.97	1.11	1.26	1.4	1.55	1.7	1.85	2	2.16	2.32	2.48	2.64	2.81	2.99	3.16	3.35	3.53	3.73	3.93	4.14
2	1.43	1.59	1.75	1.91	2.07	2.24	2.41	2.58	2.75	2.93	3.11	3.29	3.48	3.68	3.88	4.09	4.3	4.53	4.76	5.01	5.26	5.53
0	5.4	5.72	6.05	6.4	6.77	7.16	7.58	8.04	8.53	9.08	9.68	10.36	11.13	12.03	13.09	14.39	16.03	18.22	21.37	26.58	38.15	87.58
	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43

Tabla 2: Valores de $[T]$, para $n_1 = 43, n_2 = 10, \alpha = 0.05$ y $d_0 = 0.1$.

6. Conclusiones

Las pruebas de no inferioridad son comúnmente utilizadas en ensayos clínicos y dada la importancia que tiene el hecho de que las regiones críticas correspondientes a dichas pruebas sean un conjunto convexo de Barnard. En este trabajo se presentan algunos ejemplos con el propósito de estudiar el comportamiento de las regiones críticas y a su vez mostrar algunas situaciones en las que la región crítica no es un CCB.

Motivados por el hecho de establecer una metodología que permita construir pruebas estadísticas cuyas regiones críticas siempre sean CCB a partir de una prueba estadística dada, se introdujo una lista de propiedades que la redefinición de una prueba estadística deberá de cumplir para garantizar que la nueva región crítica sea un CCB, más aún se estableció un método para poder redefinir una prueba estadística que cumpliera con las propiedades deseadas garantizando que la nueva prueba estadística es un CCB, con esto se obtienen beneficios para calcular el tamaño de las pruebas de no inferioridad, puesto que la nueva redefinición permite aplicar el teorema de Röhmel y Mansmann el cual reduce el tiempo de cómputo que se requiere para calcular los tamaños de prueba.

Bibliografía

Almendra-Arao, F. (2009). A study of the classical noninferiority test for two binomial proportions. *Drug Information*, 43:547–571.

Almendra-Arao, F. (2011). Barnard convex sets. *Communications in Statistics-Theory and Methods*, 40:2574–2582.

Almendra-Arao, F. (2012). Efficient calculation of test sizes for non-inferiority. *Computational Statistics and Data Analysis*, 56:4138–4145.

Almendra-Arao, F. and Sotres-Ramos, D. (2014). On the importance in clinical trials that critical regions for comparing 2 independent proportions must be barnard convex sets. *Therapeutic Innovation and Regulatory Science*, 48:208–212.

Barnard, G. (1947). Significance tests for 2x2 tables. *Biometrika*.

- Berger, R. (1982). Qualitymultiparameter hypothesis testing and acceptance sampling. *Taylor & Francis, Ltd.*
- Finner, H. and Strassburger, K. (2002). Structural properties of umpu-tests for 2x2 tables and some applications. *Journal of Statistical Planning and Inference*.
- Frick, H. (2000). Undominated p-values and property c for unconditional one-sided two sample binomial tests. *Biometrical J.*
- Phillips, K. F. (2003). A new test of non-inferiority for anti-infective trials. *Statistics in Medicine*, 22:201–212.
- Röhmel, J. (1998). Therapeutic equivalence investigations: statistical considerations. *Statistics in Medicine*.
- Röhmel, J. (2005). Problems with existing procedures to compute exact unconditional p-values for noninferiority/superiority y confidence intervals for two binomials y how to resolve them. *Biometrical J.*
- Röhmel, J. and Mansmann, U. (1999). Unconditional non-asymptotic one-sided tests for independent binomial proportions when the interest lies in showing non-inferiority and/or superiority. *Biometrical Journal*, 41:149–170.
- Zhang, Z. (2006). Non-inferiority testing with a variable margin. *Biometrical Journal*, 48:948–965.

Averaged Shifted Histograms (ASH) or Weighted Averaging of Rounded Points (WARP), Efficient Methods to Calculate Kernel Density Estimators for Circular Data^{*}

Isaías Hazarmabeth Salgado-Ugarte^a, Verónica Mitsui Saito-Quezada
*Laboratorio de Biometría y Biología Pesquera; Carrera de Biología, Facultad de Estudios
Superiores Zaragoza, UNAM, Batalla 5 de mayo S/N esq. Fuerte de Loreto, Ejército de
Oriente, Iztapalapa 09230, CdMx*

Marco Aurelio Pérez-Hernández
*Departamento de Biología, Universidad Autónoma Metropolitana, Iztapalapa. San Rafael
Atlixco 186, Vicentina, Iztapalapa 09340, CdMx*

By solving the histogram's problems of origin dependency, of discontinuity and by having guidance to choose the best bandwidth and the feasibility of variable bandwidth procedures, the Kernel density estimators (KDE's) are powerful tools to explore and analyze data distributions. However, an important drawback of these methods is that they require a considerable number of calculations, which may take a long time to obtain the result even using fast processors and moderate sample sizes. A way to overcome this problem is the average shifted histogram (ASH), procedure later recognized as being a part of the more general method named Weighted Averaging of Rounded Points (WARP). On the other side, the information with a circular measure scale commonly occurs in diverse human activities. Circular data distribution is an important characteristic that must be understood to properly interpret its message. The Rose Diagram is the histogram equivalent sharing the same drawbacks besides others derived from the circular scale. In this contribution we present a new program which permits the calculation of kernel density estimators for circular data with

* Apoyado por PAPIME PE206213 y PAPIIT IG201215, DGAPA, ICMyL, FES Zaragoza, UNAM; Posgrado en Ciencias Biológicas, UNAM y CONACyT

^aisalgado@unam.mx; ihsalgadougarte@gmail.com

different weight functions by means of the ASH-WARP procedure with an impressive calculation time saving, from several minutes with the former programs to less than a second with the new one to obtain the results.

Area-MSC: Simulación, Cómputo y Software Estadístico

Subárea-MSC: 62G07 Estimación de densidad

1. Introduction

Kernel density estimators (KDE's) are powerful tools to explore and analyze data distributions (Salgado-Ugarte, 2002; Salgado-Ugarte, et al. 1993). However, an important drawback in the application of these methods is that they require a considerable number of individual calculations, which may take a long time to obtain the result even using fast processors and moderate sample sizes. Scott (1985) suggested a way to overcome this problem: to average histograms shifted in their origin. Later, Härdle and Scott (1988) proposed the more general procedure named Weighted Averaging of Rounded Points (WARP). On the other side, the quantitative information associated with a circular measure scale occurs commonly in diverse fields of human activities. As with the linear scales, circular data distribution is an important characteristic that must be understood to properly interpret its message. The traditional procedure to investigate the distribution is the histogram or its linear interpolate, the frequency polygon, which share four problems: dependency on the origin and on the width (or number) of intervals, discontinuity and fixed interval width. For circular data, histogram variants known as Rose diagrams have been proposed, so these methods have the same problems as the linear histograms in addition with others associated with the circular representation of information. The KDE's solve the histogram's problems of origin and discontinuity besides having guidance to choose the interval (band) width and feasibility to implement KDE's with variable bandwidth. From the work of Fisher (1989; 1993) some algorithms to calculate KDE's for circular data are available. Based on these, Cox (1997; 2001; 2004) has proposed several computerized versions. Posteriorly, Salgado-Ugarte and Pérez-Hernández (2014) presented a program series to calculate KDE's for circular data. In this contribution we present a new program (in the Stata package language) "circwarp.ado" which permits the calculation of kernel density estimators with different weight functions (uniform, triangular, Epanechnikov, biweight, triweight or Gaussian) by means of the efficient aver-

ging of shifted histograms (ASH) or weighted averaging of rounded points (WARP) with an impressive calculation time saving.

2. Methods

Three steps are involved in the ASH-WARP method: a) data grouping; b) Weight values calculation, and c) weighting the bins (intervals). In the first step, a mesh of intervals is created and the number of observations in each interval is counted. The information about the data is reduced to a list of bin counts along with their midpoints. In the second step, a nonnegative, symmetric weight function is calculated. The weights are normalized to sum to M , the number of shifted histograms. Finally, the density estimate in each bin is computed as the product of the bin count and the weight (for details of this procedure see Härdle, 1991 or Scott, 1992). The new program “circwarp.ado” calculates the density of data measured in azimuthal scale by means of the ASH-WARP procedure and draws the result. It is possible to specify the weight (kernel) function, the bandwidth (smoothing parameter) and the number of histograms to average (ten as default). Besides, it can use three different types of graphic: linear interpolant (default, Figure 1.); step (histogram type, Figure 2.) and circular (Figure 3.). All these graphs may be customized (Figure 4.); moreover, it gives information on the modes (and antimodes) and, if desired, generates two variables with the resulting density and midpoints values for later use. Taking into account their efficiency and the concept of “Canonical Kernels”, even the Uniform Kernel is almost as efficient as the Epanechnikov weight function. The two weight functions traditionally more employed are the Quartic (Biweight) and the Gaussian due to algorithm availability and published research (Salgado-Ugarte, 2002; Scott, 2015).

3. Results and Conclusions

To exemplify the use of “circwarp.ado” it was applied to analyze the wind direction data registered in the weather station of the Facultad de Estudios Superiores Zaragoza, Campus II, Universidad Nacional Autónoma de México, from April 9 2012 to November 11th 2013, time lapse spanning 612 days with hourly registration that produced 13,352 individual observations. The results were presented in Salgado-Ugarte & Pérez-Hernández (2014). With

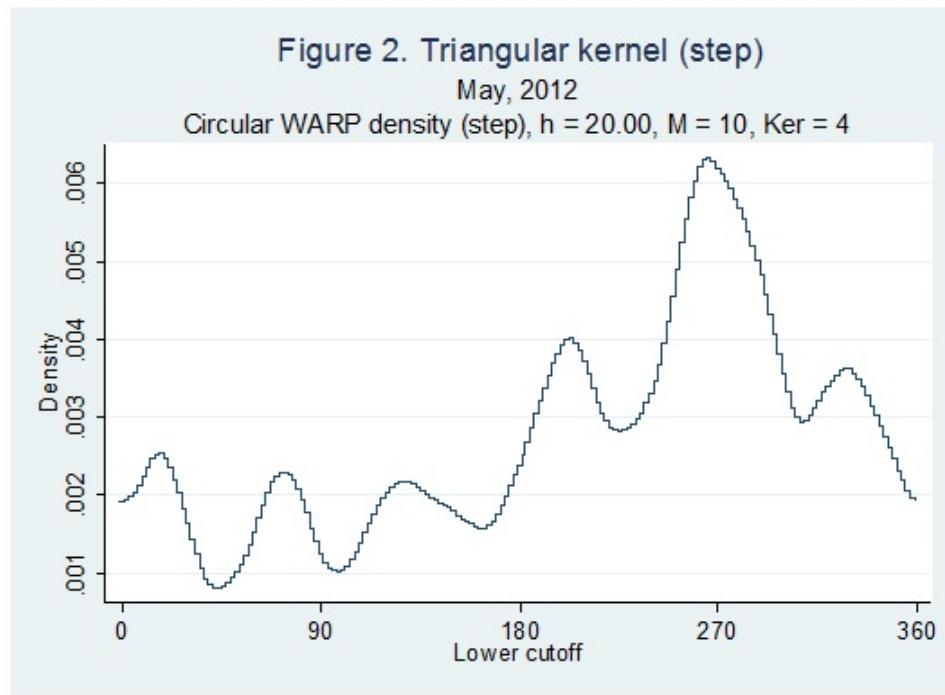
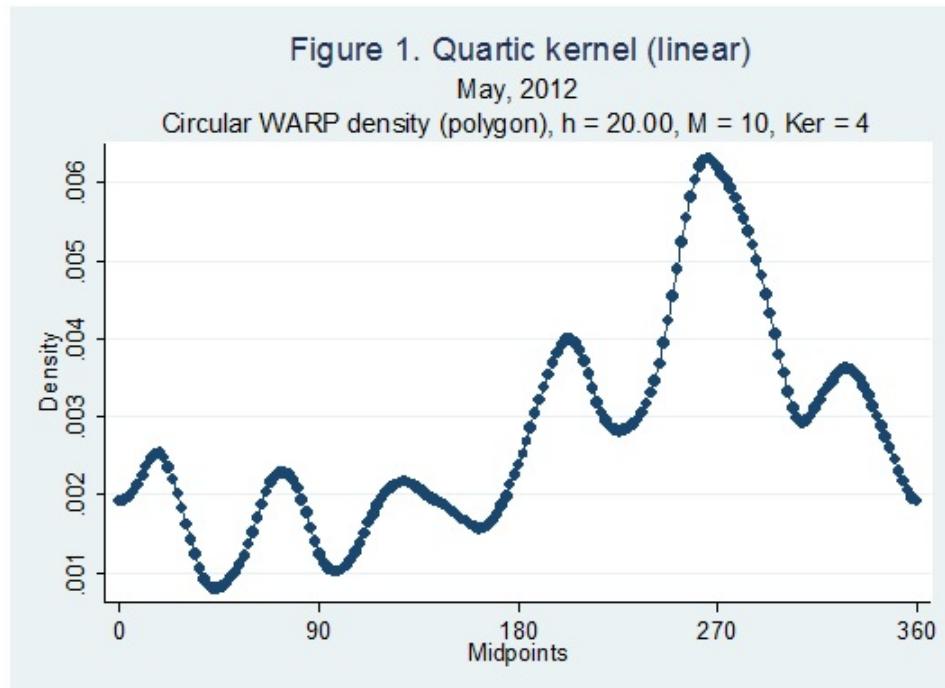
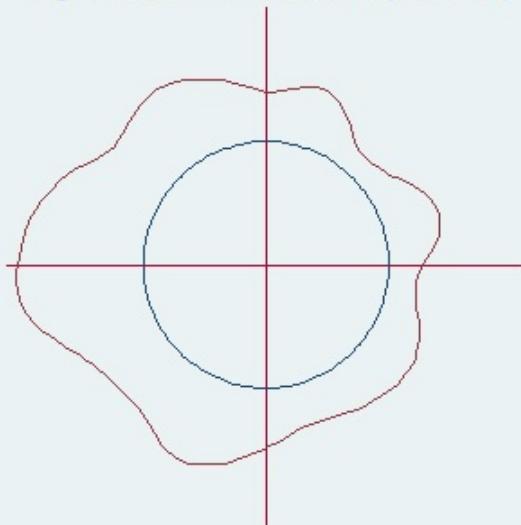


Table 1.

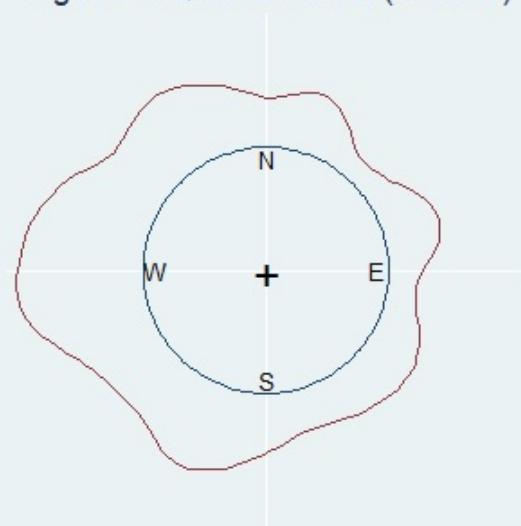
circkden.ado(quartic)			
n	bw	lineal	circular
100	87	00'0.59"	00'0.59"
1022	58	00'00.96"	00'1.12"
5007	64	00'10.17"	00'10.23"
10041	46	00'38.46"	00'38.60"
20076	32	2'31.18"	2'31.28"
36715	28	8'22.34"	8'22.71"
circwarp.ado (quartic)			
n	bw	lineal	circular
100	87	00'2.80"	00'1.03"
1022	58	00'0.59"	00'0.59"
5007	64	00'0.65"	00'0.65"
10041	46	00'0.80"	00'0.84"
20076	32	00'1.34"	00'1.35"
36715	28	00'2.18"	00'22.4"

Figure 3. Quartic kernel (circular)



May, 2012

Figure 4. Quartic kernel (circular)



May, 2012

the aim to consider not only the direction but the force of the wind, the direction data were weighted by its force and for the same data set a total of 36,715 registers were obtained. The former programs “circden.ado” and “circdevm.ado” applied to May 2012 data ($n = 2219$) with the bandwidth indicated by the “circbw.ado” program (Salgado-Ugarte & Pérez-Hernández (2014), took around 8 minutes to present the results (graph). The new program, “cirwarp.ado” applied to the same data set with the quartic (biweight) kernel and the “optimal” width of Fisher (1989; 1992) as calculated by circbw.ado (Salgado-Ugarte & Pérez-Hernández (2015) achieves the result in less than a second (Table 1) with both, linear and circular versions (Figures 1 to 3) with an Intel Xeon E5-1607 v4, 3.10GHz, 3100MHz, 4 processors. This represents an amazing computing time saving. Figure 2 shows an “averaged shifted histogram like version (Härdle & Scott, 1988)”. As with the previous versions, it is possible to customize the graphics (Figure 4). Some preliminary simulation results indicated that the circwarp.ado program recovers (as expected) the data information (modality and location) adequately. More detailed and complete research on the subject is on its way.

Again it is noted the importance of the use of the KDE's as a very powerful tool to investigate fundamental characteristic (symmetry, bias and modality) of the distribution of circular data. The efficient method of ASH-WARP permits an amazing time saving in the calculation of these intensive computing statistical procedures. The former (circden.ado, circdevm.ado, circbw.ado, circgph.ado) and the new warping program (cirwarp.ado) are available by request to the first author.

Acknowledgements

We thank Mitsui Myrna Salgado Saito for her help with the L^AT_EX version.

Bibliografía

Cox, N. (1997). Circular statistics in Stata. In *Proceedings of the 3rd UK User Group Meeting*, London, UK.

Cox, N. (2001). Analysing circular data in Stata, revisited. In *Proceedings of the 3rd North American User Group Meeting*, page 4, Boston, USA.

- Cox, N. (2004). Circular statistics in Stata. In *Proceedings of the 10th UK User Group Meeting*, page 4, London, UK.
- Fisher, N. (1989). Smoothing a sample of circular data. *Journal of Structural Geology*, 11(6):775–778.
- Fisher, N. (1993). *Statistical Analysis of Circular Data*. Cambridge University Press, Cambridge, UK.
- Härdle, W. (1991). *Smoothing Techniques. With Implementations in S*. Springer-Verlag, New York, USA.
- Härdle, W. and Scott, D. (1992). Smoothing by weighted averaging of rounded points. *Computational Statistics*, 7:129–136.
- Salgado-Ugarte, I. (2002). *Suavización no Paramétrica para Análisis de Datos*. FES Zaragoza and DGAPA, UNAM, Mexico.
- Salgado-Ugarte, I. and Pérez-Hernández, M. (2014). Estimación de densidad por núcleo (kernel) para datos circulares. In *XXIX Foro Internacional de Estadística*, Universidad Popular Autónoma del Estado de Puebla, Puebla, México.
- Salgado-Ugarte, I., Rivera-Reyes, R., Monroy-Ata, A., and Saito-Quezada, V. (2015). Distribución de la dirección del viento en la FES Zaragoza analizada mediante estimadores de densidad por kernel circulares. In *Resúmenes del 11o Congreso de Investigación de la FES Zaragoza, UNAM*, CDMX, Mexico.
- Salgado-Ugarte, I., Shimizu, M., and Taniuchi, T. (1993). Exploring the shape of univariate data using kernel density estimators. *Stata Technical Bulletin*, 16:8–9.
- Scott, D. (1985). Averaged shifted histograms: effective nonparametric density estimation in several dimensions. *Annals of Statistics*, 13:1024–1040.
- Scott, D. (2015). *Multivariate Density Estimation: Theory, Practice, and Visualization*. John Wiley and Sons, New York, 2nd edition.

Pruebas de No Inferioridad Comparando Dos Distribuciones Poisson*

María de Lourdes Morales Sánchez, Hortensia Reyes Cervantes^a

Facultad de Ciencias Físico Matemáticas, Benemérita Universidad Autónoma de Puebla

Félix Almendra Arao

UPIITA del Instituto Politécnico Nacional

Con frecuencia los eventos raros se describen mediante una distribución Poisson. En investigación clínica el análisis de tales eventos puede ser la base de una prueba de no inferioridad. Para planear tales ensayos se requiere de la potencia de una prueba estadística para comparar las medias de dos distribuciones Poisson, particularmente se requiere calcular los tamaños de prueba. Este cálculo es complicado y para realizarlo conviene contar con herramientas teóricas que permitan su simplificación. En este trabajo demostramos que el cálculo de los tamaños de prueba puede restringirse a la curva frontera bajo la hipótesis nula, esto representa un avance fundamental pues facilita dicho cálculo considerablemente.

Área-MSN: Pruebas de hipótesis

Subárea-MSN: Pruebas de hipótesis

1. Introducción

Las pruebas de no inferioridad son procedimientos estadísticos utilizados principalmente en ensayos clínicos para la comparación de dos poblaciones, por ejemplo en la aplicación de un nuevo tratamiento contra uno ya existente. Estas pruebas son aplicadas con la finalidad de poder determinar si un tratamiento nuevo es superior, igual o inferior, por un margen generalmente pequeño, a uno ya existente que es considerado tratamiento estándar.

Para la comparación de dos distribuciones Poisson se pueden usar varias pruebas estadísticas, véase por ejemplo Ng y Tang (2008). En el desarrollo de este trabajo analizaremos las

* El tercer autor agradece el apoyo de SNI-CONACYT, COFAA-IPN y el proyecto SIP-IPN 20160687

^alourde_1991@hotmail.com

siguientes pruebas:

- Prueba de razón de verosimilitud.
- Prueba score
- Prueba exacta condicional.

Las primeras dos pruebas se basan en la distribución asintótica de sus estadísticas de prueba.

2. Marco Teórico

Supongamos que un nuevo tratamiento es comparado con un tratamiento de control en un estudio de grupos paralelos con n_1 individuos en el grupo del tratamiento de control y n_2 individuos en el grupo del nuevo tratamiento. La variable observada en cada individuo es el número de ocurrencias de cierto evento no deseado, por ejemplo el número de ataques de asma de un paciente. Así la superioridad o no inferioridad del nuevo tratamiento se basa en la reducción del número de eventos. Asumimos que el número de eventos en cada individuo sigue un proceso Poisson con media μ_1 para el grupo de control y μ_2 para el grupo del nuevo tratamiento. Este número de eventos es considerado respecto a una cierta unidad de tiempo, por ejemplo cada año. Sea $t_{i,j}$ el tiempo observado del individuo j -ésimo en el grupo i , entonces el número total de eventos Y_i en el grupo i tiene distribución Poisson con media

$$\lambda_i = m_i \mu_i \quad i = 1, 2. \quad (1)$$

donde:

$$m_i = \sum_{j=1}^{n_i} t_{i,j} \quad i = 1, 2.$$

m_i es el tiempo total de observación en el grupo i .

La función de distribución de probabilidad para el número total de eventos en el grupo i es

$$P[Y_i = y] = \frac{\exp(-\lambda_i) \lambda_i^y}{y!} \quad i = 1, 2.$$

Un procedimiento para demostrar que el nuevo tratamiento es no inferior o superior al tratamiento de control se basa en la comparación de dos poblaciones mediante el contraste de las hipótesis

$$H_0 : \mu_2 \geq g(\mu_1) \quad vs. \quad H_1 : \mu_2 < g(\mu_1). \quad (2)$$

Donde $g : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ una función no acotada, creciente y derivable tal que $\lim_{\mu \rightarrow 0^+} g(\mu) = 0$. En particular cuando $g(\mu) = \rho\mu$, para este caso notemos que:

Si ρ es menor o igual que 1, el objetivo de la prueba es mostrar la superioridad del nuevo tratamiento.

Si ρ es mayor o igual que 1, el objetivo de la prueba es mostrar la no inferioridad del nuevo tratamiento.

2.1. Pruebas de Razón de Verosimilitudes, Score y Exacta Condicional

Denotemos por y_i el número total de eventos observados en el grupo i . Sean:

$$\begin{aligned} y_0 &= y_1 + y_2 \\ \gamma &= \frac{m_2}{m_1}\rho \\ \psi &= \frac{\mu_2}{\mu_1}. \end{aligned} \tag{3}$$

Denotamos el espacio paramétrico como $\Omega = \{(\mu_1, \mu_2) | 0 < \mu_1 < \infty, 0 < \mu_2 < \infty\}$.

Y al espacio bajo la hipótesis nula como $\Omega_0 = \{(\mu_1, \mu_2) | \frac{\mu_2}{\mu_1} \geq \rho, \mu_1 \in \mathbb{R}^+, \mu_2 \in \mathbb{R}^+\}$
 $= \{(\mu_1, \mu_2) | \frac{\mu_2}{\rho} \geq \mu_1, \mu_1 \in \mathbb{R}^+, \mu_2 \in \mathbb{R}^+\}$.

También denotemos a Ω_0^c como Ω_1 , que es el espacio bajo la hipótesis alternativa.

Proposición 2.1. *El estadístico de razón de verosimilitud para el contraste de hipótesis $H_0 : (\mu_1, \mu_2) \in \Omega_0$ vs. $H_1 : (\mu_1, \mu_2) \in \Omega_1$ es*

$$G^2(y_1, y_2) = 2[y_1 \ln(y_1) + y_2 \ln(y_2/\gamma) - y_0 \ln(y_0/(1 + \gamma))]. \tag{4}$$

donde \ln denota el logaritmo y además $y \ln(y)$ es definido como cero si y es igual a cero.

Demostración. La función de verosimilitud es:

$$L(\Omega) = \frac{e^{-(m_1\mu_1+m_2\mu_2)}(m_1\mu_1)^{y_1}}{y_1!} \frac{(m_2\mu_2)^{y_2}}{y_2!}.$$

Así tenemos que:

$$\frac{\partial(\ln(L(\Omega)))}{\partial\mu_1} = -m_1 + \frac{y_1 m_1}{m_1 \mu_1} \quad \frac{\partial(\ln(L(\Omega)))}{\partial\mu_2} = -m_2 + \frac{y_2 m_2}{m_2 \mu_2}.$$

igualando a cero obtenemos los estimadores de máxima verosimilitud, los cuales son:

$$\hat{\mu}_1 = \frac{y_1}{m_1} \quad \hat{\mu}_2 = \frac{y_2}{m_2}.$$

Bajo la hipótesis nula:

$$L(\Omega_0) = \frac{e^{-m_1\frac{\mu_2}{\rho}}(m_1\frac{\mu_2}{\rho})^{y_1}}{y_1!} \frac{e^{-m_2\mu_2}(m_2\mu_2)^{y_2}}{y_2!} = \frac{e^{-\mu_2(\frac{m_1}{\rho}+m_2)}(\mu_2)^{y_1+y_2}\left(\frac{m_1}{\rho}\right)^{y_1}(m_2)^{y_2}}{y_1!y_2!}.$$

Así:

$$\frac{\partial(\ln(L(\Omega_0)))}{\partial\mu_1} = -\left(\frac{m_1}{\rho} + m_2\right) + \frac{y_1 + y_2}{\mu_2}$$

entonces los estimadores bajo la hipótesis nula son:

$$\tilde{\mu}_2 = \frac{y_1 + y_2}{\frac{m_1}{\rho} + m_2} = \rho \frac{y_1 + y_2}{m_1(1 + \gamma)} \quad \tilde{\mu}_1 = \frac{\tilde{\mu}_2}{\rho} = \frac{y_1 + y_2}{m_1(1 + \gamma)}.$$

Entonces:

$$\frac{L(\hat{\Omega}_0)}{L(\hat{\Omega})} = \frac{\left(\frac{y_0}{m_1(1+\gamma)}\right)^{y_0} m_1^{y_1} (\rho m_2)^{y_2}}{(y_1)^{y_1} (y_2)^{y_2}} = \frac{\left(\frac{y_0}{(1+\gamma)}\right)^{y_0} (\rho \frac{m_2}{m_1})^{y_2}}{(y_1)^{y_1} (y_2)^{y_2}} = \frac{\left(\frac{y_0}{(1+\gamma)}\right)^{y_0}}{(y_1)^{y_1} (\frac{y_2}{\gamma})^{y_2}}.$$

$$G^2(y_1, y_2) = 2\ln\left(\frac{(y_1)^{y_1} (\frac{y_2}{\gamma})^{y_2}}{\left(\frac{y_0}{(1+\gamma)}\right)^{y_0}}\right) = 2\left[y_1\ln(y_1) + y_2\ln\left(\frac{y_2}{\gamma}\right) - y_0\ln\left(\frac{y_0}{1+\gamma}\right)\right].$$

Se sabe que bajo la hipótesis nula el estadístico de razón de verosimilitud y el estadístico score tienen asintóticamente una distribución Ji-cuadrada con un grado de libertad una prueba de este hecho se puede encontrar en Gu et al. (2008).

Proposición 2.2. *El estadístico score para el contraste de hipótesis $H_0 : (\mu_1, \mu_2) \in \Omega_0$ vs. $H_1 : (\mu_1, \mu_2) \in \Omega_1$ está dado por:*

$$X^2(y_1, y_2) = \gamma \frac{(y_1 - y_2/\gamma)^2}{y_0}. \quad (5)$$

Demostración. La función de verosimilitud se puede reescribir de la siguiente manera:

$$L(\Omega) = \frac{e^{-\lambda_1}(\lambda_1)^{y_1+y_2} e^{-\frac{m_2}{m_1}\lambda_1\psi} (\frac{m_2}{m_1}\psi)^{y_2}}{y_1!y_2!}.$$

La función score es:

$$U(\psi, \lambda_1) = \frac{\partial \ln(L(\mu_1, \mu_2; y_1, y_2))}{\partial \psi} = \frac{y_2}{\psi} - \frac{m_2}{m_1} \lambda_1.$$

Así obtenemos que:

$$U(\tilde{\psi}, \tilde{\lambda}_1) = \frac{\gamma(\frac{y_2}{\gamma} - y_1)}{\rho(1 + \gamma)}.$$

Además la varianza es:

$$I(\psi, \lambda_1) = \text{Var}(U(\tilde{\psi}, \tilde{\lambda}_1)) = \frac{\gamma(y_0)}{\rho^2(1 + \gamma)^2}.$$

Por tanto el estadístico score está dado por:

$$X^2(y_1, y_2) = \frac{(U(\tilde{\psi}, \tilde{\lambda}_1))^2}{I(\tilde{\psi}, \tilde{\lambda}_1)} = \gamma \frac{\left(y_1 - \frac{y_2}{\gamma}\right)^2}{y_0}.$$

Como la hipótesis nula es unilateral y la hipótesis alternativa es $\frac{\mu_2}{\mu_1} < \rho$, es decir $\frac{\lambda_2}{\lambda_1} < \gamma$ entonces las pruebas de razón de verosimilitudes y score deben ser aplicadas cuando

$$y_2 < \gamma y_1. \quad (6)$$

Condicionando el número total de eventos observados y_0 entonces el número de eventos en cada grupo posee una distribución binomial veáse Miede y Mueller-Cohrs (2005), digamos

$$y_2 \sim \text{Bin}(\theta, y_0) \quad \theta = \frac{\gamma}{1 + \gamma}.$$

El p -valor de la prueba exacta condicional es

$$p = F[y_2; \theta, y_0] = T(y_1, y_2).$$

Donde F denota la función de distribución acumulada de la distribución binomial con probabilidad de éxito es θ y tamaño muestral y_0 .

3. Región Crítica

La función potencia de las pruebas mencionadas puede ser calculada sumando las probabilidades de todas las observaciones en la región crítica (C), es decir

$$\beta(\mu_1, \mu_2) = \sum_{(y_1, y_2) \in C} P[Y_1 = y_1 | \mu_1] P[Y_2 = y_2 | \mu_2].$$

Diremos que una región crítica cumple la propiedad de monotonía si dado un punto (y_1, y_2) en la región crítica, entonces los puntos $(y_1 + 1, y_2)$ y $(y_1, y_2 - 1)$ también pertenecen a la región crítica.

Las regiones críticas de las tres pruebas que consideramos en este trabajo cumplen la propiedad de monotonía.

Proposición 3.1. *La región crítica de la prueba de razón de verosimilitud con un nivel de significancia α es*

$$C_{rv} = \{(y_1, y_2) | G^2(y_1, y_2) > \chi^2_{(1, 1-2\alpha)}\}. \quad (7)$$

C_{rv} satisface la propiedad de monotonía siempre que $y_2 < \gamma y_1$.

Demostración. Veamos que el estadístico es creciente respecto a la variable y_1

$$\frac{\partial G^2(y_1, y_2)}{\partial y_1} = 2[\ln(y_1) + 1 - \left(\ln\left(\frac{y_1 + y_2}{1 + \gamma}\right) + 1\right)] = 2\ln\left(\frac{y_1(1 + \gamma)}{y_1 + y_2}\right).$$

Y como $y_2 < \gamma y_1$, entonces:

$$\frac{y_1 + \gamma y_1}{y_1 + y_2} > \frac{y_1 + y_2}{y_1 + y_2} > 1$$

así:

$$\frac{\partial G^2(y_1, y_2)}{\partial y_1} > 0.$$

Veamos que el estadístico es decreciente respecto a la variable y_2

$$\frac{\partial G^2(y_1, y_2)}{\partial y_2} = 2[\ln\left(\frac{y_2}{\gamma}\right) + \gamma - \left(\ln\left(\frac{y_1 + y_2}{1 + \gamma}\right) + 1\right)] = 2[\ln\left(\frac{(1 + \gamma)y_2}{\gamma(y_1 + y_2)}\right) + \gamma - 1].$$

Además:

$$\begin{aligned} 1 \leq y_2 < \gamma y_1 &\Rightarrow \frac{1}{y_2} > \frac{1}{y_1 \gamma} \\ \frac{(1 + \gamma)y_2}{\gamma y_1 + \gamma y_2} &< \frac{(1 + \gamma)y_2}{(1 + \gamma)y_2} < 1. \end{aligned}$$

así

$$\frac{\partial G^2(y_1, y_2)}{\partial y_2} < 0.$$

En la Figura 1 se visualiza que la región crítica cumple con la propiedad de monotonía para un ejemplo específico de la prueba razón de verosimilitud.

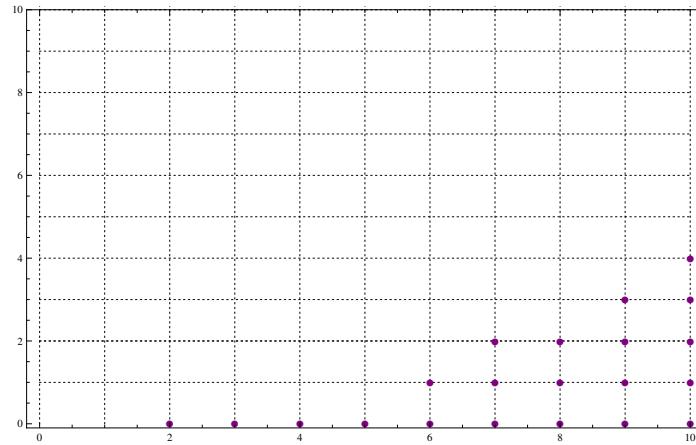


Figura 1: Región crítica para $n_1 = n_2 = 10 = m_1 = m_2$, $\gamma = \rho = 1.5$, prueba razón de verosimilitudes, $\alpha = 0.025$.

Proposición 3.2. *La región crítica de la prueba de Score con un nivel de significancia α es:*

$$C_{rs} = \left\{ (y_1, y_2) \mid X^2(y_1, y_2) > \chi^2_{(1, 1-2\alpha)} \right\}. \quad (8)$$

Si $y_2 < \gamma y_1$ entonces la región crítica C_{rs} satisface la propiedad de monotonía.

Demostración.

$$\frac{\partial X^2(y_1, y_2)}{\partial y_1} = \frac{2\gamma \left(y_1 - \frac{y_2}{\gamma} \right) (y_1 + y_2) - \gamma (y_1 - \frac{y_2}{\gamma})^2}{(y_1 + y_2)^2} = \frac{\gamma \left(y_1 - \frac{y_2}{\gamma} \right) (\gamma y_1 + y_2(2\gamma + 1))}{(y_1 + y_2)^2}.$$

Además:

$$y_2 < \gamma y_1 \Rightarrow y_1 - \frac{y_2}{\gamma} > 0.$$

Así:

$$\frac{\partial X^2(y_1, y_2)}{\partial y_1} > 0.$$

Ahora:

$$\frac{\partial X^2(y_1, y_2)}{\partial y_2} = \frac{2\frac{-\gamma}{\gamma} \left(y_1 - \frac{y_2}{\gamma} \right) (y_1 + y_2) - \gamma(y_1 - \frac{y_2}{\gamma})^2}{(y_1 + y_2)^2} = \frac{-\left(y_1 - \frac{y_2}{\gamma} \right) (y_1(2 + \gamma) + y_2)}{(y_1 + y_2)^2}.$$

Así:

$$\frac{\partial X^2(y_1, y_2)}{\partial y_2} < 0.$$

En la Figura 2 se visualiza que la región crítica cumple con la propiedad de monotonía para un ejemplo específico de la prueba score.

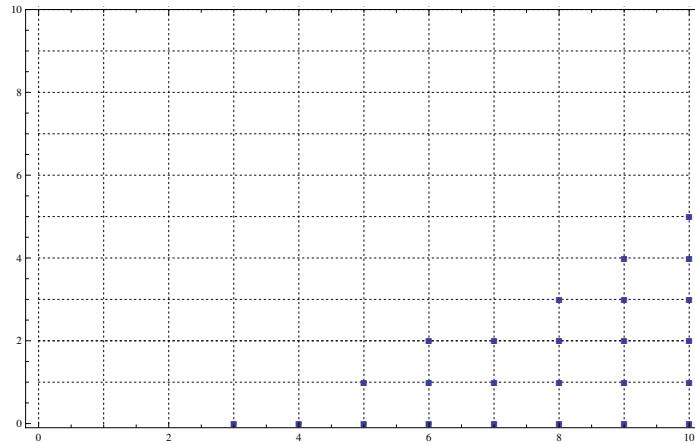


Figura 2: Región crítica para $n_1 = n_2 = 10 = m_1 = m_2$, $\gamma = \rho = 1.5$, prueba Score, $\alpha = 0.025$.

Proposición 3.3. La región crítica de la prueba de exacta condicional con un nivel de significancia α es

$$C_{re} = \{(y_1, y_2) | T(y_1, y_2) < \alpha\}. \quad (9)$$

Si $y_2 < \gamma y_1$ entonces la región crítica C_{re} satisface la propiedad de monotonía.

Demostración. Si $(y_1, y_2) \in C_{re}$ tenemos que:

$$T(y_1, y_2) = F[y_2; \theta, y_0] = F[y_2 - 1; \theta, y_0] + \binom{y_0}{y_2} \theta^{y_2} (1 - \theta)^{y_0 - y_2} < \alpha.$$

Así:

$$T(y_1, y_2 - 1) = F[y_2 - 1; \theta, y_0] < \alpha.$$

Es decir $(y_1, y_2 - 1) \in C_{re}$.

Notemos que:

$$T(y_1 + 1, y_2) = \sum_{k=0}^{k=y_2} \binom{y_0 + 1}{k} \theta^k (1 - \theta)^{y_0 + 1 - k} = \sum_{k=0}^{k=y_2} \binom{y_0}{k} \theta^k (1 - \theta)^{y_0 - k} \frac{(1 - \theta)(y_0 + 1)}{y_0 + 1 - k}.$$

Además $0 \leq k \leq y_2$ entonces $1 \leq y_1 + 1 = y_0 + 1 - y_2 \leq y_0 + 1 - k \leq y_0 + 1$ y así

$$\frac{(1 - \theta)(y_0 + 1)}{y_0 + 1 - k} < 1.$$

Y obtenemos que:

$$T(y_1 + 1, y_2) \leq T_3(y_1, y_2) < \alpha.$$

Por tanto $(y_1 + 1, y_2) \in C_3$.

En la Figura 3 se visualiza que la región crítica cumple con la propiedad de monotonía para un ejemplo específico de la prueba exacta condicional.

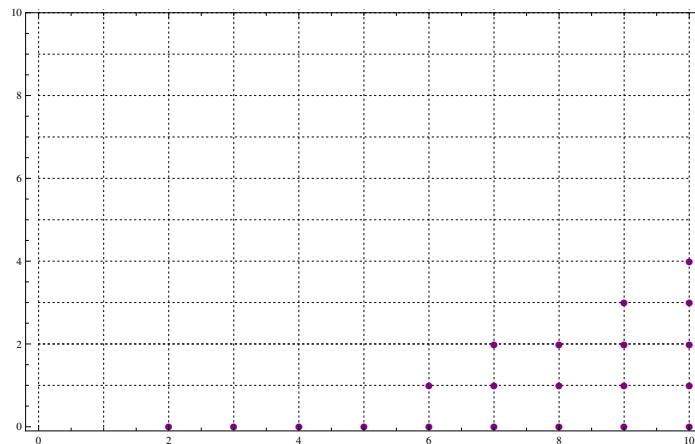


Figura 3: Región crítica para $n_1 = n_2 = 10 = m_1 = m_2$, $\gamma = \rho = 1.5$, prueba exacta condicional, $\alpha = 0.025$.

Denotemos $\mathbb{N}^* = \{0, 1, 2, 3, \dots\}$. Dado $y_1 \in \mathbb{N}^*$ definimos la *sección* de C determinada por y_1 como el conjunto $C_{y_1} = \{y_2 \in \mathbb{N}^* \mid (y_1, y_2) \in C\}$.

Similarmente, para $y_2 \in \mathbb{N}^*$ definimos la *sección* de C determinada por y_2 como $C^{y_2} = \{x_1 \in \mathbb{N}^* \mid (y_1, y_2) \in C\}$. Supongamos que la región crítica C cumple con la propiedad de monotonía. Si para $y_1 \in \mathbb{N}^*$ fijo la sección C_{y_2} es diferente del vacío, definimos $M(y_1) =$

$\max C_{y_1}$ y como $M(y_1) = \infty$ si tal máximo no existe, así $C_{y_1} = \{0, 1, 2, \dots, M(y_1)\}$ o $C_{y_1} = \{0, 1, 2, \dots\}$, respectivamente. Si para $y_2 \in \mathbb{N}^*$ fijo la sección C^{y_2} es no vacía, definimos $m(y_2) = \min C^{y_2}$ y notamos que $C^{y_2} = \{m(y_2), m(y_2) + 1, m(y_2) + 2, \dots\}$.

Así obtenemos los siguientes resultados.

Proposición 3.4. *Si C cumple la propiedad de monotonía, entonces*

- (1) $y_1 < y'_1$ implica $C_{y_1} \subset C_{y'_1}$ y $M(y_1) < M(y'_1)$.
- (2) $y_2 < y'_2$ implica $C^{y'_2} \subset C^{y_2}$ y $m(y_2) < m(y'_2)$.

Demostración. (1) Sean $y_1, y'_1 \in \mathbb{N}^*$ tal que $y_1 < y'_1$ entonces $\exists n \in \mathbb{N}$ que cumple que $y_1 + n = y'_1$. Veamos que $C_{y_1} \subset C_{y'_1}$. Si $y_2 \in C_{y_1}$ entonces $(y_1, y_2) \in C$ y por la monotonía de C se tiene que $(y_1 + n, y_2) = (y'_1, y_2) \in C$, es decir $y_2 \in C_{y'_1}$. Y como $C_{y_1} \subset C_{y'_1}$ entonces $\max C_{y_1} \leq \max C_{y'_1} \Rightarrow M(y_1) \leq M(y'_1)$.

(2) Sean $y_2, y'_2 \in \mathbb{N}^*$ tal que $y_2 < y'_2$ entonces $\exists n \in \mathbb{N}$ que cumple que $y_2 = y'_2 - n$. Veamos que $C^{y'_2} \subset C^{y_2}$. Si $y_1 \in C^{y'_2}$ entonces $(y_1, y'_2) \in C$ y por la monotonía de C se tiene que $(y_1, y'_2 - n) = (y_1, y_2) \in C$, es decir $y_1 \in C^{y_2}$. Así $C^{y'_2} \subset C^{y_2} \Rightarrow \min C^{y_2} \geq \min C^{y'_2} \Rightarrow m(y_2) \geq m(y'_2)$.

4. Tamaños de Prueba

Consideremos

$$f(y, \lambda) = \frac{\exp(-\lambda)\lambda^y}{y!}.$$

La función potencia esta dada por:

$$\beta(\mu_1, \mu_2) = \sum_{(y_1, y_2) \in C} f(y_1, m_1 \mu_1) f(y_2, m_2 \mu_2).$$

El tamaño de la prueba es definido por Hogg et al. (2005) de la siguiente manera:

$$\sup_{(\mu_1, \mu_2) \in \Omega_0} \beta(\mu_1, \mu_2).$$

Notemos que:

$$P(C^{y_2}) = \sum_{y_1=m(y_2)}^{\infty} f(y_1, m_1 \mu_1). \quad (10)$$

y

$$P(C_{y_1}) = \sum_{y_2=0}^{M(y_1)} f(y_2, m_2\mu_2). \quad (11)$$

Así podemos reescribir la función potencia como:

$$\beta(\mu_1, \mu_2) = \sum_{y_2=0}^{\infty} \sum_{y_1=m(y_2)}^{\infty} f(y_1, m_1\mu_1) f(y_2, m_2\mu_2) = \sum_{y_2=0}^{\infty} \left(f(y_2, m_2\mu_2) \sum_{y_1=m(y_2)}^{\infty} f(y_1, m_1\mu_1) \right).$$

y por (10):

$$\beta(\mu_1, \mu_2) = \sum_{y_2=0}^{\infty} f(y_2, m_2\mu_2) P(C^{y_2}). \quad (12)$$

de otra manera:

$$\beta(\mu_1, \mu_2) = \sum_{y_1=m(y_2)}^{\infty} \sum_{y_2=0}^{M(y_1)} f(y_1, m_1\mu_1) f(y_2, m_2\mu_2) = \sum_{y_1=m(y_2)}^{\infty} \left(f(y_1, m_1\mu_1) \sum_{y_2=0}^{M(y_1)} f(y_2, m_2\mu_2) \right).$$

y por (10):

$$\beta(\mu_1, \mu_2) = \sum_{y_1=m(y_2)}^{\infty} f(y_1, m_1\mu_1) P(C_{y_1}). \quad (13)$$

Proposición 4.1. Si g es una función real y diferenciable entonces:

$$\frac{\partial f(y, g(\mu))}{\partial \mu} = g'(\mu)(f(y-1, g(\mu)) - f(y, g(\mu))).$$

Liu y Hsueh (2013) probaron que para $g(\mu_1) = \mu_1$, cuando la región crítica cumple la propiedad de monotonía, entonces el tamaño de prueba se obtiene simplemente como el máximo sobre la curva frontera. El siguiente teorema generaliza dicho resultado.

Teorema 4.1. Sea $g : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ una función no acotada, creciente y derivable tal que $\lim_{\mu \rightarrow 0^+} g(\mu) = 0$. Si $C \neq \emptyset$ y cumple la propiedad de monotonía, entonces el tamaño de prueba está dado por:

$$\sup_{\mu_2 \geq g(\mu_1)} \beta(\mu_1, \mu_2) = \sup_{\mu_2=g(\mu_1)} \beta(\mu_1, \mu_2).$$

Demostración. Por la ecuación (12) y usando la proposición 4.1 tenemos que:

$$\begin{aligned}
 \frac{\partial \beta(\mu_1, \mu_2)}{\partial \mu_1} &= m_1 \sum_{y_1=m(y_2)}^{\infty} (f(y_1 - 1, m_1 \mu_1) - f(y_1, m_1 \mu_1)) P(C_{y_1}) \\
 &= m_1 \left[\sum_{y_1=m(y_2)}^{\infty} f(y_1 - 1, m_1 \mu_1) P(C_{y_1}) - \sum_{y_1=m(y_2)}^{\infty} f(y_1, m_1 \mu_1) P(C_{y_1}) \right] \\
 &= m_1 f(m(y_2) - 1, m_1 \mu_1) P(C_{m(y_2)}) + \\
 &\quad m_1 \left[\sum_{y_1=m(y_2)}^{\infty} f(y_1, m_1 \mu_1) P(C_{y_1+1}) - \sum_{y_1=m(y_2)}^{\infty} f(y_1, m_1 \mu_1) P(C_{y_1}) \right] \\
 &= m_1 \left[f(m(y_2) - 1, m_1 \mu_1) P(C_{m(y_2)}) + \sum_{y_1=m(y_2)}^{\infty} f(y_1, m_1 \mu_1) [P(C_{y_1+1}) - P(C_{y_1})] \right].
 \end{aligned}$$

y por la proposición 3.4 concluimos que $\frac{\partial \beta(\mu_1, \mu_2)}{\partial \mu_1} > 0$.

Similarmente, usando (13):

$$\begin{aligned}
 \frac{\partial \beta(\mu_1, \mu_2)}{\partial \mu_2} &= m_2 \sum_{y_2=0}^{\infty} (f(y_2 - 1, m_2 \mu_2) - f(y_2, m_2 \mu_2)) P(C^{y_2}) \\
 &= m_2 \left[\sum_{y_2=0}^{\infty} f(y_2 - 1, m_2 \mu_2) P(C^{y_2}) - \sum_{y_2=0}^{\infty} f(y_2, m_2 \mu_2) P(C^{y_2}) \right] \\
 &= m_2 \left[\sum_{y_2=0}^{\infty} f(y_2, m_2 \mu_2) P(C^{y_2+1}) - \sum_{y_2=0}^{\infty} f(y_2, m_2 \mu_2) P(C^{y_2}) \right] \\
 &= m_1 \left[\sum_{y_2=0}^{\infty} f(y_2, m_2 \mu_2) [P(C^{y_2+1}) - P(C^{y_2})] \right].
 \end{aligned}$$

y por la proposición 3.4, $\frac{\partial \beta(\mu_1, \mu_2)}{\partial \mu_2} < 0$. Por lo tanto tenemos que:

$$\sup_{\mu_2 \geq g(\mu_1)} \beta(\mu_1, \mu_2) = \sup_{\mu_2 = g(\mu_1)} \beta(\mu_1, \mu_2) = \sup_{\mu > 0} \beta(\mu).$$

5. Conclusiones

Para la comparación de poblaciones con distribuciones Poisson y bajo el supuesto de que la correspondiente región crítica cumpla la propiedad de monotonía, el teorema que probamos

en el presente trabajo garantiza que el tamaño de prueba se encuentra en la frontera del espacio nulo para una amplia variedad de funciones.

Este teorema generaliza el teorema principal de Liu y Hsueh (2013) a cualquier función frontera g no acotada, creciente y derivable tal que $\lim_{\mu \rightarrow 0^+} g(\mu) = 0$. Esto contribuye a la reducción de la carga computacional con la respectiva reducción de tiempo para una amplia variedad de funciones.

Bibliografía

- Gu, K., Ng, H., Tang, M., and Schucany, W. (2008). Testing the ratio of two poisson rates. *Biometrical Journal*, 50:283–298.
- Hogg, R., McKean, J., and Craig, A. (2005). *Introduction to Mathematical Statistics*. Pearson, New Jersey.
- Liu, M. and Hsueh, H. (2013). Exact tests of the superiority under the poisson distribution. *Statistics and Probability Letters*, 83:1339–1345.
- Miede, C. and Mueller-Cohrs, J. (2005). *Power calculation for non-inferiority trials comparing two Poisson distributions*. <http://www.lexjansen.com/phuse/2005/pk/pk01.pdf>.
- Ng, H. and Tang, M. (2008). Testing the equality of two poisson means using the rate ratio. *Statistics in Medicine*, 24:955–965.

Pruebas de Correlación Máxima, Correlación de Distancia y Covarianza para Optimización en el Problema de Selección de Variable.

Avance de Investigación

Yamil Burguete Fourzali^a, Gustavo Ramírez Valverde, David Sotres Ramos,
Benito Ramírez Valverde

Colegio de Postgraduados

El problema de selección de variable se refiere a la elección del mejor conjunto de predictores que expliquen a la variable respuesta. Varios grupos de investigación han propuesto métodos de selección de variable que operan con mayor o menor eficiencia dependiendo de las condiciones de las variables predictoras. La presente investigación pretende contrastar los métodos de correlación de distancia, correlación máxima, LASSO y LASSO adaptativo en diferentes condiciones de simulación. Se presentan de forma parcial los resultados dado que es una investigación en proceso.

Área-MSC: Estadística.

Subárea-MSC: Estimadores de encogimiento, Regresión Lineal.

1. Introducción

Con el desarrollo tecnológico y el crecimiento de los grandes grupos de datos, existen estudios que analizan o pretenden manejar miles o millones de variables. Lo cual complica la forma de establecer relaciones adecuadas para generar predicciones (Guyon y Elisseeff, 2003). Por ello, una de las metas de la selección de variable es elegir las variables predictoras X_1, \dots, X_p con mayor influencia a la variable respuesta Y (Miller, 2002).

La selección de variable pretende resolver problemas que también tienen relevancia en otras áreas de conocimiento. Como se observa en biología para la clasificación y predicción del

^aburguete.yamil@colpos.mx

comportamiento animal (Ducci et al., 2015); en química se ha utilizado para resolver el problema de señal-ruido en el manejo de datos quimométricos (Gerretzen et al., 2016), además de casos de información metabolómica que continuamente maneja sobreparametrización (van Reenen et al., 2016).

La pregunta natural es ¿Cuántas variables son las adecuadas? Sin embargo, la respuesta no es simple, este número está sujeto al balance entre sesgo y varianza (Miller, 2002). Como tal, no existe un método que funcione mejor en todas las condiciones, por lo que se han generado diversos métodos para hacer predicciones eficientes.

La presente investigación pretende comparar el comportamiento de cinco métodos de selección de variable en diferentes condiciones de simulación, para determinar cuál tiene mejor desempeño para las condiciones de los datos. Esta propuesta está basada en el trabajo de Yenigün y Rizzo (2015), aumentada con dos pruebas a la comparación además de realizar el análisis del error de predicción. Los métodos a comparar son: correlación máxima (MC), correlación de distancia (DC), LASSO, LASSO adaptativo y la prueba de covarianza (covTest).

2. Marco Teórico

Para una revisión complementaria del problema de selección de variable y algunos de sus puntos importantes, se recomienda revisar el trabajo de Miller (2002). Cada una de las herramientas elegidas será brevemente descrita en esta sección.

La MC se define como la medida de asociación con la siguiente representación:

$$S(X, Y) = \sup \rho(f(X), g(Y)) \quad (1)$$

Donde $\rho(U, V)$ es el coeficiente de correlación. La ventaja más importante de este estadístico es que cumple con todos los postulados de Rényi (1959) sobre la medida de la fuerza de asociación entre variables. Por lo que puede detectar relaciones aún cuando éstas no sean lineales. Una desventaja de este método es que no siempre existe un supremo, por lo que se puede utilizar la aproximación de Breiman y Freidman (1985) para tratar de obtener un máximo a través del algoritmo ACE.

La DC, al igual que la MC, es una medida de la dependencia entre variables se define como:

$$R_n^2(X, Y) = \frac{V_n^2(X, Y)}{\sqrt{V_n^2(X)V_n^2(Y)}} \quad (2)$$

Aunque la DC cumple parcialmente con los postulados de Rényi, tiene la propiedad de detectar dependencia no lineal entre las variables (Székely, Rizzo y Bakirov, 2007).

Por su parte, LASSO es un método que define una operación de reducción continua que puede producir coeficientes que son exactamente 0. Bajo ciertas condiciones posee propiedades oráculo (Tibshirani, 1996). El estimador LASSO queda definido como:

$$\hat{\beta}_{(LASSO)} = \arg \min_{\beta} \|y - \sum_{j=1}^p x_j \beta_j\|^2 + \lambda \sum_{j=1}^p |\beta_j| \quad (3)$$

El caso de LASSO adaptativo es una modificación de LASSO que tiene las propiedades oráculo para selección de modelo y consistencia en la estimación. Este método elige adaptativamente el grado de penalización que utiliza LASSO (Caner y Fan, 2010). Es importante destacar que bajo ciertas circunstancias, el LASSO presenta sesgo en la estimación y puede ser inconsistente en la selección del modelo (Zou, 2006). El estimador se define como:

$$\tilde{\beta}^{*(n)} = \arg \min_{\beta} \|y - \sum_{j=1}^p x_j \beta_j\|^2 + \lambda \sum_{j=1}^p \hat{w}_j |\beta_j| \quad (4)$$

Finalmente la prueba de covarianza (covTest) sirve para probar la significancia de las variables predictoras al entrar en el set activo. Asumiendo en la hipótesis nula que todas las variables de influencia ya se encuentran contenidas en el set previo de variables elegidas (Lockhart et al., 2014). Sea A el set previo a λ_k , cuando el predictor j entra a λ_k ,

$$\tilde{\beta}_A(\lambda_{k+1}) = \arg \min_{\beta_A} \frac{1}{2} \|y - X_A \beta_A\|_2^2 + \lambda_{k+1} \|\beta_A\|_1 \quad (5)$$

Donde se define el estadístico de prueba de covarianza como:

$$T_k = \frac{\langle y, X \hat{\beta}(\lambda_{k+1}) \rangle - \langle y, X_A \tilde{\beta}_A(\lambda_{k+1}) \rangle}{\sigma^2} \quad (6)$$

3. Método

3.1. Condiciones de Simulación

Las condiciones elegidas están basadas en el trabajo de Yenigün y Rizzo (2015). Para cada caso se realizaron 100 repeticiones con muestras tamaño $n = 100$. Se analizaron tres condiciones diferentes de simulación.

La primera condición es un modelo lineal: $Y = X\beta + \epsilon$, donde $X_1, \dots, X_p \sim N(0, 1)$, con $\epsilon \sim N(0, 2)$. El número de variables se fijó como $p = 8$ y $\beta = [1, 1, 1, 0, 0, 0, 0, 0]$.

La segunda condición es con colinealidad constante entre los predictores, con $p = 8$, de una distribución normal multivariada $X \sim N_p(0, \Sigma)$, donde:

$$\Sigma = \begin{bmatrix} 1 & \theta & \cdots & \theta \\ \theta & 1 & \cdots & \theta \\ \vdots & \vdots & \ddots & \vdots \\ \theta & \theta & \cdots & 1 \end{bmatrix}$$

Al igual que en la primera condición, las tres primeras variables son las que tienen influencia sobre la variable respuesta con un modelo $Y = X\beta + \epsilon$ con $\beta = [1, 1, 1, 0, 0, 0, 0, 0]$ y $\epsilon \sim N(0, 2)$. En este caso, se realizó en tres ocasiones esta simulación fijando el valor de θ a 0.6, 0.8 y 0.9.

La última condición es con colinealidad tipo Toeplitz entre los predictores. Con las mismas condiciones del caso anterior, salvo que el valor de $\theta = 0.6$ y la matriz de varianza y covarianza se define de la siguiente manera:

$$\Sigma = \begin{bmatrix} 1 & \theta & \theta^2 & \cdots & \theta^{p-1} \\ \theta & 1 & \theta & \cdots & \theta^{p-2} \\ \theta^2 & \theta & 1 & \cdots & \theta^{p-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \theta^{p-1} & \theta^{p-2} & \theta^{p-3} & \cdots & 1 \end{bmatrix}$$

3.2. Procedimiento

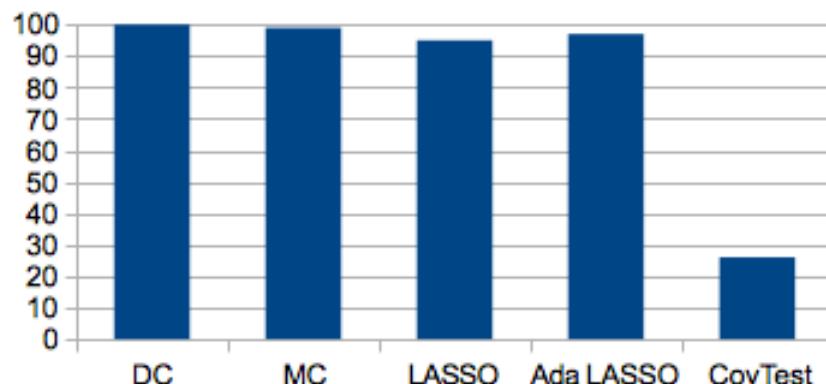
Para simplificar la explicación, se puntualizan los pasos que se siguieron:

- Se genera la simulación bajo cada una de las condiciones mencionadas.
- Cada procedimiento de selección de variable se aplica como una función. La salida de cada procedimiento se guarda en una matriz para su posterior análisis.
- Se analizaron los resultados de la siguiente manera:
 1. Se hace un conteo del número de veces que entró cada una de las variables relevantes al modelo estimado.
 2. La cantidad de veces que al menos las tres variables entraron al modelo.
 3. La cantidad de veces que entraron únicamente las variables relevantes al modelo estimado (llamado “hit rate”).
 4. Se calculó el sesgo y la varianza por cada uno de los β estimados.
 5. Se calculó el error de predicción.

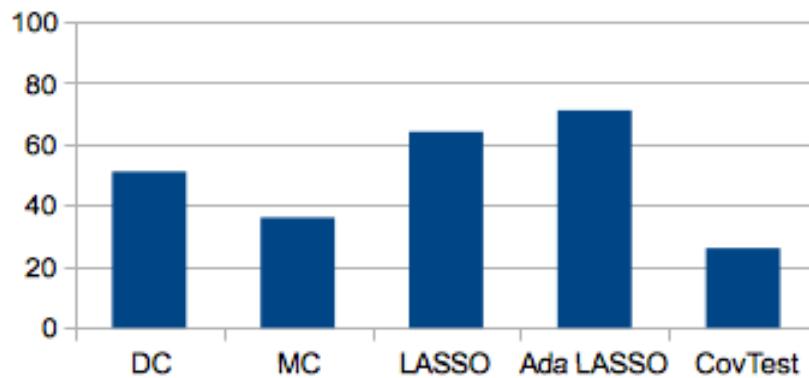
4. Resultados

4.1. Caso 1. Modelo Lineal Simple

A continuación se muestran las figuras de los condiciones de al menos las tres variables y de hit rate (sólo las tres variables).



Se puede observar cómo el comportamiento de todas las pruebas, excepto covTest, parecen ser muy buenas para “adivinar” el modelo. Sin embargo, al ver la figura del hit rate, es posible

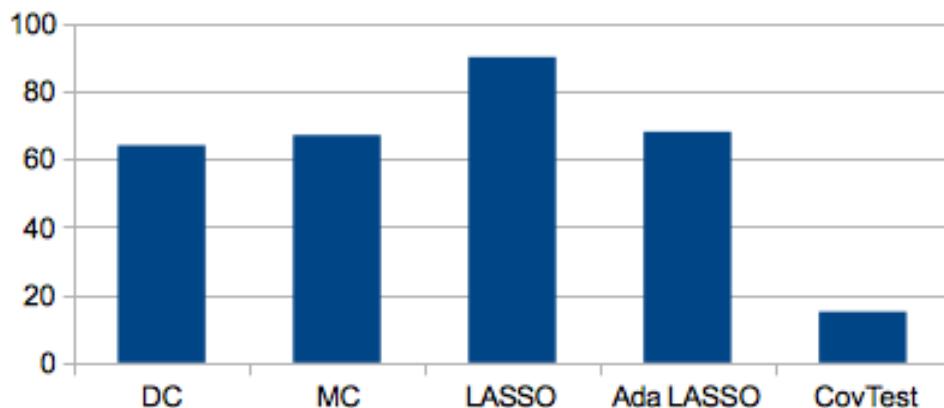


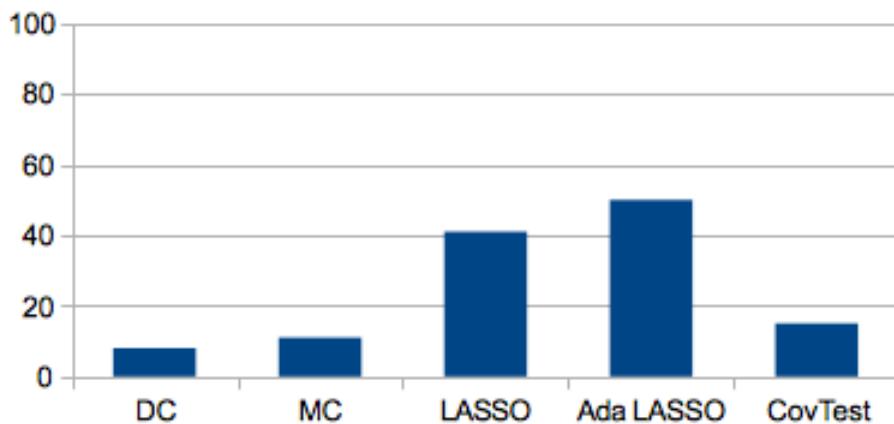
ver que las medidas de DC y MC no parecen tener tan buen rendimiento a comparación de LASSO y LASSO adaptativo. A continuación se muestra la tabla de resultados para los otros análisis.

Modelo Lin.	1 st Var	2 nd Var	3 rd Var	β_1 Var	β_1 Sesgo	β_2 Var	β_2 Sesgo	β_3 Var	β_3 Sesgo	Error Pred.
DC	100	100	100	0.03533	0.0016	0.0622	-0.0132	0.0388	0.0016	0.0018
MC	99	100	100	0.04392	-0.0078	0.0621	-0.0148	0.0388	0.0009	0.0017
LASSO	97	99	99	0.04972	-0.5181	0.0627	-0.4305	0.0463	-0.4650	0.6081
Ada LASSO	100	98	99	0.04498	-0.0875	0.0862	-0.1037	0.0501	-0.0720	0.0219
CovTest	35	46	42	0.07466	-0.8206	0.1438	-0.7080	0.1119	-0.7631	

4.2. Caso 2. Colinealidad Constante entre Predictores

Igual que en caso anterior, primero se muestran las dos figuras de las condiciones al menos y hit rate. Después de esto se muestra la tabla complementaria cuando $\theta = 0.6$.





En estos casos es mucho más notoria la capacidad de las pruebas LASSO y LASSO adaptativo para poder evaluar y tomar las variables correctas, sin la necesidad de agregar más variables al modelo estimado.

Col 6 = 0.6	1 st Var	2 nd Var	3 rd Var	β_1 Var	β_1 Sesgo	β_2 Var	β_2 Sesgo	β_3 Var	β_3 Sesgo	Error Pred.
DC	88	92	84	0.2380	0.0191	0.1852	0.0567	0.2259	-0.0784	8.9976
MC	90	95	82	0.2237	0.0174	0.1561	0.0620	0.2344	-0.1272	8.9259
LASSO	96	100	94	0.0939	-0.3440	0.0940	-0.2428	0.0962	-0.4081	7.6448
Ada LASSO	89	94	85	0.2142	-0.0886	0.1789	0.0160	0.2125	-0.1767	8.6885
CovTest	47	71	39	0.1817	-0.6613	0.1636	-0.5298	0.1340	-0.7446	

Se muestran a continuación únicamente las tablas de resultados para θ fijado en los valores de 0.8 y 0.9.

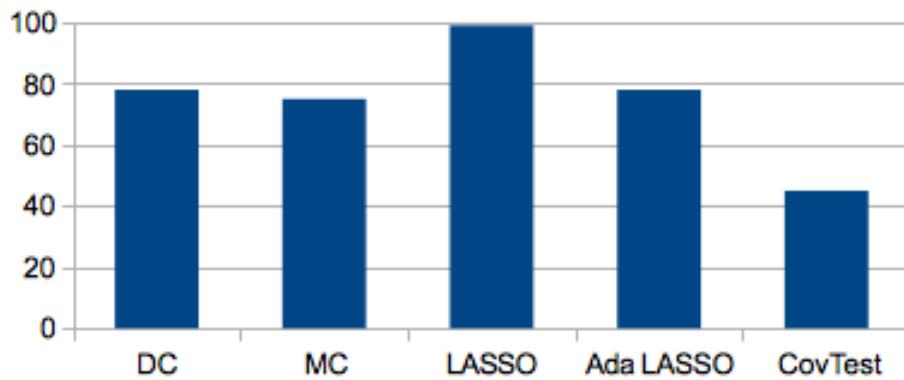
Col $\theta = 0.8$	1 st Var	2 nd Var	3 rd Var	Al menos	Hit Rate	β_1 Var	β_1 Sesgo	β_2 Var	β_2 Sesgo	β_3 Var	β_3 Sesgo	Error Pred.
DC	75	72	64	22	3	0.4578	0.0156	0.4824	-0.0674	0.6289	-0.0341	11.2436
MC	81	79	67	37	5	0.4060	0.0268	0.4330	-0.0738	0.6064	-0.0675	11.1651
LASSO	94	92	91	77	23	0.1569	-0.3172	0.1442	-0.4109	0.2078	-0.2866	9.3667
Ada LASSO	73	71	71	26	21	0.4696	-0.0799	0.4100	-0.1796	0.6316	-0.0123	10.8720
CovTest	48	37	45	1	1	0.1655	-0.6764	0.1572	-0.7456	0.1974	-0.6533	

Col $\theta = 0.9$	1 st Var	2 nd Var	3 rd Var	Al menos	Hit Rate	β_1 Var	β_1 Sesgo	β_2 Var	β_2 Sesgo	β_3 Var	β_3 Sesgo	Error Pred.
DC	65	64	61	12	0	0.7317	-0.0252	0.5559	-0.0736	0.6166	-0.0804	10.4196
MC	63	64	64	22	2	0.8360	-0.0349	0.5583	-0.1076	0.7091	-0.0723	10.4006
LASSO	91	87	83	62	18	0.2379	-0.3294	0.1646	-0.4057	0.2282	-0.3850	9.2402
Ada LASSO	61	62	60	21	13	0.7538	-0.1248	0.5964	-0.1484	0.7252	-0.1198	10.1785
CovTest	35	30	39	0	0	0.2070	-0.7397	0.1419	-0.8021	0.1825	-0.7447	

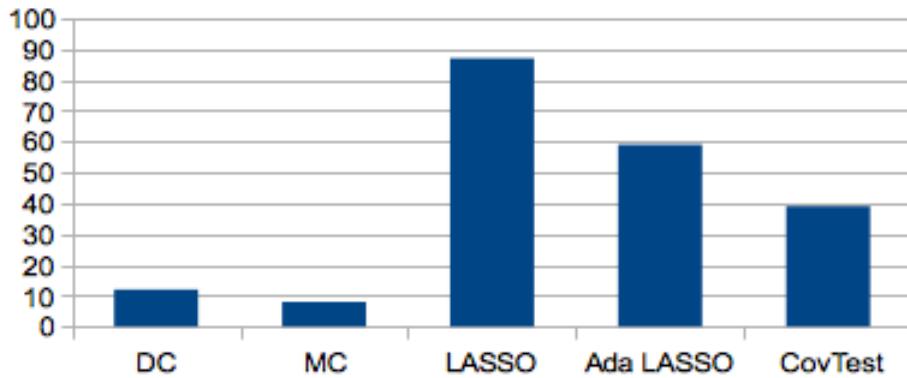
En ambos casos, se muestra un comportamiento similar al caso de $\theta = 0.6$, sólo con menor prevalencia de estimar el modelo correcto.

4.3. Caso 3. Colinealidad Tipo Toeplitz entre Predictores

Finalmente se muestran los resultados obtenidos de la tercera condición de simulación.



Nuevamente se puede observar cómo los métodos de DC y MC parecen comportarse mucho peor en cuanto a la elección del modelo correcto. Sin embargo, se puede observar que el error de predicción es menor entre esos dos métodos sobre el caso de LASSO.



Col. Toeplitz	1 st Var	2 nd Var	3 rd Var	β_1 Var	β_1 Sesgo	β_2 Var	β_2 Sesgo	β_3 Var	β_3 Sesgo	Error Pred.
DC	98	93	87	0.0721	0.0526	0.2110	0.0398	0.1859	-0.0928	0.0112
MC	97	91	87	0.0875	0.0528	0.2396	0.0301	0.1951	-0.0860	0.0083
LASSO	100	100	99	0.0462	-0.2522	0.1043	-0.0940	0.0833	-0.3760	0.4804
Ada LASSO	100	87	91	0.0725	0.0373	0.2610	-0.0354	0.1714	-0.1113	0.0150
CovTest	61	91	46	0.2138	-0.5047	0.1706	-0.2269	0.1993	-0.6261	

5. Conclusiones

Este es un proyecto en desarrollo por lo que las conclusiones todavía están siendo analizadas. Como se puede observar, parte de los resultados siguen incompletos, específicamente los valores de errores de predicción del covTest.

Los resultados muestran que DC y MC no eligen mejor que LASSO o LASSO adaptativo, sin embargo, en los errores de predicción tienen valores reducidos en el modelo lineal y en el caso de colinealidad tipo Toeplitz. Esto se debe a la sobreparametrización y al utilizar el mismo set de entrenamiento como el set de prueba (Miller, 2002).

Algunos de los resultados esperados son el mejor comportamiento de las pruebas de LASSO y LASSO adaptativo para selección del modelo, además de un mejor desempeño en la reducción del error de predicción. Esto contrasta con los resultados presentados por Yenigün y Rizzo (2015), donde presentaron ventajas al utilizar DC y MC.

Inicialmente se esperaba que la prueba de covarianza tuviera resultados aún mejores que LASSO y LASSO adaptativo. Sin embargo, los resultados parciales muestran lo contrario. Aún así, no se descarta la importancia de la prueba. Lockhart et al. (2014) mencionan que todavía se deben realizar ajustes a la misma, como la asignación del valor de α .

Bibliografía

- Breiman, L. and Friedman, J. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, 80:580–598.
- Caner, M. and Fan, Q. (2010). The adaptative lasso method for instrumental variable selection. Technical report, North Carolina State University. Working Paper. URL: <http://apps.olin.wustl.edu/MEGConference/Files/pdf/2010/1.pdf>.
- Ducci, L., Agnelli, P., Di Febbraro, M., Frate, L., Russo, D., Loy, A., Carranza, M. L., Santini, G., and Roscioni, F. (2015). Different bat guilds perceive their habitat in different ways: a multiscale landscape approach for variable selection in species distribution modelling. *Landscape Ecology*, 30(10):2147–2159.
- Gerretzen, J., Szymańska, E., Bart, J., Davies, A. N., van Manen, H., van den Heuvel, E. R., Jansen, J. J., and Buydens, L. M. (2016). Boosting model performance and interpretation by entangling preprocessing selection and variable selection. *Analytica Chimica Acta*, 938:44–52.
- Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182.
- Lockhart, R., Taylor, J., Tibshirani, R., and Tibshirani, R. (2014). A significance test for the lasso. *The Annals of Statistics*, 42:413–468.
- Miller, A. (2002). *Subset selection in regression*. CRC Press.
- Rényi, A. (1959). On measures of dependence. *Acta Mathematica Hungarica*, 10:441–451.
- Székely, G., Rizzo, M., and Bakirov, N. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35:2769–2794.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58:267–288.
- van Reenen, M., Reinecke, C. J., Westerhuis, J. A., and Venter, J. H. (2016). Variable selection for binary classification using error rate p-values applied to metabolomics data. *BMC Bioinformatics*, 17(1).

- Yenigün, C. and Rizzo, M. (2015). Variable selection in regression using maximal correlation and distance correlation. *Journal of Statistical Computation and Simulation*, 85:1692–1705.
- Zou, H. (2006). The adaptative lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429.

La Teoría Estable Acotada. Una Alternativa para Predecir el Estado Estable del Saldo Neto Migratorio en México

Javier González Rosas^a, Iliana Zárate Gutiérrez^b
CONAPO, México

De acuerdo con información difundida por la Secretaría de Gobernación en 2011, y el Pew Hispanic Center en 2012, así como por datos de la Encuesta sobre Migración en la Frontera Norte de México, en el periodo 2005-2010 la salida de mexicanos se compensó con el regreso de connacionales de Estados Unidos y de inmigrantes estadounidenses. Es decir, el Saldo Neto Migratorio (SNM) México-Estados Unidos fue aproximadamente cero en este periodo. Este hecho fue avalado además por el Census Bureau de Estados Unidos y expertos demógrafos nacionales e internacionales. En este contexto, el artículo tiene dos objetivos, primero, probar que la media del SNM en México tiende a estabilizarse en el valor cero, y en segundo lugar, elaborar proyecciones del SNM para el periodo 2011-2020. La metodología utilizada se basa en aplicar la *Teoría Estable Acotada* a datos del periodo 2004-2010 calculados con base en estimaciones de expertos demógrafos nacionales y la Sociedad Mexicana de Demografía. Los principales resultados del artículo, indican que la media del SNM en México actualmente no es cero pero tiende a estabilizarse en este valor. También muestran que la pérdida de población por migración en México continuará a la baja, lo que implicará que en 2016 saldrán del país en promedio 2,323 personas, para 2018 saldrán 924 y para 2020 serán 368 personas, es decir, la media del SNM estará cada vez más y más cerca del cero absoluto.

Área-MSN: Estadística aplicada

Subárea-MSN: Modelos no lineales, regresión lineal

^axavier.gonzalez@conapo.gob.mx

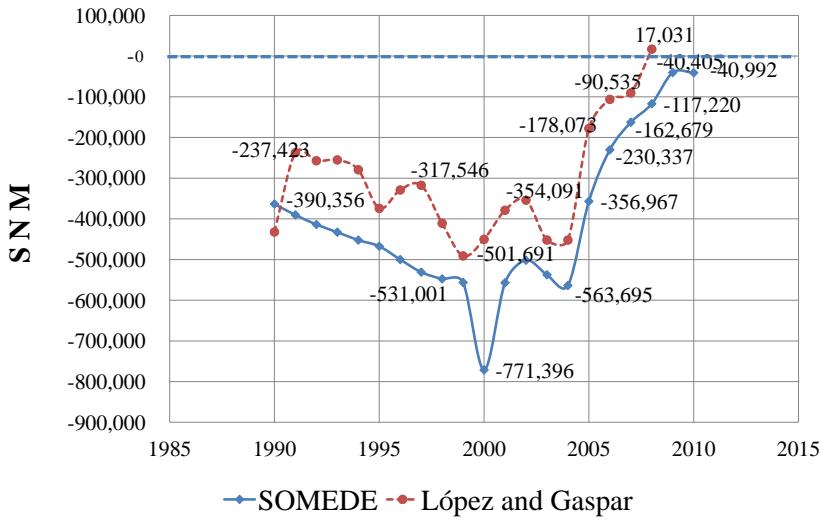
^bizarate@conapo.gob.mx

1. Introducción

En 2011, López y Gaspar y la Sociedad Mexicana de Demografía (SOMEDE), usando fuentes de información mexicanas y estadounidenses, realizaron ejercicios de estimación de la migración internacional, que sirvieron como base para la conciliación censal de ese año. Las fuentes mexicanas que usaron López y Gaspar fueron los censos de población de 1990, 2000 y 2010, en tanto que las fuentes estadounidenses fueron la muestra del 5 % del Censo de Población de 2000 y la American Community Survey (ACS) del mismo año. Por su parte, entre las fuentes mexicanas que utilizó la SOMEDE para su ejercicio están los censos de población de 1970, 1980, 1990 y 2010, y las fuentes estadounidenses que utilizó fueron las ACS de 2000 a 2009 (CONAPO, 2015, p. 24). López y Gaspar estimaron el *Saldo Neto Migratorio*¹ (SNM) para el periodo 1990-2008, en tanto que la SOMEDE lo hizo para el periodo 1990-2010.

Los datos de estos dos ejercicios indican que en los últimos años (1990-2010) el SNM en México tuvo cambios muy importantes. En el periodo de 1990 a 2004, ambas series de tiempo muestran que la tendencia de la pérdida de población por migración en el país iba en aumento. Pero a partir de 2004 la tendencia cambió de manera radical. Según ambas series, la pérdida de población empezó a descender e incluso López y Gaspar terminan estimando para 2008 un SNM positivo de poco más de 17 mil personas. Es decir, en ese año por primera vez en mucho tiempo la población mexicana se incrementó debido a la migración. La SOMEDE por su parte termina estimando para 2009 una pérdida de población de 40,405 y para 2010 de 40,992, es decir, el SNM prácticamente se mantuvo sin cambio, algo que tampoco había sucedido en muchos años. Estos resultados sugieren la hipótesis de que el SNM en México tiende a estabilizarse en el valor cero. El objetivo de este artículo es probar esta hipótesis y además elaborar proyecciones del SNM de México para el periodo 2011-2020, dado que en 2020 el censo de población de ese año proporcionará nueva información que permitirá actualizar las tendencias del SNM.

¹El Saldo Neto Migratorio se define como la entrada menos la salida de personas a un país. Cuando es negativo se dice que el país pierde población por migración internacional y cuando es positivo se dice que gana población.



Fuente: CONAPO (2015)

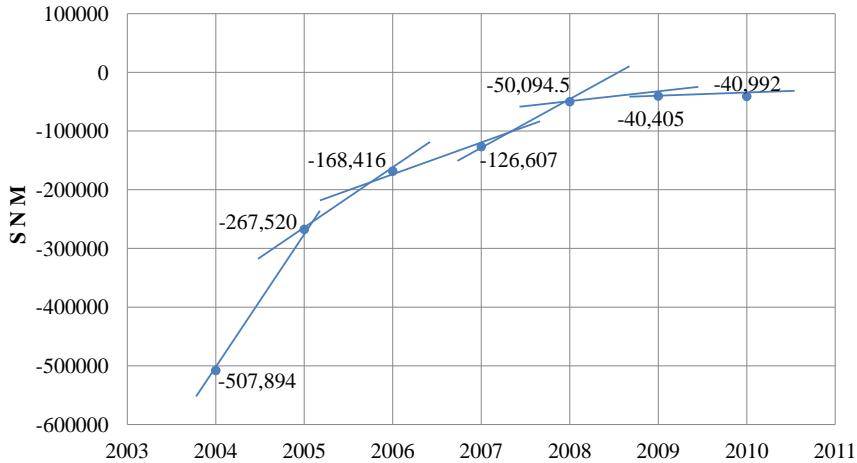
Figura 1: Saldo neto migratorio en México según fuente, 1990-2010.

2. Marco Teórico

2.1. La Cantidad de Cambio del SNM

En la Figura 1, se pueden observar las estimaciones de López y Gaspar y de la SOMEDE. Como se puede constatar, los volúmenes estimados son muy diferentes, sin embargo, en general muestran un comportamiento muy similar. De 1990 a 2004 ambas series muestran una tendencia decreciente, pero a partir de 2004 la tendencia cambió a la alza. Con el fin de tener una sola serie de tiempo, los datos usados en este artículo fueron el promedio de ambas series (véase Figura 2).

En la Figura 2, se puede observar la evolución del promedio del SNM en México a través del tiempo y líneas rectas que unen los puntos. Obsérvese que la tendencia es creciente y además que conforme los datos del SNM avanzan en el tiempo las rectas se inclinan cada vez más y más, de tal manera que, bajo el supuesto de que los datos del SNM tienden a estabilizarse, entonces las rectas tenderán a ser paralelas al eje del tiempo y por lo tanto su inclinación tenderá al valor cero. Para probar esta hipótesis se usó la *Teoría Estable Acotada* (González-Rosas, 2012). De acuerdo con esta teoría se calcularon las pendientes entre los



Fuente: cálculos propios

Figura 2: Saldo neto migratorio en México, 2004-2010.

puntos (t_i, y_i) y (t_{i+1}, y_{i+1}) y los valores medios² entre los datos y_i y y_{i+1} de la siguiente manera:

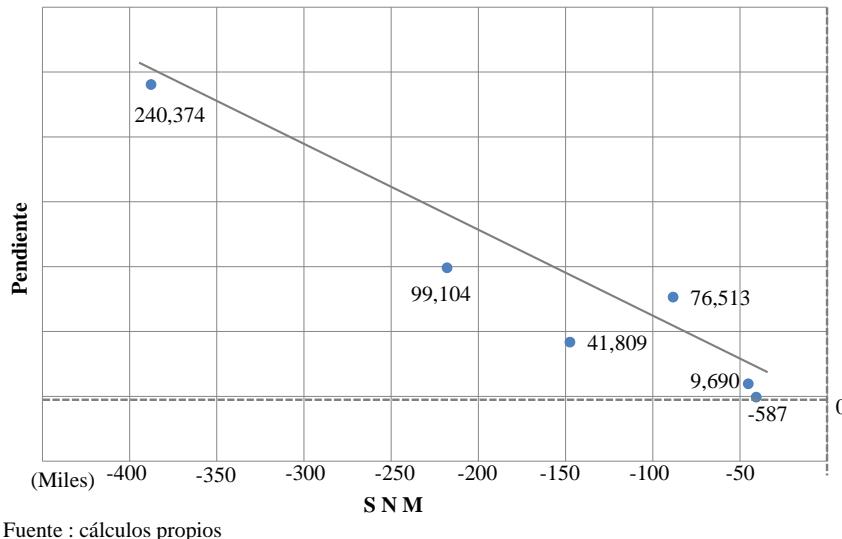
$$\Delta_i = \frac{y_{i+1} - y_i}{t_{i+1} - t_i}$$

$$VM_i = y_i + \frac{y_{i+1} - y_i}{2}.$$

Donde: Δ_i denota la pendiente entre los puntos (y_i, t_i) y (y_{i+1}, t_{i+1}) del espacio bidimensional definido por el SNM y el tiempo (Leithold, 1973, p. 137), y VM_i denota el valor medio entre y_i y y_{i+1} .

En la Figura 3, se graficaron en el eje X los valores medios y en el eje Y los valores de las pendientes. Como se puede ver, las variables están inversamente relacionadas, de tal manera que cuando el saldo neto migratorio se acerca a cero por la izquierda las pendientes se acercan también a cero. Estos resultados prueban las predicciones de la *Teoría Estable*

²La *Teoría Estable Acotada* demuestra que existen tres estimadores del valor de la estabilidad. Uno asociado con el valor y_i , otro con el valor y_{i+1} y otro más con el valor medio entre los dos. La teoría también demuestra que el estimador usando el valor medio es el mejor, ya que cuando se utiliza el punto y_i el valor de la estabilidad se subestima y cuando se utiliza el valor y_{i+1} se sobreestima.



Fuente : cálculos propios

Figura 3: Pendientes y valores medios del SNM en México, 2004-2010.

Acotada. De acuerdo con la Figura 3, la relación entre la cantidad de cambio medida por la pendiente y el SNM está dada por la siguiente ecuación:

$$\Delta_i = \alpha + \beta S_i + \theta_i. \quad (1)$$

Donde: Δ_i denota el valor i de la pendiente, α y β son constantes desconocidas, S_i es el valor i del SNM, y las θ_i son variables aleatorias que se suponen independientes con distribución normal, media $\mu_\theta = 0$ y varianza σ_θ^2 constante.

Desde el enfoque geométrico, el valor de la estabilidad es el punto donde la recta de la Figura 3 intersecta el eje X (SNM). En la figura se observa que la recta parece que se intersecta en el valor cero. Desde el punto de vista matemático el valor de la estabilidad es el número donde la cantidad de cambio es cero. Según la *Teoría Estable Acotada* el valor numérico de la estabilidad denotado como K corresponde al valor del SNM en el cual la pendiente Δ_i de la ecuación (1) es cero, por lo tanto, para calcular K primero se iguala a cero la parte determinística de (1),

$$0 = \alpha + \beta S_i$$

y después se despeja S_i quedando que:

$$K = \frac{-\alpha}{\beta}. \quad (2)$$

Por lo que, al estimar α y β de la ecuación (1) se obtiene que la fórmula (2) es un estimador del valor de la estabilidad. La *Teoría Estable Acotada* prueba que si α y β se estiman por mínimos cuadrados ordinarios o por mínimos cuadrados generalizados, entonces K calculado como indica la fórmula (2) es un estimador insesgado y consistente.

Para calcular el valor de la estabilidad del SNM en México, se ajustó un modelo de regresión lineal simple³ a los datos de la Figura 3. La estimación de los parámetros se hizo por el método de mínimos cuadrados ordinarios. Los resultados fueron $\alpha = -20,797.77$ y $\beta = -0.6379$ y los p-valores para probar la significancia de los parámetros fueron para α , $p = 0.333$ y para β , $p = 0.003$, lo que quiere decir que β es significativamente diferente de cero pero α no, por lo que, se concluye que $\alpha = 0$ (Montgomery y Peck, 1982, p. 21). Esto demuestra matemáticamente lo que la Figura 3 sugiere, es decir que, el valor de la estabilidad del SNM en México es cero, ya que de acuerdo con 2,

$$K = \frac{0}{-0.6379} = 0.$$

2.2. La Ecuación del SNM y el Tiempo

Con base en la tendencia de los datos observados del SNM a través del tiempo de la Figura 2 y bajo el supuesto de que existe un valor K en donde se estabilizará el SNM, la *Teoría Estable Acotada* prueba que una ecuación que explica la evolución del SNM a través del tiempo es:

$$S_t = K - \gamma e^{r t} + \epsilon_t \quad ; \quad r < 0. \quad (3)$$

Donde: S_t denota la variable aleatoria del SNM en el tiempo t , t es la variable tiempo, K , $-\gamma$ y r son parámetros desconocidos, y las ϵ_t son variables aleatorias que se suponen independientes, con distribución probabilística normal, media $\mu_\epsilon=0$ y varianza σ_ϵ^2 constante.

³La validación del modelo indica que el p-valor de la prueba F es 0.0003, que el coeficiente de determinación es 94.2 por ciento y que los errores θ_i tienen distribución normal, son independientes y con varianza constante, por lo que no existe evidencia estadísticamente significativa para rechazar los resultados del modelo.

Obsérvese que en la ecuación (3), cuando t tiende a infinito, dado que r es negativa, la expresión e^{rt} tiende a cero y por lo tanto S_t tiende a:

$$S_t = K + \epsilon_t.$$

Y por el supuesto de que la media o valor esperado de las ϵ_t es cero se tiene entonces que:

$$E(S_t) = K.$$

Lo que indica que K es una cota para la media o valor esperado del SNM pero no para las observaciones, las cuales, al distribuirse alrededor de la media podrán estar por arriba o por debajo de K , y su ocurrencia estará gobernada por una ley probabilística.

Nótese que el cumplimiento de los supuestos de las variables ϵ_t de la ecuación (3), resulta absolutamente necesario, por un lado, para tener los mejores estimadores de los parámetros K , $-\gamma$ y r , y por otro, para garantizar la convergencia del SNM al valor K , por lo que, en cualquier aplicación es muy importante comprobar el cumplimiento de dichos supuestos.

El parámetro K se conoce también como la *cota superior* del SNM, porque su media o valor esperado no podrá en cualquier momento del tiempo estar por arriba de este valor (Leithold, 1973, p. 663, Lehmann, 2000, p. 41). Los parámetros $-\gamma$ y r son *los parámetros de la rapidez*, porque determinan qué tan rápido la parte determinística de la ecuación (3) se acercará a la estabilidad. El parámetro $-\gamma$ junto con K representan las condiciones iniciales de la media del SNM en el tiempo cero, en tanto que r determina la cantidad de reducción de la pérdida de población por unidad de tiempo.

2.3. Estimación de los Parámetros de la Rapidez

Según Draper y Smith (1966) la parte determinística de la ecuación (3) es no lineal en los parámetros K , $-\gamma$ y r , por lo que, no pueden estimarse por mínimos cuadrados. Sin embargo, si en la parte determinística, K se pasa del lado izquierdo de la igualdad y si se multiplica por -1 y se aplica logaritmo natural se obtiene que,

$$\ln(K - S_t) = \ln\gamma + rt. \quad (4)$$

Es decir, el resultado es una ecuación lineal en los parámetros $\ln\gamma$ y r los cuales pueden estimarse por el método de mínimos cuadrados ordinarios o mínimos cuadrados generalizados. Esto sugiere que la estimación de los parámetros de la parte determinística de (3) puede

hacerse en dos etapas. La *Teoría Estable Acotada*, prueba que existe un procedimiento en el que primero se estima K y después tomando en cuenta la estimación de K se estiman los parámetros $-\gamma$ y r . La teoría prueba además que los estimadores de la segunda etapa son insesgados y consistentes. A la variable $\ln(K - S_t)$ se le conoce como la transformada del SNM (González-Rosas, 2010, p. 74).

Ahora bien, dado que para el caso de México $K = 0$ entonces al sustituir en la ecuación (4) y considerar un error aleatorio se tiene que:

$$\ln(-S_t) = (\ln\gamma + rt) + \xi_t. \quad (5)$$

Donde: $\ln(-S_t)$ denota la transformada del SNM en el tiempo t , $\ln\gamma$ y r son parámetros desconocidos, y las ξ_t son variables aleatorias que se suponen independientes, con distribución probabilística normal, media $\mu_\xi=0$ y varianza σ_ξ^2 constante.

En la Figura 4, se puede comprobar que la relación entre la transformada del SNM $\ln(-S_t)$ y el tiempo t en México, está dada efectivamente por una recta como lo predice la teoría y por lo tanto es lineal en los parámetros $\ln\gamma$ y r . Para estimarlos se ajustó un modelo de regresión lineal simple a los datos de la Figura 4.

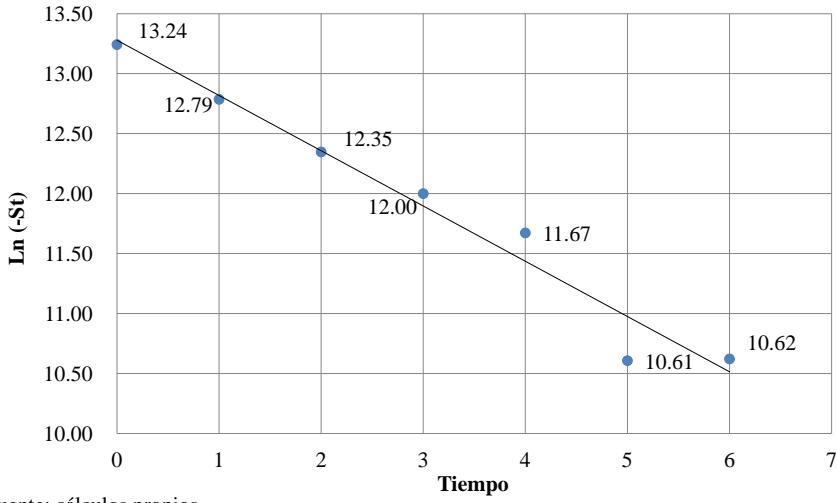
La estimación de mínimos cuadrados ordinarios de $\ln\gamma$ fue 13.279 y de r fue -0.4607, y sus p-valores para probar su significancia fueron $p=0.001$ para ambos, por lo que ambos parámetros son significativamente diferentes de cero⁴. Finalmente, para obtener la estimación de $-\gamma$ se aplicó primero la función exponencial a $\ln\gamma = 13.279$ y después se multiplicó por -1, obteniéndose que $-\gamma = -584,785.27$.

2.4. Proyección del SNM en México, 2011-2020

Considerando los resultados del proceso de estimación de los parámetros del componente determinístico, así como la validación de los supuestos, la ecuación del comportamiento del SNM a través del tiempo en México quedó como,

$$S_t = -584,785.27e^{-0.4607 t} + \epsilon_t. \quad (6)$$

⁴La validación del modelo indica que el p-valor de la prueba F es 0.0001, que el coeficiente de determinación es 96.55 por ciento y que los errores ξ_t tienen distribución normal, son independientes y con varianza constante, por lo que no existe evidencia estadísticamente significativa para rechazar los resultados del modelo.



Fuente: cálculos propios

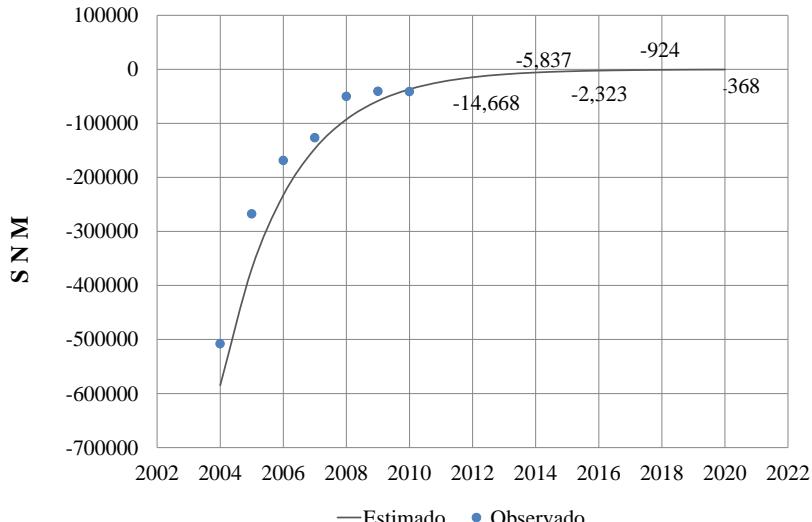
Figura 4: Transformada del SNM y el tiempo en México, 2004-2010.

Donde: S_t es la variable del SNM en el tiempo t , t es la variable tiempo, la constante -584,785.27 es la media del SNM en el tiempo cero, -0.4607 es la constante de reducción de la pérdida de población, y las ϵ_t son variables aleatorias independientes, con distribución normal, media $\mu_\epsilon = 0$ y varianza σ_ϵ^2 constante.

Asignando valores a la variable tiempo en la ecuación (6), se obtuvieron pronósticos puntuales de la media del SNM en México para el periodo 2011-2020. En la Figura 5 se observa que el modelo se ajusta adecuadamente a los datos observados y que conforme el tiempo crece el SNM se acerca cada vez más a cero. De acuerdo con los resultados del modelo, se tiene que en 2012 México perdió por migración en promedio 14,608 habitantes, para 2014 fueron 5,837, para 2016 se espera que la pérdida de población sea de 2,323 personas, en 2018 se reducirá a 924 y para 2020 se espera que se reduzca a 368 habitantes.

3. Conclusiones

En México, en el periodo 2004-2020, el comportamiento del SNM en términos del tiempo está gobernado por una ecuación matemática que tiene dos componentes, uno predecible y otro impredecible.



Fuente: cálculos propios

Figura 5: SNM observado y proyección puntual en México, 2004-2020.

La parte predecible indica que el valor esperado del SNM se podrá pronosticar en cualquier momento del tiempo y que está determinado por tres parámetros, la cota $K = 0$, el valor inicial en el tiempo cero $-\gamma = -585$ mil, y la constante de reducción de la pérdida de población $r = -0.4607$.

La parte impredecible indica que el valor observado del SNM en el tiempo no se puede predecir, y que está gobernado en cada momento del tiempo por una ley probabilística que se distribuye como una variable aleatoria normal, con media cero, varianza constante a través del tiempo, y entre un año y otro las variables aleatorias son independientes.

De acuerdo con la parte predecible, se espera que en 2016 la pérdida promedio de población en México por migración sea de 2,323 personas, para 2018 será de 924 y para 2020 la pérdida será de 368 personas que saldrán del país.

Es necesario advertir que, los resultados de este artículo descansan en el supuesto de que las condiciones políticas, económicas y sociales que afectan la migración entre México y Estados Unidos principalmente, continuarán sin cambio. Si este supuesto no es cierto, las predicciones no se cumplirán.

También es necesario advertir que, la modelación matemática de la realidad se basa en supuestos, y que los resultados teóricos que se obtienen se basan en el cumplimiento de éstos, por lo que, al aplicar modelos es necesario probar el cumplimiento de los supuestos y tener en cuenta que si éstos no son ciertos, las conclusiones que se hagan con base en los modelos estarán erróneas.

Por último, todo ejercicio de predicción del futuro está expuesto a varias fuentes de error: datos erróneos, hipótesis inapropiadas, supuestos que no son ciertos, modelos equivocados, etc., por lo que, es necesario identificar todas las posibles fuentes de error, y en consecuencia utilizar metodologías que minimicen los errores, la *Teoría Estable Acotada* es un ejemplo de ello.

Bibliografía

- CONAPO (2015). *Modelación de fenómenos demográficos: Análisis de las fuentes de información sobre mortalidad, fecundidad y migración internacional en México para determinar las bases empíricas de la Conciliación y Proyecciones de Población*. CONAPO, México.
- Draper, J. and Smith, W. (1966). *Applied Regression Analysis*. Wiley, New York.
- González-Rosas, J. (2010). Teoría estadística y probabilística de los fenómenos estable acotados. Master's thesis, Universidad Nacional Autónoma de México. Tesis de maestría.
- González-Rosas, J. (2012). *La Teoría Estable Acotada: Fundamentos, conceptos y métodos, para proyectar los fenómenos que no pueden crecer o decrecer indefinidamente*. Académica Española, Saarbrucken, Alemania.
- Lehmann, C. (2000). *Geometría Analítica*. Limusa, México.
- Leithold, L. (1973). *El Cálculo: Con geometría analítica*. Harla, México.
- Montgomery, D. and Peck, E. (1982). *Introduction to Linear Regression Analysis*. Wiley, New York.

Inferencia sobre Modelos Epidemiológicos en Redes de Contactos

Rocío M. Ávila Ayala^a, J. Andrés Christen Gracia, L. Leticia Ramírez
Ramírez

Centro de Investigación en Matemáticas

Este trabajo propone un método de inferencia bayesiana sobre modelos epidemiológicos, en particular en el SEIR estocástico disperso sobre una red social que modela las relaciones entre los individuos de una población. Se plantea un escenario donde se cuenta con reportes agregados de nuevos infectados en vez de la información completa. Se hace uso de técnicas como *Approximate Bayesian Computation* (ABC) para aproximar la densidad posterior y *Markov Chain Monte Carlo* (MCMC) para simularla.

Clasificación: Tesis de Maestría.

Área-MSC: Análisis Estadístico de Datos (62-06)

Subárea-MSC: Inferencia Bayesiana (62F15), Cómputo Estadístico (65C60) Métodos Monte Carlo (65C05), Epidemiología (92D30), Redes Sociales (92D30).

1. Introducción

Un paso importante en el método científico es la experimentación; sin embargo, en el contexto de epidemiología no es concebible (ni sería ético) experimentar sobre una población para ver cómo se dispersa un agente infeccioso dentro de la misma. Esto hace que en esta área sea sumamente importante el planteamiento de modelos matemáticos que intenten describir la evolución de una enfermedad y que al mismo tiempo puedan incorporar resultados de laboratorio o estimaciones originadas de brotes anteriores.

Una gran parte de los modelos epidemiológicos considera que un individuo puede transitar por diversos estatus al ser infectado durante un brote. Esta clasificación individual lleva naturalmente a dividir a la población de estudio en grupos o categorías disjuntas de acuerdo

^arocio.avila@cimat.mx

a su estado respecto a la enfermedad (susceptibles e infectados, por ejemplo). Este tipo de modelos epidémicos se denominan modelos compartimentales.

En los modelos epidemiológicos compartimentales existe un parámetro umbral que determina si el número de infectados decrece rápidamente hasta desaparecer, o la enfermedad se propaga en una gran parte de la población y se presenta un brote. Este parámetro umbral se conoce como R_0 y es una función de las tasas de transferencia de los individuos entre los compartimentos.

Dado un conjunto de observaciones, por ejemplo pensemos en reportes de nuevos infectados cada cierto intervalo de tiempo, y un modelo epidemiológico que describa adecuadamente los datos observados, es relevante poder realizar inferencia sobre los parámetros que rigen el modelo para los datos, ya que esto permitiría un mejor entendimiento del comportamiento de la epidemia, y además el establecimiento de políticas públicas que ayuden a controlar el brote que permitan evitar un impacto mayor en la salud de la población y en la economía.

El objetivo principal de este trabajo es presentar un proceso de inferencia en un modelo epidemiológico simple llamado SIR (Susceptibles - Infectados - Removidos), el cual se planteará bajo el enfoque estocástico. Se supondrá que los datos observados son el número de nuevos infectados reportados en un intervalo de tiempo y se trabajará con modelos de tipo bayesiano. El agente infeccioso se dispersará en una red que representa la estructura y las interacciones entre los individuos de una población.

La implementación computacional del método de inferencia se programó usando el software estadístico R R Development Core Team (2008).

2. Marco Teórico

En esta sección se presentan las bases teóricas necesarias para la implementación de la metodología de inferencia propuesta.

2.1. Modelo SIR Estocástico

Uno de los modelos epidemiológicos más sencillos es el SIR. Éste describe la dinámica de enfermedades en que los individuos susceptibles son infectados, pero posteriormente desarrollan una inmunidad a la enfermedad o mueren (y son removidos del sistema). En la práctica

ha sido utilizado para modelar enfermedades comunes en la niñez como sarampión, varicela y paperas, que son originadas por virus y de los cuales se suele desarrollar inmunidad.

El modelo SIR consta entonces de tres compartimentos: susceptibles (S), infectados (I) y removidos (R), y considera que el agente infeccioso se transmite por contacto entre susceptibles e infectados. La dinámica de la enfermedad sigue el esquema mostrado en la Figura 1.

Al tratarse de una población cerrada, los parámetros involucrados en el modelo son la tasa de contagio $\beta > 0$ y la tasa de recuperación o remoción $\gamma > 0$.

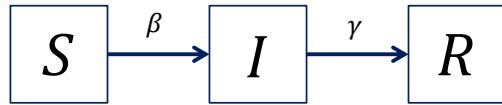


Figura 1: Dinámica del modelo SIR.

Existen condiciones adicionales, como efectos demográficos, efectos de vacunación, etc. que pueden ser sumamente relevantes en el proceso infeccioso. Aunque varios de estos cambios pueden introducirse a su vez como nuevas categorías de los individuos en el modelo compartmental, en este trabajo suponemos que la evolución del brote epidémico es rápida (semanas) y que cambios demográficos pueden ser omitidos. En este sentido asumimos que la población es cerrada. Esto es, libre de nacimientos, muerte natural de los individuos de la población y migración, de manera que los parámetros del modelo serán únicamente aquellos que determinan el tiempo que los individuos pasan en cada compartimento.

Para evitar soluciones triviales es necesario establecer un conjunto de condiciones iniciales para el sistema dinámico al tiempo t_0 . Denotemos $S(t)$, $I(t)$ y $R(t)$ al número de individuos susceptibles, infectados y removidos al tiempo t , respectivamente. Típicamente se toman como condiciones iniciales $t_0 = 0$ y $S(t_0) > 0$, $I(t_0) > 0$ para que pueda haber contagio. Además, como se supone que el brote inicia en t_0 , suele tomarse $R(t_0) = 0$.

El modelo SIR estocástico puede plantearse bajo distintos enfoques. Allen (2008) presenta tres enfoques distintos, los cuales se distinguen de acuerdo a los supuestos acerca del tiempo y el espacio de estados. El que será utilizado en este trabajo considera el modelo SIR definido sobre una escala de tiempo continua $t \in [0, \infty)$, donde los estados son variables aleatorias

discretas. Por la forma de las transiciones entre los posibles compartimentos, este modelo puede verse como un proceso de nacimiento y muerte, donde las tasas asociadas dependen directamente de las tasas de flujo entre compartimentos. La aleatoriedad del modelo proviene de asignar a los tiempos que los individuos pasan en cada compartimento una distribución de probabilidad. Si ésta se considera exponencial, el proceso puede verse como una Cadena de Markov a tiempo continuo.

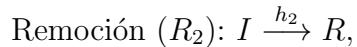
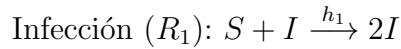
Denotemos por $N = N(t)$ al tamaño de la población, constante para cualquier tiempo $t > 0$ por ser una población cerrada. Obsérvese que el sistema puede ser monitoreado con los estados $S(t)$ e $I(t)$, ya que $R(t)$ puede obtenerse directamente a partir de estos últimos: $R(t) = N - S(t) - I(t)$.

Así, el proceso bivariado a monitorear es $\{S(t), I(t)\}$, el cual tiene asociado una función de probabilidad conjunta dada por $p_{(s,i)}(t) = \mathbb{P}[\{S(t), I(t)\} = (s, i)]$, donde (s, i) es un vector en el espacio de estados posibles del proceso $\{S(t), I(t)\}$.

Del proceso bivariado anterior se deduce el sistema de ecuaciones diferenciales *forward* de Kolmogorov o *Chemical Master Equation*:

$$\begin{aligned} \frac{dp_{(s,i)}(t)}{dt} = & \frac{\beta}{N}(s+1)(i-1)p_{(s+1,i-1)} + \gamma(i+1)p_{(s,i+1)} \\ & - \left[\frac{\beta}{N}si + \gamma i \right] p_{(s,i)}. \end{aligned} \quad (1)$$

Usando la notación de reacciones químicas planteada en Boys *et al.* (2008), la evolución del sistema que describe la ecuación anterior consta de tres especies $(S(t), I(t), R(t))$ y dos reacciones posibles que se dan cuando hay una nueva infección o una remoción, y puede representarse de la siguiente forma:



donde los coeficientes que multiplican a las especies del lado izquierdo de la ecuación corresponden a la cantidad de cada especie que entra en la reacción, y los coeficientes del lado derecho son la cantidad de la misma especie que se obtiene como producto de la reacción. Así, en una infección entran un susceptible y un infeccioso, y al contagiarse el individuo susceptible, el producto de la reacción son dos infecciosos.

Denotemos al sistema total como $\mathbf{X}(t) = (S(t), I(t), R(t))$. Para un intervalo de tiempo Δt suficientemente pequeño, ocurre una y sólo una de las reacciones anteriores. Cada reacción R_k , $k = 1, 2$ tiene asociada una tasa $h_k(\mathbf{X}(t), \beta, \gamma)$ que depende del estado del sistema al tiempo t y de las tasas de infección y remoción. En el modelo estocástico subyacente, los tiempos en que sucede la k -ésima reacción, para $k = 1, 2$ se distribuye exponencialmente con tasa h_k , y en consecuencia el tiempo de la primera reacción (el mínimo entre los tiempos de la reacción 1 y la 2), se distribuye exponencial con tasa $h_0(\mathbf{X}(t), \beta, \gamma) = h_1(\mathbf{X}(t), \beta, \gamma) + h_2(\mathbf{X}(t), \beta, \gamma)$. Además, la k -ésima reacción ocurre con probabilidad $h_k(\mathbf{X}(t), \beta, \gamma)/h_0(\mathbf{X}(t), \beta, \gamma)$. Esto constituye una cadena de Markov de saltos puros y hace que el proceso sea fácil de simular utilizando técnicas de simulación discreta. El método utilizado en este trabajo hace uso de lo descrito en este párrafo y se le conoce como algoritmo Gillespie Gillespie (2007).

2.2. Fundamentos de Redes

Se planteará la dispersión del agente infeccioso sobre una población, donde las interacciones entre los individuos de la población de estudio se modelan con una red. La Teoría de Redes proporciona un marco teórico útil para modelar las interacciones entre los individuos y así plantear un escenario más realista de la dispersión de un agente infeccioso en una población pequeña. En este enfoque se modela la población como una estructura espacial donde los miembros de la misma son nodos de una red, y las aristas de la red representan interacciones entre los individuos que potencialmente pueden llevar a la transmisión de la enfermedad Newman (2010); Kolaczyk (2009).

A la clase de redes que modelan interacciones entre los miembros de una población (entidades sociales) se le denomina redes sociales Scott (2000); Wasserman and Faust (1994). Una red social puede ser representada mediante una gráfica.

Una gráfica $\mathcal{G} = (V, E)$ es una estructura matemática que consta de un conjunto finito de nodos o vértices V y un conjunto de aristas E . El conjunto de aristas está conformado por pares $\{u, v\}$ de vértices distintos $u, v \in V$. Una gráfica donde el orden de los vértices que conforman una arista es ordenado, es decir, si $\{u, v\}$ es distinto de $\{v, u\}$, se denomina *gráfica dirigida*. A las aristas de una gráfica dirigida se les llama *arcos* o *aristas dirigidas* y usualmente se representan con flechas. En el caso en que el orden de los vértices en una

arista es irrelevante, se tiene una *gráfica no dirigida* y la representación común es como una recta que conecta un vértice con otro.

Una gráfica se denomina simple si no es dirigida y el conjunto de aristas conecta siempre dos diferentes vértices. Esto es, si la gráfica no tiene aristas que conecten a un nodo consigo mismo (bucles). Se denotará al número total de vértices en la gráfica como $N_V = |V|$ y al total de aristas como $N_E = |E|$. Por simplicidad se etiquetará a los vértices con los enteros $1, \dots, N_V$.

La conectividad de la gráfica puede determinarse por las adyacencias que existen en la misma. Se dice que dos vértices $u, v \in V$ son *adyacentes* si existe una arista en E que conecte u con v . Se dice que una arista $e \in E$ es *incidente* en un vértice $v \in V$ si v es elemento del par de vértices a los que conecta e . De aquí surge la noción de *grado* d_v de un vértice, el cual se define como el número de aristas incidentes a dicho vértice. El grado d_v de un vértice $v \in V$ nos brinda una cuantificación de la medida en la que v está conectado con los otros vértices de la gráfica.

2.2.1. Simulación de Redes Aleatorias

En este trabajo se considera como *red aleatoria* a una gráfica no dirigida donde el grado de sus nodos sigue cierta distribución de probabilidad. La distribución de los grados de los nodos debe ser una densidad discreta definida sobre los enteros no negativos.

Para simular una red aleatoria se utiliza el algoritmo propuesto por Molloy and Reed (1995), el cual se explica brevemente a continuación. Supóngase que se desea simular una red aleatoria con N_v nodos. En primer lugar se generan los grados de los nodos de la red $\{d_1, \dots, d_{N_v}\}$. El grado de cada nodo puede ser: a) fijo y especificado por el usuario, b) $N_v - 1$ para generar una gráfica completa, o c) seguir cualquier distribución de probabilidad discreta y sobre los enteros no negativos.

El siguiente paso en el algoritmo es generar la arista $\{u, v\}$, con $u, v \in V$ con probabilidad proporcional al producto de los grados d_u y d_v . Posteriormente se actualiza el grado de los vértices (restando uno a los nodos que se unieron), para considerar las conexiones que aún se pueden establecer, a éste nuevo grado se le llama *grado disponible*. Este proceso se continúa iterativamente, seleccionando en cada paso una arista con probabilidad proporcional al producto de los grados disponibles de los nodos que la conforman.

2.2.2. Simulación del SIR Estocástico en una Red

Se planteará el modelo SIR estocástico en una red social, donde cada nodo corresponde a un individuo y tiene un atributo asignado que describe su estado respecto a la enfermedad, es decir, el comportamiento dentro del cual se encuentra dicho individuo.

Supóngase que se tiene una red \mathcal{G} con N_v nodos etiquetados como $1, \dots, N_v$. A continuación se muestra el pseudo-algoritmo para simular en \mathcal{G} la dispersión de un agente infeccioso que se modela con un SIR estocástico de parámetros β y γ . Se supondrá que el número inicial de individuos infectados es i_0 y que al tiempo cero no existen aún individuos recuperados, por lo que el sistema inicial es $(N_v - i_0, i_0, 0)$. También se supondrá que un individuo infeccioso únicamente es capaz de infectar a sus vecinos inmediatos susceptibles.

El tiempo que los individuos pasan en el estado infeccioso puede ser fijo o tener una distribución de probabilidad de rango positivo. En este caso nos enfocaremos en el caso exponencial.

Algoritmo 1: Simulación del modelo SIR estocástico en una red.

1. Elegir i_0 nodos de la red, ya sea de manera determinista o aleatoria, los cuales serán etiquetados como infecciosos al tiempo 0. Los demás nodos se etiquetan como susceptibles en esta etapa.
 2. Determinar el tiempo del siguiente cambio, y a qué tipo de reacción corresponde (infección o recuperación) mediante el algoritmo Gillespie (1977), el cual permite una simulación discreta de una Cadena de Markov a tiempo continuo. Posteriormente actualizar el estado de cada nodo.
 3. Iterar el proceso hasta un tiempo de observación máximo (si existe) o hasta que no haya más individuos infecciosos.
-

Se considera que las transmisiones de las aristas son independientes para cada conexión de un susceptible con un infeccioso. El algoritmo para simular el agente infeccioso es análogo al SIMID (SIMulation of Infectious Diseases) implementado en Ramírez-Ramírez *et al.* (2013), y difieren en que con este último se pueden implementar algunos esquemas de vacunación como políticas de control de la epidemia.

2.3. Algoritmo ABC-MCMC

El algoritmo ABC-MCMC (Algoritmo 2) es un método de simulación de una densidad posterior que incorpora dos enfoques, *Approximate Bayesian Computation* Del Moral *et al.* (2012); Marin *et al.* (2012), que permite obtener una aproximación de la densidad posterior en casos donde la verosimilitud no puede obtenerse de manera explícita, y un método *Markov Chain Monte Carlo* Robert and Casella (2013) mediante el cual se simulará dicha aproximación de la densidad posterior.

Algoritmo 2: ABC-MCMC.

Dado un parámetro $\theta^{(t)}$ y una simulación $\mathbf{x}_{\theta^{(t)}}$ del modelo $f(\cdot | \theta^{(t)})$,

1. Generar ν_t de la densidad propuesta $q(\cdot | \theta^{(t)})$
2. Simular pseudo-observaciones \mathbf{x}_{ν_t} provenientes del modelo $f(\cdot | \nu_t)$.
3. Hacer

$$\theta^{(t+1)} = \begin{cases} \nu_t & \text{con probabilidad } \alpha(\theta^{(t)}, \nu_t), \\ \theta^{(t)} & \text{con probabilidad } 1 - \alpha(\theta^{(t)}, \nu_t). \end{cases}$$

donde:

$$\alpha(a, b) = \min \left\{ \frac{\pi_h(b | \mathbf{y})}{\pi_h(a | \mathbf{y})} \frac{q(a | b)}{q(b | a)}, 1 \right\}$$

4. Si se acepta ν_t , guardar $\mathbf{x}_{\theta^{(t+1)}} = \mathbf{x}_{\nu_t}$; en caso contrario, $\mathbf{x}_{\theta^{(t+1)}} = \mathbf{x}_{\theta^{(t)}}$.
-

Marjoram *et al.* (2003) propone este método de simulación MCMC de una distribución posterior considerando que se desconoce el modelo a partir del cual fueron generados los datos, pero puede simularse de éste. Este algoritmo se conoce como ABC-MCMC, ya que, a pesar de la ausencia de la verosimilitud, se genera una cadena de Markov que converge a la aproximación de la distribución objetivo.

3. Inferencia sobre el Modelo SIR en una Red Social

Supóngase que se modela la evolución de una epidemia con un SIR estocástico con tasas de transmisión y recuperación por hora β y γ , respectivamente, y que se tiene una red social que describe los contactos entre los individuos de una población, la cual puede ser representada con una gráfica. Además, supóngase que se cuenta con n reportes de nuevos infectados agregados diariamente y_1, \dots, y_n .

A partir de los reportes anteriores se desea hacer inferencia acerca de los parámetros que rigen la evolución de la epidemia en cuestión. Al tener una forma de simular el modelo en la red para un vector fijo de parámetros (β, γ) , puede utilizarse el Algoritmo 2 (ABC-MCMC) para obtener simulaciones de la densidad posterior de los parámetros. Es necesario agrupar en las simulaciones del modelo el número de reportes de nuevos infectados por día para hacer los datos simulados comparables con los observados.

Este procedimiento de inferencia puede generalizarse para cualquier modelo compartimental, únicamente se requiere poder simular dicho modelo en una red con un algoritmo análogo al Algoritmo 1. Por ejemplo, para el modelo SEIR (Susceptibles - Expuestos - Infectiosos - Removidos) únicamente habría que agregar el periodo de exposición, al cual se le puede asignar una distribución. Sin embargo, si únicamente se tiene información de los nuevos infectados del proceso como aquí se supone, si el número de parámetros del modelo aumenta, las estimaciones puntuales se vuelven menos precisas y los intervalos de probabilidad más amplios.

3.1. Experimentos Computacionales

Para ilustrar el algoritmo de inferencia del modelo SIR estocástico en una red, se simuló una red aleatoria con 500 nodos con distribución de grado Poisson(2.42). Supondremos que esta red simulada modela los contactos entre los individuos de una población hipotética. Fijando como estado inicial a dos individuos infectados, se simuló la dispersión de un agente infeccioso en la red con un modelo SIR estocástico, considerando $\beta = 0.03$ y $\gamma = 0.01$ como las tasas por hora de infección y recuperación, respectivamente.

Se considera que se tienen reportes diarios de nuevos infectados $\mathbf{y} = y_1, \dots, y_n$. Es decir, se supondrá que no se cuenta con los tiempos exactos de cada cambio en el sistema, sino con

el número de nuevos individuos infectados acumulados en períodos de 24 horas, con lo cual se obtienen 35 reportes que se observa en la gráfica 2.

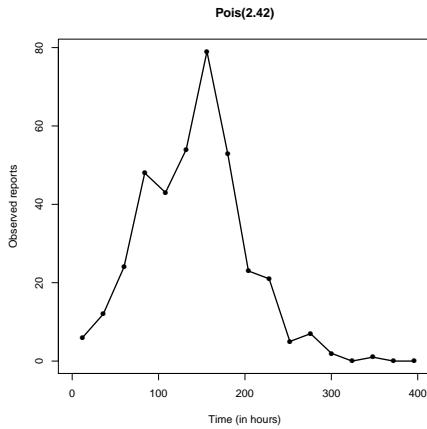


Figura 2: Reportes observados (Poisson).

Para medir la diferencia entre los datos observados \mathbf{y} y una simulación \mathbf{x} dada, se considera el siguiente estadístico univariado:

$$t(\mathbf{x}, \mathbf{y}) = \sum_i \sqrt{\frac{(x_i - y_i)^2}{y_i + \varepsilon}},$$

donde $0 < \varepsilon \ll 1$.

Para la estimación de la densidad del error de aproximación de la densidad posterior (ver Algoritmo 2) se considera un kernel normal con ancho de banda proporcional al número de reportes observados. Además, para la implementación del Algoritmo ABC-MCMC, se utilizó como densidad propuesta una mezcla de densidades normales, una con varianza más pequeña que la otra para permitir una mejor exploración del espacio paramétrico. Además se utilizaron densidades *a priori* Gamma(1,10) para ambos parámetros, que se supondrá que reflejan la información previa que se tiene acerca de las tasas de infección y recuperación del modelo SIR estocástico. El algoritmo de simulación es un algoritmo de aceptación y rechazo, el cual acepta valores de los parámetros que produzcan simulaciones *cercanas* en algún sentido a la curva de reportes observados.

Después de analizar la convergencia de la cadena simulada y eliminar el periodo de *burn-in*, y considerando los rezagos correspondientes, se obtienen simulaciones de la densidad

posterior que dan lugar a las estimaciones puntuales y por intervalos de 95 % de probabilidad que se muestran en la Tabla 1.

Cuantil	β	γ
2.5 %	0.0286	0.0072
50 %	0.0440	0.0214
97.5 %	0.0681	0.0520

Tabla 1: Estimaciones puntuales y por intervalos de probabilidad (Red Poisson).

Se observa que los intervalos contienen a los parámetros con que se simularon los datos (0.03, 0.01), y que son aproximadamente simétricos.

4. Conclusiones

Se planteó el modelo epidemiológico SIR bajo el enfoque estocástico (estocasticidad demográfica), donde se asigna a los tiempos que los individuos pasan en cada compartimento una distribución de probabilidad, lo cual se refleja en la aleatoriedad de la solución del sistema en un tiempo fijo. En este caso se supuso que el sistema se desarrollaba en una escala de tiempo continua, pero que los posibles estados del sistema eran variables aleatorias discretas. Si se asigna una distribución exponencial a los tiempos de cambio en el sistema, se trata de un Proceso Poisson no homogéneo, donde las tasas al tiempo $t > 0$ dependen del número de individuos que en ese instante se encuentren en cada uno de los compartimentos. Adicionalmente, se agregó al modelo una estructura relacional de los individuos mediante una red social, de tal forma que un individuo únicamente fuera capaz de contagiar a sus vecinos inmediatos. Este planteamiento permite un enfoque más realista, sin embargo las simulaciones y la inferencia son intensivos computacionalmente.

Se supuso un escenario apegado a la realidad, en el cual se contaba con una serie de reportes de nuevos infectados y_1, \dots, y_n , a partir de los cuales se pretendía realizar inferencia sobre los parámetros del modelo.

Se plantearon esquemas y modelos bayesianos de inferencia estadística, ya que en el contexto de epidemiología generalmente se cuenta con información previa acerca de los paráme-

tros (tasas de transferencia entre los compartimentos). La información previa puede ser obtenida analizando enfermedades similares, por ejemplo.

A partir de la densidad posterior es posible obtener estimaciones puntuales e intervalos de probabilidad para los parámetros, y los algoritmos MCMC nos hacen más sencilla la simulación de dicha densidad.

Una posible deficiencia del método computacional de inferencia sobre redes grandes es baja la eficiencia de las simulaciones. En estudios futuros esto podría mejorarse utilizando algún método de procesamiento en paralelo, o migrando el código a algún otro lenguaje de programación más eficiente como C o Python.

Varias generalizaciones pueden sugerirse a partir de este trabajo. Algunas relacionadas con el modelo, tal como considerar escenarios más realistas, como un horizonte de tiempo más amplio donde se permita la migración de los individuos. Otras mejoras pueden apuntar a la eficiencia de los métodos de cómputo, como ya se mencionó. Otra dirección a explorar es sobre variantes de los métodos de inferencia usados y el análisis de sus propiedades.

Bibliografía

- Allen, L. J. (2008). An introduction to stochastic epidemic models. In *Mathematical epidemiology*, pages 81–130. Springer.
- Boys, R. J., Wilkinson, D. J., and Kirkwood, T. B. (2008). Bayesian inference for a discretely observed stochastic kinetic model. *Statistics and Computing*, 18(2):125–135.
- Del Moral, P., Doucet, A., and Jasra, A. (2012). An adaptive sequential monte carlo method for approximate bayesian computation. *Statistics and Computing*, 22(5):1009–1020.
- Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *The journal of physical chemistry*, 81(25):2340–2361.
- Gillespie, D. T. (2007). Stochastic simulation of chemical kinetics. *Annu. Rev. Phys. Chem.*, 58:35–55.
- Kolaczyk, E. D. (2009). *Statistical Analysis of Network Data: Methods and Models*. Springer Publishing Company, Incorporated, 1st edition.

- Marin, J.-M., Pudlo, P., Robert, C. P., and Ryder, R. J. (2012). Approximate bayesian computational methods. *Statistics and Computing*, 22(6):1167–1180.
- Marjoram, P., Molitor, J., Plagnol, V., and Tavaré, S. (2003). Markov chain monte carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26):15324–15328.
- Molloy, M. and Reed, B. (1995). A critical point for random graphs with a given degree sequence. *Random structures & algorithms*, 6(2-3):161–180.
- Newman, M. (2010). *Networks: an introduction*. Oxford university press.
- R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
- Ramírez-Ramírez, L. L., Gel, Y. R., Thompson, M., de Villa, E., and McPherson, M. (2013). A new surveillance and spatio-temporal visualization tool simid: Simulation of infectious diseases using random networks and gis. *Computer methods and programs in biomedicine*, 110(3):455–470.
- Robert, C. and Casella, G. (2013). *Monte Carlo statistical methods*. Springer Science & Business Media.
- Scott, J. (2000). *Social Network Analysis: A Handbook*. SAGE Publications.
- Wasserman, S. and Faust, K. (1994). *Social network analysis: Methods and applications*, volume 8. Cambridge university press.

La Trata de Personas en México: Un Modelo para Identificar Patrones de Conducta de Posibles Tratantes

Paulina Martínez Rosas^a, Blanca Rosa Pérez Salvador^b

Universidad Autónoma Metropolitana, Unidad Iztapalapa

Por lo complejo del problema de trata de personas es imposible identificar a probables sospechosos de esta práctica usando métodos automatizados aplicados a bases de datos existentes. Una alternativa es utilizar métodos no supervisados en los que se puede aplicar la experiencia y la observación humana. En este sentido, un grupo interdisciplinario de científicos en la ciudad de Amsterdam propone utilizar el análisis de los conceptos formales para identificar a posibles sospechosos de este delito. En este trabajo se propone mostrar la problemática, y un modelo que utiliza patrones, tendencias o tipologías de comportamiento para ayudar a identificar a posibles sospechosos, usando bases de datos de la policía, de la Secretaría de Hacienda y Crédito Público, de servicio de aduanas, entre otras. En particular se muestra el ejemplo exitoso de esta metodología utilizado en los Países Bajos.

Clasificación: Trabajo de Tesis

Área-MSC: Estadística.

Subárea-MSC: Análisis de datos.

1. Introducción

La Comisión Nacional de los Derechos Humanos (CNDH) menciona que la trata de personas es un fenómeno delictivo que se encuentra extendido por todo el mundo; miles de personas

^ampaulina17@live.com.mx

^bpsbr@xanum.uam.mx

son víctimas de este delito, particularmente mujeres, niños y niñas son capturados, trasladados, vendidos y comprados con fines de explotación. La forma más conocida de la trata de personas es la explotación sexual. La trata de personas ha estado presente desde tiempos muy remotos, por ejemplo, en las civilizaciones: mesopotámica, egipcia, griega y romana, existió la exclavitud que es un tipo de trata de personas. Con el descubrimiento de América por los occidentales, se establece la trata de negros a gran escala. Los españoles y los portugueses, que se reparten el Nuevo Mundo ya en 1493, desean explotar estas regiones. La explotación de las minas de oro y de plata requerían una mano de obra abundante, robusta y, de ser posible, barata. Entonces los españoles tomaron a los indígenas como esclavos para que trabajaran las tierras, las minas y trabajos domésticos. La esclavitud de indios y negros se extendió por todo México durante la época colonial. Sin embargo fue a fines del siglo XIX y principios del siglo XX que se reconoció a la trata de personas como un problema social producto de secuestros, engaños o coacciones sobre mujeres vulnerables que se denominó Trata de Blancas (según Fondo de las Naciones Unidas para la Infancia (UNICEF, 2010)).

La trata de personas se considera un delito y es perseguido por las autoridades. Uno de los objetivos de la ley es identificar a posibles sospechosos de este delito, una manera de hacerlo es con el análisis de conceptos formales, que será analizado en las dos secciones siguientes.

2. Aprendizaje No Supervisado

El problema de trata de personas por ser muy complejo, se ha estudiado con modelos no supervisados, en particular con el modelo del análisis de conceptos formales (FCA, por sus siglas en inglés).

2.1. Análisis de Conceptos Formales

El Análisis de Conceptos Formales fue introducido por Rudolf Wille en 1984, empleando tanto la Teoría de Retículos como la Teoría del Orden, desarrollada por Birkhoff y otros en la década de los treintas. Aquí comenzaremos dando las definiciones básicas.

Definición 2.1. (Contexto Formal). Un contexto formal, es una terna (X, Y, I) donde X es un conjunto cuyos elementos son llamados objetos, Y es un conjunto cuyos elementos se denominan atributos e I es una relación binaria entre X e Y , es decir, $I \subseteq X \times Y$. La relación $(x, y) \in I$ significa "el objeto x tiene el atributo y ".

Entonces, un contexto formal (X, Y, I) consiste en un conjunto $X = \{x_1, \dots, x_n\}$, un conjunto $Y = \{y_1, \dots, y_m\}$, y una relación definida por: $(x_i, y_j) \in I$. El contexto formal se puede representar mediante una tabla, donde cada objeto x_i tiene asignado un renglón y cada atributo y_j tiene asignada una columna, y si el vector $(x_i, y_j) \in I$ entonces la tabla tiene una \times en la celda correspondiente a la fila i y columna j .

Definición 2.2. Siendo (X, Y, I) un contexto formal, definimos la operación de derivación $('')$ para $A \subseteq X$ y $B \subseteq Y$ como:

$$A' = \{y \in Y | (x, y) \in I \text{ para cada } x \in A\}, B' = \{x \in X | (x, y) \in I \text{ para cada } y \in B\}.$$

La definición anterior significa que dado un subconjunto A de objetos en X , A' es el conjunto de todos los atributos del conjunto Y que se aplican sobre todos y cada uno de los objetos en A . De la misma manera, dado un subconjunto B del conjunto de atributo Y , B' es el conjunto de objetos pertenecientes a X sobre los que se aplican todos los atributos de B .

Definición 2.3. (Concepto Formal). Un concepto formal en (X, Y, I) , es una pareja (A, B) con $A \subseteq X$ y $B \subseteq Y$ tal que $A' = B$ y $B' = A$, los conjuntos A y B se denominan la **extensión** y la **intención** de un concepto formal (A, B) , respectivamente.

La siguiente proposición muestra algunas propiedades importantes del operador $('')$.

Proposición 2.1. Para todos subconjuntos $A, A_1, A_2 \subseteq X$ y $B, B_1, B_2 \subseteq Y$ se satisfacen las siguientes propiedades:

- $A''' = A''$,
- Si $A_1 \subseteq A_2 \implies A'_1 \subseteq A'_2$,
- $A \subseteq A''$,
- $B'''' = B''$,
- Si $B_1 \subseteq B_2 \implies B'_2 \subseteq B'_1$,
- $B \subseteq B''$.

Definición 2.4. (Cerrado) Sea $A \subseteq X$, A (un subconjunto de objetos) es cerrado si cumple que $A'' = A$. Análogamente, para B (el subconjunto de atributos), es cerrado si cumple que $B'' = B$.

Definición 2.5. (Orden Parcial) Sea (A_1, B_1) y (A_2, B_2) conceptos formales, se define un orden parcial \leq de la siguiente manera, $(A_1, B_1) \leq (A_2, B_2)$ si y sólo si $A_1 \subseteq A_2$ o $B_2 \subseteq B_1$.

La colección de todos los conceptos formales de un contexto formal (X, Y, I) , se llama un concepto de retículo, noción fundamental en FCA.

Definición 2.6. (Concepto de Retículo) El conjunto de todos los conceptos formales obtenidos a partir de un contexto (X, Y, I) , junto con el orden parcial \leq , se llama concepto de retículo de (X, Y, I) y se denota como $\mathcal{B}(X, Y, I)$.

Definición 2.7. Dado el concepto de Retículo, se puede dar una definición alterna de los conceptos de Extensión y de Intensión:

$$Ext(X, Y, I) = \{A | (A, B) \in \mathcal{B}(X, Y, I) \text{ para alguna } B\} \quad (1)$$

$$Int(X, Y, I) = \{B | (A, B) \in \mathcal{B}(X, Y, I) \text{ para alguna } A\} \quad (2)$$

3. Aplicación del Método en los Países Bajos

En los Países Bajos la prostitución está legalizada, sin embargo es un crimen que una persona obtenga beneficios económicos obligando a otra, u otras personas a ejercer la prostitución. Por lo tanto, uno de los objetivos de la justicia Holandesa es perseguir a este tipo de delincuentes.

Poelmans et al. (2012) desarrollaron un modelo para identificar a posibles sospechosos de trata usando el FCA sobre los reportes de la policía de Amsterdam llenados durante sus rondines, esta metodología permite a los investigadores obtener información útil de un cúmulo enorme de información no estructurada.

Una vez que los investigadores encuentran indicios de que alguien es obligado a ejercer la prostitución, se hace una investigación más a fondo para determinar si existe realmente el delito.

Un ejemplo de un FCA con 5 objetos (reportes) y 6 atributos se representa en la Tabla 1.

		y_1	y_2	y_3	y_4	y_5	y_6
Reporte 1	x_1	×	×				×
Reporte 2	x_2			×	×	×	
Reporte 3	x_3	×	×	×	×	×	
Reporte 4	x_4						×
Reporte 5	x_5				×	×	

Tabla 1

Donde los atributos en la tabla son: y_1 : se detectaron prácticas de prostitución, y_2 : se obtuvo indicios de que la persona puede ser tratante, y_3 : se detectó violencia física, y_4 : se reportaron la propiedad de autos caros, y_5 : posesión de gran cantidad de dinero, y_6 : Nacionalidad Búlgara.

En esta tabla podemos encontrar los siguientes conceptos formales:

$$C_1 = (\{x_1, x_3\}, \{y_1, y_2\}), C_2 = (\{x_2, x_3, x_5\}, \{y_4, y_5\}), C_3 = (\{x_2, x_3\}, \{y_3, y_4, y_5\}), \\ C_4 = (\{x_1, x_4\}, \{y_6\}) \text{ y } C_5 = (\{x_3\}, \{y_1, y_2, y_3, y_4, y_5\}).$$

La relación de orden entre estos 5 conceptos formales esta dada por $C_5 \leq C_3 \leq C_2$. Esta tabla es una pequeña fracción que se utilizó en la investigación. De los conceptos formales se puede elaborar el siguiente retículo:

Los nodos del retículo en la Figura 1 representan a los informes policiacos y su etiqueta se encuentra en las casillas no sombreadas, y las casillas sombreadas representa a los atributos. Los reportes 3 y 2 están asociados con los atributos ‘violencia’, ‘autos caros’ y ‘gran cantidad de dinero’, el reporte 5 está asociado con los atributos ‘autos caros’ y ‘gran cantidad de dinero’, los reportes 1 y 3 están relacionados con los atributos ‘tratante’ y ‘prostitución’, el reporte 4 está asociado con el atributo ‘Búlgaros’.

El FCA fue capaz de ofrecer a la policía un enfoque ideal para una búsqueda más accesible de los sospechosos. A partir de este análisis la policía pudo darse cuenta que aquellos individuos cuyo nombre estaba en una posición más baja en el retículo tienen más probabilidad de ser tratante pues estos individuos poseen mayores atributos asociados a los tratantes.

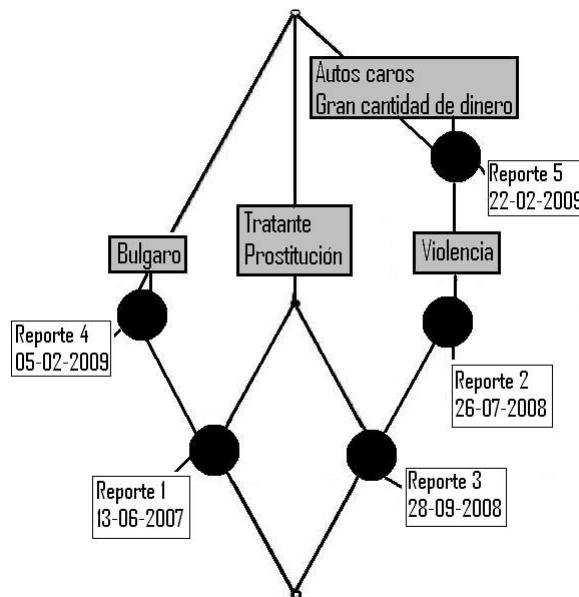


Figura 1: Diagrama de concepto de retículo.

Para obtener información de los reportes policiacos, Poelmans et al (2012) determinaron cuáles son los atributos asociados a los tratantes a partir de los casos positivos conocidos, con estos atributos se identificaron los conceptos formales y con ellos se elaboraron los perfiles de las personas que aparecieron en los reportes, para ello tuvieron que unificar la nomenclatura utilizada en los informes policiacos ya que una misma acción o concepto los policías lo escribían con palabras sinónimas, con los retículos formados se encontró, para algunos individuos, evidencias suficientes para iniciar una investigación profunda, usando FCA se lograron efectuar arrestos reales y el cierre de locales donde se practicaban la prostitución forzada.

4. Conclusiones

El problema de trata de personas es muy complejo y difícil de estudiar debido a que se realiza clandestinamente. Gracias a Poelmans et al. (2012) que realizaron un análisis de este delito, se pudo identificar a posibles sospechosos, usando bases de datos de la policía de Amsterdam. Nosotros pensamos que la misma metodología se puede utilizar en México con las bases de la Secretaría de Hacienda y Crédito Público, de los servicios de aduanas, entre otras. Para

realizar los análisis se tiene conocimiento que existe softwares (comercial y libre) tales como: ConExp, ToscanaJ, Lattice Miner, FcaBedrock.

El FCA es una técnica que se puede utilizar para exponer, investigar o relacionar grandes cantidades de información. Si el número de objetos es muy grande los retículos obtenidos se vuelven complejos, sin embargo, para los análisis se pueden considerar solamente algunos subconjuntos de objetos en el retículo para obtener conclusiones, dado que los punto de interés se conforman de un subconjunto pequeño de todos lo reportes.

Bibliografía

Belohlávek, R. (2008). *Introduction to Formal Concept Analysis*. Department of computer Science, Faculty of Science Palacký University, Olpmouc.

Comisión Nacional de los Derechos Humanos (CNDH) (2013). *Diagnóstico sobre la Situación de la Trata de Personas en México*.

Fondo de las Naciones Unidas para la Infancia (UNICEF) (2010). *Trata de personas. Una forma de esclavitud moderna. Un fenómeno mundial que afecta principalmente a niños, niñas y adolescentes*. Argentina, primera edición edition.

Ganter, B., Stumme, G., and Wille, R. (2005). *Formal Concept Analysis. Foundations and Applications*. Springer.

Petko Valtchev, R. J. (2011). Formal concept analysis. In *9th International Conference, ICFCA 2011*, pages 107–118. Nicosia, Cyprus.

Poelmans, J., Elzinga, P., Ignato, D. I., and Kuznetsov, S. O. (2012). Semi-automated knowledge discovery: identifying and profiling human trafficking. *Int. J. Gen. Syst.*, 4(8):774–804.06B23.

Proyecciones Aleatorias Tipo Random Fourier Features Basadas en Información Distribucional para Kernel PCA

Flor de María Martínez Sermeño^a

Numérika, Centro de Investigación en Matemáticas

Johan Van Horebeek

Centro de Investigación en Matemáticas

Recientemente, los métodos aleatorizados han demostrado su utilidad en técnicas de análisis de datos basados en transformaciones implícitas (métodos kernel) para resolver problemas de visualización, predicción y agrupamiento de conjuntos de datos multivariados de gran tamaño. En este trabajo se estudia una variante del método aleatorio Random Fourier Features (RFF) que saca provecho de cierta información distribucional de los datos; se muestra con ejemplos su potencial para aumentar la eficiencia de RFF en el contexto de Kernel PCA.

Área-MSN: Análisis Multivariado, Algoritmos Aleatorizados

Subárea-MSN: Análisis de Componentes Principales

1. Introducción

En general, los métodos aleatorizados en álgebra lineal se remontan al teorema de Johnson Lindenstrauss. Este teorema demuestra que un conjunto de puntos en un espacio de suficientemente alta dimensión puede ser mapeado a otro de menor dimensión preservando casi completamente las distancias entre los puntos, utilizando proyecciones aleatorias (una demostración en Dasgupta et al. (2003)). Así, la idea de los métodos aleatorizados es transformar los datos a un espacio de menor dimensión manteniendo la mayor cantidad de información posible.

^aflower10@cimat.mx

En primer lugar su finalidad es obtener algoritmos más rápidos. Además, como lo menciona Mahoney (2012), se tienen otras ventajas entre las que se puede resaltar: obtener algoritmos que sean más fáciles de analizar, cuyos resultados se puedan interpretar de manera más intuitiva y que permiten aprovechar técnicas computacionales modernas, por ejemplo: el cálculo en paralelo.

Como se explicará en la Sección 2, el método *Random Fourier Features* (RFF) es un método aleatorizado particular cuya idea básica es expresar cierta clase de matrices K de dimensión $n \times n$ como una esperanza matemática y luego aproximarlas mediante una media muestral. De esta manera se aproxima K con una factorización de la siguiente manera:

$$\widehat{K} = \widehat{Z}\widehat{Z}^t, \quad (1)$$

donde \widehat{Z} es una matriz de rango l y de dimensión $n \times l$ y con $l \ll n$.

Lo anterior es útil cuando se requiere calcular vectores propios de K . Como se demuestra en Cristianini et al. (2004), se pueden obtener los vectores propios de K directamente a partir de los vectores propios de la matriz $\widehat{Z}^t\widehat{Z}$ que es de dimensión $l \times l$, mucho menor que la de K .

En la Sección 3, se presentará una modificación a RFF en la cual las proyecciones ya no son independientes de los datos, lo cual permite aprovechar cierta información distribucional y aumentar su eficiencia.

En la Sección 4, se muestra con dos ejemplos el potencial de la propuesta para aumentar la eficiencia de RFF en el contexto de Kernel PCA.

2. Random Fourier Features

El método RFF fue propuesto por Rahimi y Recht (2007). Fue ideado específicamente para aproximar una matriz K de la forma:

$$K_{k,j} = k(x_k, x_j)$$

donde $\{x_i\}_1^n$ es un conjunto de datos en \mathcal{R}^d y $k(\cdot, \cdot)$ es una función kernel que es continua, positiva definida e invariante bajo traslaciones.

La idea es expresar la función kernel como una esperanza matemática:

$$k(x, y) = E_\theta(z_\theta(x) z_\theta(y)) \quad (2)$$

Usando el teorema de Bochner, la proyección aleatoria z_θ que permite escribir el kernel como una esperanza es:

$$z_\theta(x) = \sqrt{2} \cos(w^t x + b), \quad (3)$$

donde $\theta = (b, w)$, b se genera de una distribución uniforme sobre $[0, 2\pi]$ y w de una distribución p que depende del kernel. Por ejemplo en el caso de un kernel de base radial $k(x, y) = \exp(-||x - y||/\sigma^2)$, p es $N(\mathbf{0}, \sigma^{-2}I)$.

Se estima la esperanza (2) mediante la media muestral; para esto se generan l proyecciones aleatorias $z_{\theta_1}, \dots, z_{\theta_l}$ y se define $\vec{z}_\theta(x) = \frac{1}{\sqrt{l}}(z_{\theta_1}(x), \dots, z_{\theta_l}(x))$, aproximando cada entrada de la matriz K de la siguiente manera:

$$\begin{aligned} k(x_k, x_j) &= E_\theta(z_\theta(x_k) z_\theta(x_j)) \\ &\approx \frac{1}{l} \sum_{i=1}^l z_{\theta_i}(x_k) z_{\theta_i}(x_j) \\ &= \langle \vec{z}_\theta(x_k), \vec{z}_\theta(x_j) \rangle. \end{aligned}$$

Por lo anterior, se aproximarán la matriz K mediante:

$$\hat{K}^{rff} = Z^{rff} (Z^{rff})^t,$$

$$\text{con } Z^{rff} = \begin{pmatrix} \vec{z}_\theta(x_1)^t \\ \vdots \\ \vec{z}_\theta(x_n)^t \end{pmatrix}_{n \times l}.$$

3. RFF PCA: Una Modificación del Método RFF

Como puede observarse, el método RFF está basado en la proyección de los datos sobre direcciones aleatorias w generadas por la densidad p , para calcular la variable aleatoria $z_\theta(x)$. La distribución de las direcciones solamente depende del kernel que se desea aproximar y no de los datos. Yang et al. (2012) mencionaron que probablemente esta característica del método es lo que lo hace un poco ineficiente. Tomando en cuenta dicha observación, se presenta en esta sección un cambio al método RFF que se llamó RFF PCA, con el cual se pretende introducir información de los datos al método.

Idea detrás de la modificación No es muy útil proyectar los datos en una dirección en la cual no tienen (mucha) variabilidad. Considérese el caso extremo en que los datos viven en un subespacio. Si se elige w perteneciente al espacio ortogonal a este subespacio, entonces $w^t x = 0$ y la columna de Z^{rff} correspondiente será constante ($\approx \sqrt{2} \cos(b)$) por lo que tendrá un peso (“loading”) cero en los componentes principales de Z^{rff} . Para evitar que $w^t x \approx 0$ se decidió dar preferencia a direcciones en las cuales la proyección tenga mayor variabilidad. Lo anterior se hará realizando PCA sobre la matriz de datos X , tomando los primeros d^* componentes principales y forzando a que w pertenezca al espacio generado por dichos componentes.

Supóngase que $\{vp_1, \dots, vp_{d^*}\}$, $d^* \leq d$, son las componentes principales que se obtienen de hacer PCA sobre X . La nueva dirección w^* , obtenida restringiendo a que w esté en el espacio generado por $\{vp_1, \dots, vp_{d^*}\}$, está dada por:

$$w^* = \sum_{i=1}^{d^*} \langle w, vp_i \rangle vp_i = \sum_{i=1}^{d^*} w^t vp_i vp_i. \quad (4)$$

Se puede utilizar (4) para generar las nuevas direcciones de proyección. No obstante, en el caso de un kernel de base radial es fácil obtener la distribución de $w^t vp_i$ la cual se presenta a continuación y hace más fácil la generación de w^* .

Como ya se mencionó, para aproximar un kernel de base radial con parámetro σ cuando se utiliza RFF, se genera $w \sim N(\mathbf{0}, \sigma^{-2} I)$. Como los $\{vp_i\}_{i=1}^{d^*}$ están dados, entonces $w^t vp_i \sim N(vp_i \mathbf{0}, vp_i^t \sigma^{-2} I vp_i)$, para $i = 1, \dots, d^*$. Puesto que los $\{vp_i\}_{i=1}^{d^*}$ son ortonormales se tiene que $vp_i^t vp_i = 1$ y por lo tanto:

$$\begin{aligned} N(vp_i \mathbf{0}, vp_i^t \sigma^{-2} I vp_i) &= N(0, \sigma^{-2} vp_i^t vp_i) \\ &= N(0, \sigma^{-2}). \end{aligned}$$

Es decir $w^t vp_i \sim N(0, \sigma^{-2})$. Por lo tanto, las nuevas direcciones de proyección se generarán calculando

$$w^* = \sum_{i=1}^{d^*} n_i vp_i. \quad (5)$$

donde n_1, \dots, n_{d^*} son una muestra de $N(0, \sigma^{-2})$.

El procedimiento a seguir para obtener las aproximaciones de la matriz kernel, de sus eigenvalores y de sus eigenvectores es el mismo que para RFF simplemente cambiando las w por w^* .

Para entender un poco mejor el efecto de limitarse a proyecciones w^* , hay que recordar que para cualquier subespacio S y su normal S° :

$$\|x_i - x_j\|^2 = \|\mathcal{P}_S(x_i - x_j)\|^2 + \|\mathcal{P}_{S^\circ}(x_i - x_j)\|^2,$$

donde \mathcal{P}_S es el operador de proyección sobre S .

Entonces la entrada i, j de la matriz kernel aproximada mediante la modificación es de la forma:

$$\widehat{K}^{rff\ pca}[i, j] = K[i, j]\alpha_{i,j}, \quad (6)$$

donde:

$$\alpha_{i,j} = \begin{cases} 1, & \text{si } i = j \\ \exp(-\|\mathcal{P}_{S^\circ}(x_i - x_j)\|^2/\sigma^2) \geq 1, & \text{si } i \neq j \end{cases}.$$

Si $\alpha_{i,j} = c$ para $i \neq j$, entonces:

$$\widehat{K}^{rff\ pca} = c\widehat{K}^{rff} + (1 - c)I$$

Es fácil ver que los valores propios son afectados por la aproximación pero los vectores propios, no.

Aunque en la práctica $\alpha_{i,j}$ no es constante, diversos experimentos confirmaron que la aproximación afecta en primer lugar a los valores propios pero no a los vectores propios. Como se verá en la Sección 4, en muchos métodos kernel es mucho más importante aproximar bien los vectores propios que los valores propios o la matriz K .

En la búsqueda de un mejor método de aproximación se debe tomar en cuenta que calcular $\alpha_{i,j}$ explícitamente es muy costoso cuando el número de datos es alto. Como alternativa, se decidió estimar los $\alpha_{i,j}$ tomando un subconjunto pequeño de los datos y calculando la media de los $\alpha_{i,j}$ para ese subconjunto. Las fórmulas para obtener las diferentes estimaciones dadas por RFF PCA pueden revisarse a detalle en Martínez (2015).

El método RFF PCA pierde la propiedad de convergencia a la verdadera matriz kernel. Sin embargo, se desea ver si para cuando el número de variables aleatorias l que se genera es pequeño, el hecho de que la elección de w depende de los datos ayuda a capturar mejor la esencia de estos.

4. Experimentos

En esta sección, se muestran algunas aplicaciones de lo anterior para el caso en que la matriz K proviene de los llamados métodos kernel (Cristianini et al. (2004)). La idea de los métodos kernel es primero mapear los datos a un (otro) espacio dotado de producto punto, en el cual los datos tengan una estructura más sencilla, para después analizarlos mediante técnicas clásicas de análisis de datos. Por ejemplo, mapearlos a un espacio en donde las observaciones presenten una estructura aproximadamente lineal. Lo particular de los métodos kernel es que en lugar de transformar los datos explícitamente, se trabaja con los productos punto entre las observaciones en el nuevo espacio. La matriz que contiene dichos productos punto es llamada matriz kernel o matriz de Gram K ; la función $k(\cdot, \cdot)$ que calcula el producto punto entre los datos transformados se llama la función kernel.

Así, un método kernel puede verse como un método que resulta de reescribir (*kernelizar*) una técnica de análisis de datos en función de productos punto y donde se cambia la distancia Euclídea entre observaciones por la inducida por $k(\cdot, \cdot)$.

A continuación nos enfocamos en Kernel PCA (KPCA): el método que resulta de kernelizar el análisis de componentes principales (PCA). El interés de aproximar K por la factorización (1) radica en la relación entre los vectores y valores propios de $\widehat{Z}^t \widehat{Z}$ y $\widehat{Z} \widehat{Z}^t$, y el hecho que los *loadings* de una son los *scores* de la otra (Cristianini et al. (2004)). Si $l \ll n$ y se estima la matriz de covarianzas con $\widehat{Z}^t \widehat{Z}$, se obtiene una simplificación substancial de KPCA.

4.1. Ejemplo 1

El primer experimento está inspirado en el realizado por Yang et al. (2012) y pretende investigar qué tan buenas son las estimaciones de los eigenvectores de K , dados por los métodos RFF y RFF PCA. Ellos usan el signo del segundo vector propio de K para construir un clasificador (ver Yang et al. (2012) para mayores detalles).

Para este ejemplo se generaron $n = 5,000$ observaciones de dimensión $d = 102$. Las primeras dos dimensiones separarán las observaciones en dos grupos como se muestra en la Figura 1; las últimas 100 dimensiones corresponden a ruido generado de una distribución uniforme en $(0, 1)$. Los grupos fueron creados de la siguiente manera: la mitad de los puntos se generaron de una distribución uniforme en un círculo de radio 0.5 centrado en $(0.5, 0.5)$ y

la otra mitad, en un círculo de radio 0.5 y centrado en $(-0.5, 0.5)$. Así, la matriz generada X es de dimensión $5,000 \times 102$ y de rango 102.

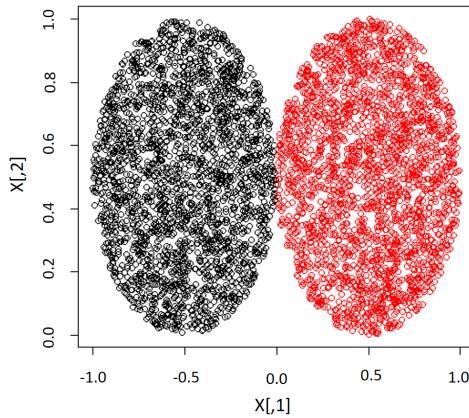


Figura 1: Gráfica de las primeras dos dimensiones de los datos, las cuales forman dos grupos que son diferenciados mediante el color.

La matriz de Gram para estos datos se produjo utilizando un kernel de base radial centrado de parámetro $\sigma \approx 8.69$, el cual se estimó mediante la función `sigest` de R. Se calcularon las aproximaciones por descomposición espectral de rango $k = 60$ dadas por el método RFF y de rango $k = d^* = 50$ para la modificación RFF PCA. Yang et al. (2012) mostraron que el método RFF no se desempeñaba bien para clasificación cuando se generan solamente $l = 100$ variables aleatorias. Se desea observar si la modificación RFF PCA tiene un mejor desempeño que RFF en los casos en que se generan pocas variables aleatorias. Por lo anterior, el número de variables aleatorias generadas se eligió como un porcentaje del número de columnas de K : 2 %, 3 %, 5 % y 10 %, de forma tal que $l = 100, 150, 250$ y 500. Se estimó el múltiplo α con la media de los múltiplos $\alpha_{i,j}$ calculados para el subconjunto dado por los primeros 101 renglones y la primer columna de X .

Por la estructura creada en los datos, se tiene una matriz de Gram con una gran brecha entre los eigenvalores. Los primeros dos eigenvalores son 1847.11 y 65.89 respectivamente. Se tiene interés en estimar el segundo vector propio de K porque el grupo al que se asigna un dato, se verá reflejado en el signo de la entrada correspondiente a dicho dato del segundo

vector propio de K . Es decir, la finalidad es construir un clasificador basado en el segundo vector propio.

Para estimar el error de clasificación, se repitió el experimento 100 veces. En la Figura 2, se presentan las gráficas del segundo vector propio de K y sus estimaciones obtenidas mediante RFF y RFF PCA utilizando solamente el 2 % de las columnas de K , de forma tal que $l = 100$ y el diagrama de caja para el número de observaciones mal clasificadas con base en dicho vector.

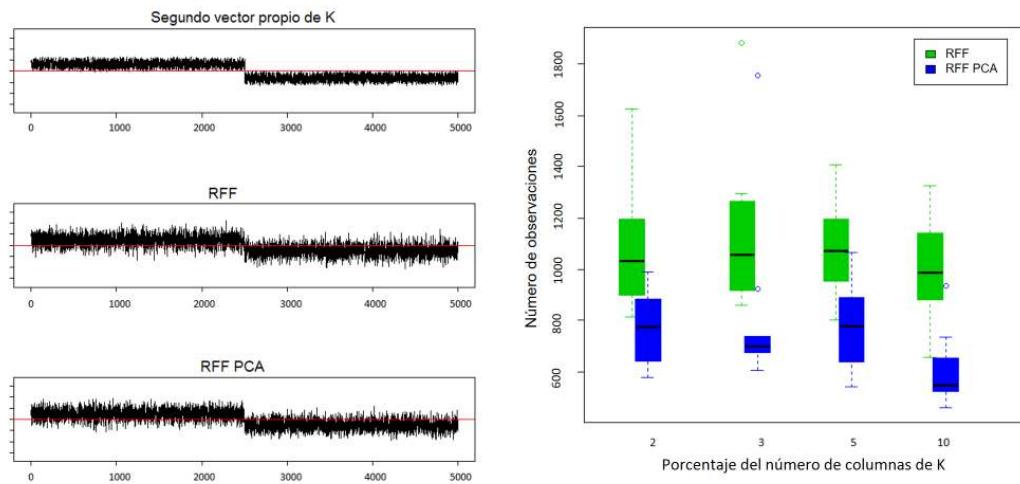


Figura 2: Izquierda: Segundo vector propio de K y sus estimaciones obtenidas con RFF y RFF PCA. Derecha: Observaciones mal clasificadas con base en el signo del segundo vector propio estimado por RFF y RFF PCA.

Se puede observar que los dos grupos creados o la brecha en los eigenvalores se refleja en el segundo vector propio como un “salto”. El grupo al que se asigna la j -ésima observación, está dado por el signo de la entrada j -ésima del segundo vector propio. En este caso simple, ambos métodos parecen rescatar el salto con tan sólo el 2 % de columnas de K como número de variables aleatorias generadas. Sin embargo, el método RFF no capta muy bien el salto, lo cual provoca malas clasificaciones. Se puede observar que, en este caso, la modificación RFF PCA sí produce mejores resultados que el método RFF, ya que clasifica bien un mayor número de observaciones que el RFF.

4.2. Ejemplo 2

Este ejemplo está inspirado en el realizado por Lopez-Paz et al. (2014) donde se usan las proyecciones dadas por KPCA como un método de compresión (autoencoder). Los datos que se utilizarán son los del conjunto MNIST LeCun *et al.* (2013); cada dato es la digitalización de una imagen de un dígito manuscrito del 0 al 9. La idea es mapear los datos con KPCA a un espacio de menor dimensión que los caracterice bien para posteriormente reconstruirlos (llamadas preimagen), como se muestra en la Figura 3.

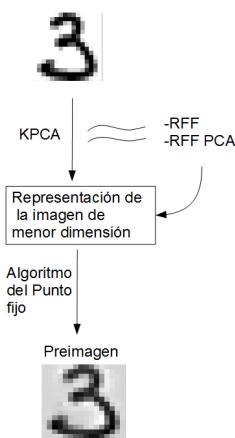


Figura 3: Diagrama de randomized autoencoders.

Como se muestra en la Figura 3, el resultado de KPCA se puede aproximar utilizando los métodos aleatorizados de aproximación de matrices. Al procedimiento por el cual la reconstrucción de la imagen en el espacio original se obtiene a partir de las proyecciones aproximadas mediante los métodos aleatorizados de aproximación de matrices, se le conoce en Ciencias de la computación como “*randomized autoencoders*”. En este ejemplo se pretende mostrar el desempeño de RFF y RFF PCA cuando se realiza *randomized autoencoders* y además se reduce bastante algunas variables que afectan la calidad de la imagen.

Se consideró un subconjunto de entrenamiento que cuenta con 7,291 observaciones. Así, la matriz de datos X es de dimensión 7291×256 y cada renglón de dicha matriz corresponde a la imagen de un dígito entre el cero y el nueve. El rango de la matriz X es 256. Se utilizó un kernel de base radial de parámetro $\sigma = 11$. El porcentaje de variables aleatorias generadas

fue 2% que equivale a $l = 146$ y el rango de la matriz de aproximación se tomó como $k = 100$, lo cual es menos de la mitad del rango original. Para ambos parámetros del procedimiento, lo anterior representa una gran reducción. El número de vectores propios sobre el que se proyecta se tomó como 50 y para RFF PCA se tomó $d^* = 50$.

En la Figura 4 se presenta la imagen original que se desea estimar, la preimagen que se obtendría si se hiciera Kernel PCA sin utilizar aproximaciones y los resultados obtenidos para ambos métodos de aproximación. Se debe señalar que se realizó el procedimiento con diferentes dígitos y los resultados son similares a los aquí presentados.

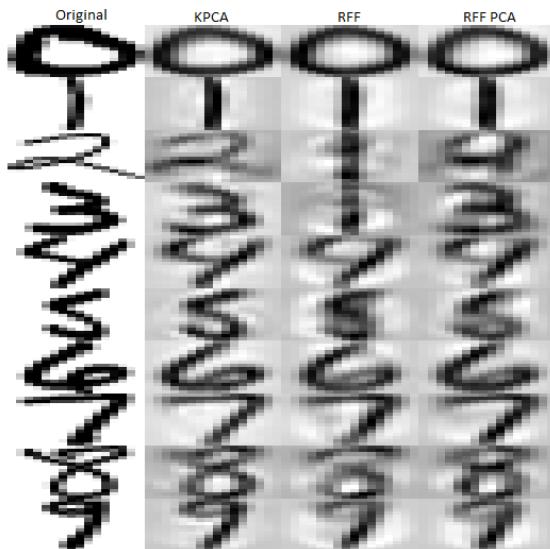


Figura 4: Aproximaciones obtenidas con porc = 2, k = 100 y proyectando sobre 50 v.p.

Se puede observar que, aunque las preimágenes han perdido calidad, el desempeño de los métodos es muy bueno y similar. Sin embargo, por ejemplo para los dígitos 2, 3 y 5, se obtiene una mejor aproximación con RFF PCA que con RFF. Se debe recalcar que el rango de la matriz de aproximación utilizada para el método RFF PCA es 50, lo cual es la mitad del rango que se utilizó para RFF. Se puede percibir muy poca diferencia entre el desempeño de Kernel PCA y las aproximaciones de su resultado dadas por ambos métodos. Lo anterior es muy bueno ya que se obtiene un resultado similar al de KPCA utilizando aproximaciones que implican una gran reducción de costos computacionales.

En los diferentes experimentos realizados se observó que algunos datos eran más difíciles de distinguir debido a su forma. En este caso extremo, se puede notar que los métodos de aproximación parecen captar la esencia de cada dígito. Se debe enfatizar que se está utilizando solamente el 1% del número de columnas para generar la aproximación de K , una matriz de aproximación cuyo rango es menor que la mitad del rango de K y se proyecta sobre el 0.5% del número de vectores de K .

5. Conclusiones

La modificación propuesta a RFF muestra que el uso de información distribucional de los datos permite reducir el número de proyecciones requeridas en RFF sin una pérdida inmediata en la calidad de la aproximación. Lo anterior, es de gran utilidad en métodos kernel cuando el número de observaciones es muy alta. Queda por explorar el potencial de otras variantes para estimar el subespacio de los datos, buscando un equilibrio entre tiempo de cómputo requerido y la calidad de la estimación.

Bibliografía

Cristianini, N. and Shawe-Taylor, J. (2004). *Kernel Methods for Pattern Analysis*. Cambridge University Press.

Dasgupta, S. and Gupta, A. (2003). An elementary proof of a theorem of johnson and lindenstrauss. *Random Structures & Algorithms*, 22(1):60–65.

Izenman, A. (2008). *Modern multivariate statistical techniques: regression, classification and manifold learning*. Springer Texts in Statistics. Springer Verlag.

LeCun, Y., Cortes, C., and Burges, C. J. C. (2013). The mnist database of handwritten digits. <http://yann.lecun.com/exdb/mnist/>.

Lopez-Paz, D., Sra, S., Smola, A., Ghahramani, Z., and Scholkopf, B. (2014). Randomized nonlinear component analysis. *International Conference on Machine Learning*. arXiv:1402.0119.

- Mahoney, M. (2012). Randomized algorithms for matrices and data. *Advances in Machine Learning and Data Mining for Astronomy*, pages 647–672.
- Martínez, F. (2015). Proyecciones aleatorias para aproximar métodos kernel. Tesis de maestría, CIMAT.
- Rahimi, A. and Recht, B. (2007). Random features for large-scale kernel machines. *Neural Information Processing Systems*, pages 1177–1184.
- Yang, T., Li, Y., Mahdavi, M., Jin, R., and Zhou, Z. (2012). Nyström method vs random fourier features: a theoretical and empirical comparison. *Neural Information Processing Systems*, pages 485–493.

Pronóstico del ITAEE del Estado de Veracruz a través de la Metodología Box-Jenkins

Ángel Luis López Morales^a, Juan Ruiz Ramírez

Universidad Veracruzana

Edson Valdés Iglesias

Universidad Autónoma Metropolitana

Los agentes económicos que interactúan en la economía nacional (principalmente empresas y el Estado) bajo la premisa de tomar decisiones adecuadas que les permitan tener un crecimiento, se ven inmersos en la necesidad de buscar información que les ayude hacerlo. Para el caso de México, la institución que se encarga de dar a conocer el resultado de las estadísticas oficiales es el INEGI. Sin embargo presenta atrasos en las publicaciones que van de los 53 días hasta los 100 días para el caso del PIB estatal. Situación por la cual se elaboró un modelo de pronóstico para el Índice Trimestral de la Actividad Económica de los Estados (ITAEE), a partir de la metodología propuesta por Box y Jenkins, la cual tiene una ventaja importante al solo necesitar alrededor de 50 datos para hacer un pronóstico confiable y los resultados tienen un alto grado de significancia. El pronóstico elaborado se realizó con un nivel de confianza del 95 %, por lo cual los resultados de este modelo pueden ser tomados en cuenta por cualquier agente económico para el proceso de toma de decisiones.

Clasificación: Tesis de Especialización.

Área-MSC: Series de Tiempo.

Subárea-MSC: Modelos ARIMA.

1. Introducción

La elaboración de modelos de pronóstico es una técnica que facilita tener conocimientos de acontecimientos futuros con datos históricos; estos modelos son usados en diferentes campos

^aangelopez_m@hotmail.com

que van desde el área de salud (Jara et al., 1998; Coutin, 2007; Collantes, 2001), minería (Aranibar y Humérez, 1996) y electricidad (Castaño, 2007) por nombrar algunos, hasta la economía donde los pronósticos juegan un papel muy importante, ya sea en la micro o macroeconomía

En economía hay dos divisiones básicas para el estudio de los fenómenos que ocurren, microeconomía y macroeconomía; en ambas divisiones hay una constante interacción de agentes económicos que buscan que sus decisiones sean las mejores y que les generen los mayores beneficios.

La macroeconomía se centra en el estudio de los grandes agregados económicos de un país en un periodo determinado, para su cuantificación destaca el uso de un índice el Producto Interno Bruto (PIB), el cual es muy importante, debido a que refleja el crecimiento de la actividad de un país, al mostrar cual ha sido el comportamiento de la actividad económica de la Nación (Elizondo, 2012; Cuadrado, 2002).

En México, la Institución encargada en realizar y dar a conocer los cálculos referentes al Sistema de Cuentas Nacional de México (SCNM), es el INEGI, en el SCNM se presenta la información referente a la producción, consumo, ahorro, etc. de los grandes agregados económicos además de los indicadores macroeconómicos como el PIB y el Índice Trimestral de la Actividad Económica de los Estados (ITAEE). La publicación de los diferentes productos del SCNM se realiza en fechas previamente establecidas, pero que en ocasiones tienen un retraso de consideración al cierre del periodo a contabilizar; para el PIB nacional la información del periodo se presenta hasta después de 42 días de finalizado el periodo, mientras que el PIB de los estados es presentado con una posterioridad de más de 100 días al cierre del periodo. Los retrasos presentados en las publicaciones afectan la toma de decisiones de los agentes económicos que dependen de esta información (principalmente a las empresas y al gobierno) (Ruiz *et al.*, 2014).

Debido a los atrasos en las publicaciones del PIB nacional y estatal, en el presente, se ofrece un modelo predictivo basado en el Índice Trimestral de la Actividad Económica Estatal (ITAEE) con el cual se busca dar un adelanto de la información dada por INEGI de forma oficial, esperando ayude a los agentes económicos en la toma de decisiones. Para la elaboración del modelo predictivo se siguió la metodología Box y Jenkins.

2. Marco Teórico

Para la elaboración del modelo predictivo se siguió la metodología presentada por Box y Jenkins (1970), en la cual se usan los datos históricos propios de la variable a predecir, omitiendo el uso de variables explicativas en la creación del modelo. Por sus siglas en inglés (Autorregresive Integrated Moving Average), esta metodología también es conocida como ARIMA.

La metodología Box y Jenkins sigue cuatro etapas: Identificación, estimación, diagnóstico y predicción (Gujarati, 2010). Cabe mencionar que antes de la etapa de identificación, la serie de tiempo debe de cumplir un requerimiento esencial, el cual es que la misma sea estacionaria, sino presentara esta característica se puede correr el riesgo que la predicción sea espuria (Gujarati, 2010), si la serie de tiempo no cumple este requerimiento se le debe de realizar cierto número de diferenciaciones, este es el caso, principalmente, de las series macroeconómicas (Kennedy, 1997).

Los modelos ARIMA son realizados de una serie de tiempo que no es estacionaria de inicio, por lo cual constan de dos procesos uno autorregresivo (AR) y otro proceso de promedios móviles (MA) y un proceso de integración (I) para estacionarizarla.

Los procesos de la metodología ARIMA están representados por la forma general (p,d,q) , donde p representa el proceso de medias móviles, q las medias móviles y d es el orden de integración; si d es igual a cero indica que la serie es estacionaria de inicio tomando la forma ARMA, pero si d es mayor que cero, entonces será un modelo ARIMA con d diferenciaciones.

2.1. Macroeconomía

La ciencia económica hace una división en dos grandes grupos los fenómenos económicos con el fin de facilitar su estudio, esta división es: microeconomía y macroeconomía. De momento se dejará de lado la microeconomía y se centrará en la macroeconomía, la cual estudia los fenómenos económicos en conjunto.

En macroeconomía, para la medición del crecimiento económico se usa el Producto Interno Bruto (PIB), el cual es el valor monetario de bienes y servicios producidos en una nación en un periodo, resumidos en una sola cifra (Samuelson y Nordhaus, 2001; Parkin, 2009; Mankiw, 1997).

El PIB es el resultado de la suma del consumo, inversión privada, gasto gubernamental más las exportaciones netas (exportaciones netas = exportaciones - importaciones), la medición de este indicador es principalmente realizada de forma anual y trimestral; a la medición anual se le considera de baja frecuencia mientras que a la trimestral de alta frecuencia (Klein y Coutiño, 2004).

La periodicidad con la que INEGI presenta el PIB estatal es de baja frecuencia, por lo cual, si se desea estudiar el cambio trimestral de la actividad económica esta información no sería la adecuada; el INEGI publica en el SCNM, de forma análoga al PIB, un indicador que presenta la misma información referente al panorama general de la situación y evolución de la economía de cada una de las entidades federativas con la diferencia en la periodicidad, éste es el Índice de la Actividad Económica Estatal (ITAEE).

El ITAEE puede ser utilizado como un sustituto del PIB trimestral por estado, debido a que sigue con los mismos principios y normas contables que éste, con la diferencia que el INEGI solo lo considera como un adelanto del PIB, por lo cual no lo presenta de forma nominal sino únicamente como un índice; los resultados al ser preliminares, se sugiere considerarlo como un indicador de tendencia o dirección de la actividad económica estatal (Elizondo, 2012).

3. Resultados

En la Figura 1 se grafica el comportamiento de la serie de tiempo original, en este gráfico se puede observar como la serie tiene una tendencia positiva por lo que visualmente se puede decir que la serie es no estacionaria.

Por lo tanto se diferencia la serie de tiempo, se realiza la prueba de Dickey-Fuller aumentada a la serie transformada, esta prueba volvió a tener las misma hipótesis estadísticas. Los resultados obtenidos, mostrados en la Tabla 1, indican en esta ocasión, la serie diferenciada es estacionaria ya que el valor de p-value es menor al valor crítico de 0.05 presentados en la tabla de MacKinnon.

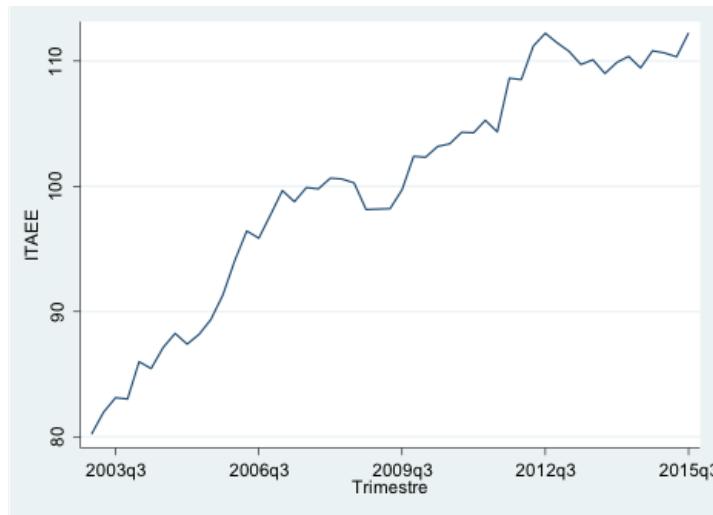


Figura 1: Serie de tiempo del ITAE del estado de Veracruz, 2003q1–2015q3.

Dickey-Fuller test for unit root		Number of obs = 49		
Test Statistic	Interpolated Dickey-Fuller			10% Critical Value
	1% Critical Value	5% Critical Value	10% Critical Value	
Z(t)	-7.773	-4.159	-3.504	-3.182

MacKinnon approximate p-value for Z(t) = 0.0000

D.Diferencia	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
Diferencia					
L1.	-1.15392	.1484553	-7.77	0.000	-1.452745 -.8550955
_trend	-.02088	.0134835	-1.55	0.128	-.0480208 .0062608
_cons	1.235367	.4144131	2.98	0.005	.4011969 2.069537

Tabla 1: Resultados de la prueba de Dickey-Fuller aumentada realizada a la serie diferenciada.

Ya teniendo estacionaria la serie, se elaboró el autocorrelograma y el autocorrelograma parcial, presentados en la Figura 2, para poder identificar el orden de los procesos y determinar el modelo que presentara mejores resultados para el pronóstico del ITAEE.

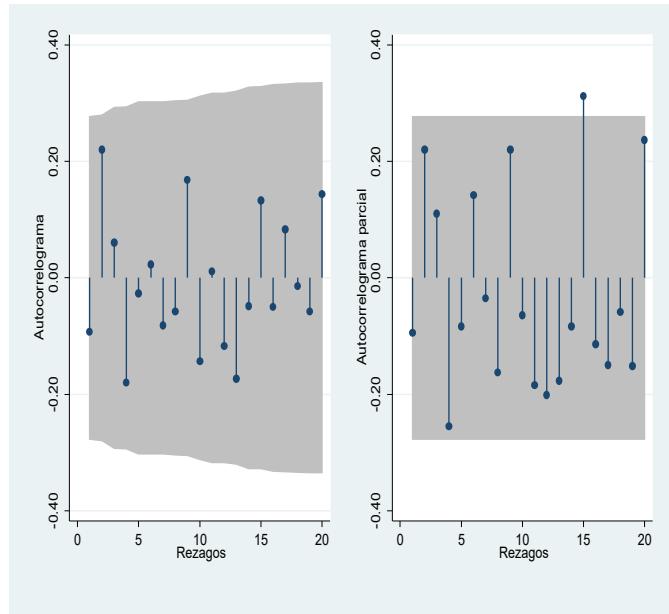


Figura 2: Autocorrelogramas de la serie diferenciada.

Con los autocorrelogramas de la serie diferenciada se identificó que el rezago 15 del autocorrelograma parcial salía de las bandas de confianza, por lo cual se realizó el modelo AR(15) I(1) del cual se graficaron los autocorrelogramas de sus residuos, presentados en la Figura 3, dónde se observa que, en esta ocasión, el rezago 4 sale de la banda confianzas al 95 %.

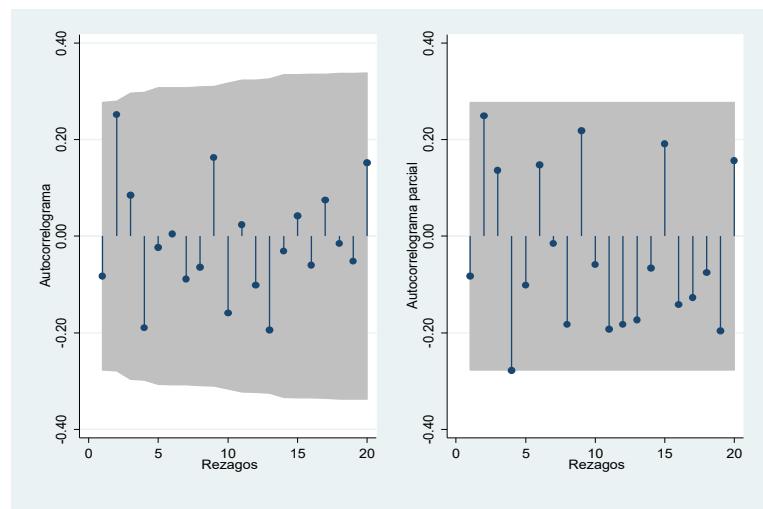


Figura 3: Residuos del modelo AR(15) I(1).

Al agregar el rezago 4 al modelo, se graficaron los autocorrelogramas de los residuos correspondientes al modelo AR (4, 15)¹ I(1) , los cuales se presentan en la Figura 4.

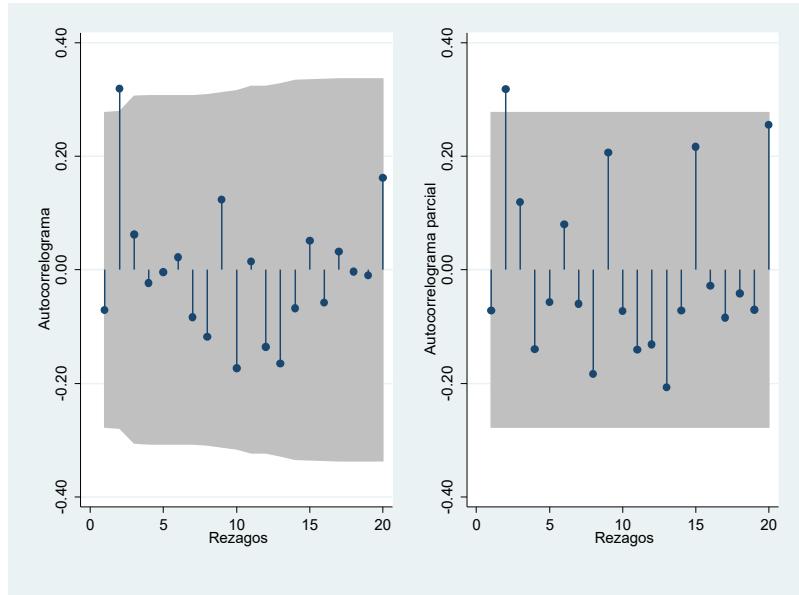


Figura 4: Autocorrelogramas de los residuos del modelo AR(4, 15) I(1).

En los autocorrelogramas de los residuos del modelo AR(4, 15) I(1) se identificó que el rezago dos, ambos autocorrelogramas, salía de la banda de confianza al 95 % por lo que se propusieron tres modelos, en el primero de ellos se agrega un término autoregresivo AR(2,4,15) I(1), en el segundo modelo se agregó un proceso autoregresivo y de media móvil AR(2,4,15) I(1) MA(2) y en el último se agregó un proceso de media móvil AR(4,15) I(1) MA(2); para evaluar cuál de los tres modelos propuestos se ajusta mejor se utilizó el criterio de información de Akaike y Schwarz para poder elegir el modelo que mejores resultados presentase al tener un error cuadrático menor (Tabla 2).

¹Esta notación señala los dos rezagos que se incluyen en el modelo a estimar su forma funcional es: $\nabla Y_t = \delta + \phi_4 \nabla Y_{t-4} + \phi_{15} \nabla Y_{t-15} + e_t$.

		Modelo		
		AR(2,4,15)	AR(2, 4, 15) MA(2)	AR(4, 15) MA(2)
AR	L2	0.33976	-0.2521	
	L4	-0.26757	-0.08701	-0.18329
	L15	0.25766	0.29218	0.28018
MA	L2		0.69279	0.43638
Statistic	AIC	168.1343	168.2888	166.8021
	BIC	177.6944	179.7609	176.3622

Tabla 2: Criterio de Akaike y Schwarz para los tres modelos propuestos.

Con el criterio de Akaike y Schwarz se identificó que el modelo que mejores resultados mostraba era el que incluía el rezago 2 en el proceso de medias móviles, por lo cual se obtuvieron sus autocorrelogramas; en sus autocorrelogramas, presentados en la Figura 4, se puede observar que el rezago 13 del autocorrelograma parcial sale de las bandas de confianza al 95 %.

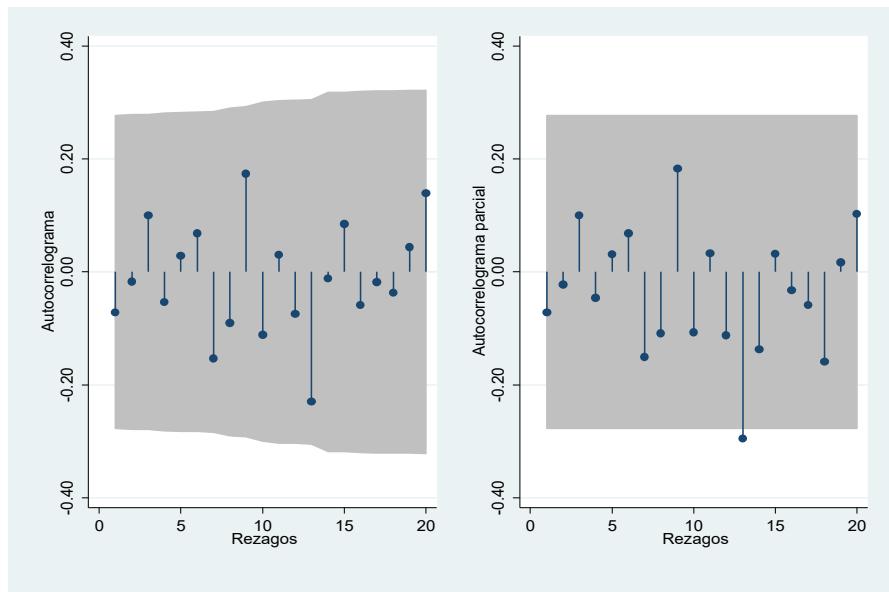


Figura 5: Autocorrelogramas de los residuos del modelo AR (4,15) I(1) MA(2).

Tomando en cuenta el rezago 13 del autocorrelograma parcial que sale de la banda de confianza se creó el modelo AR(4,13,15) I(1) MA(2), del cual se obtuvieron sus autocorrelogramas presentados en la Figura 6, donde se puede observar que ya no hay rezagos que salgan de la banda de confianza, por lo cual se consideró como un modelo admisible al cual se le realizaron las pruebas correspondientes sugeridas por Guerrero (2003) de las cuales cumplió las mismas. Para corroborar si los valores de los residuos tienden a una distribución normal se utiliza la prueba de normalidad, al obtener un valor de probabilidad mayor al valor crítico de 0.05 se puede concluir que los residuos del modelo tienden a comportarse de forma normal, los resultados se presentan a continuación en la Tabla 3.

Variable	Skewness/Kurtosis tests for Normality				
	Obs	Pr(Skewness)	Pr(Kurtosis)	adj chi2(2)	Prob>chi2
residuo	50	0.2622	0.8394	1.36	0.5073

Tabla 3: Prueba de normalidad de los residuos.

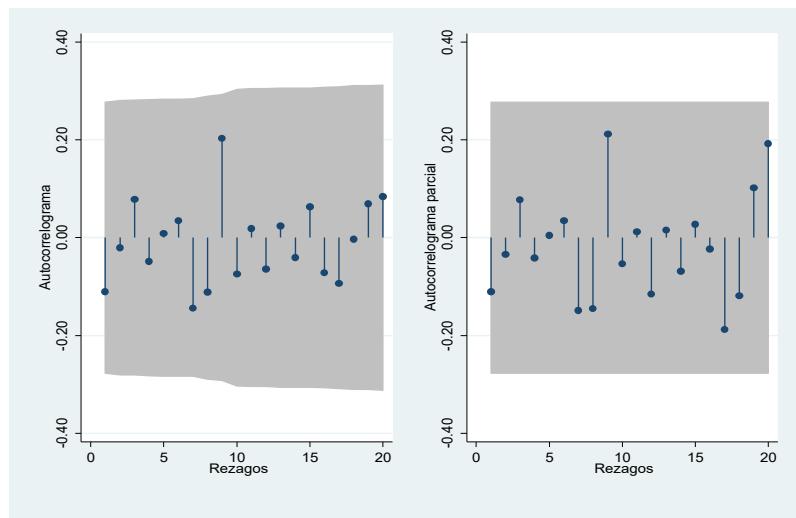


Figura 6: Autocorrelogramas de los residuos del modelo AR (4, 13, 15) I (1) MA(2).

El modelo elaborado puede ser expresado de la siguiente manera:

$$\nabla ITAEE = \delta + \phi_4 \nabla ITAEE_{t-4} + \phi_{13} \nabla ITAEE_{t-13} + \phi_{15} \nabla ITAEE_{t-15} + \theta_2 \varepsilon_{t-2} + \varepsilon_t$$

$$\begin{aligned} \nabla ITAEE = & 0.6420 - 0.113 \nabla ITAEE_{t-4} - 0.2937 \nabla ITAEE_{t-13} + 0.2809 \nabla ITAEE_{t-15} \\ & + 0.4902 \varepsilon_{t-2} + \varepsilon_t \end{aligned}$$

Al contar con un modelo que tuviera estabilidad en sus residuos y cumpliera los supuestos, se usó para hacer el pronóstico de la actividad económica del estado de Veracruz, a partir del tercer trimestre del 2015 al tercero del 2016; los resultados del pronóstico se presentan en la Tabla 4 así como en la Figura 7 donde se muestra la serie original y la serie pronosticada a través del modelo ARIMA.

PRONÓSTICO		
Trimestre	ITAEE	Tasa de crecimiento
2015/4	112.0508	-0.2013
2016/1	113.8453	1.6015
2016/2	115.0992	1.1014
2016/3	115.7128	0.5331

Tabla 4: Pronóstico del ITAEE usando el modelo AR(4,13,15) I(1) MA(2).

4. Conclusiones

Al seguir adecuadamente la metodología de Box y Jenkins se obtuvo un modelo que se ajustó a los requerimientos de la serie de tiempo presentada, este modelo al cumplir con los supuestos metodológicos se puede tomar como aceptable; cabe mencionar que si no hubiese un shock exógeno en la economía o una coyuntura económica, se puede decir de forma reservada, que el modelo presentado cuenta con una precisión del 95.



Figura 7: Serie original y pronosticada del ITAEE.

Bibliografía

- Aranibar, J. and Humérez, J. (1996). Modelos de series de tiempo para el pronóstico de precios de minerales. *Revista de Análisis Económico (UDAPE)*, 14:24–71.
- Box, G. and Jenkins, G. (1970). *Time series Analysis, Forecasting and Control*. Holden-Day, San Francisco.
- Castaño, E. (2007). Reconstrucción de datos de series de tiempo: una aplicación a la demanda horaria eléctrica. *Revista Colombiana de Estadística*, 30(2):247–263.
- Collantes, J. V. (2001). *Predicción con redes neuronales: Comparación con las metodologías de Box y Jenkins*. Universidad de los Andes, Mérida, Venezuela.
- Coutin, G. (2007). Utilización de modelos arima para la vigilancia e enfermedades transmisibles. *Revista Cubana de Salud Pública*, 33(2):1–11.

- Cuadrado, E., Lorenzo, F., and Quijo, V. (2002). *Prediciendo el pbi: ¿qué aportan los métodos cuantitativos?* Banco Central de Uruguay.
- Elizondo, R. (2012). *Estimaciones del PIB Mensual Basado en el IGAE*. Banco de México, D.F.
- Guerrero, V. (2003). *Análisis estadístico de series de tiempo económicas*. International Thompson, México.
- Gujarati, D. (2010). *Econometría*. McGraw Hill, México, quinta edición edition.
- Jara, P., Rosel, J., and Francesc, P. (1998). Análisis de la evolución de la sintomatología del ciclo menstrual mediante modelos arima. *Psicothema*, 10(1):183–195.
- Kennedy, P. (1997). *Introducción a la econometría*. Fondo de Cultura Económica, México.
- Klein, L. R. and Coutiño, A. (2004). Enfoque metodológico para un modelo de pronósticos de alta frecuencia para la economía mexicana. *Investigación Económica*, LXVIII(250):47–58.
- Mankiw, N. G. (1997). *Macroeconomía*. Antoni Bosch, Barcelona, tercera edición edition.
- Parkin, M. (2009). *Economía*. Pearson Educación, México, octava edición edition.
- Ruiz, J., Hernández, G., and Díaz, M. (2014). Importancia del modelo arima en el pronóstico del producto interno bruto trimestral de méxico. *Observatorio de la Economía Latinoamericana*, 201:1–15.
- Samuelson, P. and Nordhaus, W. (2001). *Macroeconomía*. McGraw-Hill, Madrid.

Visualización de Datos Espaciales en R: Elecciones Gubernamentales 2016 en Zacatecas

Iván Pacheco Soto^a

Instituto Tecnológico y de Estudios Superiores de Monterrey

La visualización de resultados electorales en mapas o imágenes satelitales permite entender en forma clara y sencilla las preferencias o tendencias de los votantes. El software estadístico R además de facilitar el manejo de las bases de datos y el análisis estadístico de los mismos, también incluye algunas librerías como `ggplot2` y `ggmap` que facilitan considerablemente la visualización de información geoelectoral. Utilizamos estas herramientas para realizar un análisis descriptivo geoespacial de las elecciones gubernamentales 2016 del estado de Zacatecas. Finalmente desarrollamos algunas aplicaciones web usando la librería `shiny` que nos permite visualizar geográficamente los resultados electorales a niveles de sección o regional en forma interactiva y dinámica.

Clasificación: Trabajo de divulgación

Área-MSC: 62-XX: Estadística

Subárea-MSC: 62-09: Métodos Gráficos

1. Introducción

La visualización de resultados provenientes de elecciones o de encuestas de opinión se limitan regularmente a diagramas barras, pastel o líneas, y hacen poco uso de mapas o diagramas dinámicos e interactivos para comunicar los resultados electorales a la población. Por otro lado, cuando se desean construir estos gráficos espaciales, generalmente se usa software especializado o costoso como ESRI ArcGIS que pueden combinar puntos, polígonos o diagramas básicos con mapas o imágenes de satélites, pero que en general no está disponible para la

^aipacheco@itesm.mx

mayoría de las empresas o universidades. Aunque el Instituto Nacional de Estadística y Geografía (INEGI) ha desarrollado el *Mapa Digital de México*, que es una aplicación libre que nos permite realizar consultas, visualizaciones y análisis similares a los que se realizan en ArcGIS de manera simple y sin costo adicional, no permite el nivel deseado de interacción con el usuario. En este trabajo proponemos el uso de R y sus librerías, que además de facilitar el manejo de las bases de datos, también nos permiten generar visualizaciones llamativas e interactivas para el análisis de la información.

2. Visualización de Datos Espaciales en R

R es un lenguaje de programación enfocado al análisis estadístico que es de libre acceso. R ha sido el resultado del esfuerzo colaborativo de muchas personas, que desde mediados de 1977 se han integrado en el R Core Team (2016) para seguir mejorando y desarrollando este lenguaje. La curva de aprendizaje para R se puede considerar un poco empinada, y tal vez un poco más empinada para la visualización de datos espaciales, pero la gran cantidad y diversidad de librerías integradas nos facilitan la solución de problemas. Actualmente podemos accesar más de 9,000 librerías de R sobre una gran cantidad de temas o áreas.

RStudio es un entorno de programación integrado (IDE) para R. El entorno de RStudio facilita considerablemente la programación y el aprendizaje de R. Además contiene otras herramientas que facilitan la depuración del código generado, la generación o modificación de librerías o la publicación de las aplicaciones.

RStudio se puede usar en una computadora personal, *RStudio Desktop*, o desde un servidor en Linux, *RStudio Server*. RStudio Server nos permite acceder remotamente a la interface de RStudio por medio de un navegador y también compartir y ejecutar nuestras aplicaciones usando alguna computadora o hasta un teléfono inteligente. Además, RStudio provee una plataforma de autoservicio, *Shinyapps.io*, para compartir o publicar las aplicaciones desarrolladas.

Aunque en R hay varias librerías o funciones para visualizar datos espaciales, como `maptools`, `googleVis` or `RgoogleMaps`, nos enfocaremos básicamente a las librerías `ggplot` y `ggmap`.

2.1. Bases de Datos

En este trabajo usamos los resultados de las elecciones gubernamentales realizadas el 5 de Junio de 2016 en el estado de Zacatecas. El desglose por casillas de las elecciones gubernamentales lo obtuvimos de la página del Programa de Resultados Electorales Preliminares (PREP) de Zacatecas. El código en R para obtener la base es el siguiente.

```
fUrl <- "http://resultadospreliminares.ieez.org.mx/PREP20152016/20160606_2144_BD_
ZACATECAS.zip"
download.file(fUrl, destfile = "elecciones_2016.zip")
unzip("elecciones_2016.zip", files = "ZACATECAS_GOBERNADOR_2016.csv")
```

El archivo ZACATECAS_GOBERNADOR_2016.csv contiene el resultado desglosado por casilla para cada partido, alianza y candidato independiente, la cual filtramos con la librería `dplyr`.

En los mapas generados incluimos las regiones geográficas de las secciones electorales. Estas regiones las construimos con un archivo shapefile que obtuvimos en el sitio de INEGI. El código R para la obtener el archivo shapefile de las secciones electorales de Zacatecas y convertirlo a un objeto de tipo data frame que se muestra abajo. Usamos las funciones `readOGR` y `spTransform` de la librería `rgdal` para leer los datos geoespaciales del archivo shapefile, y la función `tidy` de la librería `broom` para convertir la información a un data frame. Además, fue necesaria la función `gpclibPermit` de la librería `maptools` para lograr identificar los polígonos por sección electoral.

```
fUrl <- "http://cartografia.ife.org.mx/descargas/distritacion2017/federal/32/32.
zip"
download.file(fUrl, destfile = "estado_32.zip")
unzip("estado_32.zip")
seccs_utm <- readOGR("32", layer = "SECCION", stringsAsFactors = FALSE)
seccs_ll <- spTransform(seccs_utm, CRS("+proj=longlat+datum=WGS84"))
gpclibPermit(); secciones <- tidy(seccs_ll, region = "seccion")
```

Una vez obtenidas las bases de datos en objetos data frame, usamos programación clásica en R para incluir identificadores de región como localidad, municipio o distrito y así visualizar los resultados de las elecciones en las diferentes regiones del estado de Zacatecas. El código R

que nos permite generar completamente estas bases o data frames, que denominamos *dfelec* para los resultados de las elecciones y *secciones* para las polígonos, se puede descargar de la liga:

<https://github.com/ivanps/Mapas-Electorales/blob/master/Data-Frames-Mapas.R>.

2.2. Librería ggplot2

La librería ggplot2 está basada en la *gramática de gráficas* de Wilkinson (2006) e implementada por Wickham (2009). ggplot2 es una herramienta muy popular y útil para la exploración y visualización de datos. Su enfoque de manejar los componentes de un gráfico por partes y la forma de construir el gráfico en serie de capas permite construir gráficos más sofisticados y presentables. Por ejemplo el código para generar el gráfico de la Figura 1 se muestra abajo.

```
mapsecc <- filter(secciones, id_dtofed == 4)
colores <-c("#FFFF00", "#339900", "#FF0033", "#660099", "#990000", "#999999", "#FFFFFF")
ggplot(data=mapsecc, aes(map_id=id)) +
  geom_map(map=mapsecc, fill="white", size=0.2) +
  expand_limits(x = mapsecc$long, y = mapsecc$lat) +
  geom_map(map=mapsecc, aes(fill=partido), size=0.2) +
  scale_fill_manual(values=colores) + coord_equal() + xlab("Longitud") + ylab("Latitud")
```

2.3. Librería ggmap

La librería ggmap extiende las capacidades de ggplot2 e integra la información de mapas estáticos de Google Maps, OpenStreetMap, Stamen Maps and CloudMade Maps. Esta librería fue desarrollada por Kahle y Wickman (2003). Usando esta librería podemos integrar las imágenes de caminos, calles o detalles topográficos a las imágenes generadas con ggplot2. Por ejemplo el código para generar el gráfico de la Figura 2 se muestra abajo. Primero se obtiene el mapa de Google con `get_map` (librería ggmap) y similar a ggplot se usa la función `ggmap`.

```
mapsecc <- filter(secciones, id_dtofed == 4)
```

```

colores <-c("#FFFF00", "#339900", "#FF0033", "#660099", "#990000", "#999999", "#FFFFFF")
region_map <- get_map(location = c(mean(mapsecc$long), mean(mapsecc$lat)), zoom =
  9, maptype = "roadmap")
ggmap(region_map) +
  geom_polygon(aes(x=long, y=lat, group=id, fill=partido), data=region, color="black",
  size = .1, alpha = .3) +
  scale_fill_manual(values=colores[table(mapsecc$partido) > 0]) +
  xlab("Longitud") + ylab("Latitud") + theme(legend.position="top")

```

2.4. Librería shiny

La librería **shiny** nos permite hacer aplicaciones web interactivas (app) que el usuario puede utilizar sin necesidad de modificar o programar el código de R. Además es posible publicar estas aplicaciones con RStudio Server para que el usuario las pueda usar en forma remota.

La dificultad de entender o programar el código R puede ser un impedimento para que el usuario utilice o adopte R como una herramienta en la visualización de la información, pero con **shiny** esta dificultad se reduce considerablemente. Aunque para ello, el programador primero debe generar dos archivos R, uno que controla la interface del usuario y otro que contiene el código de la aplicación. Posteriormente se necesita un servidor en Linux que ejecute la aplicación remotamente.

Las aplicaciones que desarrollamos se pueden ejecutar de dos maneras: localmente desde RStudio usando el repositorio de github.com o usando el servidor de RStudio en [Shinyapps.io](https://shinyapps.io) (de uso limitado).

Para ejecutar localmente las aplicaciones en RStudio es necesario instalar previamente todas las librerías que se indiquen, y posteriormente ejecutar los comandos que se indican abajo. Probamos este código usando la versión 3.4.1 de R, y las versiones de las librerías **shiny** 1.0.3, **ggplot2** 2.2.1, **ggmap** 2.6.1 y **dplyr** 0.7.2.

```

library(shiny)
runGitHub("Mapas-Electorales", "ivanps", subdir = "shape-voto")
runGitHub("Mapas-Electorales", "ivanps", subdir = "ggmap-voto")
runGitHub("Mapas-Electorales", "ivanps", subdir = "ggmap-partidos")

```

El código se muestra en la misma ejecución de la aplicación.

También es posible ejecutar las aplicaciones desarrolladas desde el servidor de RStudio en Shinyapps.io (plan gratuito de 25 horas mensuales) o instalando su propio servidor con RStudio Server. No es necesario instalar software o ejecutar algún código o comando, solamente requerimos acceder a las siguientes ligas desde cualquier navegador o dispositivo conectado a la internet: <https://pach.shinyapps.io/shape-voto/>, <https://pach.shinyapps.io/ggmap-voto/>, y <https://pach.shinyapps.io/ggmap-partidos/>.

3. Elecciones Gubernamentales

El objetivo de analizar los resultados de las elecciones gubernamentales es visualizar de forma interactiva las preferencias de los votantes por región e identificar aglomeraciones partidistas donde el usuario tenga la posibilidad que *jugar* con la información y obtener sus propias conclusiones.

La Figura 1 muestra el gráfico coloreado del shapefile del Distrito Federal IV y a la izquierda se observa el menú del usuario en la aplicación shiny. El usuario puede seleccionar si desea visualizar algún otro distrito federal, pero también puede seleccionar una localidad, municipio, distrito local o todo el estado. Cada sección está coloreada por el partido ganador en esa sección.

También usamos la librería ggmap para incluir información topográfica al gráfico para que el usuario pueda relacionar mejor los resultados visualizado con la región correspondiente. Podemos comparar la salida que se obtiene de esta manera en la Figura 2.

La Figura 3 muestra un diagrama de burbujas de la votación por partido, donde el color de la burbuja representa el partido ganador en esa sección. Podemos observar que al norte y suroeste del estado existe una baja densidad poblacional, que corresponden a la zona desértica y montañosa del estado, respectivamente. Con respecto a la votación notamos que las zonas más pobladas o centros urbanos tuvieron una alta votación por el partido MORENA, aunque decayó considerablemente en las zonas rurales, donde hubo una mayor preferencia por el partido del PRI. Con la aplicación shiny desarrollada el usuario puede además maximizar alguna región de interés seleccionando la longitud o latitud deseada.

La Figura 4 muestra donde se concentró la votación por el partido PT, aquí el tamaño de la burbuja indica el total de votos en la sección. Para identificar esta concentración se

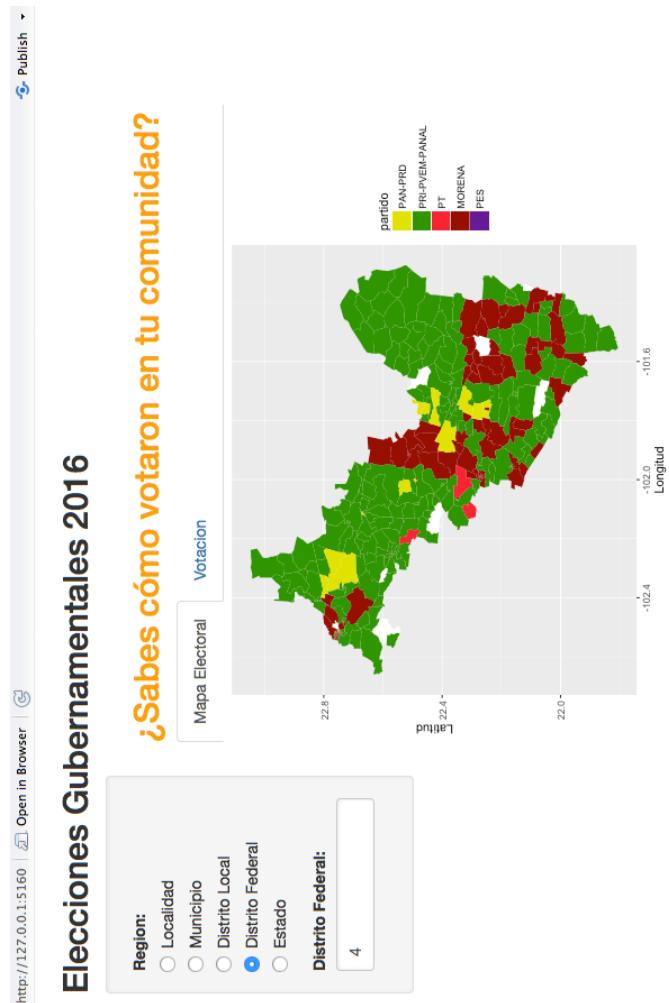


Figura 1: Resultados de la elección a gobernador en el Distrito Federal IV.

visualizan solamente aquellas secciones donde al menos el 20 % de los votantes prefirieron al PT. Podemos identificar una mayor preferencia por este partido en los municipios Loreto, Juan Aldama, Nochistlán y en menor grado en Zacatecas.

4. Conclusiones

Sin duda, el uso de herramientas llamativas e interactivas facilitan la visualización de la información y la posible identificación de patrones, y podrían ser un catalizador para que las personas en general se involucren en el análisis de los datos y a la postre tomemos mejores

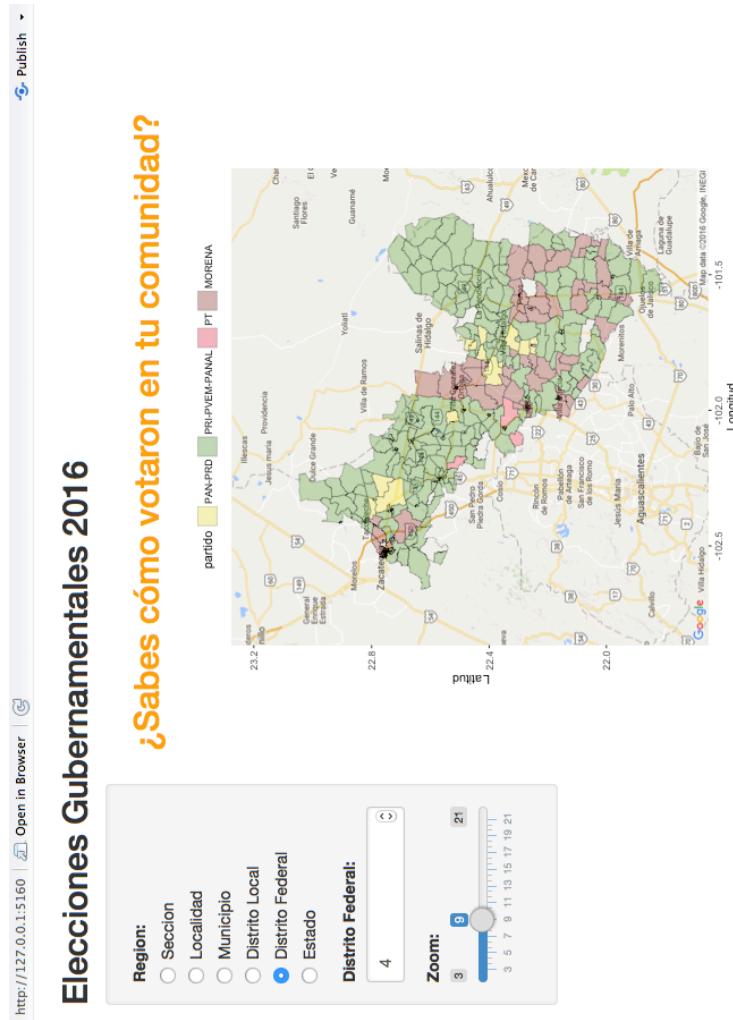


Figura 2: Visualización de la votación en el Distrito Federal IV con ggmap.

decisiones. El lenguaje R puede considerarse un poco técnico y académico, pero también incluye herramientas que facilitan su uso y su divulgación.

Bibliografía

Kahle, D. and Wickman, H. (2013). Spatial visualization with ggplot. *The R Journal*, 5(1):144–161. URL <http://journal.r-project.org/archive/2013-1/kahle-wickham.pdf>.

R Core Team (2016). *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

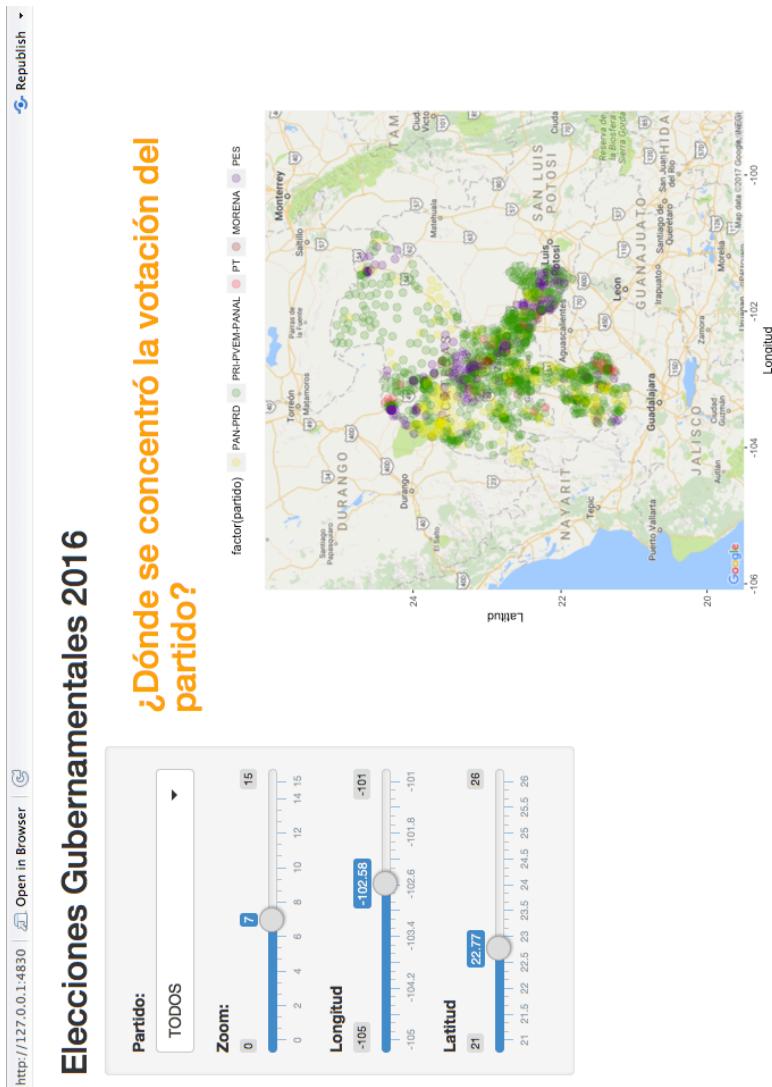


Figura 3: Diagrama de burbuja de la votación por partido político.

Wickham, H. (2009). *Elegant Graphics for Data Analysis*. Springer, New York. URL <http://ggplot2.org>.

Wilkinson, L. (2006). *The grammar of graphics*. Springer Science & Business Media.

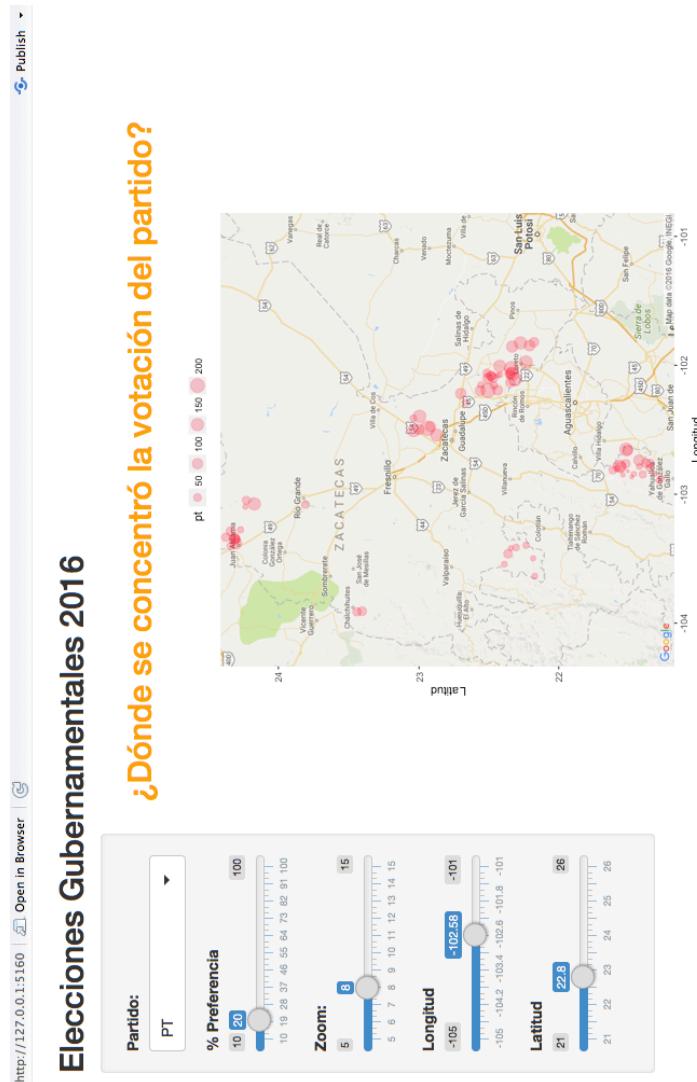


Figura 4: Diagrama de burbuja de la votación por el PT.

**Trabajos presentados en el
XXXII Foro Nacional de Estadística**

Análisis Estadístico de Trayectorias sobre la Esfera: un Caso de Estadística sobre Variedades*

Lilia Karen Rivera Escovar^a

Centro de Investigación en Matemáticas

El análisis estadístico sobre variedades es un tema de actualidad que se encuentra en la frontera de la estadística moderna. Ejemplos diversos se han desarrollado recientemente en el área de medicina y de biología, así como en otras ramas de la ciencia. Sin embargo, el asunto presenta ciertas dificultades teóricas, en virtud de que la metodología de \mathbb{R}^n no es aplicable. Esto es consecuencia de la estructura del espacio en donde se encuentran los datos de interés. No obstante se han logrado extender varias nociones estadísticas, un ejemplo de ello es la media muestral. La exposición planteada en el presente trabajo se especializará en el análisis estadístico de trayectorias sobre variedades Riemannianas, con un enfoque desarrollado principalmente sobre la esfera. Para fines de incursionar en la temática, se adoptó el artículo de Su *et al.* (2014) titulado *Statistical Analysis of Trajectories on Riemannian Manifolds: Bird Migration, Hurricane Tracking and Video Surveillance*. Finalmente, el presente trabajo proporcionará una breve motivación y una introducción al análisis estadístico sobre variedades, con la finalidad de afianzar la noción e importancia de esta temática. Aunado a lo anterior, conceptualizará un breve resumen del artículo base, complementado con un ejemplo de trayectorias de huracanes.

Área-MSN: Estadística.

Subárea-MSN: Variedades.

*Este trabajo fue realizado con el auspicio del Centro de Investigación en Matemáticas.

^arelk280988@gmail.com

1. Introducción

A lo largo de la historia el ser humano ha intentado entender el entorno que le rodea, con la finalidad de poder hacer pronósticos y contar con herramientas para la mejora de toma de decisiones desde ámbitos sociales hasta ambientales. Es por lo anterior que se ha dado a la tarea, particularmente en los últimos años, de analizar datos “comunes” con otras perspectivas, pues se ha percatado que hay datos que en sí mismos poseen cierta complejidad y por ende ha visto la necesidad de tratarlos con teoría distinta a la que se conoce para \mathbb{R}^n . Estos datos se caracterizan por ser elementos de espacios más abstractos que el n -dimensional, los cuales son conocidos como variedades no lineales.

Definición 1.1 (Variedad). *Una variedad topológica M es un espacio Hausdorff, segundo numerable que localmente es un espacio euclídeo. Se dice que es de dimensión n si localmente es un espacio Euclídeo de dimensión n .*

Intuitivamente una variedad podría entenderse como un espacio conformado exclusivamente por “parches” de \mathbb{R}^n . Tómese como ejemplo el toro, Figura 1, el cual tiene dimensión dos porque localmente se puede aproximar con “parches” de \mathbb{R}^2 .

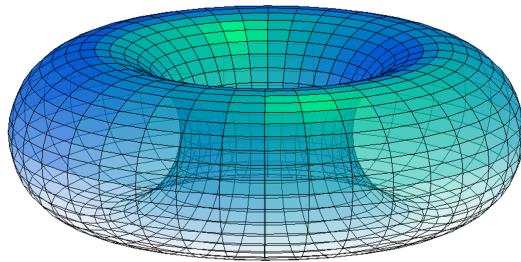


Figura 1: Toro.

Hacer estadística sobre espacios no lineales posee complejidad, debido a que no son espacios vectoriales. Por lo tanto, las herramientas estadísticas desarrolladas para \mathbb{R}^n no funcionan. A continuación se muestra el concepto de media muestral, el cual se ha logrado extender a variedades.

Definición 1.2 (Media de Fréchet o Karcher). *Sea M una variedad y sea $\{x_i\}_{i=1}^n$ una*

colección de puntos tales que $x_i \in M$ para $i = 1, \dots, n$. La media de Fréchet se define como:

$$\mu = \arg \min_{p \in M} \sum_{i=1}^n d(p, x_i)^2, \quad (1)$$

donde $d(\cdot, \cdot)$ representa la distancia definida en M .

En otras palabras, $p \in M$ es el punto que minimiza la distancia entre todos los datos $x_i \in M$. Esta noción de media surge con Fréchet (1948), que es el artículo pionero en definir el concepto de media sobre variedades, mientras que Karcher (1977) es el primero en ofrecer un estudio acerca de sus propiedades.

La incursión de la estadística en el marco de geometría diferencial, ha sido abordada por algunos libros. Uno de ellos es Amari (2012), quien ofrece una de las primeras referencias en tratar esta sinergia. No obstante, a pesar de la existencia de libros como el ya comentado, todavía no existe una cantidad significativa de libros que aborden el análisis estadístico sobre variedades. En este trabajo dicho análisis estará particularizado al estudio de trayectorias sobre la esfera. La motivación para trabajar con trayectorias es la diversidad de aplicaciones que engloba; por ejemplo, análisis de imágenes, reconocimiento de patrones, animación del cuerpo humano, así como otros ejemplos que se pueden encontrar en Joshi *et al.* (2016). La restricción de trabajar en la esfera se debe a que los resultados del proceso estadístico se pueden visualizar, por lo cual son más sencillos de interpretar y entender. Por otra parte, se facilitará el cómputo, pues al ser la esfera una de las variedades más estudiadas, se cuenta con expresiones analíticas cerradas para algunas nociones geométricas de interés.

2. Análisis Estadístico de Trayectorias sobre la Esfera

El análisis estadístico de trayectorias tiene su origen con Trouvé y Younes (2000). Sin embargo, es hasta Su *et al.* (2014) con *Statistical Analysis of Trajectories on Riemannian Manifolds: Bird Migration, Hurricane Tracking and Video Surveillance*, que surge el primer artículo en abordar un estudio estadístico de trayectorias sobre variedades Riemannianas. El presente artículo se caracteriza por usar nociones maduras de probabilidad y estadística, así como por concebir a la trayectoria como un dato. Además, logra una conjunción del marco teórico de geometría diferencial con el de probabilidad y estadística. Lo anterior se traduce en la implementación de la teoría abordada y con ello en el estudio de algunos casos, tales como

el análisis de trayectorias de vehículos y de actividad humana. En otras palabras, Su *et al.* (2014) es un artículo que innovó la representación y estudio de trayectorias sobre variedades. Por consiguiente y después de una extensa búsqueda bibliográfica, se adoptó esta referencia como base para el desarrollo del presente trabajo.

El análisis estadístico de trayectorias ha sido emprendido con diferentes perspectivas. Una de ellas versa en el estudio del tiempo con el que fue recorrida la trayectoria. Este enfoque, a su vez, se divide en dos vertientes: considerar tiempos aleatorios o no aleatorios en el estudio estadístico. La segunda vertiente es la más común y se clasifica dentro del análisis estadístico tradicional de trayectorias; sin embargo presenta ciertas deficiencias como no representatividad en la trayectoria media. De acuerdo con lo anterior surgió el estudio de trayectorias ocupando tiempos aleatorios. Dicho planteamiento se puede motivar con la migración de aves y el seguimiento de huracanes. En el caso de la migración de aves, a pesar de que una parvada siga la misma curva, no necesariamente vuela con la misma velocidad. Lo mismo ocurre con los huracanes; dos huracanes pueden tener la misma curva, y sin embargo pueden estar asociados a diferentes intensidades de recorrido y corresponder a diferentes años de registro. Esto quiere decir que se involucra cierta aleatoriedad temporal al observar las trayectorias. En consecuencia, al incorporarla en un estudio estadístico, se observan resultados que hacen más sentido con la intuición. No obstante, dado el reciente desarrollo de esta teoría, presenta algunas dificultades. Es importante mencionar que esta perspectiva de estudio constituye una de las principales aportaciones del artículo Su *et al.* (2014).

A continuación se muestra un análisis gráfico comparativo de los dos enfoques expuestos. Se tomó como caso de estudio un conjunto conformado por 35 trayectorias que representan la migración del halcón de Swainson. Estas trayectorias fueron observadas durante el período que comprende de 1995 a 1997. La Figura 2 muestra el conjunto de trayectorias de esta especie durante su período de migración. La Figura 3 muestra la trayectoria media bajo las dos perspectivas de análisis: en la izquierda se muestra el resultado con el enfoque usual, mientras que en la derecha el resultado considerando tiempos aleatorios. La Figura 4 representa, mediante elipses y círculos (análisis tradicional y análisis con tiempos aleatorios, respectivamente), las varianzas puntuales.



Figura 2: Conjunto de trayectorias del halcón de Swainson durante su época de migración.



Figura 3: Trayectoria media del halcón Swainson.



Figura 4: Varianzas puntuales asociadas al conjunto de trayectorias del halcón Swainson.

Como se puede apreciar, para esta muestra de trayectorias, el análisis propuesto por Su *et al.* (2014) arroja resultados que concuerdan con la intuición estadística. Lo anterior es en el sentido de que la curva o traza asociada a la trayectoria media es compatible con la curva de las trayectorias individuales, a diferencia de la media que se obtuvo vía el análisis clásico. De esa misma forma, las varianzas puntuales crecen conforme las trayectorias se van

desfasando entre sí, contrariamente a las varianzas que se obtienen con el enfoque tradicional. Estos resultados muestran que en un estudio estadístico de trayectorias (las cuales poseen variabilidad temporal y una forma particular) el desarrollo de la teoría propuesta por el artículo citado es pertinente.

2.1. Trayectoria Media

El algoritmo con el cual se obtendrá dicha trayectoria estará basado principalmente en la siguiente función objetivo:

$$h_\mu = \arg \min_{[h_\alpha] \in \mathcal{H}/\sim} \sum_{i=1}^n d_s([h_\alpha], [h_{\alpha_i}])^2. \quad (2)$$

La función (2) es análoga a la función (1), que es la media de Karcher para datos puntuales que se encuentran en una variedad M . Las piezas que cambian, en esta nueva función, son la distancia y los elementos sobre los cuales se realizará el proceso de minimización. Por tanto, la intuición de esta media sigue siendo encontrar aquel elemento $[h_\alpha]$ en \mathcal{H} , bajo la relación de equivalencia \sim , que minimice la distancia entre los elementos $[h_{\alpha_i}]$ que pertenecen a dicho espacio. Es valioso comentar que $[h_{\alpha_i}]$ representa la clase de equivalencia de los TSRVFs asociados a la trayectoria $\alpha_i(t)$. El TSRVF, o *Transported Square Root Vector Field*, es un concepto de geometría diferencial que ayudará a representar las trayectorias de interés de un espacio a otro, por ejemplo, de la esfera unitaria S^2 a un espacio lineal, como el plano $T_c S^2$. El TSRVF tiene la siguiente definición.

Definición 2.1. *Para cualquier trayectoria $\alpha(t) \in M$, el TSRVF es el transporte paralelo del campo vectorial de velocidades escaladas de una trayectoria $\alpha(t)$ a un punto de referencia $c \in M$ de acuerdo con $h_\alpha(t) = \left((\dot{\alpha}(t)_{\alpha(t)} \rightarrow c) / \left(\sqrt{|\dot{\alpha}(t)|} \right) \right) \in T_c M$. En este caso $|\cdot|$ denota la norma relacionada con la métrica intrínseca de la variedad M , $\dot{\alpha}(t)$ denota la derivada de la curva $\alpha(t)$ con respecto a t , $\alpha(t) \rightarrow c$ repesenta la geodésica que va de $\alpha(t)$ a c y $T_c M$ es el espacio tangente a M en un punto c .*

El siguiente algoritmo explica el procedimiento para encontrar la trayectoria media de un conjunto de trayectorias.

Trayectoria media de un conjunto de trayectorias $\{\alpha_i(t)\}_{i=1}^n$

Datos de entrada:

- El conjunto de trayectorias observadas $\{\alpha_i(t)\}_{i=1}^n$.
- Un punto de referencia c .

Las trayectorias $\{\alpha_i(t)\}_{i=1}^n$ deben de ser suaves y no pasar por el punto antípodo a c .

Datos de salida:

- Trayectoria media $\mu(t)$.
- El conjunto de trayectorias $\{\alpha_i(t)\}_{i=1}^n$ alineadas.

Pasos:

1. Encontrar la media de Fréchet de los puntos $\{\alpha_i(0)\}_{i=1}^n$. A este punto se le denominará como $\mu(0)$. Es fundamental aclarar que únicamente para este paso será usada la métrica de la variedad M con la que se esté trabajando. En este caso la distancia de la esfera unitaria.
2. Del conjunto de trayectorias $\{\alpha_i(t)\}_{i=1}^n$ seleccionar una trayectoria como $\mu(t)$. Posteriormente hallar $h_\mu(t)$, es decir el TSRVF de $\mu(t)$. En este paso es que se requiere el punto de referencia c , pues es el lugar donde se hará el TSRVF de las trayectorias $\{\alpha_i(t)\}_{i=1}^n$ es $T_c S^2$. Donde $T_c S^2$ representa el plano tangente a la esfera en el punto c .
3. Obtener $h_{\alpha_i}(t)$ para $i = 1, \dots, n$.
4. Alinear cada $h_{\alpha_i}(t)$ con base en h_μ . Para el desarrollo de este paso se requerirá encontrar la función de deformación temporal, $\gamma_i^*(t)$, que satisfaga la siguiente igualdad $\gamma_i^* = \arg \min_{\gamma_i \in \Gamma} \left(\int_0^1 |h_\mu(t) - h_{\alpha_i}(\gamma_i(t)) \sqrt{\dot{\gamma}_i(t)}|^2 dt \right)^{\frac{1}{2}}$.

En general γ_i^* se puede entender como una reparametrización de $h_{\alpha_i}(t)$. De acuerdo con lo anterior es que este paso requiere de una mayor capacidad de cómputo, pues se busca una reparametrización de $h_{\alpha_i}(t)$ que aproxime a $h_\mu(t)$. Por lo tanto se necesita el uso de algoritmos numéricos.

5. Obtener $\tilde{\alpha}_i = \alpha_i \circ \gamma_i^*$, tal que $i = 1, \dots, n$. En este caso $\{\tilde{\alpha}_i(t)\}_{i=1}^n$, representará el conjunto de trayectorias alineadas. También se aclara que en el caso de la trayectoria α_i que fue elegida como la trayectoria media se tiene que $\tilde{\alpha}_i = \alpha_i(\text{Id}(t))$; es decir $\gamma_i^* = \text{Id}(t)$.
6. Hallar $h_{\tilde{\alpha}_i}(t)$, donde $i = 1, \dots, n$.
7. Actualizar $h_\mu(t)$, como una curva en $T_c S^2$, de acuerdo con $h_\mu(t) = \frac{1}{n} \sum_{i=1}^n h_{\tilde{\alpha}_i}(t)$.

Nótese que en este paso es donde se aprovecha al máximo que $T_c S^2$ es un espacio vectorial, ya que la media $h_\mu(t)$ se calcula igual que una media muestral en \mathbb{R}^n .

8. Regresar la trayectoria media a la variedad S^2 , vía la ecuación diferencial $d\mu(t)/dt = |h_\mu(t)| h_\mu(t)_{c \rightarrow \mu(t)}$, con condición inicial $\mu(0)$.
9. Encontrar $E = \sum_{i=1}^n d_s([h_\mu], [h_{\alpha_i}])^2 = \sum_{i=1}^n d_h(h_\mu, h_{\tilde{\alpha}_i})^2$ y revisar su convergencia. Si ésta no existe regresar al paso tres del presente algoritmo. Para mayores referencias de cómo se define la distancia $d_s(\cdot, \cdot)$ consultar Su *et al.* (2014).

Es relevante comentar que la función (2) decrece iterativamente hacia cero. Por lo tanto siempre convergerá, con lo cual se puede asegurar la existencia de una trayectoria media.

En términos sencillos, para el caso de la esfera, lo que plantea el algoritmo es representar las trayectorias que están en S^2 a un espacio lineal $T_c S^2$. Una vez que las trayectorias están en el plano $T_c S^2$ se realiza un proceso de alineación entre las trayectorias, tomando como base la trayectoria que se eligió como media. Dicha alineación se aplica a las trayectorias originales de la esfera, con la finalidad de hacerlas comparables. Estas trayectorias alineadas se representan nuevamente en $T_c S^2$ para obtener la trayectoria media, usando la definición de media muestral, y posteriormente mandarla a la esfera. Finalmente se calcula la suma de la distancia entre la trayectoria media y cada una de las trayectorias alineadas. Si la suma converge se detiene el algoritmo; de lo contrario se repite el proceso descrito tomando como referencia la trayectoria media que ya se tiene calculada.

2.2. Análisis Estadístico de Trayectorias de Huracanes

Con la finalidad de materializar y ejemplificar la utilidad de la teoría desarrollada, es que se decidió hacer una breve aplicación. En ese mismo sentido, se planteó para mostrar el transporte paralelo y la trayectoria media de datos reales. La aplicación será sobre ocho trayectorias de huracanes, pertenecientes al Océano Atlántico y que se obtuvieron del siguiente sitio de Internet: <http://weather.unisys.com/hurricane/atlantic/>. Dichas trayectorias corresponden a un huracán seleccionado de los años de 1857, 1887, 1892, 1909, 1910, 1917, 1933 y 1944. Éstas se pueden observar en la Figura 5. De acuerdo con lo anterior se procedió con la implementación del algoritmo para obtener la trayectoria media de los ocho huracanes. Los resultados se pueden observar en las Figuras 6, 7 y 8.

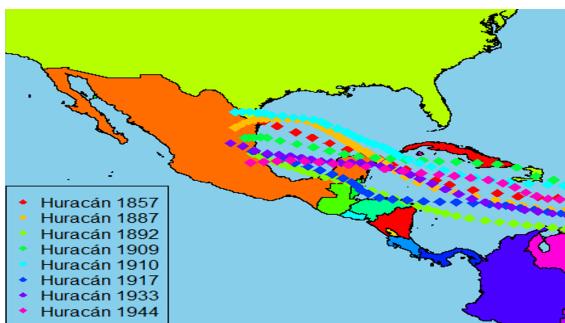


Figura 5: Trayectorias de ocho huracanes.

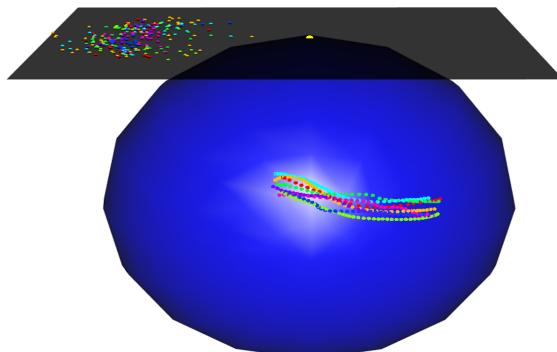


Figura 6: Trayectorias de los ocho huracanes en la esfera y TSRVF de éstas.

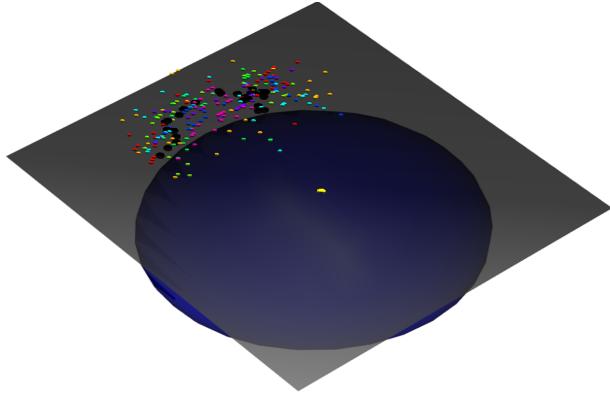


Figura 7: Puntos negros que representan la trayectoria media del conjunto de TSRVFs.

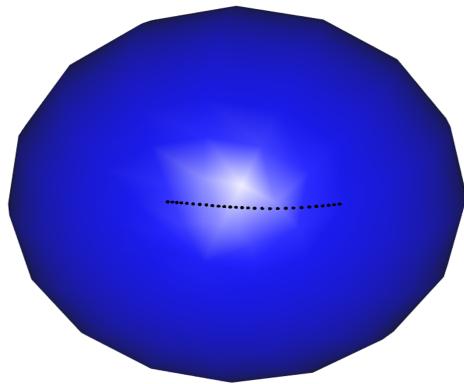


Figura 8: Trayectoria media de los ocho huracanes.

3. Conclusiones

La inserción en la temática de estadística sobre variedades presentó varios desafíos. Uno de ellos fue la labor computacional. En el área de estadística sobre variedades se carece de riqueza en cuanto a *software* implementado y accesible. De acuerdo con lo anterior uno de los mayores retos computacionales fue implementar desde principios básicos los conceptos de geometría diferencial, principalmente el transporte paralelo y aquellos que derivaron de éste.

Por otra parte, es importante mencionar que el concepto de transporte paralelo jugó un rol esencial en el algoritmo de la trayectoria media, pues gracias a éste se logró representar las trayectorias en un espacio lineal y por ende hacer uso de las herramientas que se conocen para \mathbb{R}^n , como es el caso de la media muestral.

Finalmente, el análisis estadístico sobre variedades es una área multidisciplinaria que posee diversas aplicaciones. En el presente texto se mostró una de ellas. Por lo tanto esta materia representa una área de oportunidad para estadísticos, computólogos, geométricas y todo aquél científico que desee realizar análisis estadístico con datos más complejos que aquellos producidos en el espacio n -dimensional.

Bibliografía

- Amari, S. I. (2012). *Differential-Geometrical Methods in Statistics*. New York: Springer.
- Fréchet, M. (1948). Les éléments aléatoires de nature quelconque dans un espace distancié. *Ann. Inst. H. Poincaré*, 10(3):215–310.
- Joshi, S. H., Su, J., Zhang, Z., and Amor, B. B. (2016). Elastic shape analysis of functions, curves and trajectories. In Turaga, P. and Srivastava, A., editors, *Riemannian Computing in Computer Vision*, pages 211–231. Springer, Cham.
- Karcher, H. (1977). Riemannian center of mass and mollifier smoothing. *Communications on Pure and Applied Mathematics*, 30(5):509–541.
- Su, J., Kurtek, S., Klassen, E., and Srivastava, A. (2014). Statistical analysis of trajectories on Riemannian manifolds: Bird migration, hurricane tracking and video surveillance. *The Annals of Applied Statistics*, 8(1):530–552.
- Trouvé, A. and Younes, L. (2000). Diffeomorphic matching problems in one dimension: Designing and minimizing matching functionals. In Vernon, D., editor, *Computer Vision-ECCV 2000*, pages 573–587. Springer.

Desempeño de Intervalos de Confianza para una Proporción y Criterios para su Aplicación

Marcos Morales Cortés^a, Hortensia J. Reyes Cervantes^b

Facultad de Ciencias Físico Matemáticas, Benemérita Universidad Autónoma de Puebla

Félix Almendra Arao^c

UPIITA, Instituto Politécnico Nacional

En la mayoría de las disciplinas del conocimiento es frecuente la realización de experimentos Bernoulli. La probabilidad de éxito (p) en ocasiones es desconocido, una forma de estimarlo es usando intervalos de confianza. El intervalo de confianza de Wald es el más utilizado para este tipo de estimaciones, sin embargo, diversos autores han demostrado que tiene serios problemas especialmente cuando p está cerca de 0 o de 1 así también cuando n es pequeño, e incluso se ha observado que las deficiencias continúan cuando p está cerca de 0.5 o cuando n es grande. Por estas razones diversos textos, al presentarlo le incluyen alguna condición para su uso. En este trabajo se analizó su desempeño con base en las condiciones más recurrentes a través de su probabilidad de cobertura y de su longitud esperada. Debido al amplio conocimiento del mal desempeño del intervalo de Wald, se han propuesto diversos intervalos que presentan un mejor desempeño. En este trabajo se consideran tres de ellos; el intervalo de Agresti-Coull, el intervalo de Wilson y el intervalo arcoseno. Tales intervalos alternos continúan presentando inconsistencias, en especial para p cerca de 0 y de 1, por esta razón se analizaron sus comportamientos tomando en cuenta los mismos criterios. A manera de conclusión se tiene que el intervalo de confianza de Wald es inestable y poco confiable en términos de probabilidades de cobertura, y lo mejor será usar un intervalo alternativo siendo el de Agresti-Coull para $n \geq 200$ y el de Wilson para $n < 200$.

Área-MSC: Estadística Inferencial

Subárea-MSC: Pruebas de hipótesis

^aaverandmeph@gmail.com.mx

^bhreyes@fcfm.buap.mx

^cfalmendra@ipn.mx

1. Introducción

En el presente trabajo se revisa el problema de la estimación por intervalo de la probabilidad de éxito p en una distribución binomial. En Khurshid y Ageel (2010) se presenta una amplia y detallada revisión de la literatura sobre el cálculo de intervalos de confianza de las distribuciones binomial y Poisson, disponible hasta el año 2010. Sin embargo, de acuerdo con Schilling y Doi (2014), la obtención de un intervalo de confianza óptimo para p continúa sin resolverse.

En la mayoría de libros elementales de estadística, se presenta al intervalo de confianza de Wald, el cual ha adquirido aceptación casi universal en la práctica. Sin embargo, diversos autores, como Agresti y Coull (1998), Agresti y Caffo (2000) y Brown *et al.* (2001), han demostrado que dicho intervalo presenta serios problemas cuando p está cerca de 0 o 1 y cuando n es pequeño. Sin embargo, la probabilidad de cobertura del intervalo de Wald aún puede estar muy por debajo del coeficiente de confianza incluso si p está cercano a 0.5 y también cuando n es grande. Por estas razones, en diversos textos al presentarlo le incluyen una condición, esto con la finalidad de mejorar su desempeño.

Se ha realizado una gran cantidad de propuestas de intervalos de confianza para un parámetro binomial en los más de 80 años desde el desarrollo original de los intervalos de confianza, que continúan incluso en el siglo XXI. En este trabajo se analizó el desempeño de algunas propuestas.

1.1. Revisión de la Literatura

En Agresti y Coull (1998) se muestra que los intervalos exactos además de ser conservadores, sus probabilidades de cobertura pueden ser bastante mayores al nivel de confianza nominal $1 - \alpha$. Y que los intervalos de Agresti-Coull y de Wilson en ocasiones pueden tener probabilidades de cobertura inferiores a $1 - \alpha$, pero la probabilidad de cobertura es cercana a ese nivel, concluyendo, para la mayoría de las aplicaciones se deben preferir los intervalos aproximados (Wilson y de Agresti-Coull) en lugar de los intervalos exactos. Brown *et al.* (2001) sugiere que el intervalo de Wilson o de Jeffreys sean usados para n pequeños ($n \leq 40$). En el caso de n más grandes recomiendan para $n \geq 40$ al intervalo de Agresti-Coull. Måns (2013) muestra que el costo de usar un intervalo exacto en lugar de un intervalo aproximado está dado en términos de su longitud esperada, siendo ésta más grande para los intervalos exactos. En

Schilling y Doi (2014) se menciona que una de las razones por la cual aún no se ha resuelto el problema es que se han usado dos estándares distintos, uno requiere que el ínfimo de las probabilidades de cobertura sea mayor o igual que $1 - \alpha$ y el otro permite que eso se cumpla sólo aproximadamente. Al mismo tiempo presentan el nuevo método LCO (Longitud/Cobertura Optima) que es óptimo con respecto a la longitud y la cobertura cuando se cumpla que el ínfimo de las probabilidades de cobertura sea mayor o igual que $1 - \alpha$, y proporcionan una versión aproximada que supera los procedimientos existentes para el criterio de cobertura aproximado.

El intervalo de Wald, proviene de la prueba de Wald para muestras grandes para el caso binomial y tiene la siguiente expresión:

$$[\hat{p} - \kappa n^{-1/2}(\hat{p}\hat{q})^{1/2}, \hat{p} + \kappa n^{-1/2}(\hat{p}\hat{q})^{1/2}], \quad (1)$$

donde X es el número de éxitos en n realizaciones, $\hat{p} = \frac{X}{n}$, $\kappa = z_{\frac{\alpha}{2}} = \Phi^{-1}(1 - \frac{\alpha}{2})$ y $\hat{q} = 1 - \hat{p}$, $\Phi(z)$ es la función de distribución de una variable aleatoria con distribución normal estándar.

Realizando manipulaciones algebraicas a la expresión (1) se obtiene, que la probabilidad de cobertura del intervalo de Wald para n y p ($\mathbf{PC}(n, p)$) puede ser calculada de la siguiente forma:

$$\mathbf{PC}(n, p) = \sum_{i=\lceil x \rceil}^{\lfloor y \rfloor} \binom{n}{i} p^i (1-p)^{n-i}, \quad (2)$$

donde $x = \frac{n}{1+\frac{k^2}{n}} \left(p + \frac{\frac{k^2}{n}}{2} - \sqrt{-\frac{k^2}{n}p^2 + \frac{k^2}{n}p + \frac{(\frac{k^2}{n})^2}{4}} \right)$, $y = \frac{n}{1+\frac{k^2}{n}} \left(p + \frac{\frac{k^2}{n}}{2} + \sqrt{-\frac{k^2}{n}p^2 + \frac{k^2}{n}p + \frac{(\frac{k^2}{n})^2}{4}} \right)$ y $\lceil x \rceil$, $\lfloor y \rfloor$ son las funciones techo y piso, respectivamente.

Las figuras siguientes muestran la probabilidad de cobertura del intervalo de Wald para casos específicos.

En la Figura 1(a), se muestra la probabilidad de cobertura del intervalo de Wald cuando $\alpha = 0.05$, para $p = 0.001$ y $n \in [1, 10000]$. Se observa que la oscilación es considerable y que la probabilidad de cobertura se acerca a 0.95 pero no de manera monótona, se presentan saltos extremos en su probabilidad de cobertura, por ejemplo, en 2959 y 4771. Se aprecia que la probabilidad de cobertura es deficiente para $p = 0.001$, aún con tamaños de muestra (n) bastante grandes. En (b) se muestra la probabilidad de cobertura del intervalo de Wald con $\alpha = 0.01$, para $n = 20$ y $p \in \{0.001, 0.002, \dots, 0.999\} = \mathcal{P}$. Se observa que no hay algún p ($p \in \mathcal{P}$) para el cual $\mathbf{PC}(n, p) \geq 0.99$, el valor más cercano a 0.99 se alcanza en $p = 0.272$.

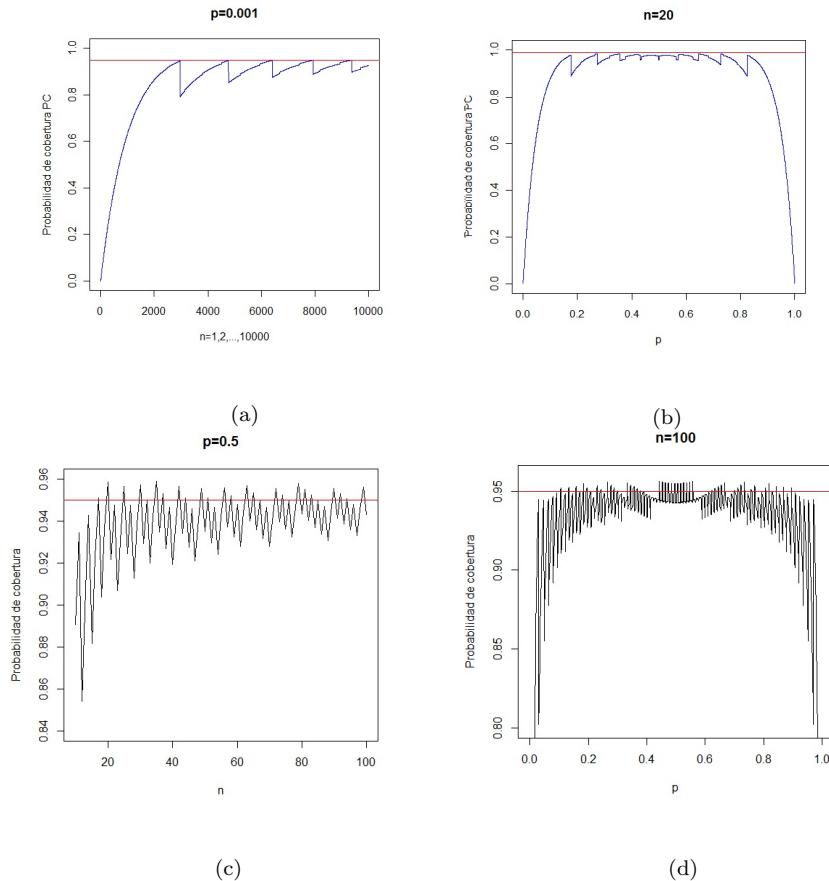


Figura 1: Probabilidad de cobertura del intervalo de Wald, (a) $p = 0.001$ y (b) $n = 20$.

(por simetría también en $p = 0.728$) y es 0.9831, su valor promedio es sólo de 0.8835. En (c) se muestra la probabilidad de cobertura del intervalo de Wald con $\alpha = 0.05$, para $p = 0.5$ y $n \in [1,100]$. Se observa que su desempeño mejora notablemente para $p = 0.5$ y que el grado de oscilación disminuye, aún continúan presentándose inconsistencias, por ejemplo, $n = 17$, $\mathbf{PC}(17, 0.5) = 0.951$ pero en $n = 40$, $\mathbf{PC}(40, 0.5) = 0.9193$. En (d) se muestra la probabilidad de cobertura del intervalo de Wald cuando $\alpha = 0.05$, para $n = 100$ y $p \in \mathcal{P}$. Se observa que para la mayoría de p , con $p \in \mathcal{P}$, $\mathbf{PC}(100, p) < 0.95$ y su comportamiento es bastante inestable, además aún se siguen presentando probabilidades de cobertura pequeñas para p cercanos a 0 o a 1, por ejemplo $\mathbf{PC}(100, 0.001) = 0.0952$. La probabilidad de cobertura solo es razonable cuando p está cerca de 0.5.

El comportamiento errático de la probabilidad de cobertura del intervalo de Wald es ampliamente conocido en la literatura, debido a esto, en diversos textos al presentarlo agregan un criterio para mejorar su desempeño. De una muestra de 11 textos populares sobre estadística, Brown *et al.* (2001) obtuvieron una lista de criterios, entre ellas destacan:

- **Criterio 1** $np \geq 5$ y $n(1 - p) \geq 5$;
- **Criterio 2** $np \geq 10$ y $n(1 - p) \geq 10$;
- **Criterio 3** $np(1 - p) \geq 5$;
- **Criterio 4** $np(1 - p) \geq 10$;
- **Criterio 5** $n \geq 100$;
- **Criterio 6** $n \geq 50$ y $0.2 < p < 0.8$.

2. Marco Teórico

Con base en los criterios anteriores, se definen las siguientes *variables de comparación* mediante las cuales el desempeño de los intervalos de confianza es analizado.

Sea \mathbb{N} el conjunto de números naturales y sea $\mathcal{N} = [k, k+m]$ con $k, m \in \mathbb{N}$. Para $n \in \mathcal{N}$ y $p \in \mathcal{P}$, se definen los siguientes conjuntos:

$\mathcal{M}_i = \{(n, p) \in \mathcal{N} \times \mathcal{P} \mid (n, p) \text{ cumple el criterio } i\}$, $\mathcal{A}_i = \{(n, p) \in \mathcal{M}_i \mid \mathbf{PC}(n, p) \geq 1 - \alpha\}$. Entonces el *porcentaje de puntos adecuados para el criterio i* ($\mathcal{P}\mathcal{A}_i$) se define de la siguiente manera:

$$\mathcal{P}\mathcal{A}_i = 100 * \frac{\text{Card}(\mathcal{A}_i)}{\text{Card}(\mathcal{M}_i)}. \quad (3)$$

Sea $\mathcal{I}_i = \{(n, p) \in \mathcal{M}_i \mid \mathbf{PC}(n, p) < 1 - \alpha\}$, la *probabilidad de cobertura promedio por defecto* para el criterio i ($\mathcal{P}\mathcal{CPD}_i$) se define como:

$$\mathcal{P}\mathcal{CPD}_i = \sum_{(n,p) \in \mathcal{I}_i} \mathbf{PC}(n, p). \quad (4)$$

La *probabilidad de cobertura promedio por exceso* para el criterio i ($\mathcal{P}\mathcal{CP}\mathcal{E}_i$) se define como:

$$\mathcal{P}\mathcal{CP}\mathcal{E}_i = \sum_{(n,p) \in \mathcal{A}_i} \mathbf{PC}(n, p). \quad (5)$$

La longitud esperada **LE** para n y p ($\mathcal{L}\mathcal{E}(n, p)$) es:

$$\mathcal{L}\mathcal{E}(n, p) = \sum_{x=0}^n (U(x, n) - L(x, n)) \binom{n}{x} p^x (1-p)^{n-x},$$

$U(x, n)$ y $L(x, n)$ son los límites superior e inferior del intervalo de confianza, respectivamente. La *longitud media esperada* para el criterio i se define como:

$$\mathcal{LME} = \sum_{(n,p) \in \mathcal{N} \times \mathcal{P}} \mathcal{LE}(n, p). \quad (6)$$

Agresti y Min (2001) concluyen que al usar variables discretas se obtienen comportamientos inesperados en los intervalos de confianza y esto es independiente del método usado para construirlos. Por esta razón en este trabajo se considera que un intervalo de confianza tiene un buen desempeño cuando: (i) las probabilidades de cobertura promedio por defecto y por exceso estén cercanas al nivel de confianza nominal y (ii) cuando la longitud media esperada sea pequeña.

3. Resultados

A continuación se presentan otros intervalos para una proporción que resaltan en la literatura debido a que se ha demostrado que tienen un mejor desempeño que el intervalo de Wald.

El intervalo de Wilson, se obtiene también al invertir la región de aceptación de la prueba de Wald para muestras grandes solo que en lugar de usar $(\hat{p}\hat{q})^{1/2}n^{-1/2}$ se utiliza $(pq)^{1/2}n^{-1/2}$,

$$\left[\frac{X + \kappa^2/2}{n + \kappa^2} - \frac{\kappa n^{1/2}}{n + \kappa^2} \left(\hat{p}\hat{q} + \frac{\kappa^2}{4n} \right)^{1/2}, \frac{X + \kappa^2/2}{n + \kappa^2} + \frac{\kappa n^{1/2}}{n + \kappa^2} \left(\hat{p}\hat{q} + \frac{\kappa^2}{4n} \right)^{1/2} \right].$$

El intervalo de Agresti-Coull, tiene una forma familiar al intervalo de Wald usando un nuevo estimador \tilde{p} en lugar de \hat{p} . Sea $\tilde{X} = X + \frac{\kappa^2}{2}$, $\tilde{n} = n + \kappa^2$, $\tilde{p} = \frac{\tilde{X}}{\tilde{n}}$ y $\tilde{q} = 1 - \tilde{p}$,

$$\left[\tilde{p} - \kappa \tilde{n}^{-1/2} (\tilde{p}\tilde{q})^{1/2}, \tilde{p} + \kappa \tilde{n}^{-1/2} (\tilde{p}\tilde{q})^{1/2} \right].$$

El intervalo Arcoseno, está basado en la estabilización de varianza para la distribución binomial $T(\hat{p}) = \text{arcoseno}(\hat{p}^{1/2})$, basada en el método delta. Si se remplaza \hat{p} por $\check{p} = (X + 3/8)/(n + 3/4)$ se obtiene una mejor estabilización de varianza y conduce al intervalo

$$[\sin^2(\arcsin(\check{p}^{1/2}) - \frac{1}{2}\kappa n^{-1/2}), \sin^2(\arcsin(\check{p}^{1/2}) + \frac{1}{2}\kappa n^{-1/2})].$$

A continuación se analizará el comportamiento de los intervalos de Wald, Wilson, Agresti-Coull y Arcoseno con base en los criterios y en las expresiones 3, 4, 5 y 6. Para esto, sea

$\kappa = 0.05$, $i \in [1,6]$, y considere los siguientes conjuntos $A_i = [100i - 99, 100i]$, los resultados se presentan en las Tablas 1 y 2.

	Criterio 1 ($np \geq 5$ y $n(1-p) \geq 5$)						Criterio 2 ($np \geq 10$ y $n(1-p) \geq 10$)						Criterio 3 ($np \geq 10$ y $n(1-p) \geq 5$)						$\mathcal{P}_{\mathcal{A}}$					
	Intervalo de Wald			Intervalo de Wilson			Intervalo de Agresti-Coull			Intervalo Arcoseno			$\mathcal{P}_{\mathcal{A}}$			\mathcal{PCPD}			\mathcal{PCPE}			\mathcal{LME}		
	$\mathcal{P}_{\mathcal{A}}$	\mathcal{PCPD}	\mathcal{PCPE}	\mathcal{LME}	$\mathcal{P}_{\mathcal{A}}$	\mathcal{PCPD}	\mathcal{PCPE}	\mathcal{LME}	$\mathcal{P}_{\mathcal{A}}$	\mathcal{PCPD}	\mathcal{PCPE}	\mathcal{LME}	$\mathcal{P}_{\mathcal{A}}$	\mathcal{PCPD}	\mathcal{PCPE}	\mathcal{LME}	$\mathcal{P}_{\mathcal{A}}$	\mathcal{PCPD}	\mathcal{PCPE}	\mathcal{LME}	$\mathcal{P}_{\mathcal{A}}$	\mathcal{PCPD}	\mathcal{PCPE}	\mathcal{LME}
A_1	10.69	0.9326	0.9527	0.2383	54.83	0.9439	0.9567	0.2307	67.22	0.9451	0.9582	0.2332	55.46	0.9436	0.9568	0.2331	52.82	0.9458	0.9543	0.1618	52.16	0.9466	0.9534	0.1244
A_2	17.71	0.9399	0.9522	0.1329	53.41	0.9462	0.9543	0.1322	64.9	0.9468	0.9558	0.1331	52.82	0.9458	0.9543	0.1618	51.15	0.9471	0.9533	0.1049	51.22	0.9474	0.9527	0.0925
A_3	21.66	0.9424	0.9519	0.1005	52.88	0.9469	0.9555	0.1003	63.06	0.9474	0.9547	0.1008	52.16	0.9466	0.9534	0.1244	50.84	0.9485	0.9536	0.074	51.35	0.9476	0.9524	0.0836
A_4	24.31	0.9437	0.9517	0.0842	52.48	0.9473	0.953	0.084	62.05	0.9478	0.954	0.0843	51.15	0.9471	0.9533	0.1049	50.84	0.9482	0.9532	0.0667	51.35	0.9476	0.9524	0.0836
A_5	25.9	0.9446	0.9516	0.0739	52.18	0.9476	0.9527	0.0738	60.85	0.948	0.9536	0.074	51.22	0.9474	0.9527	0.0925	50.84	0.9482	0.9532	0.0667	51.35	0.9476	0.9524	0.0836
A_6	27.64	0.9451	0.9515	0.0666	52.4	0.9478	0.9525	0.0665	60.45	0.9482	0.9532	0.0667	51.15	0.9471	0.9528	0.1049	50.84	0.9482	0.9532	0.0667	51.35	0.9476	0.9524	0.0836

Tabla 1: Variables de comparación, $\kappa = 0.05$, criterios 1, 2 y 3.

De la Tabla 1, para el criterio 1 se tiene que para todo A_i los *porcentajes de puntos adecuados* mayores se logran con el intervalo de Agresti-Coull así también las *probabilidades de cobertura promedio por defecto* más cercanas a 0.95, pero éste provee las *probabilidades de cobertura promedio por exceso* más alejadas de 0.95, las más cercanas a 0.95 son obtenidas por los intervalos de Wilson y arcoseno. Sin embargo las *longitudes medias esperadas* del intervalo Arcoseno son demasiado grandes comparadas con las del intervalo de Wilson, las cuales son las mínimas. La diferencia entre las *probabilidades de cobertura promedio por*

exceso de los intervalos de Agresti-Coull y de Wilson son 0.0015 en A_1 y decrecen hasta 0.0007 en A_6 . Para el criterio 2, se obtienen las mismas observaciones que en el criterio 1, y las diferencias entre las *probabilidades de cobertura promedio por exceso* de los intervalos de Agresti-Coull y de Wilson varian entre 0.0007 y 0.0006. Para el criterio 3 se obtienen las mismas observaciones que en el criterio 1, la diferencia entre las *probabilidades de cobertura por exceso* de los intervalos de Agresti-Coull y de Wilson varian entre 0.0014 y 0.0007.

	Criterio 4 ($n\hat{p}(1-\hat{p}) \geq 10$)						Criterio 5 ($n \geq 100$)						Criterio 6 ($n \geq 50$ y $0.2 < p < 0.8$)							
	Intervalo de Wald			Intervalo de Wilson			Intervalo de Agresti-Coull			Intervalo Arcoseno			\mathcal{P}_{CPD}			$\mathcal{P}_{CP\bar{E}}$			\mathcal{LME}	
	\mathcal{P}_A	\mathcal{P}_{CPD}	$\mathcal{P}_{CP\bar{E}}$	\mathcal{LME}	\mathcal{P}_A	\mathcal{P}_{CPD}	$\mathcal{P}_{CP\bar{E}}$	\mathcal{LME}	\mathcal{P}_A	\mathcal{P}_{CPD}	$\mathcal{P}_{CP\bar{E}}$	\mathcal{LME}	\mathcal{P}_A	\mathcal{P}_{CPD}	$\mathcal{P}_{CP\bar{E}}$	\mathcal{LME}	\mathcal{P}_A	\mathcal{P}_{CPD}	$\mathcal{P}_{CP\bar{E}}$	\mathcal{LME}
A_1	15.13	0.9394	0.9527	0.2122	53.59	0.9449	0.9552	0.2075	61.41	0.9455	0.9557	0.2085	54.57	0.9449	0.9553	0.2316				
A_2	19.23	0.9421	0.9522	0.1384	52.95	0.9463	0.954	0.1373	62.26	0.9468	0.9547	0.1379	52.54	0.9461	0.954	0.1615				
A_3	23.62	0.9438	0.9519	0.1033	52.75	0.947	0.9532	0.1029	61.3	0.9474	0.954	0.1033	52.62	0.9468	0.9532	0.1244				
A_4	25.07	0.9448	0.9517	0.0859	52.27	0.9474	0.9528	0.0857	60.82	0.9478	0.9535	0.086	51.5	0.9473	0.9528	0.1049				
A_5	26.53	0.9454	0.9516	0.0751	51.98	0.9476	0.9525	0.075	59.89	0.948	0.9532	0.0752	51.54	0.9475	0.9525	0.0925				
A_6	28.19	0.9459	0.9515	0.0675	52.19	0.9478	0.9523	0.0674	59.67	0.9482	0.9529	0.0676	51.63	0.9477	0.9523	0.0836				

Tabla 2: Variables de comparación, $\kappa = 0.05$, criterios 4, 5 y 6.

De la Tabla 2 se tiene para el criterio 4, las mismas observaciones que en el criterio 1 y las diferencias entre las *probabilidades de cobertura por exceso* del intervalo de Agresti-Coull y de Wilson se encuentran entre 0.0006 y 0.0005. Para el criterio 5, se obtienen las mismas observaciones que en el criterio 1 y las diferencias entre las *probabilidades de cobertura por exceso* del intervalo de Agresti-Coull y de Wilson se encuentran entre 0.0035 y 0.0011. En el criterio 6 se obtienen las mismas observaciones que en el criterio 1 y las diferencias entre las *probabilidades de cobertura por exceso* del intervalo de Agresti-Coull y de Wilson se encuentran entre 0.0004 y 0.

4. Conclusiones

Aún en estos tiempos, el intervalo de Wald continúa siendo presentado en los cursos de estadística básica y los instructores únicamente sugieren alguna condición al momento de aplicarlo. La amplia literatura demuestra que el intervalo de Wald es mucho más inconsistente de lo que se piensa. Con base en este trabajo presentamos las conclusiones más sobresalientes.

1. Con los resultados de este trabajo se refuerzan las conclusiones de que el intervalo de Wald no tiene un buen desempeño en términos de probabilidades de cobertura (ya que las $\mathbf{PC}(n, p)$ se comportan de forma errática y muy por debajo del nivel nominal en muchos casos). También se concluye que los criterios que se sugieren para su aplicación no son del todo adecuados.
2. Los intervalos alternativos producen probabilidades de cobertura mejores que el intervalo de Wald y esto es más notable cuando p está cerca de 0 o 1 o cuando el tamaño muestral n es pequeño. Los criterios sugeridos para el intervalo de Wald mejoran el desempeño de los intervalos recomendados sobre todo los criterios 4 y 6 en los cuales las probabilidades de cobertura se sitúan bastante cerca del nivel de confianza y en el criterio 6 las probabilidades de cobertura obtenidas por los intervalos de Wilson y Agresti-Coull son muy similares y podría hacerse uso de ambos.
3. Para $p \leq 0.2$ o $p \geq 0.8$ en los cuales sea difícil o costoso cumplir con alguno de los criterios 4 o 6 se recomienda usar el intervalo de Wilson para $n < 200$ y el intervalo de Agresti-Coull $n \geq 200$.

4. Para p con $0.2 < p < 0.8$ y n grande $n \geq 100$ también se podría recomendar el intervalo Arcoseno, éste provee probabilidades de cobertura aceptables, sin embargo su *longitud esperada* es superior a los demás intervalos.

Bibliografía

Agresti, A. and Caffo, B. (2000). Simple and Effective Confidence Intervals for Proportions and Differences of Proportions Result from Adding Two Successes and Two Failures. *The American Statistician*, 54(4):280–288.

Agresti, A. and Coull, B. (1998). Approximate is Better than Exact for Interval Estimation of Binomial Proportions. *The American Statistician*, 52(2):119–126.

Agresti, A. and Min, Y. (2001). On Small-Sample Confidence Intervals for Parameters in Discrete Distribution. *Biometrics*, 57:963–971.

Agresti, A. and Minon, Y. (2002). On sample Size Guidelines for Teaching Inference about the Binomial Parameter in Introductory Statistics. Technical report, Department of Statistics, University of Florida, Gainesville, Florida.

Brown, L. D., Cai, T. T., and DasGupta, A. (2001). Interval Estimation for a Binomial Proportion. *Statistical Science*, 16:101–133.

Khurshid, A. and Ageel, M. (2010). Binomial and Poisson Confidence Intervals and its Variants: A Bibliography. *Journal Statistical and Operation Research*, VI(1):75–100.

Schilling, M. F. and Doi, J. A. (2014). A coverage probability approach to finding an optimal binomial confidence procedure. *The American Statistician*, pages 133–145.

Thulin, M. (2013). The cost of using exact confidence intervals for a binomial proportion. Technical report, Department of Mathematics, Uppsala University.

JRStat: Una Plataforma de Código Abierto para Implementación de Análisis Estadístico Usando Interfaces Gráficas*

Nallely Izel Bautista Pérez, Paulino Pérez Rodríguez^a
Colegio de Postgraduados, Campus Montecillo

El presente trabajo de investigación tiene como objetivo presentar el *software* JRStat, una plataforma de código abierto para análisis estadístico que combina el poder de cómputo de R con la facilidad de uso de las interfaces gráficas de usuario (GUI). Dicha plataforma utiliza el *software* R desde una aplicación escrita en el lenguaje de programación Java y las librerías rJava, JavaGD, JFreeChart. El *software* JRStat actualmente cuenta con algunos módulos: 1. Consola interactiva para ejecución de comandos en R, 2. Explorador de objetos, 3. Editor de código, 4. Rutinas para graficación, y 5. Menús para análisis básicos de datos (pruebas de hipótesis, regresión lineal, modelo lineal generalizado, análisis multivariado, etc.). Sin embargo, cabe destacar que el *software* JRStat es extensible, ya que es posible crear nuevos módulos para análisis de datos basados en archivos xml, además es multiplataforma y está disponible para su ejecución en las plataformas de cómputo modernas (Windows, MacOS y Linux). El programa permite a los usuarios de diversas disciplinas usar las herramientas de análisis estadístico sin tener conocimientos avanzados de programación, ya que la interacción entre el usuario y R se da de una manera dinámica a través de interfaces gráficas.

Área-*MSC*: Estadística.

Subárea-*MSC*: Cálculo y programas explícitos de la máquina.

1. Introducción

El análisis de grandes volúmenes de datos con herramientas de estadística es hoy en día algo muy útil y complejo por lo que se han desarrollado diversos *softwares* para facilitar esta

*Este trabajo fue realizado como parte de la tesis de maestría.

^aperpdgo@colpos.mx

tarea, como SPSS (IBM Corp., 2013), eviews (IHS Global Inc., 2015), Stata (StataCorp LLC, 1996), Minitab (Minitab Inc., 2017), SAS (SAS Institute Inc.), R (R Core Team, 2016), entre otros. Sin embargo, algunos de los *softwares* mencionados anteriormente son comerciales por lo que el acceso a ellos puede ser limitado. Afortunadamente, existe la alternativa del *software* libre como R.

R es un entorno para computación y gráficos estadísticos altamente extensible que en los últimos años ha incrementado su popularidad frente a otros *softwares* como SPSS, Stata o eviews. Y se ha convertido en la herramienta preferida para el análisis de datos, tanto con fines científicos como comerciales (Vance, 2009).

No obstante, para poder acceder a todas sus funcionalidades se requieren conocimientos de programación. Esto resulta complicado para aquellos usuarios que no han tenido oportunidad de utilizar algún lenguaje de programación como C (Kernighan y Ritchie, 1988) o C++ (Stroustrup, 1997), por ejemplo. Por tal motivo se han desarrollado interfaces gráficas que hacen más amigable la interacción entre el usuario y R.

JRStat es una plataforma de código abierto para análisis estadísticos que combina el poder de cómputo de R con la facilidad de uso de las Interfaces Gráfica de Usuario.

2. Marco Teórico

Una interfaz gráfica de usuario o GUI (*Graphical User Interface*), por sus siglas en inglés, es el medio principal para interactuar con los entornos de escritorio. Los recursos, como los documentos, están representados por íconos gráficos y los controles de usuario se incluyen en menús desplegables, botones, barras deslizantes, etc. El usuario puede manipular las ventanas, los íconos y los menús con un dispositivo, como un *mouse* o teclado (Lawrence y Verzani, 2012).

En otras palabras, una GUI es el conjunto de componentes gráficos que posibilitan y facilitan la interacción entre el usuario y la aplicación, como son íconos, ventanas, botones, combos, listas, cajas de diálogo y campos, en lugar de líneas de comandos.

El objetivo de una GUI es facilitar la manera en que el usuario interactúa con la información. En las computadoras, la forma de interactuar con los usuarios se lleva a cabo mediante el uso de ventanas, íconos, menús y cursor, o, para abreviar, WIMP (*windows, icons, menus and pointer*). La interacción WIMP generalmente se vale de un *mouse* y organiza la

información en ventanas; las ventanas están representadas con íconos; y los comandos están arreglados en menús. Finalmente, existe un sistema que permite minimizar, maximizar o mover dichas ventanas (Lawrence y Verzani, 2012).

Actualmente se han desarrollado interfaces gráficas que hacen más amigable la interacción entre el usuario y R, por ejemplo RKward (Rodiger *et al.*, 2012), Deducer (Fellows, 2002), RCommander (Fox, 2005), o Rattle (Williams, 2009), por mencionar algunas. Desafortunadamente, estas herramientas distan mucho de las implementaciones presentes en aplicaciones comerciales, es por esto que nace la idea de JRStat.

2.1. JRStat: una Plataforma para Implementación de Algoritmos

El *software* JRStat es una plataforma de código abierto para análisis estadísticos que combina el poder de cómputo de R con la facilidad de uso de las Interfaces Gráfica de Usuario. JRStat permite usar el *software* R desde una aplicación escrita en el lenguaje de programación Java (Gosling y McGilton, 1995). Para ello, se hace uso del entorno de desarrollo integrado NetBeans (Oracle Corporation, 2017) así como las librerías rJava (Urbanek, 2016), JavaGD (Urbanek, 2012), JFreeChart (Gilbert, 2014), entre otras.

Entre las principales características del *software* se pueden listar las siguientes:

- Utiliza R para realizar cálculos y generación de gráficos.
- Interacción con usuario usando GUI.
- Código abierto, GNU-GPL v3.
- Multiplataforma (Windows, macOS, Linux).
- Extensible, se pueden agregar nuevos módulos rápidamente con *plugins* basados en archivos xml.

2.2. Arquitectura de JRStat

La Figura 1 muestra la arquitectura del programa JRStat. El *software* interactúa de manera indirecta con el programa R instalado en el equipo de cómputo local. La comunicación de JRStat con R es posible gracias a la librería rJava, misma que se realiza de manera muy

eficiente pues utiliza la Interface Nativa de Java (JNI por sus siglas en inglés) para ejecutar programas compilados utilizando otros lenguajes de programación.

Las librerías JFreeChart y JavaGD permiten la visualización de cualquier tipo de gráficos desde Java usando los comandos usuales de R.

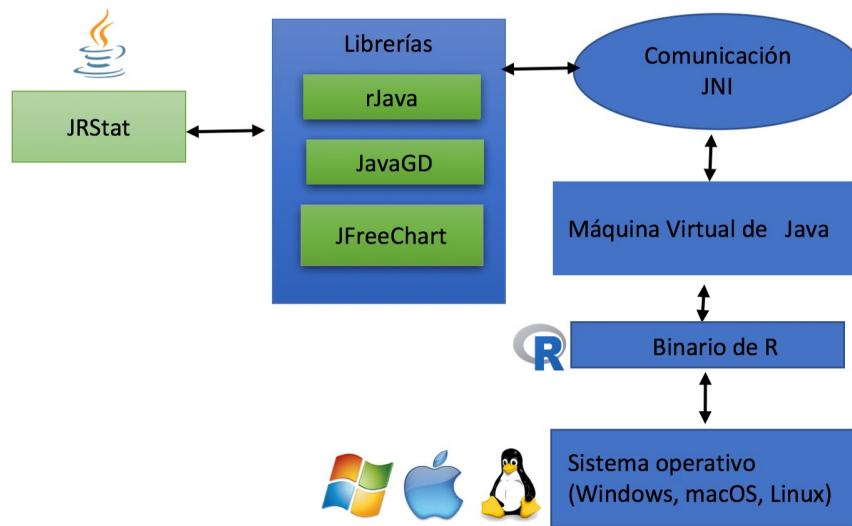


Figura 1: Arquitectura del programa JRStat.

2.3. Módulos Principales

El *software* JRStat actualmente cuenta con los siguientes módulos:

1. Consola interactiva para ejecución de comandos en R y visualización de resultados.
2. Explorador de objetos.

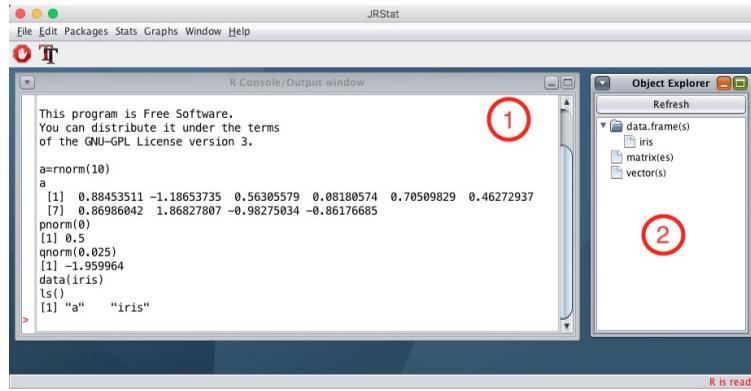


Figura 2: Consola y explorador de objetos.

3. Editor de código.

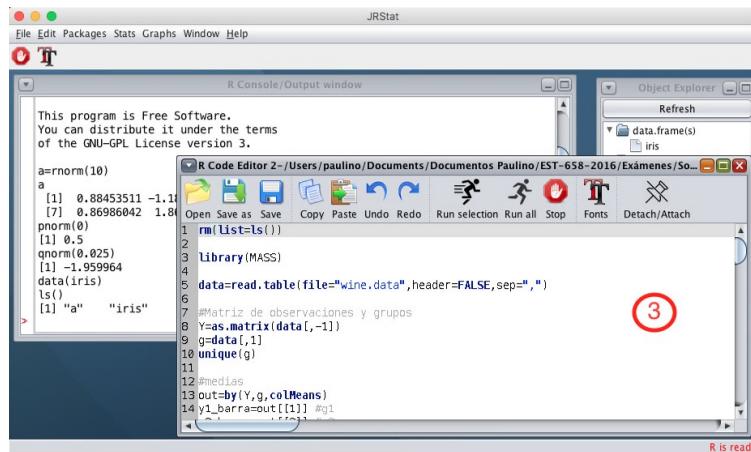


Figura 3: Editor de código.

4. Rutinas para graficación, los gráficos tradicionales creados mediante comandos de R pueden visualizarse en JRStat gracias a la librería JavaGD. Los gráficos pueden editarse (en algunos casos) y exportarse en formato pdf, png, entre otros.
5. Menús para análisis básicos de datos (pruebas de hipótesis, regresión lineal, modelo lineal generalizado, análisis multivariado, etcétera).
6. Módulo para creación de análisis personalizados.

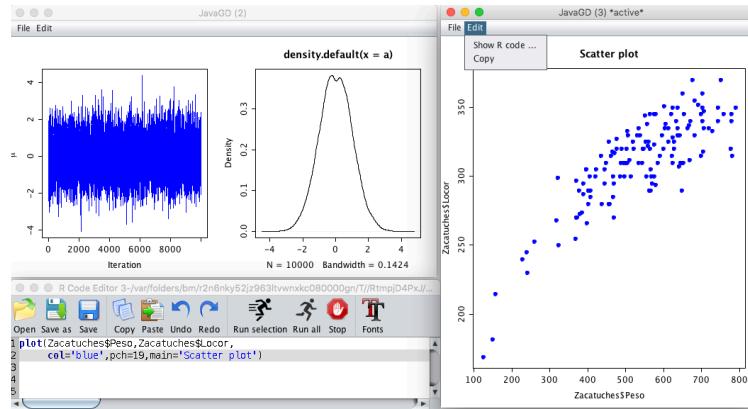


Figura 4: Ejemplos de rutinas para graficación.

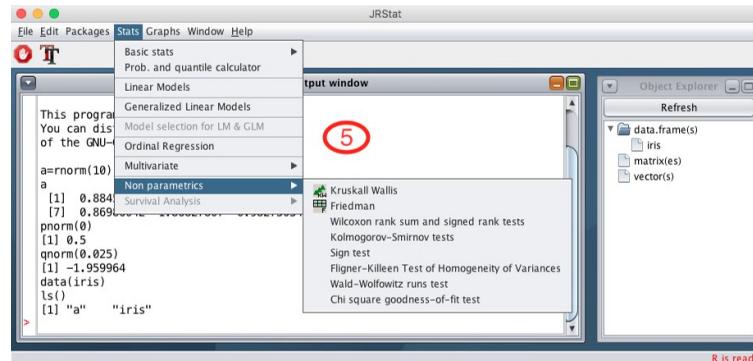


Figura 5: Menú para análisis estadísticos.

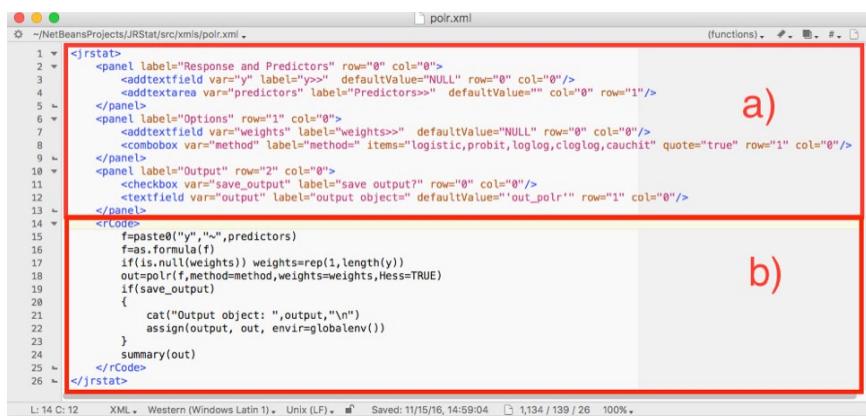


Figura 6: Archivo xml para regresión ordinal. Panel a) Cuadro de diálogo. Panel b) Código para análisis.

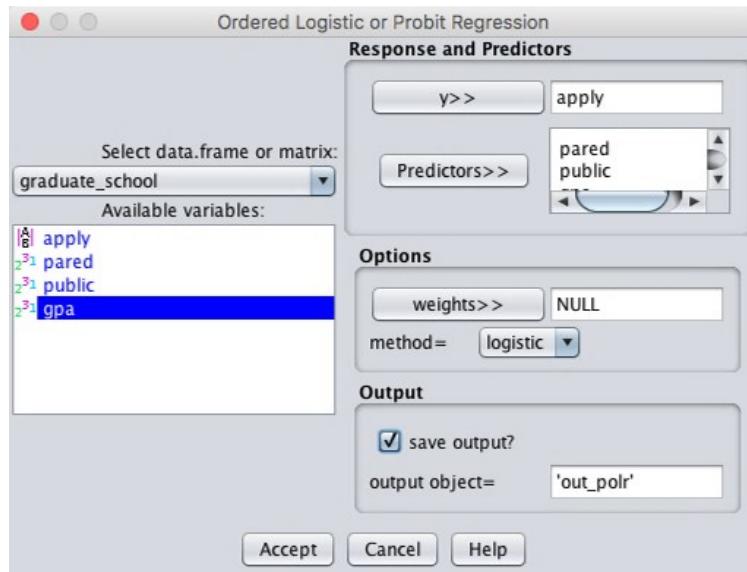


Figura 7: Cuadro de diálogo generado usando el xml para regresión ordinal.

3. Distribución de JRStat

Para poder ejecutar el programa es necesario instalar R (<http://www.r-project.org>) y la máquina virtual de Java (JRE, descargar desde <http://www.oracle.com>). El programa JRStat se puede descargar desde la página del Colegio de Postgraduados, <http://est.colpos.mx/JRStat> (ver Figura 7).

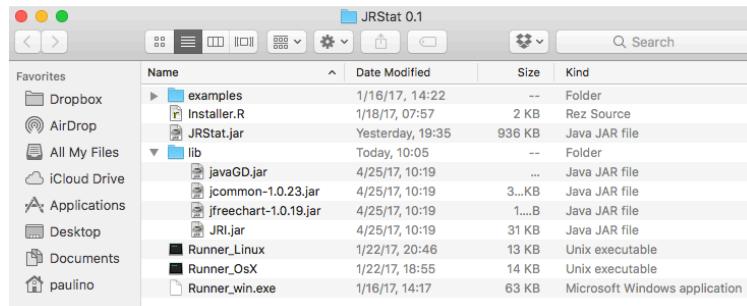


Figura 8: Folder con *software* JRStat.

La carpeta con el *software* se puede colocar en cualquier directorio en el que el usuario tenga permiso de escritura. Para ejecutar el programa hay que dar doble *click* en el archivo “Runner_win.exe” en Windows, “Runner_Linux” en Linux y “Runner_OsX” en macOS.

4. Conclusiones

- JRStat puede ser una excelente alternativa a la consola de R. Este *software* está dirigido tanto a usuarios que prefieran interfaces gráficas como a usuarios avanzados que puedan desarrollar sus propios módulos y distribuirlos.
- JRstat aún está en fase alfa, sin embargo, se pretende seguir desarrollándolo y compartir el código en plataforma de Github.
- Se planea finalizar los módulos de inferencia estadística e implementar un módulo de minería de datos, con el cual usuarios de diversas disciplinas podrán acceder a estas poderosas herramientas sin tener conocimientos avanzadas de programación.

Bibliografía

David Gilbert (2014). Jfreechart. [Internet; descargado 24-Marzo-2017].

Fellows, I. (2002). Deducer: A data analysis GUI for R. *Journal of Statistical Software*, 8(49):2–14.

Fox, J. (2005). The R commander: A basic-statistics graphical user interface to R. *Journal of Statistical Software*, 14(9):1–42.

Gosling, J. and McGilton, H. (1995). *Java Language Environment, The*. California: Sun Microsystems.

IBM Corp. (2013). *IBM SPSS Statistics for Windows, Version 22.0*. Armonk, NY.

IHS Global Inc. (2015). Eviews 10. [Internet; descargado 05-Mayo-2017].

Kernighan, B. W. and Ritchie, D. M. (1988). *The C Programming Language*. Prentice Hall.

Lawrence, M. and Verzani, J. (2012). *Programming Graphical User Interfaces in R*. Chapman and Hall.

Minitab Inc. (2017). Minitab. [Internet; descargado 05-Mayo-2017].

Oracle Corporation (2017). Netbeans ide. [Internet; descargado 24-Marzo-2017].

-
- R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Rodiger, S., Kapat, P., Friedrichmeir, T., and Michalke, M. (2012). A comprehensive graphical user interface and integrated development environment for statistical analysis with R. *Journal of Statistical Software*, 49(9):1–34.
- SAS Institute Inc. (2011). *SAS/STAT Software, Version 9.3*. Cary, NC.
- StataCorp LLC (1996). Stata. [Internet; descargado 05-Mayo-2017].
- Stroustrup, B. (1997). *C++ Programming Language, The*. Addison-Wesley Professional.
- Urbanek, S. (2012). *JavaGD: Java Graphics Device*. R package version 0.6-1.
- Urbanek, S. (2016). *rJava: Low-Level R to Java Interface*. R package version 0.9-8.
- Vance, A. (2009). Data analysts captivated by R’s power. [Internet; descargado 23-Marzo-2017].
- Williams, G. (2009). Rattle: A data mining gui for R. *The R Journal*, 1(2):45–55.

Lista de Árbitros

El Comité Editorial de las Aportaciones a la Estadística organizada por la Asociación Mexicana de Estadística (AME) agradece la valiosa colaboración de los siguientes árbitros:

1. Alamilla López, Norma Edith *CECyTE Tabasco*
2. Ariza Hernández, Francisco J. *UAGro*
3. Baltazar Larios, Fernando *Facultad de Ciencias – UNAM*
4. Barraza Barraza, Diana *UJED*
5. Batún Cutz, José Luis *UADY*
6. Burguete Hernández, Esteban *COLPOS*
7. Chávez Cano, Margarita Elvira *Facultad de Ciencias – UNAM*
8. Christen Gracia, José Andrés *CIMAT*
9. Contreras Carreto, Nilson Agustín
10. Contreras Cristán, Alberto *IIMAS – UNAM*
11. Contreras Cruz, Luis Fernando *COLPOS*
12. Díaz Avalos, Carlos *IIMAS – UNAM*
13. Díaz-Francés Murguía, Eloísa *CIMAT*
14. Dominguez Dominguez, Jorge *CIMAT*
15. Duran Fernández, Juan José *ITAM*
16. Erdely Ruiz, Arturo *FES Acatlán – UNAM*

17. Escarela Pérez, Gabriel *UAM-I*
18. Fuentes García, Ruth Selene *Facultad de Ciencias – UNAM*
19. García Banda, Agustín Jaime *UV*
20. Godímez Jaimes, Flaviano *UAGro*
21. González Farías, Graciela *CIMAT*
22. Gracia-Medrano, Leticia *IIMAS – UNAM*
23. Gutiérrez Peña, Eduardo *IIMAS – UNAM*
24. López Escobar, Emilio *ITAM*
25. Macias Moreno, Hortensia *UAM-I*
26. Martínez Ovando, Juan Carlos *ITAM*
27. Méndez Gómez-Humarán, Ignacio *CIMAT*
28. Montano Rivas, Julia Aurora *UV*
29. Naranjo Albarrán, Lizbeth *Facultad de Ciencias – UNAM*
30. Nuñez Antonio, Gabriel *UAM-I*
31. Pérez Abreu Carreón, Rafael *CIMAT*
32. Pérez Salvador, José Enrique *Facultad de Ciencias – UNAM*
33. Ramírez Ramírez, Leticia *CIMAT*
34. Ramos Quiroga, Rogelio *CIMAT*
35. Reyes Cervantes, Hortensia *BUAP*
36. Reyes Cortés, Miguel Angel *CIMAT*
37. Rivera Rosales, Elsa Edith *UAdeC*

38. Rodríguez Esparza, Luz Judith UACH
39. Romero Mares, Patricia Isabel *IIMAS – UNAM*
40. Ruiz-Velasco Acosta, Silvia *IIMAS – UNAM*
41. Silva Urrutia, Eliud *Universidad Anáhuac*
42. Soriano Flores, Antonio *IIMAS – UNAM*
43. Suárez Espinosa, Javier *COLPOS*
44. Trejo Valdivia, Belem *INSP*
45. Vázquez García, Cristina del Carmen *PEMEX*
46. Zamora Muñoz, José Salvador *Facultad de Ciencias – UNAM*
47. Zuloaga Garmendia, María Antonieta *CIMAT*