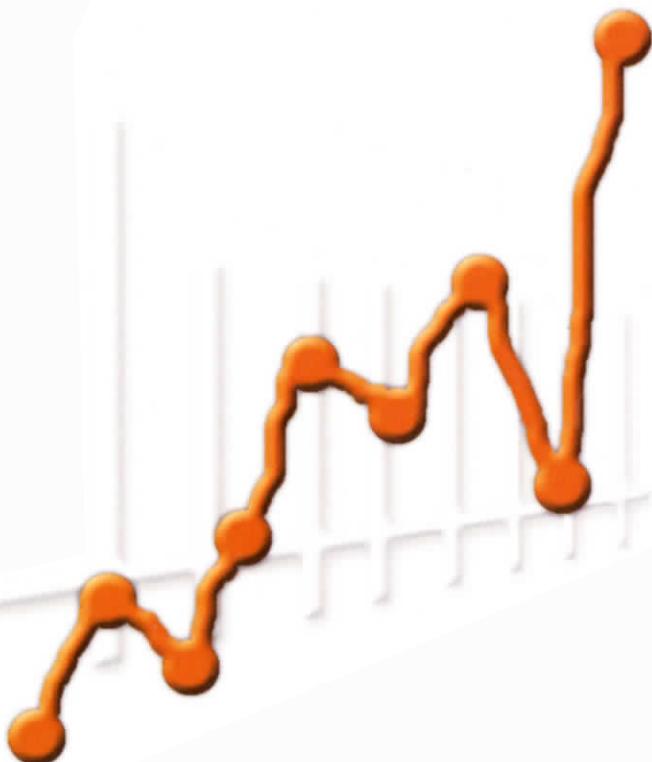


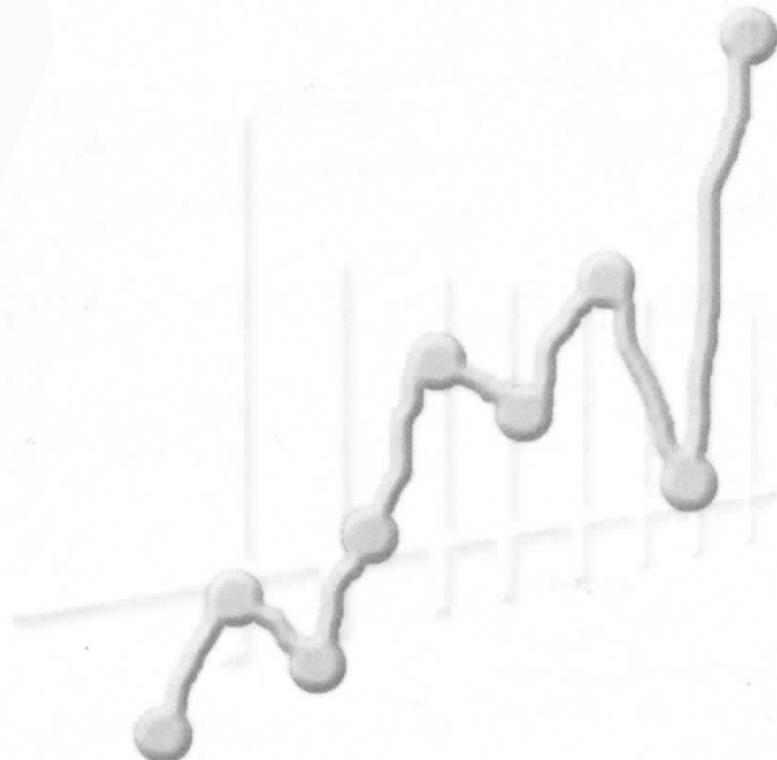
Memoria del **XXII Foro Nacional de Estadística**



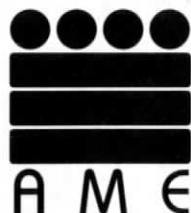
INSTITUTO NACIONAL DE ESTADÍSTICA
GEOGRAFÍA E INFORMÁTICA



Memoria del **XXII Foro Nacional de Estadística**



INSTITUTO NACIONAL DE ESTADÍSTICA
GEOGRAFÍA E INFORMÁTICA



DR © 2008, **Instituto Nacional de Estadística,
Geografía e Informática**
Edificio Sede
Av. Héroe de Nacozari Sur Núm. 2301
Fracc. Jardines del Parque, CP 20270
Aguascalientes, Ags.

www.inegi.gob.mx
atencion.usuarios@inegi.gob.mx

***Memoria del
XXII Foro Nacional
de Estadística***

**Impreso en México
ISBN 978-970-13-5055-3**

Presentación

El XXII Foro Nacional de Estadística fue organizado por el Instituto Tecnológico Autónomo de México (ITAM) y se llevó a cabo en la Ex-Hacienda de Júrica en Querétaro, Qro. del 17 al 20 de octubre de 2007. Esta memoria presenta resúmenes extendidos de algunas contribuciones libres presentadas en este evento.

Por una iniciativa de transformar nuestras memorias en una publicación con reconocimiento académico, a partir de este volumen la Mesa Directiva de la Asociación Mexicana de Estadística ha comenzado un proceso de revisión del contenido de los trabajos tomando en cuenta criterios mínimos de originalidad en los resultados y/o aplicaciones presentadas. Por lo tanto, los trabajos incluidos en esta memoria fueron sometidos a una revisión de forma y contenido.

Aprovechamos este espacio para agradecer a la comunidad estadística mexicana por su apoyo como árbitros en el proceso de revisión académica de estos trabajos. Adicionalmente, a nombre de la Asociación Mexicana de Estadística agradecemos al Instituto Tecnológico Autónomo de México por el apoyo otorgado para la realización de este foro y al Instituto Nacional de Estadística y Geografía el apoyo para la edición e impresión de esta memoria.

El Comité Editorial:

Elida Estrada Barragán

Asael F. Martínez Martínez

Ramsés H. Mena Chávez

Luis E. Nieto Barajas

Índice general

Estimación con el algoritmo EM estocástico de modelos de espacios de estados con observaciones censuradas	1
<i>Francisco J. Ariza Hernández, Gabriel A. Rodríguez Yam</i>	
Modelado conjunto de frecuencias y severidades	7
<i>Gabriel Escarela</i>	
Desempeño en muestras complejas de tres estimadores de regresión	13
<i>Flaviano Godínez Jaimes, Ignacio Méndez Ramírez</i>	
Una generalización de la prueba de Shapiro-Wilk para normalidad multivariada	19
<i>Elizabeth González Estrada, José A. Villaseñor Alva</i>	
Ecuaciones diferenciales en la modelación de datos funcionales	25
<i>María Guzmán Martínez, Eduardo Castaño Tostado</i>	
Modelo de decremento múltiple semiparamétrico para datos de supervivencia	33
<i>Angélica Hernández Quintero, Jean François Dupuy, Gabriel Escarela</i>	
Modelado atmosférico para determinar niveles máximos diarios de ozono en la ciudad de Guadalajara	39
<i>Lorelie Hernández Gallardo, Gabriel Escarela</i>	

Regresión por mínimos cuadrados parciales aplicada al estudio de emisiones de dióxido de carbono en suelos de Veracruz, México	47
<i>Gladys Linares Fleites, José Adrián Saldaña Munive, Luis G. Ruiz Suárez</i>	
Discriminación lineal y discriminación logística en estudios de calidad de suelos	55
<i>Gladys Linares Fleites, Miguel Ángel Valera Pérez, Maribel Castillo Morales</i>	
Análisis bivariado de extremos para evaluar los niveles de ozono troposférico en la zona metropolitana de Guadalajara	61
<i>Tania Moreno Zúñiga, Gabriel Escarela</i>	
Contraste de una hipótesis nula central compuesta frente una hipótesis alternativa bilateral en la distribución normal	69
<i>Leonardo Olmedo</i>	
Análisis de sendero como herramienta confirmatoria en un experimento de campo	75
<i>Emilio Padrón Corral, Ignacio Méndez Ramírez, Armando Muñoz Urbina</i>	
Comparación de poblaciones normales asimétricas.	83
<i>Paulino Pérez Rodríguez, José A. Villaseñor Alva</i>	
Análisis espectral aplicado al electroencefalograma	89
<i>Verónica Saavedra Gastélum, Thalía Fernández Harmony, Eduardo Castaño Tostado, Víctor Manuel Castaño Meneses</i>	
Software que trata las principales causas de la diabetes.	99
<i>Bárbara Emma Sánchez Rinza, Jessica Giovanna Huerta López, Jazmin Jiménez Bedolla, M. Bustillo Díaz, A. Rangel Huerta</i>	
Comparación de algunas pruebas estadísticas asintóticas de no-inferioridad para dos proporciones independientes	109
<i>David Sotres Ramos, Félix Almendra Arao</i>	

Procedimientos para analizar los datos no detectados en contaminación ambiental	115
<i>Fidel Ulín Montejo, Humberto Vaquera Huerta</i>	
Evaluating cluster solutions with reference to data generation processes - a simulation study	123
<i>Alexander von Eye, Patrick Mair</i>	

Estimación con el algoritmo EM estocástico de modelos de espacios de estados con observaciones censuradas

Francisco J. Ariza Hernández^a *Colegio de Postgraduados*
Gabriel A. Rodríguez Yam^b *Universidad Autónoma Chapingo*

1. Introducción

Los modelos de espacios de estados (SSM, por sus siglas en inglés) son una clase de modelos que permiten describir y modelar series de tiempo en una gran variedad de disciplinas. En algunas aplicaciones, las observaciones pueden estar “incompletas”, por ejemplo en ciencias ambientales cuando se monitorea algún tipo de contaminante, los equipos utilizados para registrar los valores de la variable de interés pueden presentar ciertas restricciones en la medición provocando que algunas de las observaciones estén censuradas. Además, estas observaciones, registradas de manera secuencial, pueden presentar un cierto grado de correlación. Los datos con estas características pueden ser analizados como un SSM; sin embargo, aún en el caso simple del SSM lineal gaussiano, los parámetros no se pueden estimar directamente con las recursiones de Kalman. En este trabajo se implementará el algoritmo EM estocástico para estimar los parámetros de un SSM con observaciones censuradas.

En la Sección dos se formula el SSM con observaciones censuradas y en la Sección tres se describe el algoritmo EM estocástico como una alternativa al EM clásico para calcular aproximaciones del estimador de máxima verosimilitud (EMV) del SSM con censura. En la Sección cuatro se presenta un ejemplo con datos simulados del SSM lineal gaussiano con observaciones censuradas por la izquierda, bajo diferentes porcentajes de censura.

^aarizahfj@colpos.mx

^bgrodrigu@correo.chapingo.mx

2. Modelo de espacio de estados con censura

Sea y_1, y_2, \dots, y_n una realización de un modelo de espacios de estados. Es decir,

$$p(y_t; \boldsymbol{\xi} | y_{t-1}, \dots, y_1, \alpha_t, \dots, \alpha_1) := p(y_t; \boldsymbol{\xi} | \alpha_t), \quad (1)$$

que pertenece a cierta familia de distribuciones, donde la variable de estado α_{t+1} tiene la función de densidad condicional

$$p(\alpha_{t+1}; \boldsymbol{\psi} | \alpha_t, \alpha_{t-1}, \dots, \alpha_1, y_t, \dots, y_1) := p(\alpha_{t+1}; \boldsymbol{\psi} | \alpha_t). \quad (2)$$

Se asume que Y_1, Y_2, \dots , son condicionalmente independientes dadas las variables de estado, y que algunas de las observaciones y_1, y_2, \dots, y_n están censuradas. Sea Z_t la variable ‘latente’ dada por

$$Z_t = \begin{cases} Y_t, & \text{si } \delta_t = 1, \\ \text{valor de la observación no registrada en el tiempo } t, & \text{si } \delta_t = 0, \end{cases}$$

donde δ_t es una variable ‘indicadora’ que toma el valor de 0 si la observación y_t está censurada por la izquierda y 1 de otro modo. Se sigue que

$$p(z_t; \boldsymbol{\xi} | \alpha_t, y_t, \delta_t) = \begin{cases} p_{Y_t | \alpha_t}(z_t; \boldsymbol{\xi} | \alpha_t), & \text{si } \delta_t = 1, \\ \frac{p_{Y_t | \alpha_t}(z_t; \boldsymbol{\xi} | \alpha_t)}{P[Y_t \leq y_t; \boldsymbol{\xi} | \alpha_t]} 1_{(-\infty, y_t)}(z_t), & \text{si } \delta_t = 0. \end{cases}$$

Sea $\mathbf{y}_0 := (y_1, \dots, y_{n_o})^t$ el vector de datos observados de tamaño n_o , $\mathbf{z}_m := (z_1, \dots, z_{m_o})^t$ el vector de datos censurados de tamaño m_o y $\boldsymbol{\theta} := (\boldsymbol{\xi}, \boldsymbol{\psi})$ el vector de parámetros. Entonces, la función de verosimilitud de este modelo de espacio de estados está dada por

$$L(\boldsymbol{\theta}; \mathbf{y}_0) = \int_A \int_{R^n} L(\boldsymbol{\theta}; \mathbf{y}_0, \mathbf{z}_m, \boldsymbol{\alpha}) d\boldsymbol{\alpha} d\mathbf{z}_m \quad (3)$$

donde $A = \{\mathbf{z}_m : z_j \leq y_j, \delta_m = 0\}$. Calcular explícitamente la expresión en (3) puede ser imposible, de aquí que el estimador de máxima verosimilitud de $\boldsymbol{\theta}$ sea difícil de obtener. En este trabajo, se utiliza el algoritmo EM estocástico para obtener una aproximación de este estimador.

3. Algoritmo EM estocástico

El algoritmo EM estocástico es una alternativa cuando el cálculo del paso E del algoritmo EM es difícil de obtener. Este algoritmo consta de dos pasos que se realizan de forma iterativa: el Paso-S, donde los valores perdidos son reemplazados por valores “simulados”, dados los valores observados y, el paso M, donde $\theta^{(i)}$ es el EMV del modelo completo obtenido. Este proceso alternado del paso-S y el paso-M genera una cadena de Markov, $\{\theta^{(i)}, i = 1, 2, \dots\}$, la cual converge a su distribución estacionaria bajo condiciones de regularidad. Diebolt e Ip (1996) mencionan que esta distribución estacionaria está aproximadamente centrada en el EMV de θ y que su varianza depende de la razón de cambio de $\theta^{(i)}$ en las iteraciones del algoritmo.

4. Distribución predictiva

Para llevar a cabo el Paso-S, en cada iteración se obtiene una muestra de $p(\mathbf{z}|\mathbf{y}, \theta^{(i)})$, lo que implica conocer a $p(\alpha_t|\mathbf{z}^{(t)}, \mathbf{y}^{(t)}, \boldsymbol{\delta}^{(t)})$ y $p(\alpha_t|\mathbf{z}^{(t-1)}, \mathbf{y}^{(t)}, \boldsymbol{\delta}^{(t)})$. Aplicando el teorema de Bayes y la suposición de que la distribución de Z_t dado $(\alpha_t, \mathbf{z}^{(t-1)}, \mathbf{y}^{(t)}, \boldsymbol{\delta}^{(t)})$ no depende de $(\mathbf{z}^{(t-1)}, \mathbf{y}^{(t-1)}, \boldsymbol{\delta}^{(t-1)})$ se tiene que

$$p(\alpha_t|\mathbf{z}^{(t)}, \mathbf{y}^{(t)}, \boldsymbol{\delta}^{(t)}) = \frac{p(z_t|\alpha_t, y_t, \delta_t)p(\alpha_t|\mathbf{z}^{(t-1)}, \mathbf{y}^{(t)}, \boldsymbol{\delta}^{(t)})}{p(z_t|\mathbf{z}^{(t-1)}, \mathbf{y}^{(t)}, \boldsymbol{\delta}^{(t)})}, \quad (4)$$

y

$$p(\alpha_{t+1}|\mathbf{z}^{(t)}, \mathbf{y}^{(t+1)}, \boldsymbol{\delta}^{(t+1)}) = \int_{-\infty}^{\infty} p(\alpha_{t+1}|\alpha_t)p(\alpha_t|\mathbf{z}^{(t)}, \mathbf{y}^{(t)}, \boldsymbol{\delta}^{(t)})d\alpha_t. \quad (5)$$

El denominador en (4) es una constante de normalización. Entonces la densidad condicional de Z_t dado $(\mathbf{z}^{(t-1)}, \mathbf{y}^{(t)}, \boldsymbol{\delta}^{(t)})$ puede ser calculada de (5) como

$$p(z_t|\mathbf{z}^{(t-1)}, \mathbf{y}^{(t)}, \boldsymbol{\delta}^{(t)}) = \int_{-\infty}^{\infty} p(z_t|\alpha_t, y_t, \delta_t)p(\alpha_t|\mathbf{z}^{(t-1)}, \mathbf{y}^{(t)}, \boldsymbol{\delta}^{(t)}) d\alpha_t \quad (6)$$

Finalmente, la densidad condicional de \mathbf{Z} dado $(\mathbf{y}, \boldsymbol{\delta}, \boldsymbol{\theta})$ puede ser expresada como

$$p(\mathbf{z}|\mathbf{y}, \boldsymbol{\delta}, \boldsymbol{\theta}) = \prod_{t=1}^n p(z_t|\mathbf{z}^{(t-1)}, \mathbf{y}^{(t)}, \boldsymbol{\delta}^{(t)}) \quad (7)$$

donde $p(z_1|z_0, y_1, \delta_1) = p(z_1|y_1, \delta_1)$.

5. Ejemplo

Considere el siguiente SSM lineal gaussiano

$$Y_t = \mu + \alpha_t + \varepsilon_t \quad (8)$$

donde μ es la media general y $\varepsilon_t \sim \text{iid}N(0, \sigma^2)$, $t = 1, 2, \dots, n$ representan los errores del modelo (8). Además, el proceso de estados es un modelo $AR(1)$, i. e.

$$\alpha_t = \phi\alpha_{t-1} + \eta_t \quad (9)$$

donde $\eta_t \sim \text{iid}N(0, \tau^2)$, $t = 1, 2, \dots, n$, además ε_t y η_t , $t = 1, 2, \dots, n$ son independientes. En este ejemplo $\boldsymbol{\theta} := (\mu, \sigma^2, \phi, \tau^2)$ es el vector de parámetros del modelo (8)–(9). En la Figura 1, se muestra una realización de este SSM de tamaño $n = 500$ y $\boldsymbol{\theta} := (30, 2, 0.9, 1)$. Considere una observación censurada aquella que sea menor a L_j , $j = 1, 2, 3$, donde $L_1 = 23.7$, $L_2 = 27.09$ y $L_3 = 29.77$. De esta forma, se obtienen el 5, 20 y 50 % de censura en los datos simulados.

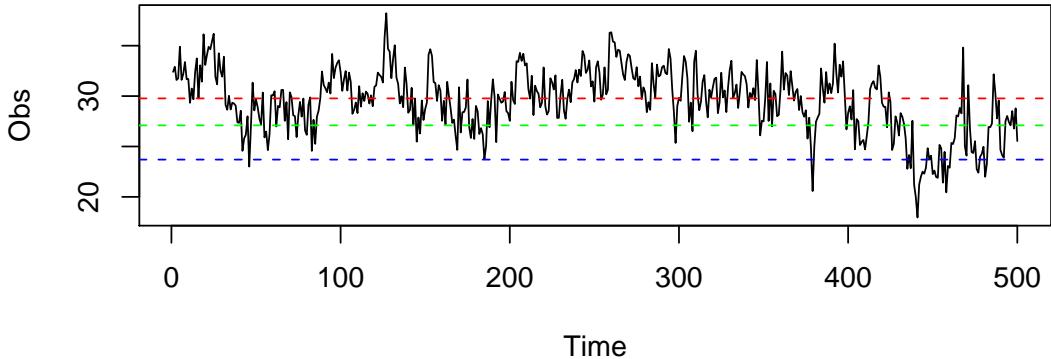


Figura 1: Serie de tiempo simulada

Note que para el modelo en (8)–(9) la log-verosimilitud de los datos pseudo-completos es

$$l(\boldsymbol{\theta}; \mathbf{z}) = -\frac{n}{2} \log(2\pi) - \frac{1}{2} \sum_{t=1}^n \log(\Omega_t + \sigma^2) - \frac{1}{2} \sum_{t=1}^n \frac{(z_t - \mu - \hat{\alpha}_t)^2}{(\Omega_t + \sigma^2)}$$

donde $\hat{\alpha}_t$, $t = 1, \dots, n$ son las predicciones de un paso y Ω_t es la varianza del error, los cuales son obtenidos a partir de las recursiones de predicción de Kalman (Brockwell y Davis, 2002).

En este trabajo se realizaron 1600 iteraciones de acuerdo con el criterio de Raftery y Lewis (Raftery y Lewis 1996) y se eliminaron las primeras 200 iteraciones (burn-in). El resto de la secuencia generada, da un comportamiento aproximadamente estacionario. En la Figura 2 se observa una convergencia rápida de la cadena generada aún con porcentajes de censura altos. En la Tabla 1 se presentan los estimadores de los parámetros y sus errores estándar (en paréntesis) bajo los diferentes porcentajes de censura.

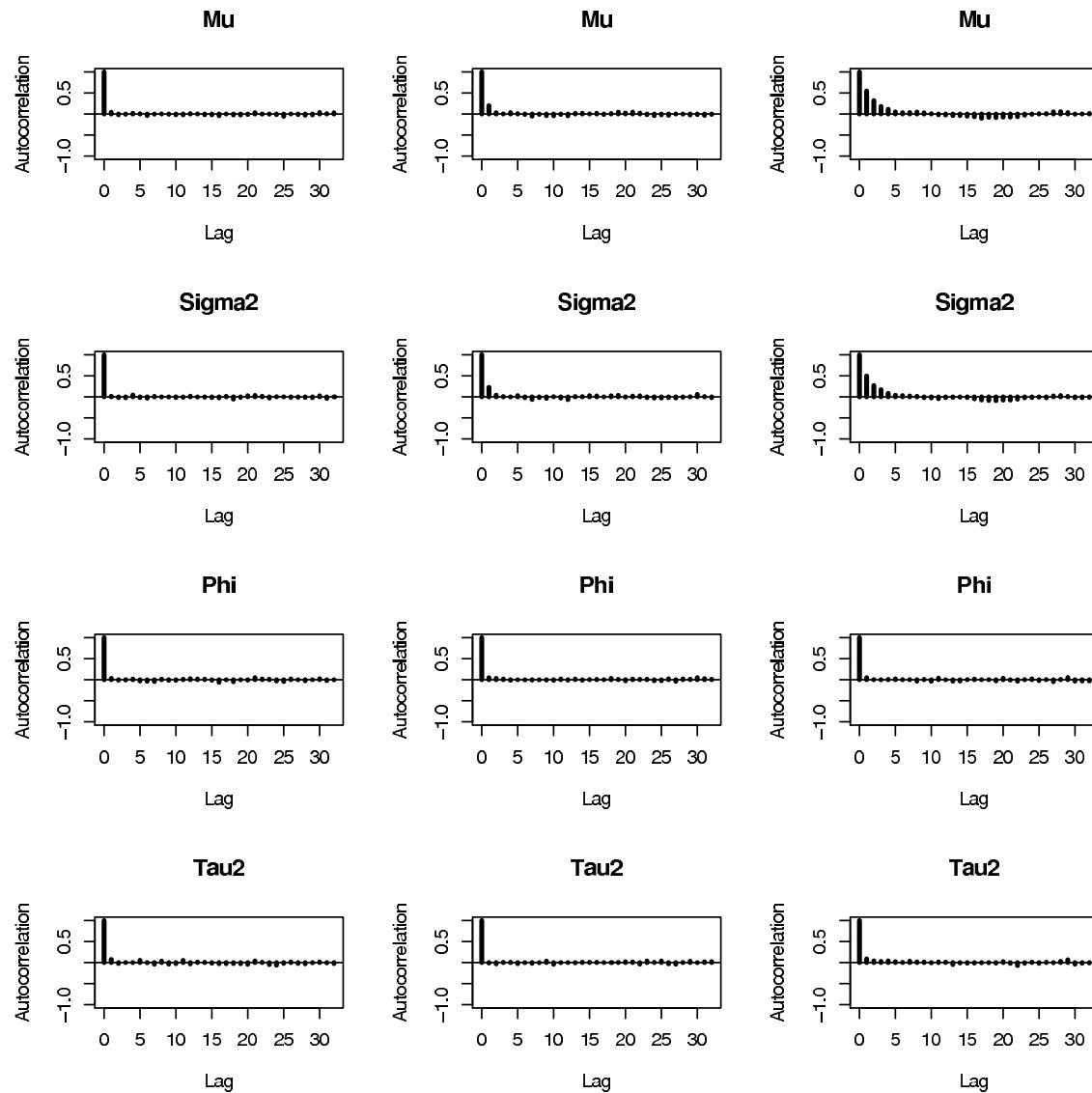


Figura 2: Gráfica de autocorrelaciones de μ, σ^2, ϕ y τ^2 (5 %, 20 % y 50 % de censura)

Censurados	μ	σ^2	ϕ	τ^2
0 %	29.503 (0.027)	2.070 (0.011)	0.921 (0.001)	1.233 (0.011)
5 %	29.573 (0.007)	1.970 (0.019)	0.919 (0.001)	1.156 (0.011)
20 %	29.877 (0.030)	1.809 (0.089)	0.910 (0.004)	0.803 (0.040)
50 %	30.434 (0.074)	1.369 (0.158)	0.863 (0.016)	0.690 (0.102)

Tabla 1: Estimadores y desviaciones estándar para una realización del SSM en (8)–(9) con porcentajes de censura en los datos de 5 %, 20 % y 50 %.

6. Conclusión

Se ha implementado el algoritmo *EM* estocástico para obtener estimadores aproximados de los estimadores de máxima verosimilitud del modelo de espacios de estados lineal gaussiano cuando se tienen observaciones censuradas por la izquierda. Con este procedimiento iterativo se obtienen buenos estimadores bajo diferentes porcentajes de censura. Las cadenas de Markov generadas convergen rápidamente aún con porcentajes de censura altos.

Referencias

- Brockwell, P. J. y Davis, R. A. (2002) *Introduction to Time Series and Forecasting*. Second Edition. Springer, NY.
- Dempster A. P., Laird N.M. y Rubin D.B. (1977) Maximun Likelihood from Incomplete Data Via EM Algorithm. *J. R. Stat. Soc. Ser. B*, **39**, 1–38.
- Diebolt, J. e Ip, E. H. S. (1996) *Stochastic EM: Method and Application*. En *Markov Chain Monte Carlo in Practice*. (eds Gilks, W.R., Richardson, S. y Spiegelhalter, D.J.) London: Chapman & Hall, pp. 259-273.
- Raftery, A.E. y Lewis, S.M. (1996) *Implementing MCMC*. En *Markov Chain Monte Carlo in Practice*. (eds Gilks, W.R., Richardson, S. y Spiegelhalter, D.J.) London: Chapman & Hall, pp. 115 - 130.

Modelado conjunto de frecuencias y severidades

Gabriel Escarela^a *Universidad Autónoma Metropolitana – Iztapalapa*

1. Introducción

En un análisis de portafolios de seguros el interés se puede centrar en la estimación de la distribución conjunta de los montos correspondientes a dos coberturas de la póliza. En situaciones prácticas estos datos - conocidos en la literatura de habla inglesa como *loss data* - vienen con información *concomitante* de la que se cree que tiene influencia sobre frecuencia y severidad de los montos.

Sean X e Y las variables aleatorias correspondientes a los montos de dos tipos de cobertura de la póliza. La función de distribución conjunta se define como

$$F(x, y) = \Pr\{X \leq x, Y \leq y\}, \quad x, y \geq 0.$$

La meta de este estudio es la de modelar esta distribución bivariada en presencia de variables explicativas.

Como los asegurados pueden no ser indemnizados por alguna de las dos coberturas o ambas, es necesario definir las siguientes *probabilidades de frecuencia*:

$$\begin{aligned} p_{00} &= \Pr\{X = 0, Y = 0\}, & p_{01} &= \Pr\{X = 0, Y > 0\}, \\ p_{10} &= \Pr\{X > 0, Y = 0\} & \text{y} & \quad p_{11} = \Pr\{X > 0, Y > 0\}. \end{aligned}$$

Nótese que $p_{00} + p_{01} + p_{10} + p_{11} = 1$. Las *distribuciones de severidad* son:

$$\begin{aligned} F_{01}(y) &= \Pr\{Y \leq y \mid X = 0, Y > 0\}, \\ F_{10}(x) &= \Pr\{X \leq x \mid X > 0, Y = 0\}, \quad \text{y} \\ F_{11}(x, y) &= \Pr\{X \leq x, Y \leq y \mid X > 0, Y > 0\} \end{aligned}$$

^age@xanum.uam.mx

En este estudio se supondrá que F_{01} , F_{10} y F_{11} son absolutamente continuas con funciones de densidad f_{01} , f_{10} y f_{11} respectivamente.

Si $x \geq 0$ e $y \geq 0$, se puede demostrar que

$$F(x, y) = p_{00} + p_{01}F_{01}(y) + p_{10}F_{10}(x) + p_{11}F_{11}(x, y).$$

De esta forma, la estimación de $F(x, y)$ se reduce a la estimación conjunta de los modelos de severidad y frecuencia. En práctica, cuando los asegurados contratan este tipo de pólizas lo hacen con la certidumbre de que recibirán cierto beneficio, i.e. hay pocas observaciones de asegurados a los que no se les indemniza en ninguna de las dos coberturas, por lo que es conveniente suponer que $p_{00} = 0$.

2. Inferencia y modelado

Sean x_k e y_k los montos indemnizados por la k -ésima póliza, $k = 1, \dots, n$, y sea $\mathbf{z}_k^T = (1, z_{k1}, z_{k2}, \dots, z_{kp})$ el vector de variables explicativas correspondiente al cual incluye a la ordenada al origen. Defínanse los *vectores indicadores de estatus* como:

$$c_{01,k} = I(x_k = 0, y_k > 0), \quad c_{10,k} = I(x_k > 0, y_k = 0) \quad \text{y} \quad c_{11,k} = I(x_k > 0, y_k > 0),$$

donde $I(A) = 1$ si A es verdad y $I(A) = 0$ de otra forma.

La función de verosimilitud es:

$$\begin{aligned} L &= \prod_{k=1}^n [p_{01,k} f_{01}(y_k)]^{c_{01,k}} [p_{10,k} f_{10}(x_k)]^{c_{10,k}} [p_{11,k} f_{11}(x_k, y_k)]^{c_{11,k}} \\ &= \prod_{k=1}^n [p_{01,k}]^{c_{01,k}} [p_{10,k}]^{c_{10,k}} [p_{11,k}]^{c_{11,k}} \times \\ &\quad \times \prod_{k=1}^n [f_{01}(y_k)]^{c_{01,k}} [f_{10}(x_k)]^{c_{10,k}} [f_{11}(x_k, y_k)]^{c_{11,k}} = L_f \times L_s. \end{aligned}$$

De esta forma es posible modelar y analizar a los modelos de frecuencia y severidad en forma separada.

La presencia de efectos de variables explicativas en las probabilidades de las frecuencias pueden modelarse usando un *modelo logístico multinomial*:

$$p_{ij,k} = \exp(\boldsymbol{\beta}_{ij}^T \mathbf{z}_k) / [\exp(\boldsymbol{\beta}_{01}^T \mathbf{z}_k) + \exp(\boldsymbol{\beta}_{10}^T \mathbf{z}_k) + \exp(\boldsymbol{\beta}_{11}^T \mathbf{z}_k)],$$

donde $ij = 01, 10, 11$, y β_{01}, β_{10} y β_{11} son los vectores de coeficientes correspondientes. Para evitar redundancia, $\beta_{11} = \mathbf{0}$.

Cuando se trata de modelar las distribuciones de severidad es importante considerar que las observaciones apareadas positivamente pueden poseer cierta dependencia. La construcción de $F_{11}(x, y)$ debe de tomar en consideración cierta asociación de las variables aleatorias.

El uso de la cópula C_θ es atractivo para modelar $F_{11}(x, y)$ (ver e.g. Klugman y Parsa, 1999). Una cópula bivariada es una función de distribución conjunta de una pareja aleatoria la cual toma valores en el cuadro unitario. La definición de $F_{11}(x, y)$ con marginales dadas $F_{21}(x) = F_{11}(x, \infty)$ y $F_{12}(y) = F_{11}(\infty, y)$ se puede dar por:

$$F_{11}(x, y) = C_\theta [F_{21}(x), F_{12}(y)].$$

Hay situaciones - como la actual - para las cuales es más fácil encontrar expresiones analíticas para la *función de supervivencia bivariada*:

$$\begin{aligned} S_{11}(x, y) &= \Pr\{X > x, Y > y \mid X > 0, Y > 0\} \\ &= C_\theta [S_{21}(x), S_{12}(y)]. \end{aligned}$$

Aquí $S_{21}(x) = 1 - F_{21}(x)$ y $S_{12}(y) = 1 - F_{12}(y)$ son las funciones de supervivencia marginales. La función de densidad conjunta $f_{11}(x, y)$ correspondiente puede obtenerse fácilmente al diferenciar $S_{11}(x, y)$ con respecto a x e y .

Una cópula conveniente para el presente estudio es la descrita por Frees y Valdez (1998) como de *cola derecha pesada*, cuya representación está dada por:

$$C_\theta(u, v) = u + v - 1 + [(1 - u)^{-1/\theta} + (1 - v)^{-1/\theta} - 1]^{-\theta}, \quad \theta > 0. \quad (1)$$

La aplicación de ésta cópula es atractiva pues tiene poca correlación en la cola izquierda pero alta en la derecha; intuitivamente se esperaría que entre más severa sea la indemnización de un tipo de cobertura, la indemnización correspondiente al otro tipo también.

Para evaluar el grado de concordancia entre los riesgos es posible usar la τ de Kendall, cuyo valor para esta cópula es $\tau_\theta = 1/(2\theta + 1)$.

Para modelar las distribuciones univariadas F_{01}, F_{10}, F_{21} y F_{12} es conveniente usar la familia de distribuciones Burr de tres parámetros cuya función de supervivencia es $S(t) = [1 + \gamma(t\lambda)^\alpha]^{-1/\gamma}$, donde $t > 0$ y $\lambda, \gamma, \alpha > 0$. La elección de esta forma paramétrica para

este estudio se debe a su cola derecha pesada. Cuando se usa la cópula en la ecuación (1) y marginales Burr para construir la función de distribución conjunta se obtiene la distribución Burr descrita por Frees y Valdez (1998), cuya distribución condicional es Burr también.

Para permitir los efectos de las variables explicativas y para asegurar que los parámetros λ , γ y α permanezcan positivos, es posible usar una liga log, en forma análoga a como se hace con los modelos lineales generalizados, de manera tal que $\lambda = \exp\{\mathbf{a}^T \mathbf{z}\}$, $\gamma = \exp\{\mathbf{b}^T \mathbf{z}\}$ y $\alpha = \exp\{\mathbf{c}^T \mathbf{z}\}$; aquí, $\mathbf{z}^T = (1, z_1, z_2, \dots, z_p)$ es el vector de variables explicativas la cual incluye a la ordenada, y \mathbf{a} , \mathbf{b} y \mathbf{c} son los vectores de parámetros correspondientes.

Los modelos de frecuencia y severidad no son lineales, por esto es necesario aplicar técnicas numéricas para encontrar los estimadores de máxima verosimilitud.

Para criticar el ajuste de los modelos de severidad univariados y los condicionales de F_{11} cuando ésta distribución es Burr bivariada es posible usar métodos empleados en el análisis de supervivencia. Si T se distribuye Burr entonces $-\log S(T) = (1/\gamma) \log[1 + \gamma(T\lambda)^\alpha]$ se distribuye exponencial con parámetro igual a 1; de esta forma, es posible ordenar los valores de $u = (1/\gamma) \log[1 + \gamma(t\lambda)^\alpha]$ en la forma de $u_{(j)}$ y entonces graficar

$$\frac{m+1-j}{m+1} \text{ contra } \exp(-u_{(j)}), \quad \text{para } j = 1, \dots, m,$$

donde m es el número de observaciones en cada una de las especificaciones de severidad. Básicamente esta es una gráfica de cuantil contra cuantil.

En general, las severidades que son simultáneamente positivas pueden ser evaluadas al considerar

$$\begin{aligned} G(v, w) &= \Pr\{X > v \mid Y < w\} \\ &= \frac{S_{21}(v) - S_{11}(v, w)}{1 - S_{12}(w)}, \end{aligned}$$

y entonces ordenar v para obtener los estadísticos de orden $v_{(j)}$ y así graficar

$$\frac{m+1-j}{m+1} \text{ contra } \frac{1}{n} \sum_{i=1}^m G(v_{(j)}, w_i), \quad \text{para } j = 1, \dots, m,$$

donde m es el número de observaciones donde ambas severidades son estrictamente positivas.

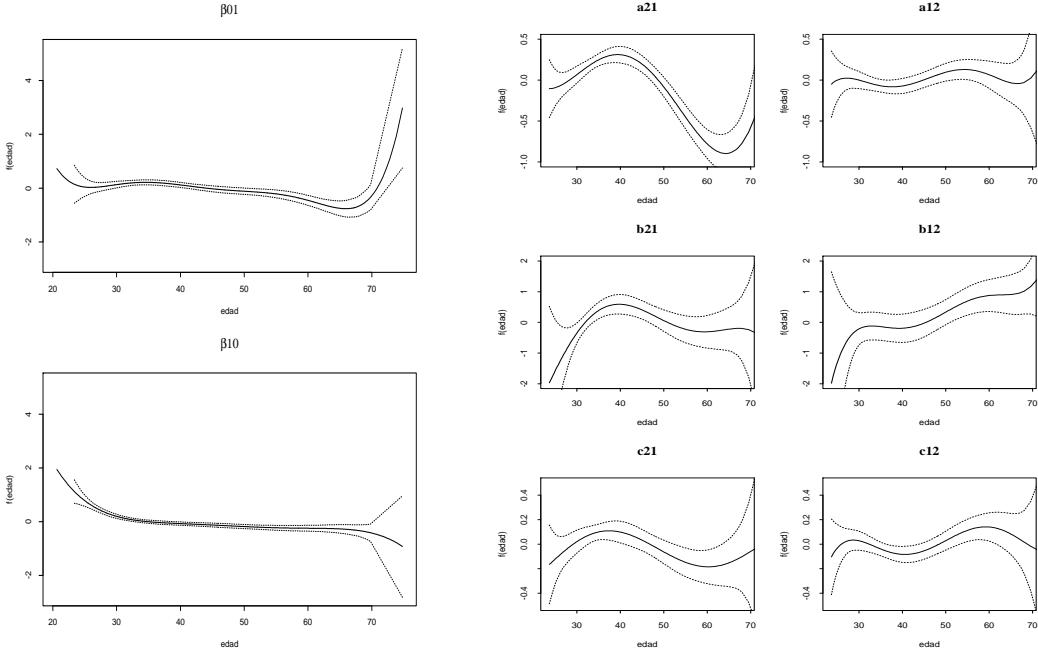


Figura 1: Estimadores de $f(\text{edad})$ para el modelo de frecuencia y para las marginales de S_{11} .

3. Ilustración

En este estudio se analizó una serie de $N = 19827$ pólizas de gastos médicos. Los beneficios se clasificaron en *medicina* y *otros gastos*. Las variables explicativas son **edad** y **GÉNERO** (masculino=0; femenino=1). Para visualizar y evaluar los efectos de edad en los dos modelos de frecuencia y severidad se usaron bases de polinomios ortogonales de edad y así estimar $f(\text{edad})$ en vez de un sólo parámetro el cual supone linealidad.

Se encontró que el género es importante y que la regresión polinomial de edad mejora considerablemente el ajuste en ambos modelos. La Figura 1 muestra las curvas estimadas. A pesar de que el grado de concordancia en S_{11} es modesto ($\tau_\theta = 0.1$), la inclusión de θ es estadísticamente significativa y el ajuste resultante mejora al de un modelo más simple como el que supone independencia. Los diagnósticos indicaron un ajuste relativamente bueno.

Referencias

Frees, E.W. y Valdez, E.A. (1998). Understanding relationships using copulas, *North American Actuarial Journal*, **2**, 1-25.

Klugman, S.A. y Parsa, R. (1999). Fitting bivariate loss distributions with copulas, *Insurance: Mathematics and Economics*, **24**, 139-148.

Desempeño en muestras complejas de tres estimadores de regresión del total

Flaviano Godínez Jaimes^a *Unidad Académica de Matemáticas, Universidad Autónoma de Guerrero*

Ignacio Méndez Ramírez^b *Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, UNAM*

1. Introducción

En muchas encuestas la variable de interés, Y , depende de las características de las unidades individuales, X_1, X_2, \dots, X_p . Es natural considerar el modelo de regresión lineal (MRL) para explicar la relación entre Y y X_1, X_2, \dots, X_p el cual se expresa matricialmente:

$$E_M(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}, \quad \text{var}_M(\mathbf{Y}) = \mathbf{V}$$

donde \mathbf{X} , \mathbf{Y} , \mathbf{V} y $\boldsymbol{\beta}$, son matrices de Nxp , $Nx1$, NxN y $px1$ de los valores poblacionales de las variables explicatorias, de la variable respuesta, varianzas y parámetros. Los estimadores de regresión del total aprovechan la relación lineal entre Y y X_1, X_2, \dots, X_p y que se conoce \mathbf{X} o al menos los totales poblacionales, $\mathbf{1}^T\mathbf{X}$. Los principales enfoques de inferencia en muestreo son: Basado en Modelo (BM) y Asistido por Modelo (AM). En el enfoque BM, el MRL se usa para motivar el estimador y evaluar sus propiedades (Valliant *et al.* 2000). En el enfoque AM, el MRL se usa para motivar el estimador pero sus propiedades se evalúan con respecto al diseño (Särndal, *et al.* 1992).

Consideremos una población formada por L estratos y cada estrato esta formado por N^h conglomerados con N_i^h elementos $h = 1, \dots, L$ e $i = 1, \dots, N^h$. Raj (1968) propuso dos esquemas de muestreo apropiados para esta población. En el Esquema A de Raj las unidades

^afgodinezj@gmail.com

^bimendez@servidor.unam.mx

primarias de muestreo (UPM) se toman dentro de cada estrato por muestreo aleatorio simple sin reemplazo (MASSR) y dentro de las UPM seleccionadas se usa cualquier forma de muestreo incluso diferente. Es decir, se eligen con MASSR n^h de las N^h UPM ($\pi_{Ii} = \frac{n^h}{N^h}$ y $\pi_{Iij} = \frac{n^h(n^h-1)}{N^h(N^h-1)}$ $i, j = 1, \dots, n^h$) y en cada UPM en la muestra se seleccionan con MASSR n_i^h de los N_i^h individuos ($\pi_{k|i} = \frac{n_i^h}{N_i^h}$ y $\pi_{kl|i} = \frac{n_i^h(n_i^h-1)}{N_i^h(N_i^h-1)}$ $k, l = 1, \dots, n_i^h$). En el estrato h , el vector de probabilidades de inclusión $\boldsymbol{\pi}^h = (\pi_i)^h = (\pi_{Ii}\pi_{k|i})^h$ de $n = n^h n_i^h$ es $[\frac{n^h}{N^h} \frac{n_i^h}{N_i^h} \dots \frac{n^h}{N^h} \frac{n_i^h}{N_i^h}]^T$.

En el Esquema B de Raj las UPM dentro de los estratos se seleccionan con probabilidad proporcional al tamaño con reemplazo (PPT) y cada vez que se extrae una UPM se realiza el muestreo dentro de ella. Es decir, se seleccionan n^h de las N^h UPM con PPT del conglomerado con reemplazo y en cada UPM en la muestra se seleccionan n_i^h de los N_i^h individuos con MASSR ($\pi_{Ii} = \frac{N_i^h}{N^{(h)}}$ y $\pi_{Iij} = \frac{N_i^h(N_j^h)}{N^{(h)}(N^{(h)})}$ con $N^{(h)} = \sum_{i=1}^{N^h} N_i^h$). En el estrato h , el vector $\boldsymbol{\pi}^h$ de $n^h n_i^h$ es $[\frac{n_i^h}{N^{(h)}} \dots \frac{n_i^h}{N^{(h)}}]^T$.

1.1. Estimador de regresión del total basado en modelo

El predictor óptimo para β bajo este modelo es $\hat{\beta} = (\mathbf{X}_s^T \mathbf{V}_s^T \mathbf{X}_s)^{-1} \mathbf{X}_s^T \mathbf{V}_s^T \mathbf{Y}_s$, donde el subíndice s hace referencia a las n observaciones en la muestra correspondientes a \mathbf{X} , \mathbf{V} y \mathbf{Y} . El total estimado es $\hat{T}_{BM} = \mathbf{1}_s^T \mathbf{Y}_s + \mathbf{1}_r^T \mathbf{X}_r \hat{\beta} = \mathbf{g}_s^T \mathbf{Y}_s$, donde r indica las $N - n$ observaciones de la población que no están en la muestra y $\mathbf{g}_s^T = \mathbf{1}_s^T + \mathbf{1}_r^T \mathbf{X}_r (\mathbf{X}_s^T \mathbf{V}_s^{-1} \mathbf{X}_s)^{-1} \mathbf{X}_s^T \mathbf{V}_s^{-1}$ es el vector de pesos. Valliant *et al.* (2000) presentan siete aproximaciones de $\hat{V}(\hat{T}_{BM})$, pero en este trabajo solo se estudian dos: $V_1(\hat{T}_{BM}) = \sum_{i=1}^{n^h} \mathbf{g}_i^T (\mathbf{r}_i \mathbf{r}_i^T) \mathbf{g}_i$ y $V_2(\hat{T}_{BM}) = \sum_{i=1}^{n^h} \mathbf{a}_i^T (\mathbf{r}_i \mathbf{r}_i^T) \mathbf{a}_i$, donde \mathbf{g}_i , \mathbf{r}_i y \mathbf{a}_i son la parte correspondiente al cluster i en la muestra de \mathbf{g}_s , $\mathbf{r}_s = \mathbf{Y}_s - \mathbf{X}_s \hat{\beta}$ y $\mathbf{a}_s = \mathbf{V}_s^{-1} \mathbf{X}_s (\mathbf{X}_s^T \mathbf{V}_s^{-1} \mathbf{X}_s)^{-1} \mathbf{X}_s^T \mathbf{1}_r^T$. Estos estimadores se denotan por VBM1 y VBM2.

1.2. Estimador de regresión del total asistido por modelo

Dado el modelo de trabajo $E_M(\mathbf{Y}) = \mathbf{X}\beta$, $\text{var}_M(\mathbf{Y}) = \mathbf{V} = \text{diag}(\sigma_1^2 \dots \sigma_N^2)$, el estimador del total en este enfoque es el estimador de regresión generalizado (GREG) y está dado por $\hat{T}_{GREG} = \mathbf{1}_s^T \boldsymbol{\Pi}_s^{-1} \mathbf{Y}_s + (\mathbf{1}^T \mathbf{X} - \mathbf{1}_s^T \boldsymbol{\Pi}_s^{-1} \mathbf{X}_s) \hat{\mathbf{B}}$ donde $\hat{\mathbf{B}} = \mathbf{A}_{\pi s}^{-1} \mathbf{X}_s^T \mathbf{V}_s^{-1} \boldsymbol{\Pi}_s^{-1} \mathbf{Y}_s$, $\mathbf{A}_{\pi s} = \mathbf{X}_s^T \mathbf{V}_s^{-1} \boldsymbol{\Pi}_s^{-1} \mathbf{X}_s$, $\mathbf{V}_s = \text{diag}(v_{ii})$, $\boldsymbol{\Pi}_s = \text{diag}(\pi_i)$, $g_{isB} = 1 + (\mathbf{1}^T \mathbf{X} - \mathbf{1}_s^T \boldsymbol{\Pi}_s^{-1} \mathbf{X}_s)^T \mathbf{A}_{\pi s}^{-1} \mathbf{x}_i / v_{ii}$.

La varianza aproximada estimada es

$$\hat{V}(\hat{T}_{GREG}) = \sum_{i=1}^{n^h} \sum_{j=1}^{n^h} \left(\frac{\pi_{Iij} - \pi_{Ii}\pi_{Ij}}{\pi_{Iij}} \right) \frac{\hat{t}_{Ei} \hat{t}_{Ej}}{\pi_{Ii} \pi_{Ij}} - \sum_{i=1}^{n^h} \frac{1}{\pi_{Ii}} \left(\frac{1}{\pi_{Ii}} - 1 \right) \hat{V}_{BEi} + \sum_{i=1}^{n^h} \frac{\hat{V}_{BEi}}{\pi_{Ii}^2}$$

con $\hat{V}_{BEi} = \sum_{k=1}^{n_i^h} \sum_{l=1}^{n_i^h} \left(\frac{\pi_{kl|i} - \pi_{k|i} \pi_{l|i}}{\pi_{kl|i}} \right) \frac{g_{ksB} e_{ks}}{\pi_{k|i}} \frac{g_{lsB} e_{ls}}{\pi_{l|i}}$, $\hat{t}_{Ei} = \sum_{k=1}^{n_i^h} g_{ksB} e_{ks} / \pi_{k|i}$ y $e_{ks} = y_k - \mathbf{x}'_k \hat{\mathbf{B}}$. Este estimador se denominará como VAM.

1.3. Estimador de regresión cosméticamente calibrado

Los estimadores calibrados buscan asegurar la consistencia con totales de variables auxiliares especificadas por el usuario y los estimadores cosméticos se construyen de manera que posean características de los estimadores basado en diseño y asistido por modelo. El estimador cosméticamente calibrado es

$$\begin{aligned} \hat{T}_C &= \mathbf{1}_s^T \boldsymbol{\Pi}_s^{-1} \mathbf{Y}_s + (\mathbf{1}^T \mathbf{X} - \mathbf{1}_s^T \boldsymbol{\Pi}_s^{-1} \mathbf{X}_s) [\mathbf{X}_s^T \mathbf{Z}_s^{-1} (\boldsymbol{\Pi}_s^{-1} - \mathbf{I}_s) \mathbf{X}_s]^{-1} \mathbf{X}_s^T \mathbf{Z}_s^{-1} (\boldsymbol{\Pi}_s^{-1} - \mathbf{I}_s) \mathbf{Y}_s \\ &= \mathbf{1}_s^T \mathbf{Y}_s + (\mathbf{1}^T \mathbf{X} - \mathbf{1}_s^T \mathbf{X}_s) [\mathbf{X}_s^T \mathbf{Z}_s^{-1} (\boldsymbol{\Pi}_s^{-1} - \mathbf{I}_s) \mathbf{X}_s]^{-1} \mathbf{X}_s^T \mathbf{Z}_s^{-1} (\boldsymbol{\Pi}_s^{-1} - \mathbf{I}_s) \mathbf{Y}_s \end{aligned}$$

donde \mathbf{Z}_s es una matriz diagonal de $n \times n$ tal que $\mathbf{Z}_s \mathbf{1}_s = \mathbf{X}_s \boldsymbol{\beta}$. Brewer (2002) aproxima $\hat{V}(\hat{T}_c)$ mediante la varianza en diseño del estimador de Horvitz-Thompson y mediante la varianza en diseño del GREG. En ese libro se sugiere que el mejor estimador de la varianza es $\hat{V}(\hat{Y}_{GREG}) = \frac{n}{n-p} \sum_{i=1}^n \pi_i^{-1} (\pi_i^{-1} - 1) \left(Y_i - \hat{Y} \right)^2$, al cual denominamos por VCC.

2. Simulación y resultados

Los datos usados son tomados de Valliant *et al.* (2000). La variable de interés es la calificación de matemáticas en tercer grado (Total = 1159 382.6) y se consideran 8 variables auxiliares: sexo, lengua de la prueba hablada en casa (siempre, algunas veces, nunca), etnia (blanco no hispano, negro, hispano, asiático, nativo americano u otro), inscripción en la escuela. Los datos están distribuidos en cuatro estratos (Regiones: noreste, sur, centro y oeste) y cada estrato está formado por conglomerados (135 escuelas distribuidas en 24, 37, 23, 51 en los estratos). El número total de individuos (estudiantes) es de 2 427 distribuidos en las regiones en 469, 663, 438 y 857.

Se estudia el efecto del esquema de muestreo (A o B), número de UPM (C1, C2 o C3) y número de USM (2 o 6). C1, C2 y C3 se diferencian por el número de UPM que se seleccionan de cada una de las cuatro regiones, en C1 se seleccionan 2, 3, 2 y 3 UPM, mientras que en C2 se seleccionan 6, 8, 6 y 10 UPM y en C3 se seleccionan 8, 12, 8 y 17. Para comparar los estimadores se utilizan cuatro criterios: porcentaje de sesgo relativo (PSR)

$$100 * \hat{E} \left((\hat{T} - T) / T \right),$$

raíz de error cuadrático medio relativo (RECMR)

$$\left(\sqrt{\hat{E} \left((\hat{T} - T)^2 \right)} \right) / T,$$

varianza estimada y cubrimiento (bajo normalidad y $1-\alpha=0.95$).

Se obtuvieron 10 000 muestras con los esquemas estudiados y se calcularon los estimadores de regresión estudiados además de sus estimadores de varianza. En poblaciones con clusters \mathbf{V}_s no es conocida y/o es difícil de estimar. En la simulación se supuso que $\mathbf{V} = \mathbf{I}$ lo cual es incorrecto pero usual en programas estadísticos no especializados.

Los estimadores estudiados tuvieron menor porcentaje de sesgo relativo y raíz de error cuadrático medio relativo en el esquema A. En el esquema A, el estimador AM tiene menor PSR y el estimador BM en el esquema B. El estimador BM tiene ligeramente menor RECMR que los otros estimadores en ambos esquemas. Como era de esperarse, los estimadores complejos son más sesgados y tienen mayor RECMR en el Caso 1 que en los otros casos. En el mismo sentido, siempre es mayor el RECMR de los estimadores con 2 USM que con 6 USM al igual que el PSR. Los estimadores complejos AM y CC tienen menor PSR con 2 USM en C2 y C3.

Los estimadores AM y BM generalmente tienen menor varianza estimada con el esquema B. En el esquema A los mejores estimadores son VBM cuando se usan 2 USM y VCC cuando se usan 6. En el esquema B es mejor el VAM. Los estimadores tienen mayor varianza en C1. Siempre es mayor la varianza de los estimadores cuando se toman 2 USM.

Los estimadores AM y BM generalmente tienen mayor cubrimiento con el Esquema A, esto se explica porque tienen mayor varianza en este esquema. En el Esquema B, el estimador CC tiene cubrimiento de 1 en C2 y C3 pero esto se debe a que su varianza estimada es muy

grande. El mejor estimador es BM, pues tiene mayor cubrimiento y mas cercano al nominal de 0.95 (ver Tabla 1).

3. Conclusiones

Los datos usados para la simulación favorecen al esquema A pues el número de USM varia poco (entre 7 y 29), esto causa que las probabilidades de inclusión definidas por el PPT sean muy semejantes a las de MASSR. Para estos datos, se puede decir que el mejor estimador es el basado en modelo en el Esquema B donde tuvo menor sesgo relativo, menor raíz de error cuadrático medio relativo y estimación moderada de la varianza y cubrimiento. La comparación se hace suponiendo que $\mathbf{V} = \mathbf{I}$, lo cual es falso. Otras opciones son $\mathbf{V} = \mathbf{W}^{-1}$, esto es, con la matriz de factores de expansión, o estimando a \mathbf{V} mediante las ecuaciones de estimación generalizadas.

Referencias

- Brewer, K. (2002). *Combined Survey Sampling Inference, Weighing Basu's Elephants*. London: Arnold.
- Raj, D. (1968). *Sampling Theory*. McGraw-Hill: New York.
- Särndal, C.E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer Verlag.
- Valliant, R., Dorfman, A.H. and Royall, R.M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. New York: John Wiley & Sons, Inc.

Caso	USM		A	B	A/B ^a
C1	2	VAM	0.72	0.31	2.33
		VBM1	0.71	0.69	1.04
		VBM2	0.71	0.68	1.04
		VCC	0.80	0.99	0.81
	6	VAM	0.80	0.35	2.28
		VBM1	0.80	0.78	1.03
		VBM2	0.80	0.77	1.03
		VCC	0.72	0.99	0.73
C2	2	VAM	0.90	0.70	1.28
		VBM1	0.90	0.87	1.04
		VBM2	0.90	0.86	1.04
		VCC	0.90	1.00	0.90
	6	VAM	0.92	0.73	1.25
		VBM1	0.94	0.90	1.04
		VBM2	0.92	0.88	1.05
		VCC	0.80	1.00	0.80
C3	2	VAM	0.93	0.83	1.11
		VBM1	0.94	0.88	1.06
		VBM2	0.93	0.87	1.06
		VCC	0.93	1.00	0.93
	6	VAM	0.94	0.88	1.06
		VBM1	0.96	0.92	1.04
		VBM2	0.93	0.89	1.05
		VCC	0.82	1.00	0.82

^aCubrimiento A/Cubrimiento B

Tabla 1: Cubrimiento con las varianzas estudiadas en ambos esquemas

Una generalización de la prueba de Shapiro-Wilk para normalidad multivariada

Elizabeth González Estrada^a, José A. Villaseñor Alva^b

Colegio de Postgraduados

1. Introducción

En la literatura se encuentra un número considerable de formas de valorar la hipótesis de normalidad multivariada. Algunas referencias recientes son Székely y Rizo (2005) y Farrell et al. (2007) y la revisión hecha por Mecklin y Mundfrom (2005). Mecklin y Mundfrom (2005) y Farrell et al. (2007) recomiendan la prueba de Henze y Zirkler (1990) como prueba formal para normalidad multivariada. Las pruebas de Mardia (1970) son las pruebas clásicas para normalidad multivariada.

Es bien conocido que la prueba de Shapiro-Wilk es una de las mejores pruebas para normalidad univariada. Con el propósito de obtener una prueba para normalidad multivariada que herede las buenas propiedades de potencia de la prueba de Shapiro-Wilk, en este trabajo usamos la estadística de Shapiro-Wilk como base para construir una prueba de bondad de ajuste para la distribución normal multivariada después de estandarizar empíricamente las observaciones.

2. Prueba de Shapiro-Wilk (SW)

Sean x_1, \dots, x_n las observaciones de una muestra aleatoria (m.a.) de tamaño n y sean $x_{(1)} < x_{(2)} < \dots < x_{(n)}$ las estadísticas de orden correspondientes.

^aeliza_ge@yahoo.com.mx

^bjvillasr@colpos.mx

Para probar normalidad univariada, Shapiro y Wilk (1965) proponen la estadística

$$W_X = \left(\sum_{i=1}^n a_i x_{(i)} \right)^2 / \sum_{i=1}^n (x_i - \bar{x})^2 \quad (1)$$

donde $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ y a_i , $i = 1, 2, \dots, n$, son ciertas constantes.

Se rechaza normalidad con un tamaño de prueba α si $W_X < k_\alpha$, donde k_α denota el percentil $100\alpha\%$ de la distribución de W_X bajo la hipótesis nula.

Usando simulación, Royston (1992) encuentra que bajo la hipótesis de normalidad la cola superior de la distribución de $\log(1 - W_X)$ se puede ajustar con una distribución normal con parámetros

$$\mu_n = -1.5861 - .31082y - 0.083751y^2 + .0038915y^3,$$

$$\sigma_n = \exp(-.4803 - .082676y + .0030302y^2)$$

donde $y = \log n$ y el tamaño de muestra n es tal que $11 < n < 2000$.

3. Generalización de la prueba de Shapiro-Wilk para normalidad multivariada

Sea $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ una m.a. p -variada de tamaño $n > p \geq 1$. Sea $\mathbf{N}^p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ la función de densidad normal p -variada con parámetros vector de medias $\boldsymbol{\mu}$ y matriz de varianzas y covarianzas $\boldsymbol{\Sigma}$.

Aquí se propone una prueba para el juego de hipótesis

$$H_0 : \mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n \text{ es m.a. de } \mathbf{N}^p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \text{ contra } H_1 : \mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n \text{ no es m.a. de } \mathbf{N}^p(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

donde $\boldsymbol{\mu}$ y $\boldsymbol{\Sigma}$ son desconocidos, con base en la siguiente caracterización de la distribución normal multivariada.

Proposición 1.1. $\mathbf{X} \sim \mathbf{N}^p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ si y sólo si $\mathbf{Z} = \boldsymbol{\Sigma}^{-1/2}(\mathbf{X} - \boldsymbol{\mu}) \sim \mathbf{N}^p(\mathbf{0}, \mathbf{I})$, donde $\mathbf{0}$ es el vector de ceros de orden p e \mathbf{I} es la matriz identidad de orden $p \times p$.

Sean $\bar{\mathbf{X}}$ y \mathbf{S} el vector de medias y la matriz de covarianzas muestrales. Además, sea $\mathbf{S}^{-1/2}$ la matriz raíz cuadrada positiva definida simétrica de \mathbf{S}^{-1} , la matriz inversa de \mathbf{S} .

Cuando $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ es una m.a. de $\mathbf{N}^p(\mu, \Sigma)$, se espera que los vectores aleatorios

$$\mathbf{Z}_j^* = \mathbf{S}^{-1/2} (\mathbf{X}_j - \bar{\mathbf{X}}), \quad j = 1, 2, \dots, n, \quad (2)$$

tengan distribución cercana a la $\mathbf{N}^p(\mathbf{0}, \mathbf{I})$, por lo que se espera que las coordenadas de \mathbf{Z}_j^* , denotadas por Z_{1j}, \dots, Z_{pj} , serán aproximadamente independientes con distribución normal estándar univariada.

Para probar normalidad multivariada se propone la estadística

$$W^* = \frac{1}{p} \sum_{i=1}^p W_{Z_i}, \quad (3)$$

donde W_{Z_i} es la estadística de Shapiro-Wilk evaluada en la i -ésima coordenada de las observaciones transformadas en (2).

Bajo H_0 , se espera que W^* tome valores cerca de 1 ya que también se espera que cada $W_{Z_i}, i = 1, \dots, p$, tomará valores cercanos a uno.

Se rechaza H_0 con un tamaño de prueba α si $W^* < c_{\alpha;n,p}$, donde $c_{\alpha;n,p}$ es tal que:

$$\alpha = P \{ W^* < c_{\alpha;n,p} \mid H_0 \text{ es verdadera} \}.$$

La distribución de W^* bajo H_0 no depende de (μ, Σ) ya que es una función de las observaciones estandarizadas (Henze y Zirkler, 1990) y se puede obtener usando simulación de Monte Carlo.

3.1. Ajuste de la distribución de W^*

La cola superior de la distribución de $W_1^* = \log(1 - W^*)$ se puede ajustar por una distribución normal con varianza $\sigma_1^2 = \log \left\{ \frac{p-1+e^{\sigma_n^2}}{p} \right\}$ y media $\mu_1 = \mu_n + \frac{1}{2}\sigma_n^2 - \frac{1}{2}\sigma_1^2$.

Por lo tanto, $c_{\alpha;n,p} = 1 - \exp \{ \mu_1 + \sigma_1 \Phi^{-1}(1 - \alpha) \}$, en donde Φ^{-1} denota la inversa de la función de distribución normal estándar (Villaseñor-Alva y González-Estrada, 2008).

4. Estudio de potencia

La prueba propuesta se comparó con las pruebas MS y MK de Mardia (1970), $T_{.5}$ de Henze y Zirkler (1990), W_F de Mudholkar, Srivastava y Lin (1995) y M_1 Srivastava y Hui (1987).

La potencia de las pruebas se estimó usando simulación de Monte Carlo. Se simularon 5 000 muestras pseudo aleatorias de tamaño $n = 20, 50$ de dimensión $p = 2, 5$ de cada distribución alternativa. Se eligió un tamaño de prueba $\alpha = 0.05$.

En el Cuadro 1 se consideraron distribuciones alternativas p -variadas con marginales independientes con distribución f , denotadas como f^p , donde f es alguna de las distribuciones siguientes: logística, sum-estable, t de Student, beta, doble exponencial, gama y chi cuadrada las cuales se denotan como Logistic, S-Stab(α, β), t_k , $B(\alpha, \beta)$, Dexp(α, β), $G(\alpha, \beta)$, χ_v^2 , respectivamente.

También se incluyeron alternativas de la forma $f_1^{p-k} \times f_2^k$, la cual denota una distribución p -variada con $p - k$ marginales con distribución f_1 y k marginales con distribución f_2 , $k = 1, 2$. Además, se consideraron distribuciones alternativas esféricas con generador g , denotada como SPH(g), Pearson tipo VII ($PSVII_p(a)$) y mezclas de distribuciones normales p -variadas de la forma $\mathbf{N}^p(\mathbf{0}, \mathbf{R}_1)$ y $\mathbf{N}^p(\mu, \mathbf{R}_2)$ donde $\mu = \mu(1, 1, \dots, 1)'$, $\mu \in \Re$, y \mathbf{R}_i es una matriz con elementos de la diagonal iguales a uno y elementos fuera de la diagonal iguales a r_i , $i = 1, 2$ con parámetro de mezcla igual a k y se denota como NMIX(k, μ, r_1, r_2).

5. Conclusiones

La generalización propuesta de la prueba de Shapiro-Wilk (W^*) para probar normalidad multivariada es fácil de usar porque se pueden calcular los valores críticos para cualquier n y p usando la distribución univariada normal estándar.

En muchos casos W^* resulta ser más potente que las pruebas recomendadas para probar normalidad multivariada.

La distribución de W^* no depende de μ y Σ .

En dimensión 1, W^* se reduce a la estadística W de Shapiro-Wilk.

Referencias

- Farrell, P. J., Salibian-Barrera, M. and Naczk, K. (2007). On tests for multivariate normality and associated simulation studies. *Journal of Statistical Computation and Simulation* (77), 12: 1065-1080.

- Henze, N. and Zirkler, B. (1990). A class of invariant consistent tests for multivariate normality. *Communications in Statistics: Theory and Methods* **19** (10):3595-3617.
- Mardia, K.V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika* **57**: 519-530.
- Mecklin, C.J. and Mundfrom, D.J. (2005). A Monte Carlo comparison of the Type I and Type II error rates of tests of multivariate normality. *Journal of Statistical Computation and Simulation*, **75** (2): 93-107.
- Mudholkar, G., Srivastava, D. and Lin, C. (1995). Some p-variate adaptations of the Shapiro-Wilk test of normality. *Communications in Statistics: Theory and Methods*, **24** (4): 953-985.
- Royston, P. (1992). Approximating the Shapiro-Wilk W test for non-normality. *Statistics and Computing* **2**:117-119.
- Shapiro, S.S. and Wilk, M.B. (1965). An analysis variance tests for normality (complete samples). *Biometrika* **52**(3): 591-611.
- Srivastava, M.S. and Hui, T.K. (1987). On assessing multivariate normality based on Shapiro-Wilk W statistic. *Statistics and Probability Letters*, **2**, 263-267.
- Székely, G.J. and Rizzo, M.L. (2005). A new test for multivariate normality. *Journal of Multivariate Analysis*, **93** (1): 58-80.
- Villaseñor-Alva, J.A. and González-Estrada, E. (2008). A generalization of Shapiro-Wilk's test for multivariate normality. *Comm. in Stat.: Theory and Methods* (por aparecer).

Alternativa	$T_{0.5}$	MS	MK	M_1	W_F	W^*
N(0,1) ⁵	5	4	2	1	5	5
N(100,.0001)*N(100,10) ⁴	6	4	2	1	42	5
Dexp ⁵	74	68	81	38	91	93
Logistic ⁵	28	30	31	10	40	42
t_5^5	58	61	64	34	73	75
S-stab(1.75) ⁵	89	89	91	NA	93	94
B(1,1) ⁵	2	0	39	1	99	100
G(5,1) ⁵	80	68	28	30	96	98
$(\chi_{15}^2)^5$	61	47	18	17	83	86
N(0,1) ⁴ *S-Stab(1.5)	60	59	55	NA	67	71
N(0,1) ⁴ * t_5	16	16	12	12	20	23
N(0,1) ⁴ *B(1,1)	4	2	1	4	16	19
N(0,1) ⁴ * χ_5^2	33	25	10	45	48	61
N(0,1) ³ *S-Stab(1.5) ²	85	82	84	NA	91	93
N(0,1) ³ * t_5^2	27	27	24	19	37	40
N(0,1) ³ *B(1,1) ²	4	1	1	4	38	48
N(0,1) ³ * $(\chi_5^2)^2$	65	51	20	61	85	93
SPH(G(5,1))	60	57	69	21	29	28
SPH(B(1,1))	88	59	93	46	65	66
SPH(B(1,2))	100	98	100	93	97	96
PSVII(5)	88	87	92	63	62	60
PSVII(10)	28	30	30	10	14	15
NMIX(.5,2,0,0)	6	2	1	38	5	5
NMIX(.75,2,0,0)	26	7	1	67	9	5
NMIX(.75,2,.9,0)	100	100	100	96	95	98
NMIX(.75,2,0,.9)	91	89	20	94	37	39

Tabla 1: Potencia de las pruebas para NMV en % (p=5, n=50, $\alpha=0.05$)

Ecuaciones diferenciales en la modelación de datos funcionales

María Guzmán Martínez^a, Eduardo Castaño Tostado

Universidad Autónoma de Querétaro

1. Introducción

Datos funcionales son resultado de experimentos u observaciones en múltiples contextos. Un dato funcional se conceptualiza a partir de un conjunto de observaciones del que es factible suponer que surge del registro (discreto) de una función subyacente en el fenómeno de interés.

El presente trabajo expone una aplicación de la teoría y de los programas de cómputo para modelar datos funcionales, en particular en casos en que pueden ser descritos por medio de una ecuación diferencial ordinaria. Esta teoría y sus programas computacionales genéricos fueron desarrollados por Ramsay y Silverman (2005).

La aplicación se realiza sobre el *Indicador Global de la Actividad Económica Mexicana*, a partir del adecuamiento específico de los programas de cómputo disponibles.

En las secciones dos y tres se describe la metodología de utilizada y en la sección cuatro se da la aplicación de ésta. Este trabajo es parte de la tesis para obtener el título de Licenciatura en Matemáticas Aplicadas del primer autor.

^amarnezmar@yahoo.com.mx

2. Ecuaciones diferenciales ordinarias en el análisis de datos funcionales

Una ecuación diferencial ayuda a entender el comportamiento dinámico de una función dada $x(t)$; así, teoría básica puede sugerir que bajo ciertas condiciones un fenómeno debe ser aproximadamente regido por una ecuación diferencial en particular. Sin embargo en muchos casos, $x(t)$ no es conocida sino sólo a través de datos del fenómeno $x_j = x(t_j) + \epsilon(t_j)$, $j = 1, \dots, n$; en tales circunstancias, en primer lugar se debe estimar $x(t)$; en segundo lugar estimar los coeficientes de la ecuación diferencial sugerida por la teoría, y finalmente diagnosticar el grado de ajuste de $x(t)$ estimada a la ecuación diferencial. Este tipo de análisis recibe el nombre de Análisis Diferencial Principal de datos funcionales (*ADP*).

3. Estimación de operadores diferenciales en el análisis de datos funcionales

Una ecuación diferencial ordinaria de orden m , con coeficientes variables es

$$D^m x(t) + \beta_{m-1}(t) D^{m-1} x(t) + \cdots + \beta_1(t) D x(t) + \beta_0(t) x(t) = 0, \quad (1)$$

donde $D^m(x(t)) = \frac{d^m x(t)}{dt^m}$.

Si se denota por $L = D^m + \beta_{m-1}(t) D^{m-1} + \cdots + \beta_1(t) D + \beta_0(t)$, la ecuación (1) se reduce a

$$Lx(t) = 0$$

Sabemos que si $\xi_1, \xi_2, \dots, \xi_m$ son las m soluciones linealmente independientes de $Lx(t) = 0$, entonces $c_1\xi_1(t) + c_2\xi_2(t) + \cdots + c_m\xi_m(t)$ es también una solución de $Lx(t)$. El conjunto de todas las funciones ξ para el cual $L\xi_j = 0$, es llamado el espacio nulo de L y se denota con $\ker L$; de hecho las funciones ξ forman una base para este espacio.

La ecuación (1) también puede ser escrita como

$$D^m x(t) = - \sum_{j=0}^{m-1} \beta_j(t) D^j x(t)$$

En muchas circunstancias prácticas al aplicar L a $x(t)$ se obtiene $Lx(t) = f(t) \neq 0$; la función $f(t)$ recibe el nombre de *función forzada* o “*residual’ funcional*”.

En aplicaciones se cuenta con datos x_1, \dots, x_n , que bajo el paradigma del análisis de datos funcionales, provienen de un modelo subyacente

$$x_j = x(t_j) + \epsilon(t_j).$$

Entonces, antes de un *ADP* es requerido estimar $x(t)$; con tal fin es útil pensar en que

$$x(t) = \sum_{k=1}^K a_k \phi_k(t)$$

donde $\{\phi_k(t)\}$ es un conjunto dado de funciones base; entonces obtener $\hat{x}(t)$ es equivalente a estimar $\{a_k\}$. Una vez hecho lo anterior, lo que se busca es estimar un operador diferencial L de orden m de interés en la aplicación, tal que bajo condiciones ideales

$$L\hat{x}(t) = 0.$$

En general se tienen N observaciones funcionales $\hat{x}_i(t)$ generadas de un conjunto de datos x_{ij} , $i = 1, \dots, N$ y $j = 1, \dots, n$; y las funciones β_j pueden ser estimadas usando el criterio de ajuste para cada t

$$\min_L PSSE_L(t) = \sum_{i=1}^N \|L\hat{x}_i(t)\|^2 = \sum_{i=1}^N \left[\left(\sum_{j=0}^{m-1} \beta_j(t) D^j + D^m \right) \hat{x}_i(t) \right]^2 \quad (2)$$

el problema se resuelve por mínimos cuadrados. Definimos un vector de coeficientes de dimensión m

$$\beta(t) = (\beta_0(t), \dots, \beta_{m-1}(t))';$$

también se define la matriz de diseño \mathbf{Z} para cada t , de orden $N \times (m+1)$ con renglones

$$\mathbf{z}_i(t) = \{-\hat{x}_i(t), \dots, -D^{m-1}\hat{x}_i(t)\};$$

por último definimos el vector \mathbf{y} de dimensión N con elementos

$$y_i(t) = D^m \hat{x}_i(t)$$

Con estas definiciones podemos expresar el criterio (2) en términos matriciales como

$$\min_{\beta} PSSE_L(t) = [\mathbf{y}(t) - \mathbf{Z}(t)\beta(t)]'[\mathbf{y}(t) - \mathbf{Z}(t)\beta(t)].$$

Así

$$\hat{\beta}(t) = [\mathbf{Z}(t)' \mathbf{Z}(t)]^{-1} \mathbf{Z}(t)' \mathbf{y}(t).$$

4. Aplicación

El Índice Económico Global de México permite monitorear la evolución del nivel de desarrollo económico de México. Los datos mensuales de este índice de 1993 al 2005 se muestran como puntos en la gráfica de la Figura 1. De la observación de esta gráfica, se pueden apreciar un

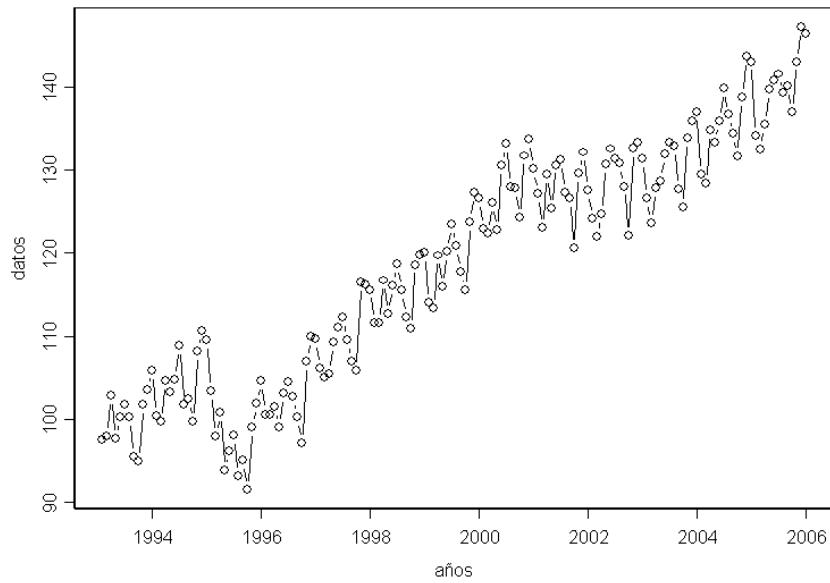


Figura 1: Gráfica de los valores discretos

componente de largo plazo, un componente estacional anual y variaciones propias de cada año observado. El *ADP* de este conjunto de datos resultó en lo siguiente:

1. $\hat{x}(t)$ estimada a partir de los datos discretos se muestra en la Figura 1. En este caso $x(t)$ fue estimado mediante un spline generado por segmentos de polinomios de grado 7.

2. Lo que se busca es una ecuación diferencial lineal de la forma

$$\beta_0(t)x(t) + \dots + \beta_{m-1}(t)D^{m-1}x(t) + D^mx(t)$$

a partir de la función $\hat{x}(t)$. Como se mencionó, del índice bajo estudio se pueden observar tres componentes de variación, por lo que se propone el siguiente operador diferencial, después de algo de experimentación

$$D^3x(t) = -\beta_0(t)x(t) - \beta_1(t)x(t) - \beta_2D^2x(t)$$

es decir con dos coeficientes variables y uno constante.

3. Se estiman los $\hat{\beta}_j$; estos se muestran en la Figura 2.

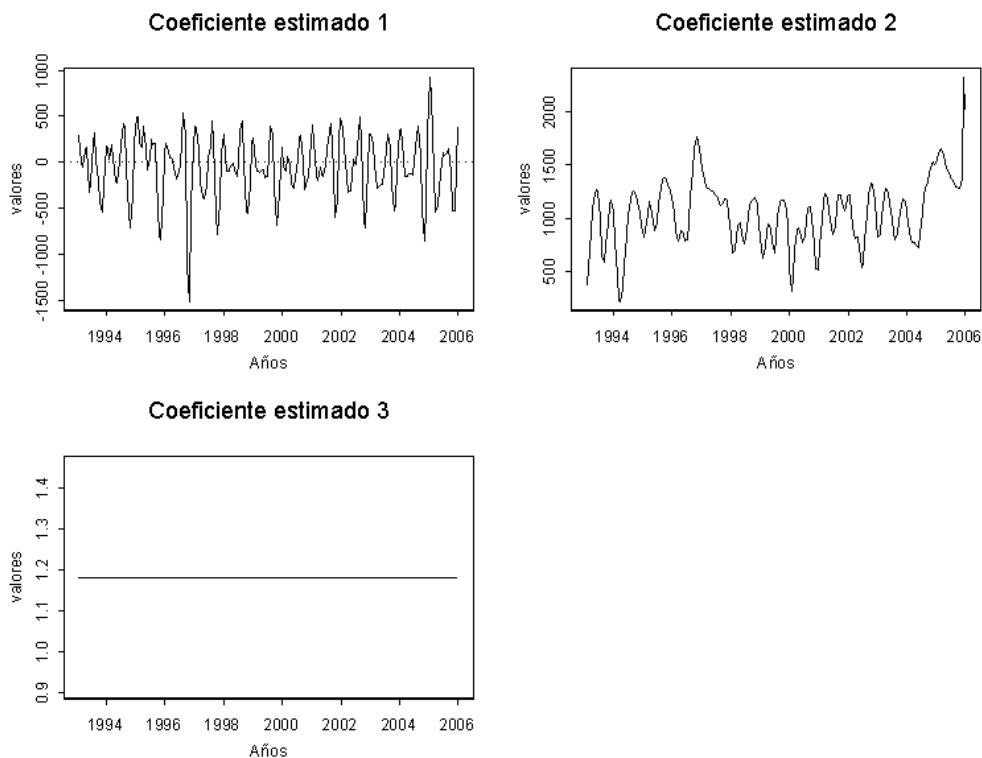


Figura 2: Coeficientes estimados

4. A partir de los $\hat{\beta}_j$, se construye la ecuación

$$L = \hat{\beta}_0(t)x(t) + \hat{\beta}_1(t)x(t) + 1.19D^2x(t) + D^3x(t).$$

Al hacer $L = 0$ y dado que este operador es de orden 3, se tienen tres soluciones linealmente independientes, que se muestran en la Figura 3.

De éstas generamos una combinación lineal, para obtener la función $\hat{x}_L(t)$ como la función estimada de la regresión lineal

$$\hat{x}_L(t) = c_1\xi_1(t) + c_2\xi_2(t) + c_3\xi_3(t) + \varepsilon;$$

así $\hat{x}_L(t)$ mejora a la función $\hat{x}(t)$, tal y como lo podemos ver en la Figura 4.

1. Dado que $L\hat{x}_L(t) = f \neq 0$, se tiene un "residual" funcional f que se muestra en la Figura 5.

Como se puede apreciar el *ADP* abre posibilidades de trabajar estadísticamente en el estudio de comportamientos dinámicos de fenómenos de interés.

Referencias

INEGI, *Indicador Global de la Actividad Económica*. Revisado Octubre 7, 2007 de <http://dgcnesyp.inegi.gob.mx/cgi-win/bdieintsi.exe/Consultar>.

Ramsay J.O. and Silverman B.W. (2005). *Functional Data Analysis*. United States of America: Springer.

Shepley R.L.(1984) *Introducción a las Ecuaciones Diferenciales*. Interamericana, tercera edición.

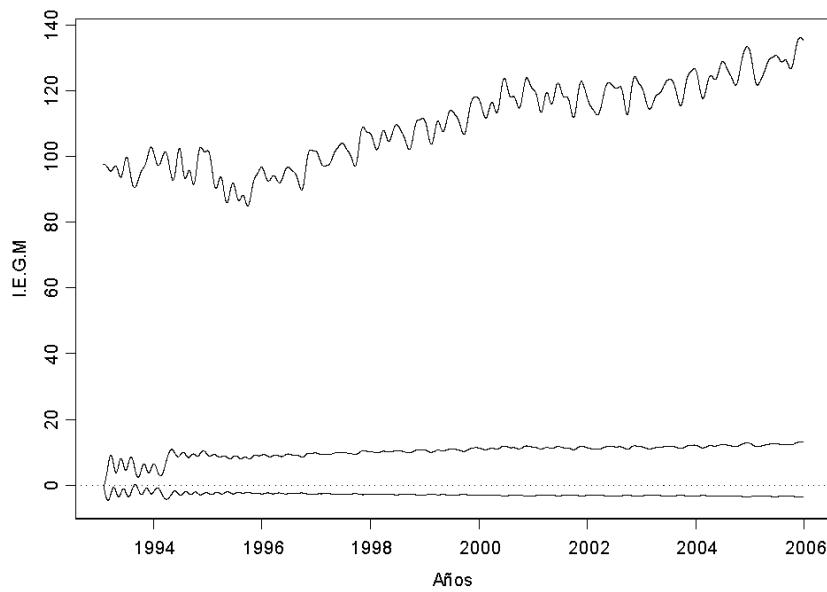


Figura 3: Soluciones estimadas

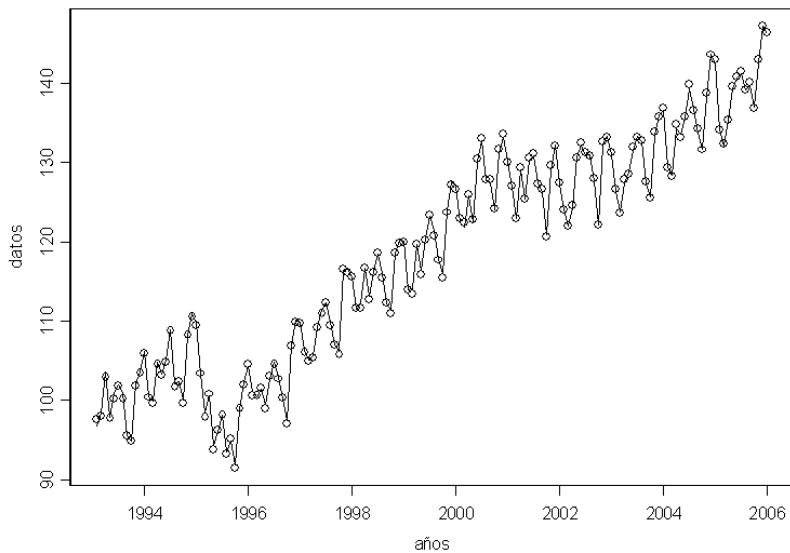


Figura 4: Modelo diferencial estimado

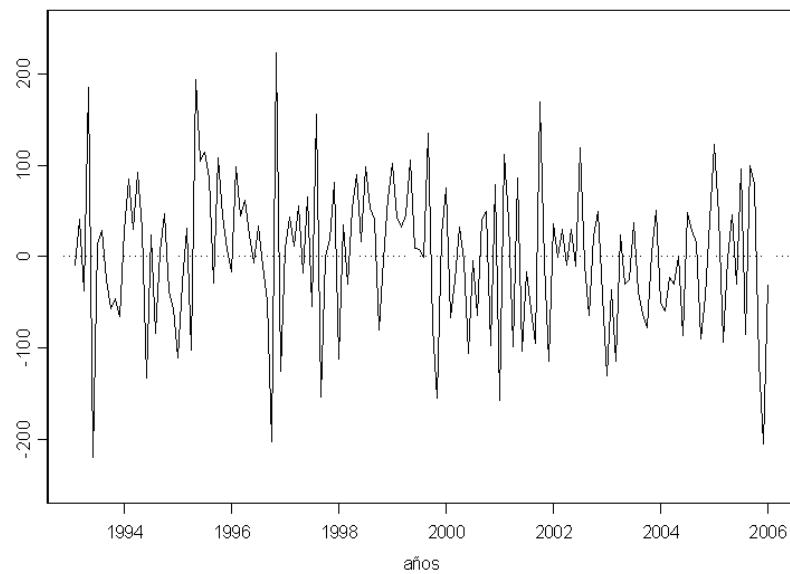


Figura 5: Residual funcional

Modelo de decremento múltiple semiparamétrico para datos de supervivencia*

Angélica Hernández Quintero^b *Universidad Autónoma Metropolitana – Iztapalapa*

Jean François Dupuy^c *Université Paul Sabatier 3, Francia*

Gabriel Escarela^d *Universidad Autónoma Metropolitana – Iztapalapa*

1. El modelo de mezclas semiparamétricas

En varias disciplinas tales como la bioestadística y la ciencia actuarial es común encontrar datos de decremento múltiple, los cuales consisten de observaciones de duraciones de un punto origen a la ocurrencia de un evento en particular que -a su vez- puede ser clasificado en J categorías mutuamente excluyentes; es posible que los datos también incluyan observaciones censuradas por la derecha, las cuales son duraciones que no han alcanzado la ocurrencia del evento al final del seguimiento. El análisis de datos de decremento múltiple provee un marco general para la predicción de la ocurrencia de cierto evento en el tiempo t en presencia de otros tipos de eventos.

El propósito de este trabajo es el estudiar las propiedades asintóticas de un modelo semiparamétrico de mezclas el cual extiende las ideas del modelo de Cox de una sola clasificación del evento para datos de supervivencia. Sea T_j el tiempo de ocurrencia del evento de tipo j , $j = 1, \dots, J$, y $T = \min(T_1, T_2, \dots, T_J)$ el tiempo de la primera ocurrencia del evento por cualquier causa. Es posible especificar a la función de supervivencia global, definida como $S_T(t) = \Pr\{T > t\}$, en términos de una mezcla de las funciones de supervivencia condicionales, las cuales pueden ser especificadas con el modelo

*Trabajo realizado con el auspicio del proyecto CONACYT-ANUIES-ECOS, No. M06-M01

^bangyka302@gmail.com

^cdupuy@cict.fr

^dge@xanum.uam.mx

de riesgos proporcionales de Cox de manera tal que $S_T(t; \mathbf{x}) = \sum_{j=1}^J p_j S_j(t; \mathbf{x})$, donde $S_j(t; \mathbf{x}) = \Pr\{T > t \mid T = T_j; \mathbf{x}\} = \exp\left\{-\Lambda_{j0}(t)e^{\beta'_j \mathbf{x}}\right\}$, $\mathbf{x} = (x_1, x_2, \dots, x_p)$ es el vector de variables explicativas, β_j su vector de coeficientes y $p_j = \Pr\{T = T_j\}$ es la probabilidad de que el evento sea de tipo j , donde $\sum_{j=1}^J p_j = 1$; una forma conveniente para modelar esta probabilidad es suponer que la causa de muerte específica tiene distribución multinomial siguiendo el modelo, $p_j(\mathbf{z}) = \exp(\boldsymbol{\delta}'_j \mathbf{z}) / \sum_{l=1}^J \exp(\boldsymbol{\delta}'_l \mathbf{z})$, donde $\mathbf{z} = (1, z_1, \dots, z_q)$ es el vector de variables explicativas y $\boldsymbol{\delta}_j$ es el correspondiente vector de coeficientes para la causa j (Cox and Snell, 1989). Para evitar redundancia se toma $\boldsymbol{\delta}_J$ igual a cero. Obsérvese que el vector \mathbf{z} puede contener algunas o todas las variables explicativas del vector \mathbf{x} , así como también otras variables que no estén incluidas en \mathbf{x} .

1.1. Inferencia a través del algoritmo EM

Sea T_{ij} el tiempo de supervivencia del j -ésimo tipo de evento para el i -ésimo individuo, $i = 1, \dots, n$. El conjunto de datos a analizar es $\{X_i, c_{ij}\}$, donde $X_i = \min(T_i, C_i)$, T_i es el tiempo de ocurrencia del primer evento para el i -ésimo individuo, C_i es el tiempo de censura, c_{ij} es la matriz indicadora de estatus definida como $c_{ij} = I(T_i = T_{ij})$ y $c_{i\cdot} = \sum_{j=1}^J c_{ij}$. Para evitar ambigüedad, se asumirá que para todo $j \neq k$, $T_{ij} \neq T_{ik}$.

Como existen observaciones incompletas, debido a que existen individuos que presenta censura, defínase una nueva variable γ_{ij} la cual toma el valor de 1 si el individuo i muere de la causa j y en cualquier otro caso toma el valor 0. Obsérvese que si $c_{i\cdot} = 1$ entonces $\gamma_{ij} = c_{ij}$ para todo j ; sin embargo, si $c_{i\cdot} = 0$, entonces γ_{ij} es indefinida para toda j . Por lo tanto $p_{ij} = p_j(\mathbf{z}_i) = \Pr\{\gamma_{ij} = 1 \mid \mathbf{z}_i\}$ y la matriz $\mathbf{G} = [\gamma_{ij}]$, son parcialmente observadas lo que permite usar el algoritmo EM, como lo propone Dempster et al., (1997). La función de verosimilitud completa para n individuos puede ser escrita como,

$$L_C = \prod_{i=1}^n \left\{ \left(\prod_{j=1}^J [p_{ij} f_j(t_i)]^{c_{ij} \gamma_{ij}} \right) \times \left(\prod_{j=1}^J [p_{ij} S_j(t_i)]^{(1-c_{i\cdot}) \gamma_{ij}} \right) \right\}.$$

El paso E, calcula la esperanza de $\log L_C$, la cual es igual a $l_p + l_S$ definidas como,

$$\begin{aligned} l_p &= \sum_{i=n}^n \sum_{j=1}^J g_{ij} \log(p_{ij}) \quad \text{y} \\ l_S &= \sum_{i=n}^n \sum_{j=1}^J c_{ij} [\log(\lambda_{j0}(t_i)) + \boldsymbol{\beta}'_j \mathbf{x}_i] - g_{ij} \Lambda_{j0}(t) \exp\{\boldsymbol{\beta}'_j \mathbf{x}_i\}, \end{aligned} \quad (1)$$

donde g_{ij} es la esperanza de γ_{ij} dadas las estimaciones de p_{ij} y S_{ij} y cuyo valor es $g_{ij} = c_{ij} + (1 - c_i)w_{ij}$; aquí, $w_{ij} = \Pr\{\gamma_{ij} \mid T > t_i\} = p_{ij}S_j(t_i) / \sum_{l=1}^J p_{il}S_l(t_i)$. Para cada causa j sean $t_{j,1} < \dots < t_{j,k_j}$ los distintos tiempos de muerte observados. Una aproximación de la ecuación (1) puede ser la propuesta por Breslow (1974):

$$l_S \approx \log \left(\prod_{j=1}^J \prod_{l=1}^{k_j} \frac{\exp(\boldsymbol{\beta}'_j \mathbf{s}_{j,l})}{\left[\sum_{m \in R_{j,l}} g_{mj} \exp(\boldsymbol{\beta}'_j \mathbf{x}_m) \right]^{\mathbf{d}_{j,l}}} \right), \quad (2)$$

donde $\mathbf{s}_{j,l} = \sum \mathbf{x}_i$ es la suma de las covariables de individuos que mueren de la causa j al tiempo $t_{j,l}$, $\mathbf{d}_{j,l}$ denota el número de muertes por la causa j al tiempo $t_{j,l}$ y $R_{j,l}$ denota el conjunto de individuos en riesgo antes de $t_{j,l}$. El paso M involucra la maximización de la función log-verosimilitud en la ecuación (2). Usando el estimador de producto límite, entonces el estimador de la función de supervivencia condicional es expresada como:

$$\hat{S}_{j0}(t) = \exp \left\{ - \sum_{m: t_{j,(m)} < \tau} \frac{\mathbf{d}_{jm}}{\sum_{m \in R_{jm}} g_{mj} \exp(\boldsymbol{\beta}'_j \mathbf{x}_m)} \right\}.$$

2. Normalidad asintótica

Sea $\hat{\theta} = \{(\hat{\boldsymbol{\beta}}_j, \hat{\delta}_j, \hat{\Lambda}_j), j = 1, \dots, J\}$ el vector de estimadores. Para demostrar la normalidad asintótica, es posible basarse en la idea propuesta por Murphy (1995) y en la teoría de procesos empíricos. En primera instancia, considérense submodelos unidimensionales respecto a los estimadores; esto es, el nuevo vector de parámetros toma la forma

$$\theta_t = \left\{ (\boldsymbol{\beta}_j + th_{\boldsymbol{\beta}_j}, \boldsymbol{\delta}_j + th_{\boldsymbol{\delta}_j}, \int_0^t (1 + th_{\Lambda_j}(s)) d\Lambda_j(s), j = 1, \dots, J \right\},$$

con $(h_{\boldsymbol{\beta}_j}, h_{\boldsymbol{\delta}_j}, h_{\Lambda_j}) \in H = \{(h_{\boldsymbol{\beta}_j}, h_{\boldsymbol{\delta}_j}, h_{\Lambda_j}) | h_{\boldsymbol{\beta}_j} \in \mathbb{R}^p, h_{\boldsymbol{\delta}_j} \in \mathbb{R}^q, \text{ y } h_{\Lambda_j} \text{ es una función de variación acotada en } [0, \tau]\}$, donde τ es el tiempo final de observación del estudio.

Proposición 2.1. *La derivada empírica puede ser expresada como*

$$S_n^{\tilde{\theta}} = \sum_{j=1}^J \left\{ S_{n\Lambda_j}^{\tilde{\theta}}(\theta)(h_{\Lambda_j}) + h'_{\beta_j} S_{n\beta_j}^{\tilde{\theta}} + h'_{\delta_j} S_{n\delta_j}^{\tilde{\theta}} \right\},$$

donde, $S_n^{\tilde{\theta}}(h) = \frac{1}{n} \frac{\partial}{\partial t} L_n^{\tilde{\theta}}(\theta_t)|_{t=0}$ y

$$\begin{aligned} S_{n\Lambda_j}^{\tilde{\theta}}(\theta)(h_{\Lambda_j})(s) &= \mathbb{P}_n \left[c_j h_{\Lambda_j}(\tau) - \int_0^\tau h_{\Lambda_j}(s) d\Lambda_j(s) e^{\beta' \mathbf{x}} E_{\tilde{\theta}}[g_j] \right], \\ S_{n\beta_j}^{\tilde{\theta}}(\theta) &= \mathbb{P}_n \left[c_j \mathbf{x} - \mathbf{x} e^{\beta' \mathbf{x}} E_{\tilde{\theta}}[g_j] \Lambda_j(\tau) \right], \\ S_{n\delta_j}^{\tilde{\theta}}(\theta) &= \mathbb{P}_n \left[E_{\tilde{\theta}}[g_j] \mathbf{z} - \sum_{k=1}^J \frac{E_{\tilde{\theta}}[g_j] \mathbf{z} e^{\delta'_k \mathbf{z}}}{\sum_{l=1}^J e^{\delta'_l \mathbf{z}}} \right]. \end{aligned}$$

Defínase la siguiente norma:

$$\|h\|_H = \sum_{j=1}^J \left\| h_{\beta_j} \right\| + \left\| h_{\delta_j} \right\| + \left\| h_{\Lambda_j} \right\|_V,$$

donde $\|h_{\Lambda_j}\|_V = |h_{\Lambda_j}(0)| + \int_0^\tau |dh_{\Lambda_j}(0)|$ y $\|\cdot\|$ es la norma Euclidiana. El siguiente resultado confirma que los estimadores son asintóticamente normales.

Teorema 2.1. *Sea $0 < r < \infty$, la sucesión*

$$\left\{ \left(\sqrt{n}(\hat{\beta}_{j,n} - \beta_j), \sqrt{n}(\hat{\delta}_{j,n} - \delta_j), \sqrt{n}(\hat{\Lambda}_{j,n} - \Lambda_j) \right), j = 1, \dots, J \right\}$$

converge en $\ell^\infty(H_r)$ a un proceso gaussiano centrado G con covarianza

$$\text{Cov}[G(g), G(g^*)] = \sum_{j=1}^J \left[\int_0^\tau g_{\Lambda_j}(u) \sigma_{\Lambda_j, \theta_0}^{-1}(g^*)(u) d\Lambda_j(u) + \sigma_{\delta_j, \theta_0}^{-1}(g^*) g_{\delta_j} + \sigma_{\beta_j, \theta_0}^{-1}(g^*) g_{\beta_j} \right],$$

donde $\sigma_{\theta_0}^{-1} = \left\{ (\sigma_{\beta_j, \theta_0}^{-1}, \sigma_{\delta_j, \theta_0}^{-1}, \sigma_{\Lambda_j, \theta_0}^{-1}), j = 1, \dots, J \right\}$ es un operador lineal continuamente invertible de σ_{θ_0} , el cual va de H_∞ a H_∞ y está definido por

$$\begin{aligned} \sigma_{\beta_j, \theta_0}(h) &= \int_0^\tau h_{\Lambda_j}(s) d\Lambda_{j,0}(s) e^{\beta'_{j,0} \mathbf{x}} \mathbf{x} E_{\theta_0}[g_j] + \mathbf{x} e^{\beta'_{j,0} \mathbf{x}} \mathbf{x}' E_{\theta_0}[g_j] \Lambda_{j,0}(\tau) h_{\beta_j}, \\ \sigma_{\delta_j, \theta_0}(h) &= \frac{\sum_{k=1}^J E_{\theta_0}[g_k] \left\{ \mathbf{z} e^{\delta'_{j,0} \mathbf{z}} \mathbf{z} \left(h'_{\delta_j} \sum_{l=1}^J e^{\delta'_{l,0} \mathbf{z}} - \sum_{l=1}^J h'_{\delta_l} e^{\delta'_{l,0} \mathbf{z}} \right) \right\}}{\left(\sum_{l=1}^J e^{\delta'_{l,0} \mathbf{z}} \right)^2}, \\ \sigma_{\Lambda_j, \theta_0}(h)(u) &= h_{\Lambda_j}(u) I_{\{u \leq T\}} e^{\beta'_{j,0} \mathbf{x}} E_{\theta_0}[g_j] + \mathbf{x} e^{\beta'_{j,0} \mathbf{x}} E_{\theta_0}[g_j] h'_{\beta_j} I_{\{u \leq T\}}. \end{aligned}$$

La demostración de este resultado está basado en el teorema dado por Van der Vaart y Wellner (1996); aquí sólo es necesario que las condiciones de éste sean satisfechas por los estimadores. Los siguientes lemas demuestran que cada una de las condiciones se cumplen.

Lema 2.1. *Para cualquier r finito existe un operador lineal continuo $\dot{S}_{\theta_0}(\theta_0) : \text{lin}\Theta \rightarrow \ell^\infty(H_r)$ tal que*

$$\left\| S_{\theta_0}(\theta) - S_{\theta_0}(\theta_0) - \dot{S}_{\theta_0}(\theta_0)(\theta - \theta_0) \right\|_r = o_r(\|\theta - \theta_0\|_r)$$

cuando $\|\theta - \theta_0\|_r \rightarrow 0$. Además, la derivada empírica $\dot{S}_{\theta_0}(\theta_0)$ puede ser expresada como:

$$\dot{S}_{\theta_0}(\theta_0)(\theta)(h) = \sum_{j=1}^J \left[- \int_0^\tau \sigma_{\Lambda_j, \theta_0}(h)(u) d\Lambda_j(u) - \boldsymbol{\beta}'_j \sigma_{\boldsymbol{\beta}_j, \theta_0}(h) - \boldsymbol{\delta}'_j \sigma_{\boldsymbol{\delta}_j, \theta_0}(h) \right].$$

La demostración del Lema 2.1 se basa en la caracterización de la diferenciabilidad de Fréchet y en el desarrollo de $S_{\theta_0}(\theta_0 + \epsilon\theta)$ en serie de Taylor de primer orden alrededor $\boldsymbol{\beta}'_0 x$.

Lema 2.2. *Para cualquier r finito $\sqrt{n} \left(S_{n, \hat{\theta}_0}(\theta_0) - S_{\theta_0}(\theta_0) \right)$ converge en ley a un proceso gaussiano centrado G sobre $\ell^\infty(H_r)$ con covarianza*

$$\text{Cov}[G(h), G(h^*)] = \sum_{j=1}^J \left\{ \int_0^\tau h_{\Lambda_j}(u) \sigma_{\Lambda_j, \theta_0}(h_{\Lambda_j}^*)(d\Lambda_j(u)) + \sigma_{\boldsymbol{\delta}_j, \theta_0}(h_{\boldsymbol{\delta}_j}^*) h'_{\boldsymbol{\delta}_j} + \sigma_{\boldsymbol{\beta}_j, \theta_0}(h_{\boldsymbol{\beta}_j}^*) h'_{\boldsymbol{\beta}_j} \right\}.$$

Para demostrar el Lema 2.2 es posible verificar que $\sqrt{n} \left(S_{n, \hat{\theta}_0}(\theta_0) - S_{\theta_0}(\theta_0) \right)$ se puede expresar como:

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \sum_{j=1}^J \left\{ h'_{\boldsymbol{\beta}_j} S_{\hat{\theta}_n \boldsymbol{\beta}_j}^{(i)}(\theta_0) + h'_{\boldsymbol{\delta}_j} S_{\hat{\theta}_n \boldsymbol{\delta}_j}^{(i)}(\theta_0) + c_{ij} h_{\Lambda_j}(\tau) + \int_0^\tau h_{\Lambda(s)} d\Lambda_j(s) e^{\boldsymbol{\beta}'_j \mathbf{x}_i} E_{\hat{\theta}_n}[g_{ij}] \right\},$$

la cual es una clase Donsker y por lo tanto $\sqrt{n} S_{n, \hat{\theta}_n}(\hat{\theta}_n)$ converge en distribución a un proceso gaussiano G con media cero en $\ell^\infty(H_r)$. Además,

$$\text{Cov}(G(h), G(h^*)) = -E_{\theta_0} \left[\frac{\partial^2}{\partial s \partial t} L_{\theta_0}(\theta_{0,s,t}) \Big|_{s=t=0} \right] = -\frac{\partial}{\partial s} S_{\theta_0}(\theta_{0,s})(h) = -\dot{S}_{\theta_0}(\theta_0)(h^*)(h),$$

donde $\theta_{0,s} = (\boldsymbol{\beta}_{0,s}, \boldsymbol{\gamma}_{0,s}, \Lambda_{0,s})$ con $\boldsymbol{\beta}_{0,s} = \boldsymbol{\beta}_0 + sh_{\boldsymbol{\beta}}^*$, $\boldsymbol{\gamma}_{0,s} = \boldsymbol{\gamma}_0 + sh_{\boldsymbol{\gamma}}^*$ y $\Lambda_{0,s}(\cdot) = \int_0^\cdot (1 + sh_{\Lambda}^*(u)) d\Lambda_0(u)$

Lema 2.3. *Para cualquier r finito $\dot{S}_{\theta_0}(\theta_0)$ es continuamente invertible sobre su rango.*

La demostración de que $\dot{S}_{\theta_0}(\theta_0)$ sea continuamente invertible equivale a demostrar que existe algún $l > 0$ tal que

$$\inf_{\theta \in \Theta} \frac{\|\dot{S}_{\theta_0}(\theta_0)(\theta)\|_r}{\|\theta\|_r} > l.$$

Como σ_{θ_0} es un operador continuamente invertible de H_∞ en H_∞ , y como σ_{θ_0} es un operador inyectivo el cual puede ser expresado como suma de un operador continuamente invertible y un operador compacto, la continuidad invertible queda demostrada.

Referencias

- Breslow, N.E.(1974). Covariance analysis of censored survival data, *Biometrics*, **30**, 89-100.
- Cox, D.R and Snell, E.J. (1989). *Analysis of Binary Data*, 2nd ed., Chapman and Hall London.
- Dempster, A.; Laird, N.M., and Rubin, D. B. (1997). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society*, **39**, 1-38.
- Murphy, S.A. (1995). Asymptotic theory for the frailty model. *The Annals of Statistics*, **23**, 182-198.
- Van der Vaart, A.W. and Wellner, J.A. (1996) *Weak convergence and empirical processes*. Springer, New York.

Modelado atmosférico para determinar niveles máximos diarios de ozono en la ciudad de Guadalajara

Lorelie Hernández Gallardo^a, Gabriel Escarela^b

Universidad Autónoma Metropolitana – Iztapalapa

1. Introducción

En términos de la Organización Mundial de la Salud, niveles de ozono superiores a $100 \mu\text{g}/\text{m}^3$ (microgramos de ozono por metro cúbico de aire) pueden ser riesgosos a la salud. Para poder predecir una concentración peligrosa de contaminantes y así proteger a la población vulnerable con cierta anticipación, es necesario crear una metodología que pueda indicar cuan probable es que ocurra dicha concentración. Los mecanismos químicos que controlan la formación del ozono troposférico son complejos y las volátiles condiciones meteorológicas contribuyen a la dificultad de predecir períodos de ozono alto con exactitud. Es bien sabido que la variación de los niveles de contaminantes corresponde a varias razones, entre las más importantes se encuentran los cambios anuales de condiciones meteorológicas y el incremento de diversas fuentes contaminantes; también se sabe que temperaturas altas junto con bajas velocidades de viento están asociadas con niveles altos de ozono. El propósito del presente estudio es proponer una técnica estadística que pueda usar tanto información no estacionaria como atmosférica para la predicción de los máximos diarios de ozono; en particular, se extienden ideas bien establecidas en la literatura de la teoría del valor extremo y se ilustra con el ajuste de datos de la ciudad de Guadalajara.

^aheilerol@yahoo.com.mx

^bge@xanum.uam.mx

2. Definición del modelo

Este estudio considera el problema de extender series de tiempo gaussianas autorregresivas a un marco de respuestas de valor extremo. Específicamente, se adopta la metodología propuesta por Zeger y Qaqish (1988), la cual consiste en especificar el modelo autorregresivo en forma de distribución condicional cuya parametrización pertenece a la familia exponencial de distribuciones; al igual que con los modelos lineales generales, a ésta distribución se le incluyen términos autorregresivos en forma de variables explicativas pasadas y presentes. De manera análoga al modelo de Zeger y Qaqish (1988), se propone definir un modelo autorregresivo de orden p para respuestas de valor extremo de tal manera que la distribución condicional pertenece a la familia de distribuciones de valor extremo cuyo parámetro de localización está ligado a una componente lineal que se forma de variables explicativas que contienen a la historia presente y pasada de los últimos p períodos y de un vector de coeficientes de regresión; esto es, la distribución condicional de cada respuesta Y_t dado el conjunto de información presente y pasada $\mathbf{H}_t = \{\mathbf{x}_t, \dots, \mathbf{x}_{t-p}, y_{t-1}, \dots, y_{t-p}\}$, donde \mathbf{x}_t es el vector de variables explicativas en el tiempo t y y_t es la respuesta observada en el tiempo t , está dada por la siguiente distribución de Valor Extremo Generalizado:

$$F(y_t | \mathbf{H}_t) = \exp \left[- \left\{ 1 + \gamma \left(\frac{y_t - \mu_t}{\sigma} \right) \right\}_+^{-1/\gamma} \right], \quad \text{para } y_t > \mu_t, \quad (1)$$

donde μ_t , σ y γ son respectivamente los parámetros de localización, escala y forma, con $-\infty < \mu_t < \infty$, $\sigma > 0$, $-\infty < \gamma < \infty$ y $h_+ = \max(h, 0)$; aquí, μ_t está relacionado con la historia presente y pasada a través de una componente lineal de manera tal que $\mu_t = \boldsymbol{\beta}^T \mathbf{z}_t$, donde \mathbf{z}_t es un vector de variables explicativas seleccionadas de \mathbf{H}_t , que incluye a la ordenada, y $\boldsymbol{\beta}$ es el vector de coeficientes correspondiente.

La especificación del modelo en la ecuación (1) es una generalización del modelo para series de máximos no estacionarias propuesto por Smith (1989), el cual sólo considera a t en \mathbf{H}_t ; en ésta formulación se propone no sólo modelar la no estacionalidad sino además incluir respuestas y variables explicativas presentes y pasadas que expliquen la dependencia entre las respuestas. Entre los beneficios de la especificación propuesta aquí, se puede mencionar que la función de verosimilitud tiene forma explícita, se pueden comparar modelos y se pueden llevar a cabo los diagnósticos correspondientes a través de los residuales estandarizados propuestos

por Dunn y Smyth (1996).

Aunque el modelo propuesto puede estar mal especificado debido a que no hay una distribución multivariada de valor extremo que tenga una distribución condicional de valor extremo, contrario a las series de tiempo gaussianas donde la condicional es gaussiana también, Dupuis y Tawn (2001) encontraron que para correlaciones de orden 1 las distribuciones ajustadas del modelo condicional correcto y del mal especificado eran casi idénticas para dependencias relativamente altas.

La función de verosimilitud de un modelo de transición de orden p para $\{y_{m+1}, \dots, y_n\}$ condicional a las primeras m respuestas puede expresarse como $L(\theta) = \prod_{k=m+1}^n f(y_k | \mathbf{H}_k)$, donde θ representa al vector de parámetros y f denota la función de densidad correspondiente a F . En el presente estudio se usó la biblioteca `evd` del paquete estadístico **R**, cuya función `fgev` proporciona el estimador de máxima verosimilitud para la distribución de la ecuación (1); el proceso optimizador que usa esta función se basa en el método quasi-Newton (también conocido como el algoritmo de métrica variable), el cual usa evaluaciones de la función objetivo y sus gradientes para generar una *fotografía* de la superficie por optimizar y así buscar el punto estacionario donde el gradiente es 0.

Como en este tipo de datos generalmente se cuenta con un tamaño de muestra grande, las pruebas de cociente de verosimilitud para encontrar un modelo parsimonioso no son muy confiables pues tienden a presentar significancias importantes a variables cuya contribución a la explicación del fenómeno es muy modesta o nula (ver Raftery, 1995), en este estudio se propone usar el **BIC** (*Bayesian Information Criterion*), pues determina con mayor grado de precisión el modelo más parsimonioso penalizando tanto al número de parámetros como al tamaño de la muestra. Si n_p es el número de parámetros en el modelo, entonces, el criterio del BIC consiste en escoger el modelo para el cual $2 \ln L(\hat{\theta}) + n_p \ln n$ tiene el valor más pequeño.

3. Los máximos de ozono en Guadalajara

En este estudio se analizan los niveles máximos diarios de ozono y_t medidos en partes por millon (ppm) registrados por siete estaciones de monitoreo en el área metropolitana de Guadalajara del 6 de enero de 1997 al 31 de diciembre de 2006. Debido a que la concentración de ozono es mayor a media tarde se tomó el valor máximo entre las 12 y 17 horas de toda

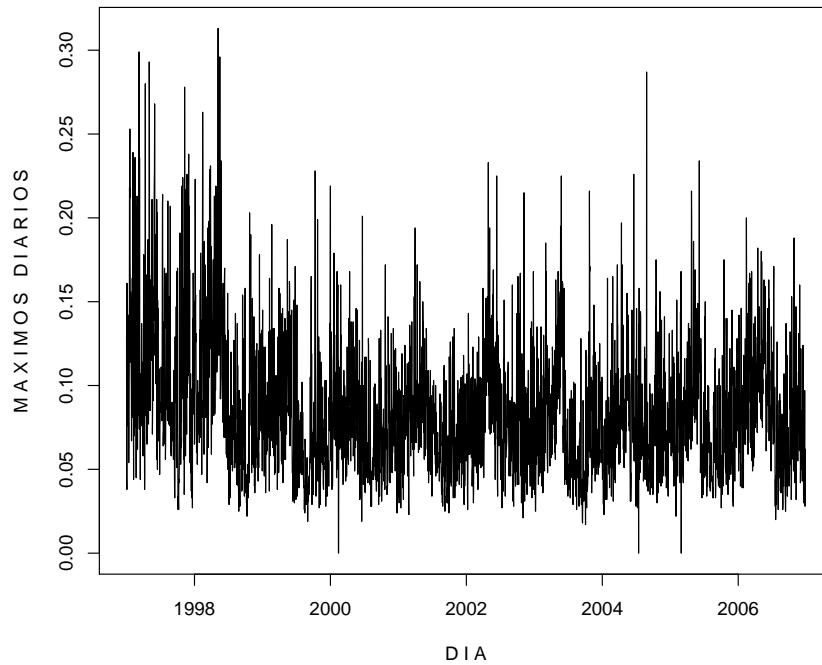


Figura 1: Máximos diarios de ozono en el área metropolitana de Guadalajara del 6 de enero de 1997 al 31 de diciembre de 2006

la red de monitoreo de cada día; de manera análoga, las variables atmosféricas empleadas se restringieron a éste horario.

La gráfica de la serie de tiempo correspondiente se muestra en la Figura 1; en ésta es posible notar un comportamiento periódico y cierta variabilidad a lo largo del tiempo, lo que sugiere un grado importante de no estacionalidad.

Las variables atmosféricas consideradas son: promedio del mínimo de velocidad de viento (vv), promedio del máximo de temperatura (tem), promedio del mínimo de humedad (h), rango de velocidad de viento (rvv), rango de temperatura ($r\text{tem}$) y rango de humedad (rh); para poder incluir a la dirección del viento (dv) se consideran los siguientes vectores de viento:

$$wu = -vv \times \sin(2\pi \times dv/360) \quad y \quad wv = -vv \times \cos(2\pi \times dv/360)$$

La variable wu es la componente este-oeste del viento, la cual es positiva cuando el viento viene del oeste; de igual forma, wv es la componente norte-sur, que es positiva cuando el viento viene del sur. Para ajustar efectos no lineales del tiempo se procedió a usar bases

de polinomios ortogonales de t ; mientras que para incluir efectos **semestrales** se incluyen los términos $\cos(2\pi t/182.5)$ y $\sin(2\pi t/182.5)$. También se incluyeron efectos anuales, sin embargo, éstos no resultaron significativos.

Al emplear la distribución de Valor Extremo Generalizado se pudo comprobar que el parámetro de forma γ no tiene efectos significativos en presencia de las variables explicativas; cuando se usó la distribución Gumbel, la cual se obtiene cuando $\gamma \rightarrow 0$ en la ecuación (1) y la cual se especifica con:

$$F(y_t | \mathbf{H}_t) = \exp \left[-\exp \left\{ -\left(\frac{y_t - \mu_t}{\sigma} \right) \right\} \right], \quad \text{para } -\infty < y_t < \infty,$$

se obtuvieron prácticamente las mismas inferencias; de esta forma, se optó por usar la distribución Gumbel para el análisis. Usando selección progresiva (*forward selection* en inglés) y el criterio BIC ya mencionado, se escogió un modelo autorregresivo de orden 6 cuya formulación es:

$$\begin{aligned} t^7 + y_{t-1} + \text{semestrales} + \text{tem}_t + \text{rtem}_t + v_t + \text{rvv}_t + wu_t + wv_t + \text{tem}_{t-1} + \\ h_{t-1} + v_{t-1} + \text{rtem}_{t-2} + \text{rvv}_{t-4} + rh_{t-5} + \text{rtem}_{t-6} + v_{t-6} + \text{rvv}_{t-6} + wu_{t-7} \end{aligned}$$

aquí, el superíndice denota el orden del polinomio.

La Figura 2 muestra el ajuste del polinomio de grado 7 del tiempo t en el mejor modelo con bandas de confianza de 95 %. Es notoria la baja en la severidad de los máximos de ozono los primeros mil días; sin embargo, el comportamiento es generalmente irregular, lo que corrobora la alta no estacionalidad de la serie de tiempo.

La Tabla 1 muestra el valor de los coeficientes lineales en el modelo de localización y el estimador del parámetro de escala σ para el modelo elegido. Es posible notar que los coeficientes correspondientes al promedio de máxima temperatura son muy significativos; estos indican que un día caluroso incrementa la severidad de los máximos de ozono; sin embargo, si el día que precede también es caluroso, los máximos pueden ser aminorados. Un efecto inverso ocurre con la velocidad de viento. Como es bien sabido, un día con viento dispersa los contaminantes; además, si el día anterior tuvo viento, es posible que sea susceptible observar un incremento en el máximo. Un incremento de humedad mínima en el día anterior también contribuye a disminuir la severidad del máximo de ozono. Como es de esperarse, los vectores de velocidad también juegan un papel preponderante tanto en la dispersión de los contaminantes como en la concentración de ellos.

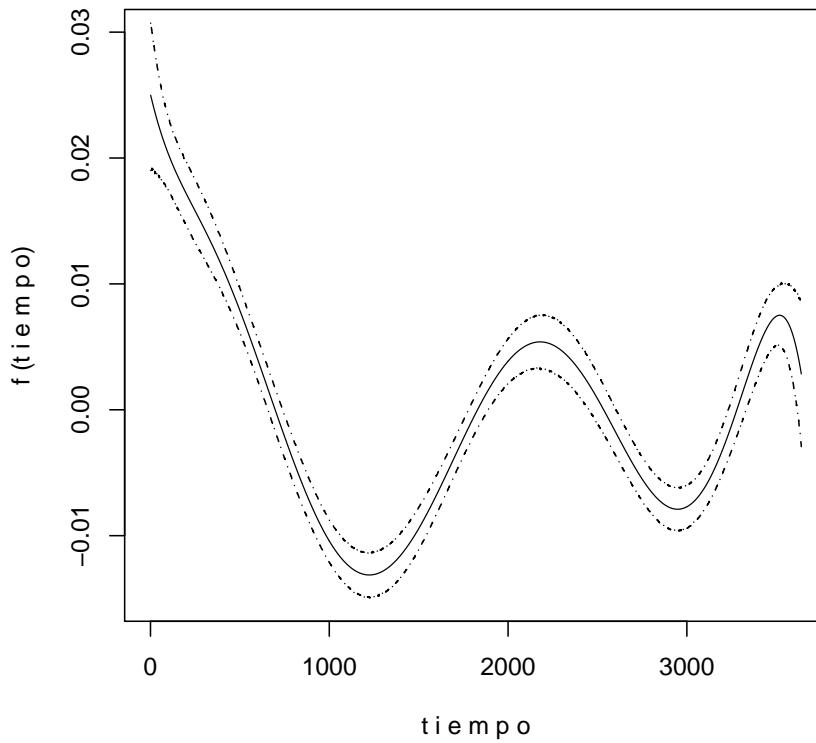


Figura 2: Efecto ajustado de tiempo en presencia de variables atmosféricas

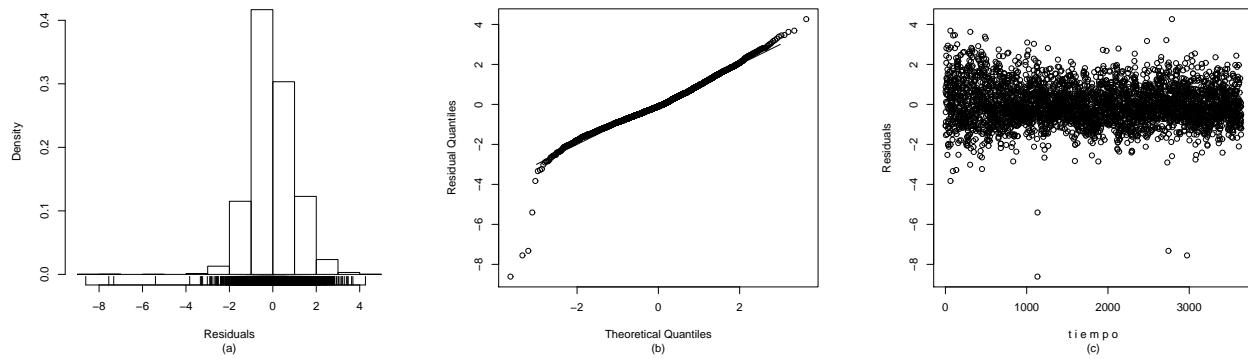


Figura 3: Gráficas de diagnósticos de los máximos diarios de ozono

La Figura 3 muestra algunas gráficas del análisis de residuales, que son: (a) densidad estimada de los residuales, (b) cuantil contra cuantil y (c) residuales contra tiempo. A pesar de que hay cinco observaciones atípicas, las cuales se ubican en la cola izquierda, las gráficas muestran que los residuales estandarizados se distribuyen aproximadamente normal, como

Parámetro	Estimador	Error Estándar
Ordenada	0.0727	0.0004
y_{t-1}	0.3859	0.0264
$\cos(2\pi t/182.5)$	0.0021	0.0006
$\sin(2\pi t/182.5)$	-0.0025	0.0006
tem_t	0.7206	0.0508
$rtem_t$	0.1567	0.0292
v_t	-0.7471	0.0354
rvv_t	-0.1820	0.0352
wu_t	0.2522	0.0331
wv_t	-0.0819	0.0230
tem_{t-1}	-0.3332	0.0502
h_{t-1}	-0.2774	0.0334
v_{t-1}	0.1981	0.0300
$rtem_{t-2}$	0.0802	0.0249
rvv_{t-4}	0.0644	0.0296
rh_{t-5}	-0.0504	0.0232
$rtem_{t-6}$	0.0941	0.0308
v_{t-6}	0.0686	0.0257
rvv_{t-6}	0.1114	0.0320
wu_{t-7}	-0.0925	0.0291
σ	0.0216	0.0003

Tabla 1: Coeficientes estimados y errores estándar del mejor modelo

era de esperarse; por tanto, el ajuste del modelo más parsimonioso es relativamente bueno.

Referencias

- Dunn, P.K. y Smyth, G.K. (1996). Randomized Quantile Residuals, *Journal of Computational and Graphical Statistics*, **5**, 236-244.
- Dupuis, D. J. y Tawn, J. A. (2001). Effects of Mis-Specification in Bivariate Extreme Value Problems, *Extremes*, **4**, 315-330.

Raftery A. E. (1995). Bayesian Model Selection in Social Research. *Sociological Methodology*, **25** 111-163.

Smith, R. L. (1989). Extreme Value Analysis of Environmental Time Series: An Application to Trend Detection in Ground-Level Ozone. *Statistical Science*, **4**, 367-393.

Zeger, S. L. y Qaqish, B. (1988). Markov Regression Models for Time Series: A Quasi-Likelihood Approach, *Biometrics*, **44**, 1019-1032.

Regresión por mínimos cuadrados parciales aplicada al estudio de emisiones de dióxido de carbono en suelos de Veracruz, México

Gladys Linares Fleites^a, José Adrián Saldaña Munive

*Departamento de Investigaciones en Ciencias Agrícolas. Instituto de Ciencias de la
Benemérita Universidad Autónoma de Puebla*

Luis G. Ruiz Suárez *Centro de Ciencias de la Atmósfera. Universidad Nacional
Autónoma de México*

1. Introducción

La Regresión Mínimo Cuadrática Parcial es una de las extensiones menos restrictivas del modelo de Regresión Lineal Múltiple. Su flexibilidad permite utilizarla en situaciones donde existen pocas observaciones y puede ser una herramienta de análisis exploratorio muy útil para seleccionar variables predictoras convenientes. Con el propósito de estudiar asociaciones entre las emisiones de dióxido de carbono (CO_2) y diversas propiedades de suelos perturbados en el Parque Ecológico Jaguarundi en Coatzacoalcos, Veracruz, México, se utiliza este modelo como herramienta exploratoria.

En la sección 2 se explican las principales características numéricas de la Regresión Mínimo Cuadrática Parcial (Partial Least Square, PLS, en inglés). En la sección 3 se muestra la aplicación de la regresión PLS al estudio de las emisiones de CO_2 en suelos de Veracruz y se comparan los resultados con la Regresión Mínimo Cuadrática Ordinaria. Finalmente se brindan conclusiones y recomendaciones.

^agladys.linares@icbuap.buap.mx

2. Regresión por mínimos cuadrados parciales

El PLS es un método para modelar relaciones entre conjuntos de variables observadas por medio de variables subyacentes o latentes (Rosipal y Krämer, (2006)). PLS puede extenderse de manera natural a problemas de regresión, aunque la regresión PLS es considerada todavía, por muchos estadísticos, como un algoritmo y no como un modelo estadístico riguroso.

La Regresión Por Mínimos Cuadrados Parciales que generaliza y combina hechos de la Regresión con Componentes Principales, puede utilizarse en regresiones donde se presenta el problema de multicolinealidad. (Abdi, H. (2003))

Para la estimación de los parámetros del modelo de regresión existen diferentes métodos (Montgomery *et al.* (2004)). Si se usa el método de los mínimos cuadrados ordinario(OLS, siglas en inglés) hay que resolver el sistema de las ecuaciones normales

$$X'X\hat{\beta}_j = X'Y_j, (1 \leq j \leq q)$$

Si en la matriz $X'X$, están presentes problemas de multicolinealidad y mal condicionamiento desde el punto de vista numérico, esto origina severos problemas de precisión en la estimación de los coeficientes β'_j s del modelo de regresión, y por ende, malas predicciones cuando la dimensión del sistema es grande. Recientemente han surgido otros métodos para la determinación del modelo de regresión que evitan los problemas antes mencionados, manejando de forma más efectiva, con respecto a la calidad de las predicciones, los problemas de multicolinealidad. Entre los métodos que logran la estabilización de los coeficientes β en la regresión, se encuentran el de mínimos cuadrados parciales (PLS).

La regresión por mínimos cuadrados parciales se basa en resolver los sistemas lineales sobre-determinados

$$X\hat{\beta}_j = Y_j, (1 \leq j \leq q),$$

evitando la formación de la matriz $X'X$, y de esta manera el mal condicionamiento de las ecuaciones normales. Aquí se usa una descomposición de X en factores T y P' que conservan el rango, los cuales están afectados por la interacción entre las matrices X y Y , que logra una mayor capacidad predictiva del modelo sin eliminar información. Se han desarrollado diferentes algoritmos para el PLS (Lohninger (1999)), pero uno de los más populares es el

llamado algoritmo NIPALS que lo ha implementado el paquete de programas MINITAB Release 15 (2005) y que utilizamos para procesar la información en el presente trabajo.

3. Aplicación al estudio de las emisiones de CO₂ en suelos de Veracruz

El dióxido de carbono (CO₂) es uno de los llamados gases de efecto invernadero. Estos gases son continuamente emitidos y removidos en la atmósfera por proceso naturales sobre la Tierra, pero las actividades antropogénicas causa cantidades adicionales de los mismos incrementando sus concentraciones en la atmósfera, lo que tiende a sobrecalentarla. El dióxido de carbono (CO₂), es el principal gas de efecto invernadero, responsable del calentamiento global. (Saldaña y Ruiz (2007).)

3.1. Caso de estudio

Con el propósito de estudiar asociaciones entre las emisiones de dióxido de carbono (CO₂) y diversas propiedades de suelos perturbados en el Parque Ecológico Jaguaroundi en Coatza-coalcos, Veracruz, México, se midieron las emisiones de CO₂ y las propiedades del suelo en determinados sitios de muestreo. Para medir las emisiones de CO₂ se utilizó el método de cámara estática. Se colocaron las cámaras en los sitios seleccionados, ubicando 3 cámaras por sitio, buscando homogeneidad en el terreno, considerando un mismo plano por altitud y cubierta vegetal. Las emisiones se analizaron con un Detector de Ionización de Flama. Las propiedades físicas y químicas de las muestras de suelo fueron analizadas por el Grupo de Edafología del Instituto de Geología de la UNAM.

3.2. Resultados

Los datos fueron analizados en dos épocas: la de lluvia y la de secas, obteniéndose los modelos de Regresión PLS para cada una de esas épocas. Previamente se obtuvieron modelos OLS, constatándose la existencia de multicolinealidad, por lo que no se consideraron recomendables.

3.2.1. Resultados: época de lluvia

La tabla 1 muestra los coeficientes de regresión del modelo de regresión con dos componentes. La prueba F ($F = 20.98$) del modelo resultó significativa con un valor de p aproximadamente igual a cero. El coeficiente determinación fue del 66.6 % y el estadístico PRESS fue 626890, más pequeño que el obtenido en la regresión OLS. En la regresión OLS se obtuvo un valor $F = 3.49$ que resultó no significativo al 5 % y un coeficiente de determinación de 88.8 %.

La primera componente destaca el hecho de que cuando se incrementa la materia orgánica disminuye la densidad aparente (D. Ap. g/cc). Esto se expresa con la oposición entre la materia orgánica, el carbono, el nitrógeno total y la conductividad eléctrica (CE) y la densidad aparente. La segunda componente expresa la oposición entre salinidad y respiración basal, esto es, indica la "potencialidad de la actividad orgánica". Los supuestos de normalidad y homogeneidad de varianza del modelo se corroboraron a través de gráficos de residuos y de normalidad.

3.2.2. Resultados: época de secas

La tabla 2 muestra los coeficientes de regresión del modelo de regresión con dos componentes para la época de secas. La prueba F ($F = 9.62$) del modelo resultó significativa con un valor de p igual a 0.001. El coeficiente determinación fue del 40.8 % y el estadístico PRESS fue 603000, más pequeño que el obtenido en la regresión OLS que alcanzó un valor de ocho cifras enteras. En esta última (regresión OLS) la prueba F ($F = 2.57$) no fue significativa al 5 % y el coeficiente de determinación fue de 87.9 %.

La primera componente, al igual que en el caso de la época de lluvia, destaca el hecho de que cuando se incrementa la materia orgánica disminuye la densidad aparente (D. Ap. g/cc). La segunda componente expresa la oposición entre salinidad y temperatura, lo que explica el fenómeno de que cuando aumenta la temperatura la actividad de los microorganismos se incrementa.

Los supuestos de normalidad y homogeneidad de varianza del modelo se corroboraron a través de gráficos de normalidad y de residuos.

	Coeficientes de regresión		
	FE ug CO2/m2h	FE ug CO2/m2h	estandarizado
Constante	-665.238	0.000000	
T. Cámara	10.016	0.135797	
T. Amb.	9.000	0.116188	
T. Suelo	5.764	0.087625	
Altitud msnm	-0.387	-0.025519	
Salinidad	-122.791	-0.140703	
N Total %	139.655	0.232672	
Carbono mg/g	1.537	0.225887	
Materia orgánica	0.892	0.226057	
P Mg/kg	-6.103	-0.025929	
pH	5.962	0.018252	
CE mS/cm	109.473	0.111550	
D. Ap. g/cc	-48.954	-0.106694	
CMRA	0.673	0.100786	
Respiración Basal	0.082	0.127452	
% arcilla	-0.643	-0.036606	
% arena	-0.854	-0.071823	

Tabla 1: Regresión PLS (con 2 componentes) para la época de lluvia

		Coeficientes de regresión FE ug CO2/m2h	FE ug CO2/m2h estandarizado
	Constante	175.719	0.000000
(1)	T. Cámara °C	-3.278	-0.115291
(2)	T. Amb. °C	-2.955	-0.047282
(3)	T. Suelo °C	-2.544	-0.072007
(4)	Humedad prom. %	0.904	0.143442
(5)	Altitud msnm	-3.057	-0.163050
(6)	Salinidad	108.244	0.066471
(7)	N Total %	85.585	0.115477
(8)	Carbono mg/g	0.263	0.031254
(9)	Materia orgánica	0.152	0.031296
(10)	P Mg/kg	13.083	0.045021
(11)	pH	31.035	0.054646
(12)	CE mS/cm	82.481	0.083419
(13)	D. Ap. g/cc	-20.653	-0.036455
(14)	CMRA	1.002	0.122284
(15)	Respiración Basal	-0.075	-0.087606
(16)	% arcilla	1.635	0.075420
(17)	% arena	1.133	0.077135

Tabla 2: Regresión PLS (con 2 componentes) para la época de seca

4. Conclusiones

Se obtuvieron dos modelos de regresión PLS, uno para la época de lluvia, con 24 muestras de suelo y 16 propiedades y, otro, para la época de secas, con igual número de muestras y 17 propiedades. Las muestras fueron tomadas en el Parque Ecológico Jaguaroundi en Coatzacoalcos, Veracruz, México.

Los modelos de regresión de las emisiones de CO₂, obtenidos por PLS mostraron mayor capacidad predictiva que los modelos estimados por la regresión OLS, para los periodos de lluvia y secas. La regresión PLS, tomada como herramienta exploratoria, permitió destacar las propiedades del suelo más importantes que explican las emisiones de CO₂ en la zona. En ambos modelos se destaca, entre otros aspectos, que el nitrógeno total del suelo tiene fuerte influencia sobre la variable respuesta (emisiones de CO₂).

Es necesario continuar profundizando en el estudio de predicciones de las emisiones de gases efecto invernadero (en particular el CO₂), por ser una de las causas del calentamiento global, fenómeno ambiental de importancia capital para la humanidad.

Referencias

- Abdi, H. (2003). Partial Least Square Regression. In M.Lewis-Bech, A. Bryman, T. Futing (Eds): Encyclopedia for research methods for the social sciences. Thousand Oaks (CA): Sage. Pp.729-795.
- Lohninger, H. (1999). *Teach/Me Data Analysis*. Libro Electrónico. Springer - Verlag. Berlin-New York-Tokyo ISBN 3-540-14743-8.
- MINITAB Release 15 (2005). Statistical Software. Minitab. Inc.
- Montgomery, D.C., Peck, E.A. and Vining G.G. (2004) *Introducción al Análisis de Regresión Lineal*. México: Compañía Editorial Continental.
- Rosipal R. and Krämer, N. (2006). Overview and Recent Advances in Partial Least Squares, *In SLSFS 2005 LNCS3940*, (eds. Saunders et al.) pp. 34-51 Springer-Verlag Berlin Heidelberg.
- Saldaña, M.J.A. y Ruiz Suárez, L. G. (2007). Emisiones de gases de efecto invernadero en suelos perturbados con diferente cobertura vegetal en Coatzacoalcos, Veracruz, México. (Artículo por publicar).

Discriminación lineal y discriminación logística en estudios de calidad de suelos

Gladys Linares Fleites^a, Miguel Ángel Valera Pérez

*Departamento de Investigaciones en Ciencias Agrícolas. Instituto de Ciencias de la
Benemérita Universidad Autónoma de Puebla*

Maribel Castillo Morales *Estudiante del Postgrado en Ciencias Ambientales, ICUAP*
Benemérita Universidad Autónoma de Puebla, México

1. Introducción

Actualmente, uno de los retos más importantes que enfrenta la ciencia del suelo es desarrollar criterios de Calidad del Suelo (CS) que puedan utilizarse en una evaluación objetiva de riesgos ambientales. Debido a la inquietud existente con respecto a la degradación del suelo y a la necesidad de un manejo sostenible de los agro-ecosistemas, ha resurgido el estudio de las propiedades del suelo enfatizándose hacia una función específica del uso del suelo (Carter *et al.* (1997)). Este enfoque ecológico del suelo reconoce las interacciones suelo - ser humano y de esta manera la CS es inseparable del concepto de sostenibilidad del sistema y de su uso. La CS se clasifica en Calidad Inherente (CI) y Calidad Dinámica (CD). La CI se conforma con las propiedades innatas del suelo, tales como textura, mineralogía, color, etc., mientras que la CD son las propiedades de la CI modificadas por las actividades antropogénicas. Es de gran importancia identificar las propiedades del suelo que establezcan la diferencia entre CI y CD y que permitan predecir la CS. Ahora bien, si desarrollar criterios de Calidad de Suelos que puedan utilizarse en una evaluación objetiva de riesgos ambientales es actualmente un reto, también lo es, buscar procedimientos de clasificación eficaces. A este último aspecto

^agladys.linares@icbuap.buap.mx

va dirigido el presente trabajo. Para desarrollar este objetivo se seleccionaron tres problemas de los propuestos por la NRSC(Natural Conservation Service) (2001), a saber:

Problema 1 Contenidos de materia orgánica y residuos en los suelos,

Problema 2 La reacción pH del suelo, y

Problema 3 La fertilidad natural del suelo.

Para cada uno de estos problemas se compararon, a través de ciertos criterios, dos enfoques de discriminación: el modelo de discriminación lineal y el modelo de regresión logístico, utilizando la información obtenida en una zona de Teziutlán, del estado de Puebla, México. A continuación se exponen brevemente los enfoques de discriminación comparados y posteriormente se muestran los resultados obtenidos en la comparación, utilizando el criterio tradicional de error de mala clasificación, y otros tres criterios obtenidos de las tablas de confusión elaboradas para cada problema.

2. Discriminación lineal y regresión logística

2.1. Discriminación lineal

Dado que existen diferentes enfoques en el problema de la discriminación, decidimos utilizar el análisis discriminante clásico, basado en la normalidad multivariada de las variables consideradas y que es óptimo bajo dicho supuesto (Peña (2002)). En este enfoque se construye una función de clasificación para cada una de las poblaciones consideradas (suelos con CI y suelos con CD) y se establece la regla de clasificación que coloca el individuo a clasificar en la población con valor máximo de la función de clasificación. Dado que la utilidad de la regla de clasificación depende de los errores esperados, estos se calcularon aplicando la regla discriminante a las 25 observaciones y clasificándolas. Este método tiende a subestimar las probabilidades de error de mala clasificación ya que los mismos datos se utilizan para estimar los parámetros y para evaluar la regla resultante. Un procedimiento mejor es clasificar cada elemento (muestra de suelo) con una regla que no se ha construido usándolo. Para ello, se construyeron $n=25$ funciones discriminantes con las muestras que resultan al eliminar uno a uno cada elemento de la población y clasificar después cada dato en la regla construida sin

él. Este método se conoce como validación cruzada y conduce a una mejor estimación del error de clasificación. En el estudio se utilizan ambos procedimientos de estimación de ese error.

2.2. Regresión logística

Una posibilidad para resolver problemas de clasificación es construir un modelo que explique los valores de clasificación. En nuestro caso, como deseamos discriminar entre suelos de CI y suelos de CD, utilizamos la variable y con los valores 1 (suelo con CI) y 0 (suelo con CD) y el problema se convierte en prever el valor de la variable y en un nuevo elemento del que conocemos el vector de variables X (propiedades de los suelos).

Para modelar este tipo de relaciones se utilizan los modelos de respuesta cualitativa, del que el modelo logístico es uno de los más utilizados ya que puede aplicarse a una amplia gama de situaciones donde las variables explicativas no tienen una distribución conjunta normal multivariada. (Linares (2007)). Las propiedades de los suelos que intervienen en la explicación de su calidad fueron exploradas previamente y tienen distribuciones muy asimétricas, por lo que la regresión logística es una buena opción para la modelación.

Obsérvese que estamos suponiendo que la variable respuesta y_i es una variable aleatoria con distribución Bernoulli con probabilidades $P(y_i = 1) = \pi_i$ y $P(y_i = 0) = 1 - \pi_i$, $0 \leq i \leq 1$

Para contrastar si una variable, o grupo de variables, de la ecuación es significativa, podemos construir un contraste de la razón de verosimilitudes comparando los máximos de la función de verosimilitud para los modelos con y sin estas variables. Sin embargo, es más habitual para comprobar si un parámetro es significativo comparar el parámetro estimado con su desviación estándar. A estos cocientes se les denomina *estadísticos de Wald* y en muestras grandes se distribuyen, si el verdadero valor del parámetro es cero, como una normal estándar.

3. Estudio comparativo

Los datos utilizados para la comparación de los enfoques de discriminación considerados, se tomaron de la zona denominada Caldera de Teziutlán, del estado de Puebla, México. (Valera et al. (2001)). Utilizaremos el problema 2 de la reacción del pH del suelo para ilustrar la metodología empleada. Las funciones discriminantes lineales son mostradas en la

Tabla 1. La proporción de clasificación correcta es sólo del 68 % con el primer procedimiento de estimación del error de mala clasificación y del 64 % con el procedimiento de validación cruzada. El modelo de regresión logística, mostrado en la Tabla 2, predijo correctamente el 72 % de los casos, luego la proporción del error de mala clasificación asciende la 28 %, lo que es un error ligeramente menor al del modelo de discriminación lineal.

	Constante	Constante
Constante	-27.105	-26.413
pH-H2O	9.991	7.936
pH-KCl	-0.053	2.199

Tabla 1: Análisis discriminante lineal de la reacción pH del suelo

Predictor	Coeficiente	SE Coeficiente	Z	P
Constante	1.45473	3.88273	0.37	0.708
pH-H2O	-2.52533	1.70054	-1.49	0.138
pH-KCl	2.65172	1.64087	1.62	0.106

Tabla 2: Regresión logística: la reacción pH del suelo

Los datos fueron procesados con MINITAB 15 (2005).

3.1. Resultados de la comparación por otros criterios

Además de la medida tradicional de error de mala clasificación dada anteriormente en cada problema, se calcularon otros criterios que surgen de las matrices de confusión o de contingencias de cada problema.(Hernández, 2004). Una matriz de confusión en estos problemas puede expresarse como:

Real		
Predicha	CI	CD
CI	TP	FP
CD	FN	TN

Los valores de TP(verdadero positivo), FN(falso negativo) y TN(verdadero negativo) son frecuencias. A partir de estos valores se definen algunos criterios, como los siguientes:

- Macro-Media: $(\text{Sensitividad} + \text{Especificidad})/2$
- Sensitividad = $P(\text{CI}_{pred}/\text{CI}_{real})$
- Especificidad = $P(\text{CD}_{pred}/\text{CD}_{real})$

El problema 1 clasificó correctamente el 100 % de los casos en ambos procedimientos de discriminación, luego el valor de los tres criterios anteriores es 1 y ambos procedimientos de discriminación se consideraron adecuados para este problema. La Tabla 3 muestra las matrices de confusión del problema 2 para ambos modelos. La parte izquierda corresponde al modelo de discriminación lineal y la derecha al modelo de regresión logística.

Predicha	CI	CD	Predicha	CI	CD
CI	9	4	CI	10	4
CD	4	8	CD	3	8

Tabla 3: Matrices de confusión para el problema 2: La reacción pH del suelo

En el modelo de discriminación lineal los criterios tomaron los valores siguientes: sensitividad 0.692, especificidad 0.666 y macro-media 0.679.

En el modelo de regresión logística los criterios fueron: sensitividad 0.769, especificidad: 0.666 y macro-media: 0.717.

Puede observarse que en dos de esos criterios, el modelo de regresión logística tuvo un comportamiento ligeramente mejor que el modelo de discriminación lineal. En el problema 3 de fertilidad natural del suelo, los criterios en el modelo de discriminación lineal tomaron los siguientes valores: sensitividad 0.769, especificidad 0.916 y macro-media 0.84.

El modelo de regresión logística predijo correctamente en el 100 % de los casos, luego obtuvo el valor 1 en sensibilidad, especificidad y macro- media y puede considerarse superior al discriminante lineal en este problema.

4. Conclusiones

En los problemas de calidad de suelo considerados, el comportamiento de la Regresión Logística, como técnica de clasificación, se comportó mejor que el Discriminante Lineal en dos de los problemas considerados: el de la reacción pH del suelo y el de la fertilidad natural del suelo. En el problema de los contenidos de materia orgánica y residuos en el suelo ambos procedimientos de clasificación fueron adecuados. Es necesario realizar otros estudios de calidad de suelos que permitan validar el resultado anterior.

Referencias

- Carter, M.R., Gregorich, E.G., Anderson, D.W., Doran, J.W., Janzen, H.H. y F.J. Pierce. (1997). Concepts of Soil Quality and their Significance , *In Soil Quality for Crop Production and Ecosystem Health. Developments in Soil Science.* , (eds. Gregorich, E.G. and Carter, M.R.) 25. Elsevier Sc.
- Hernández Orallo, J. (2004). Evaluación de Clasificadores en Minería de Datos. Universidad Politécnica de Valencia, España.
- Linares, G. (2007). *Análisis de Datos Multivariados.* México. : Editorial Benemérita Universidad Autónoma de Puebla. Facultad de Computación. 277p.
- MINITAB Release 15 (2005). Statistical Software. Minitab. Inc.
- Natural Resources Conservation Service, (NRSC) (2001)). *Guidelines for Soil Quality Assessment in Conservation Planning.* United States Department of Agriculture and Soil Quality Institute. USA.
- Peña, D. (2002). *Análisis de Datos Multivariantes.* Madrid, España. : Mc. Graw Hill/Interamericana de España, S.A.U. MADrid, España. 539p.
- Valera, M.A., Tenorio. M.G., Linares, G., Ruiz, J. y Tamariz, J.V. (2001). Aplicación de indicadores químicos de degradación para suelos ácidos de la Sierra Negra de Puebla., *En Memorias COLOQUIOS Cuba-México sobre manejo sostenible de los suelos.* , Benemérita Universidad Autónoma de Puebla. pp 57-64.

Análisis bivariado de extremos para evaluar los niveles de ozono troposférico en la zona metropolitana de Guadalajara

Tania Moreno Zúñiga^a, Gabriel Escarela^b

Universidad Autónoma Metropolitana – Iztapalapa

1. Introducción

La ciudad de Guadalajara ha experimentado una expansión en la industria y en el comercio desde 1934. Dicha expansión ha traído como consecuencia contaminación atmosférica cuyas concentraciones alcanzaron niveles riesgosos para la salud frecuentes a mediados de los 90's. Para revertir las tendencias de deterioro de la calidad del aire y así proteger la salud de la población que habita la zona metropolitana de Guadalajara, las autoridades ambientales responsables implementaron el ‘Programa para el mejoramiento de la calidad del aire en la zona metropolitana de Guadalajara 1997-2001’.

Hasta la fecha se carece de un estudio que evalúe los beneficios reales de la implementación de dicho programa el cual pueda diagnosticar la tendencia de los niveles de polución de la zona. El propósito del presente estudio es el de llenar este vacío al analizar los máximos locales semanales de ozono de las estaciones de monitoreo ambiental ubicadas en Vallarta y Tlaquepaque en el período 1997-2006 usando un modelo con la capacidad de evaluar los efectos de la tendencia en presencia de variables periódicas y atmosféricas. La justificación de usar los máximos locales de dos estaciones, en vez de encontrar el máximo global en la zona, se basa en la hipótesis de que los niveles de ozono y las tendencias pueden variar

^atania_8304@hotmail.com

^bge@xanum.uam.mx

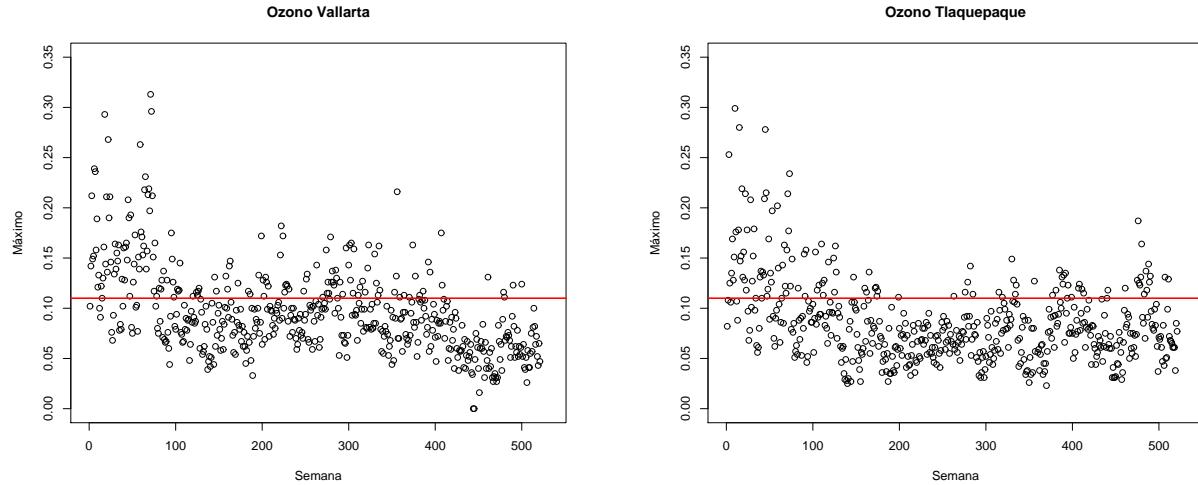


Figura 1: Máximos semanales de concentraciones de ozono del 1 de enero de 1997 al 31 de diciembre de 2006 en la estaciones de monitoreo ubicadas en Vallarta y Tlaquepaque

dependiendo de la localidad; un estudio conjunto de las dos localidades provee un análisis más informativo.

La Figura 1 muestra los máximos semanales locales para cada estación, los cuales están medidos en partes por billon (ppb). En las gráficas se sobrepone una línea horizontal que indica el límite de 110 ppb, el cual es el máximo diario permitido por la organización mundial de salud. Es posible observar en las gráficas que los máximos tienen una tendencia a la baja para las primeras 150 semanas y el comportamiento es más irregular para las semanas subsiguientes. Es entonces importante encontrar una forma funcional de la tendencia adecuada para cada localidad.

2. Modelado para valores extremos bivariados

Supóngase que (X_{1i}, X_{2i}) , para $i = 1, \dots, n$, son parejas aleatorias independientes e idénticamente distribuidas cuya función de distribución conjunta es G . Sean $Y_1 = \max(X_{11}, \dots, X_{1n})$ y $Y_2 = \max(X_{21}, \dots, X_{2n})$. La meta principal de este estudio es la de modelar la distribución conjunta de (Y_1, Y_2) denotada $F(y_1, y_2)$. De la teoría univariada del valor extremo, se sabe que Y_j tiene asintóticamente la distribución de valor extremo generalizada, para $j = 1, 2$. Bajo ciertas condiciones de regularidad (e.g Resnick, 1987), la distribución conjunta de los

máximos converge a una clase multivariada de distribuciones de valor extremo.

Una forma válida y conveniente de construir $F(y_1, y_2)$ es a través del modelo de cópula (e.g. Dupuis, 2005) de manera tal que:

$$F(y_1, y_2) \equiv C[F_1(y_1), F_2(y_2)],$$

donde C es la función cópula, la cual es una distribución bivariada con dominio en el cuadro unitario, y F_j es la función de distribución marginal de Y_j , $j = 1, 2$. En este estudio se empleará la cópula *positiva estable* la cual está dada por:

$$C_\theta(v_1, v_2) = \exp \left\{ - [(-\log v_1)^{1/\theta} + (-\log v_2)^{1/\theta}]^\theta \right\}, \quad \theta \in (0, 1).$$

Esta cópula es útil para modelar dependencias positivas; cuando $\theta \rightarrow 0$ se obtiene la cópula superior de Fréchet, mientras que valores de θ cercanos a 1 proveen estructuras de dependencia cercanas a la independencia, i.e. $\lim_{\theta \rightarrow 1} C_\theta(u_1, u_2) = u_1 u_2$.

Para medir la concordancia de Y_1 y Y_2 es posible usar la τ de Kendall cuyo valor es $\tau_\theta = 1 - \theta$ cuando se usa la cópula positiva estable. Dicha cópula exhibe dependencia en la cola superior, por lo que una forma de cuantificar a la dependencia entre eventos extremos es a través del *coeficiente de dependencia de la cola superior* dado por:

$$\lambda_u = \lim_{u \rightarrow 1^-} \Pr\{Y_2 > F_2^{-1}(u) \mid Y_1 > F_1^{-1}(u)\} = 2 - 2^\theta, \quad \theta \in (0, 1).$$

Las marginales se toman de la siguiente familia generalizada de distribuciones de valor extremo:

$$F_j(z_j) = \exp \left[- \left\{ 1 + \frac{\gamma_j(z_j - \mu_j)}{\sigma_j} \right\}_+^{-1/\gamma_j} \right], \quad \text{para } z_j > \mu_j, \quad (1)$$

donde μ_j , σ_j y γ_j son los parámetros de locación, escala y forma respectivamente, $j = 1, 2$; aquí, $-\infty < \mu_j < \infty$, $\sigma_j > 0$, $-\infty < \gamma_j < \infty$ y $h_+ = \max(h, 0)$. La clase de distribuciones dada por la ecuación (1) contiene varias distribuciones importantes útiles para ajustar máximos tales como la Gumbel, la cual se obtiene cuando $\gamma \rightarrow 0$, la Fréchet, la cual se obtiene cuando $\gamma > 0$, y la Weibull, la cual se obtiene cuando $\gamma < 0$.

Para tomar en cuenta a la tendencia y otras variables explicativas en ambas marginales, uno puede especificar al parámetro de locación de cada marginal de la siguiente forma:

$$\mu_j = \boldsymbol{\beta}_j^T \mathbf{x}_{jt}, \quad j = 1, 2, \quad (2)$$

donde \mathbf{x}_{jt} es un vector de variables explicativas del componente j , el cual incluye a la ordenada y es observado en el tiempo t , y $\boldsymbol{\beta}_j$ es el vector de coeficientes de regresión correspondientes.

La función de verosimilitud es $L = \prod_{i=1}^n f(y_{1i}, y_{2i})$, donde f es la función de densidad correspondiente a F . Para la aplicación en la siguiente sección se utilizó el paquete **evd** del lenguaje R para minimizar $-2 \times \log L$, la *deviancia*. Una característica importante del modelado bivariado es que el proceso para encontrar un modelo parsimonioso puede seguir ideas del cociente de verosimilitud, análogo a los modelos lineales generalizados.

3. Análisis de los datos de Guadalajara

La respuesta de interés es la pareja cuyas entradas son los máximos de ozono registrados por las dos estaciones de monitoreo en cada semana. La justificación de tomar máximos semanales es que éstos son aproximadamente independientes. Para capturar la dependencia remanente, la cual se debe a la no estacionalidad y a las variables atmosféricas, se incluyen combinaciones de las siguientes variables en el componente lineal de los modelos de locación especificados en la ecuación (2): **tiempo**, el número de semana; **maxTemp**, la temperatura máxima; **rangoTemp**, el rango de la temperatura; **minHum**, la humedad mínima; **rangoHum**, el rango de la humedad; **minVel**, el mínimo de velocidad; **rangoVel**, el rango de velocidad; **anual**, periodicidad anual; **semestral**, periodicidad semestral; **vientouLunes** y **vientovLunes** son vectores de viento registrados en lunes; y **vientouSabado** y **vientovSabado** son vectores de viento registrados en sábado.

Debido a que la suposición lineal de la no estacionalidad es cuestionable, se incluyó a la variable **tiempo** en términos de bases ortogonales de una regresión polinomial; la misma idea fue implementada para las variables atmosféricas. Todas las variables atmosféricas y el **tiempo** fueron incluidas en la forma de un polinomio de grado ocho y entonces se procedió a usar el algoritmo de eliminación recursiva (*backward elimination*, en inglés) para encontrar el modelo más parsimonioso. Dicho algoritmo se basó en el criterio BIC, el cual consiste en escoger el modelo para el cual $-2 \log L + n_p \log n$ es el más chico; aquí n_p representa el número de parámetros en el modelo.

Las fórmulas obtenidas en el mejor modelo para cada marginal quedaron como:

Marginal 1 (Vallarta) $\text{tiempo}^5 + \text{maxTemp}^2 + \text{rangoTemp} + \text{rangoHum}^2 + \text{anual} + \text{semestral}$.

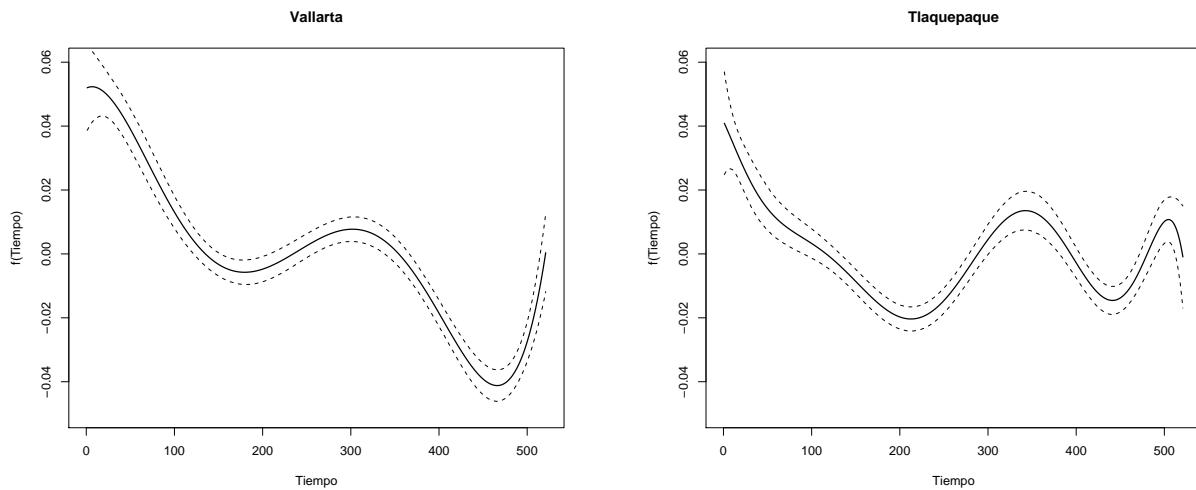


Figura 2: Ajuste de los polinomios de tiempo en el componente lineal de las marginales correspondientes a Vallarta y Tlaquepaque

Marginal 2 (Tlaquepaque) $\text{tiempo}^8 + \maxTemp^5 + \minVel + \text{rangoVel}^3 + \minHum^5 + \text{rangoHum}^5 + \text{vientouLunes}^4 + \text{vientouSabado}^4 + \text{anual} + \text{semestral}$.

La Figura 2 muestra la función de **tiempo** ajustada con las bases ortogonales de los polinomios obtenidos en el mejor modelo. Las líneas punteadas corresponden a bandas de confianza de 95 % calculadas con la aproximación $\hat{\beta}_j \sim \text{NMV}(\beta_j, \hat{\mathbf{V}}(\beta_j))$, donde NMV denota la función de distribución normal multivariada, $\hat{\beta}_j$ es el estimador de máxima verosimilitud de β_j y $\hat{\mathbf{V}}(\beta_j)$ es la matriz de covarianzas estimada correspondiente. Aunque las curvas difieren significativamente, es posible observar que en las primeras 200 semanas hay una tendencia a la baja, y que este comportamiento se repite entre las semanas 350 y 450.

La Figura 3 muestra los polinomios ajustados de las variables atmosféricas en el mejor modelo y la Tabla 1 muestra los estimadores puntuales de los coeficientes lineales y los parámetros restantes. Nótese que los límites del eje Y son los mismos para todas las gráficas y que se incluye un localizador de los datos observados. En ambas estaciones las temperaturas altas tienden a incrementar la posibilidad de que se incremente la severidad de los niveles de ozono, un hecho bien conocido en la ciencia atmosférica. Aunque en grado diferente, el incremento del rango de humedad parece favorecer la creación de ozono en ambas estaciones.

Mientras que el incremento de humedad mínima tiende a disminuir la magnitud de los máximos de ozono en Tlaquepaque, un efecto parecido se encuentra con la velocidad mínima

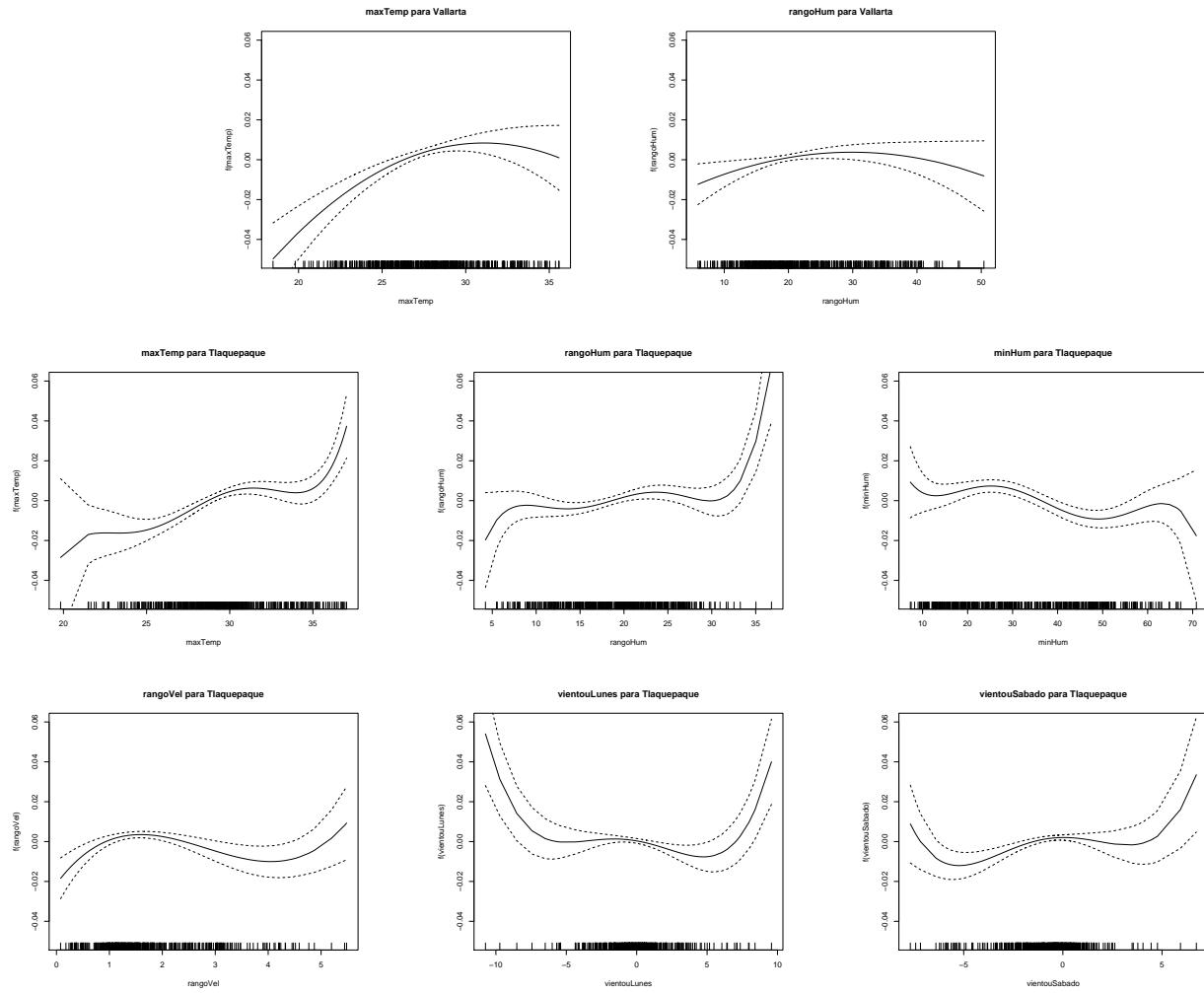


Figura 3: Polinomios con bandas de confianza de 95 % para las variables atmosféricas en el mejor modelo

Vallarta			Tlaquepaque		
Parámetro	Estimador	Error Estándar	Parámetro	Estimador	Error Estándar
Ordenada	0.0739	0.0010	Ordenada	0.0833	0.0011
MinVel	-0.1219	0.0356	RangoTemp	0.1100	0.0411
$\sin\left(\frac{\pi \times \text{tiempo}}{26}\right)$	0.0079	0.0023	$\cos\left(\frac{\pi \times \text{tiempo}}{26}\right)$	0.0118	0.0024
$\sin\left(\frac{\pi \times \text{tiempo}}{13}\right)$	-0.0060	0.0016	$\sin\left(\frac{\pi \times \text{tiempo}}{26}\right)$	0.0089	0.0020
σ_1	0.2455	0.0008	$\sin\left(\frac{\pi \times \text{tiempo}}{13}\right)$	-0.0075	0.0018
			σ_2	0.0194	0.0007
			γ_2	0.1201	0.0333

Tabla 1: Estimadores y errores estándares de los coeficientes lineales y de los parámetros de escala y forma en el mejor modelo

del viento en Vallarta. El efecto del viento en Tlaquepaque se puede valorar con las gráficas correspondientes a la dirección del viento; a diferencia de Vallarta, esta localidad tiene sus máximos de ozono influenciados por la velocidad y la dirección del viento simultáneamente. El efecto del rango de temperatura emula a la turbulencia vertical, en el caso de Talquepaque el incremento de esta variable tiende a incrementar los máximos de ozono.

La inclusión de los efectos de periodicidad resultaron ser significativos y -de hecho- absorbieron información a la que los polinomios de `tiempo` hubieran sido sensativos; esto es, en ausencia de la periodicidad se obtendría un polinomio de mayor orden en el mejor modelo. También se encontró que el parámetro de forma fue estadísticamente significativo para Tlaquepaque pero no para Vallarta; adicionalmente, el parámetro de dependencia estimado $\hat{\theta}$ es 0.723 (con error estándar 0.029), lo cual indica una moderada concordancia entre los máximos de ambas estaciones de monitoreo.

El presente estudio refuerza la conclusión de que la implementación de un programa global para reducir los niveles de contaminación en el área metropolitana de Guadalajara no se ve reflejada en una clara y estable mejoría de la calidad del aire a largo plazo.

Referencias

- Dupuis, D.J. (2005). Ozone concentrations: A robust analysis of multivariate extremes. *Tech-nometrics*, **47**, 191-201.
- Resnick, S.I. (1987). *Extreme Values, Point Processes and Regular Variation*, New York: Springer-Verlag.

Contraste de una hipótesis nula central compuesta frente una hipótesis alternativa bilateral en la distribución normal

Leonardo Olmedo^a *Universidad Autónoma Metropolitana – Iztapalapa*

1. Introducción

Una de las pruebas de hipótesis que con mayor frecuencia se presenta en situaciones de investigación reales, dada la naturaleza de los datos, y comúnmente enseñadas por su facilidad en la práctica académica es: $H_0 : \mu = K$ contra $H_a : \mu \neq K$. Sin embargo, en esta prueba existe un detalle que dificulta a un investigador aplicado, allí donde la hipótesis nula de igualdad se rechaza según el procedimiento seguido pero la significación estadística no refleja una diferencia con la igualdad que sea de interés al investigador.

La causa de la aparente contradicción entre la conclusión estadística y la recomendación del investigador no está en el procedimiento de prueba, ni en los valores que usualmente se establecen para el nivel de significación. Pudiera pensarse que se encuentra en el tamaño de la muestra, si es que éste es grande, ya que a medida que el tamaño de muestra crece también lo hace la potencia de la prueba, dando rechazos para diferencias de medias cada vez menores, pero esto también conduce a una aparente contradicción, ya que indica que conviene usar muestras pequeñas para que el resultado estadístico de lugar a una recomendación basada en él. Los estudios de tamaño de muestra producen un valor que se refiere al menor tamaño de muestra que permite rechazar cuando ocurre una diferencia mayor que un valor previamente establecido como mínimo para recomendar por las modalidades dadas por la hipótesis alternativa.

El problema radica en el establecimiento de las hipótesis nula y alternativa. Por ejemplo, es difícil pensar que el tiempo promedio en que un paciente tarda en reaccionar favorable-

^aleonardo.olmedo@hotmail.com

mente a cierta sustancia, sea exactamente el mismo para diferentes pacientes. En realidad, el investigador debería estar interesado en la prueba de hipótesis de la forma: $\{c_1 \leq \mu \leq c_2\}$ frente $\{\mu < c_1\} \cup \{\mu > c_2\}$; prueba que permitirá al investigador, determinar si la media está en el intervalo deseado o no. Si se rechaza la hipótesis nula, la media se separa del intervalo de valores aceptables en la práctica y el rechazo de la hipótesis nula permite la recomendación práctica de diferencia relevante, es decir, el investigador puede con toda naturalidad recomendar la sustancia o suspenderla.

La prueba de esta pareja de hipótesis no se presenta en los cursos porque el método requiere solución iterativa. Por ello, consideramos de interés presentar la metodología y su solución, con base en unas gráficas, que dan la pauta para la obtención de las cuantías para la prueba convencional, ; que hagan factible que esta prueba se enseñe a nivel licenciatura.

2. Contraste de hipótesis nula central compuesta y alternativa bilateral en la distribución normal

El planteamiento de hipótesis es:

$$H_0 : \{c_1 \leq \mu \leq c_2\} \leq \delta, \text{ vs } H_a : \{\mu < c_1\} \cup \{\mu > c_2\} > \delta, \text{ con } \sigma \text{ conocida.}$$

El punto angular para solucionar la prueba del tipo anteriormente descrita es, obtener el contraste *Uniformemente Más Potente Insesgado (UMPI)*.

Definición 2.1. *Un contraste será insesgado cuando la probabilidad de rechazar la hipótesis nula siendo cierta es siempre menor igual que la de rechazarla siendo falsa,*

$$\max_{\mu \in \Theta_0} P(\text{rechazar } H_0) \leq \min_{\mu \in \Theta_a} P(\text{rechazar } H_0).$$

Teorema 2.1. *Sea x_1, x_2, \dots, x_n una muestra aleatoria de $x \sim N(\mu; \sigma)$ con $\sigma > 0$ conocida y para probar $H_0 : \{c_1 \leq \mu \leq c_2\}$ frente $H_a : \{\mu < c_1 \cup \mu > c_2\}$, con nivel de significación α , la razón verosimilitud para el contraste produce la prueba UMPI*

$$ZR = \{(T(x) \leq k_1) \cup (T(x) \geq k_2)\}$$

con $k_1 < k_2$, donde $T(x) = \bar{x}$ y $P_{\mu=c_1}(ZR) = P_{\mu=c_2}(ZR) = \frac{\alpha}{2}$.

El Teorema 2.1 establece la zona de rechazo de la prueba anteriormente descrita. Para ello, utiliza los conceptos de: razón de verosimilitud para una prueba de tamaño α , da probabilidad de rechazo igual a alfa en la frontera de $\{c_1, c_2\}$, los extremos del intervalo para μ en H_0 , de donde resulta que la prueba es *UMP* de entre todas las de tamaño α , y también, el de hipótesis *insesgada*, de donde se deriva, que la potencia de la prueba será mayor o igual a α fuera de H_0 , es decir, para μ fuera de $[c_1, c_2]$ (Definición 1.1). En su prueba se usa que la función de verosimilitud de la prueba es convexa respecto a $T(x)$, con base en ello, la prueba resulta ser *Uniformemente Más Potente Insesgada (UMPI)*. La demostración de Teorema 2.1 pueden seguirse en Borovkov (1988) o Lehmann (1986).

La zona de rechazo, ZR , se compone de dos intervalos $\{(T(x) \leq k_1) \cup (T(x) \geq k_2)\}$.

3. Resultados

Para determinar $P(T(x) \leq k_1) \cup P(T(x) \geq k_2)$ se implementó un método numérico para aproximarnos a la solución de los valores, k_1 y k_2 . Para dar una solución que no dependa de los valores de K y σ , se toma $(c_2 - c_1)/\sigma = 2\delta$. La prueba se hace comparando Z calculado igual a $(\bar{x} - (K + \delta))/\sigma * \sqrt{n}$ si $\bar{x} > K$ y, con $Z = (\bar{x} - (K - \delta))/\sigma * \sqrt{n}$; si $\bar{x} < K$ se compara con el $100(\alpha_1)\%$, si $\bar{x} < K$ y con el percentil $100(1 - \alpha_2)\%$ si $\bar{x} > K$, y $\alpha_1 = \alpha - \alpha_2$. Las gráficas dan la solución para α_2 a partir de los valores de δ y n .

Se realiza la prueba hacia el lado donde quede \bar{x} usando una prueba unilateral con nivel de significación, ya sea, α_1 si $\bar{x} < K$ o α_2 si $\bar{x} > K$, esto produce α de significación, vea la figura 1(a). En la figura se muestra Alfa_2 (α_2), valor de probabilidad a la derecha, que fija al percentil $100(1 - \alpha_2)\%$ de la Normal estándar que se debe usar para realizar la prueba hacia la derecha cuando $\bar{x} > K$, δ está dado por $2\delta = (c_2 - c_1)/\sigma$, el nivel de significación es $\alpha = 0.05$ y $\bar{x} > K$, la prueba se hará hacia la derecha. Si $\bar{x} < K$ se hará prueba a la izquierda con el percentil $100(\alpha - \alpha_2)\%$ de Z como valor crítico.

Procedimiento para usar la gráfica. *Para cada valor de δ y n , tome la línea vertical que cruce a la curva de n , en el punto de cruce tome la línea horizontal que lo llevará al valor de α_2 (en la gráfica Alfa_2), que le indica que deberá comparar el valor de Z calculado con el percentil $100(1 - \alpha_2)\%$ de la Normal estándar.*

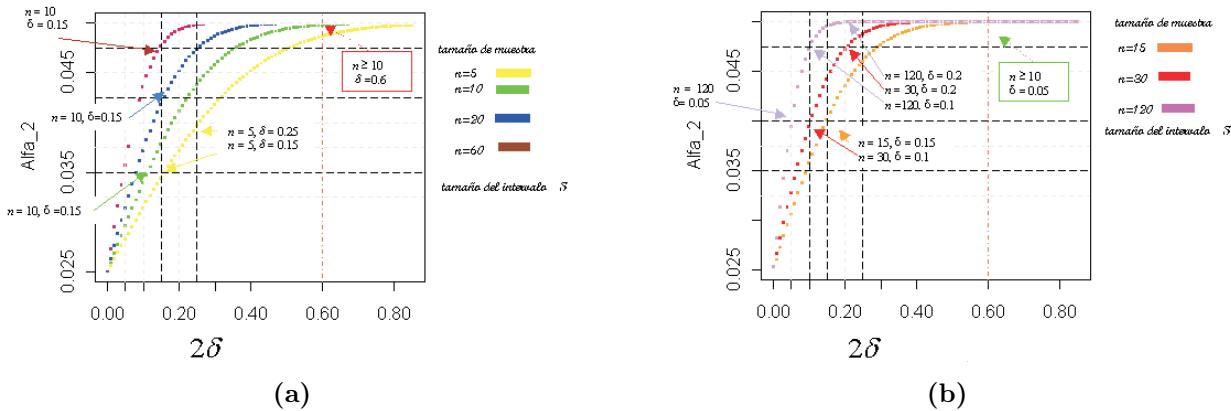


Figura 1: Método gráfico para determinar la probabilidad α_2 hacia la derecha, en la prueba con hipótesis nula compuesta $|\mu - K| < \delta$ y $\delta \in [0, 0.85]$ y tamaños de muestra de $n = \{5, 10, 15, 20, 30, 60, 120\}$

Ejemplo 3.1. Si $\bar{x} > K$ y $Z = (\bar{x} - (K + \delta)) / \sigma * \sqrt{n}$, para $\delta \geq 0.6$ y $n > 5$, se debe usar el percentil 95% de Z , como la prueba convencional de nivel 5% unilateral a la derecha. Si $n = 5$ y $\delta = 0.15$, se debe usar $\alpha_2 = 0.035$ lo que produce una prueba unilateral a la derecha con el percentil 96.5% de Z y, si $n = 5$ y $\delta = 0.25$ tenemos, $\alpha_2 = 0.04$ con la prueba unilateral a la derecha con el percentil 96% de Z .

Ejemplo 3.2. Para $n = 20$ y $\delta = 0.15$, entonces $\alpha_2 = 0.0425$, que produce una prueba unilateral a la derecha con el percentil 95.75 % de Z y, $n = 20$ y $\delta = 0.25$, entonces $\alpha_2 = 0.0475$, que produce una prueba unilateral a la derecha con el percentil 95.25 % de Z .

Ejemplo 3.3. Si $n = 10$ y $\delta = 0.1$, tenemos $\alpha_2 = 0.035$, prueba unilateral a la derecha con el percentil 96.5 % de Z , si $n = 60$ y $\delta = 0.15$, obtenemos $\alpha_2 = 0.0475$, que produce una prueba unilateral a la derecha con el percentil 95.25 % de Z ; etc.

Ejemplo 3.4. Si $\bar{x} < K$ y $Z = (\bar{x} - (K - \delta)) / \sigma * \sqrt{n}$, la prueba se hará hacia la izquierda y se usa el percentil $100(\alpha_1)\%$ de la Normal estándar. En el gráfico, si $\delta = 0.15$ y $n = 15$, la zona de rechazo sería, $\alpha_2 = 0.035$ y $\alpha_1 = 0.015$; $\delta = 0.15$ y $n = 30$ la zona de rechazo sería, $\alpha_2 = 0.045$ y $\alpha_1 = 0.005$ (ver figura 1(b)).

Ejemplo 3.5. Si $\delta = 0.2$ y $n = 30$, la zona de rechazo será, $\alpha_2 = 0.0475$ y $\alpha_1 = 0.0025$; si $\delta = 0.2$ y $n = 120$, la zona de rechazo será, $\alpha_2 = 0.05$, unilateral a la izquierda.

Ejemplo 3.6. Si fuese $\delta = 0.1$ y $n = 30$ obtendríamos una zona de rechazo de $\alpha_2 = 0.04$ y $\alpha_1 = 0.01$; y para $\delta = 0.1$ y $n = 120$, el valor de será $\alpha_2 = 0.0475$ y $\alpha_1 = 0.0025$. Para el caso cuando $\delta = 0.05$ y $n = 120$, produce una prueba unilateral a la izquierda cuyos valores de α_2 y α_1 , serán 0.04 y 0.01, respectivamente, ver figura 1(b).

4. Conclusiones

Se ha propuesto un método gráfico que da los percentiles de la normal estándar que se deben usar como valores críticos para la prueba de hipótesis nula central compuesta con alternativa bilateral en la distribución Normal, $H_0 : \{c_1 \leq \mu \leq c_2\}$ versus $H_a : \{\mu < c_1 \cup \mu > c_2\}$. El valor calculado de Z es $\bar{x} < K$ y la prueba es hacia la izquierda y es si $\bar{x} > K$ y la prueba es hacia la derecha. A diferencia de las conclusiones en prueba convencional con nula puntual, donde el rechazo de la hipótesis nula de igualdad no necesariamente lleva a una recomendación práctica, en este contraste la significación estadística refleja una diferencia con la igualdad que es de interés al investigador, adquiriendo sentido práctico, pues se enfoca en probar si la media difiere menos que delta de el valor K o la diferencia es mayor que delta por lo que, al rechazar la hipótesis nula el investigador puede con toda naturalidad recomendar que el valor de la media difiere de K más que delta, una diferencia mayor que aquella que estableció como criterio para la comparación.

Referencias

- Borovkov, A. A., Estadística Matemática; Editorial Mir, Moscú, 1988.
- Lehmann, E.L., Testing Statistical Hypotheses, 3rd.. ed., New York; John & Wiley Sons, Inc., 1986.
- Mood, A. M., The Theory of Statistics, 2nd ed., New York, McGraw Hill Book Company, Inc., 1963.

Análisis de sendero como herramienta confirmatoria en un experimento de campo

Emilio Padrón Corral^a *Universidad Autónoma de Coahuila*

Ignacio Méndez Ramírez^b *Universidad Nacional Autónoma de México*

Armando Muñoz Urbina *Universidad Autónoma Agraria Antonio Narro*

1. Introducción

La especie arbórea Ciríán (*Crescentia alata* H.B.K) se utiliza como tratamiento medicinal para controlar enfermedades, es originaria de México y se cultiva en los estados de: Michoacán, Colima, Guerrero, Jalisco y Nayarit; los datos se obtuvieron de un trabajo que se realizó en el área denominada, El Llano, Municipio de Coahuayana, Michoacán, México; Avila (1999), formando 30 cuadrantes en una superficie de 120 hectáreas, con 279 árboles muestreados. Las variables a medir fueron: altura del árbol , número de ramas, diámetro de ramas, cobertura, diámetro ecuatorial, diámetro polar, número de frutos, peso de frutos y rendimiento. El objetivo es desarrollar un análisis de coeficientes de sendero para estudiar las relaciones entre las componentes del rendimiento.

2. Metodología

Los coeficientes del análisis de sendero con efectos directos e indirectos, fueron estimados de acuerdo a Wright (1934), los que posteriormente fueron descritos por Dewey y Lu (1959) y por Li (1975). Wright, ideó la manera de interpretar ecuaciones normales para resolver coeficientes de regresión estandarizados en problemas de regresión múltiple. El análisis de sendero o método de coeficientes de sendero, es una forma de análisis de regresión estructurado, varios modelos de regresión ligados, y considerando variables estandarizadas a media

^aepadron@mate.uadec.mx

^bimendez@servidor.unam.mx

cero y varianza uno, en un sistema cerrado. Se establecen varias ecuaciones que determinan todas las correlaciones entre las variables observadas. Es prácticamente indispensable proponer un diagrama o modelo gráfico, donde se especifique las cadenas causales propuestas por el investigador. Lo que se obtiene es el grado de cercanía de las observaciones empíricas con las cadenas causales propuestas por el investigador, es decir se apoya o no la hipótesis resumida en la estructura causal propuesta, y además se evalúa el peso de cada relación, vía los llamados coeficientes de sendero. También se obtienen los efectos directos e indirectos de entre variables. Los efectos directos son los coeficientes de regresión estandarizados de una variable dependiente sobre otra dependiente. Los efectos indirectos son la influencia sumada de una variable independiente sobre otra dependiente vía las correlaciones o senderos que llevan a la dependiente de manera indirecta. Es decir, los efectos directos son coeficientes de regresión estandarizados que aplicados al mejoramiento de plantas permite un avance más rápido en la selección de genotipos sobresalientes en la variable de estudio, los cuales son denominados coeficientes de sendero (b), (denotado por una línea con una flecha); cada variable predictora tiene un efecto directo y un efecto indirecto para cada una de las otras variables asociadas. El efecto indirecto es aquel que se obtiene a través de otras variables y se estima del producto del coeficiente de correlación y su respectivo efecto directo y nos permite detectar su efecto correspondiente en la variable de estudio. (denotado por una línea con dos flechas). Figura 1, Los datos se analizaron con el paquete computacional MATLAB.

En el sistema de ecuaciones (*) el primer efecto indirecto en la primera ecuación normal [1], es dado por $r_{12}b_{25}$, es decir, la serie de expresiones que comprenden todas las vías para la primera variable predictora forman una ecuación normal [1], cuando se sumarizan igualan el coeficiente de correlación entre la variable de respuesta X_5 y la primera variable predictora X_1 , (efecto directo negreado).

$$\begin{aligned}
 \mathbf{b}_{15} + r_{12}b_{25} + r_{13}b_{35} + r_{14}b_{45} &= r_{15} \quad [1] \\
 r_{12}b_{15} + \mathbf{b}_{25} + r_{23}b_{35} + r_{24}b_{45} &= r_{25} \quad [2] \\
 r_{13}b_{15} + r_{23}b_{25} + \mathbf{b}_{35} + r_{34}b_{45} &= r_{35} \quad [3] \\
 r_{14}b_{15} + r_{24}b_{25} + r_{34}b_{35} + \mathbf{b}_{45} &= r_{45} \quad [4]
 \end{aligned} \tag{*}$$

Sustituyendo las matrices y ecuaciones siguientes en el paquete computacional MATLAB, obtenemos los coeficientes de sendero, el residual y el coeficiente de determinación correspondientes.

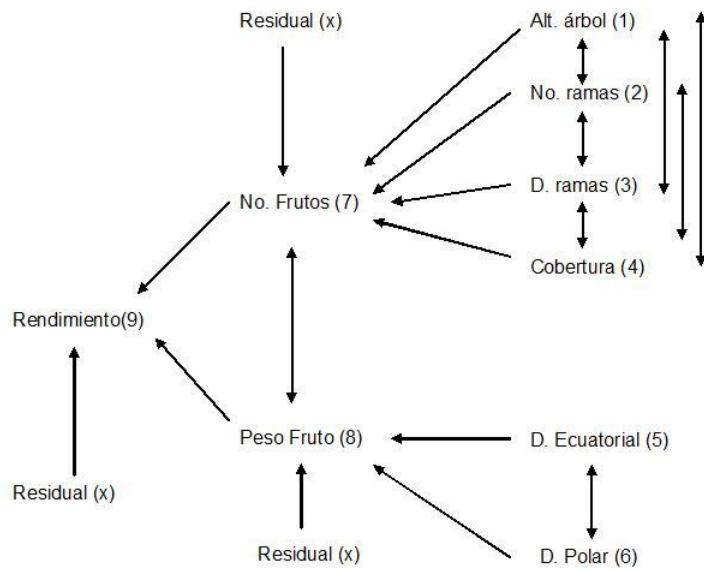


Figura 1: Diagrama o modelo gráfico de sendero mostrando las correlaciones (\longleftrightarrow) y los efectos directos (\longrightarrow) entre variables de árbol y fruto en el cultivo del Cirián.

$$A = \begin{pmatrix} 1 & r_{12} & r_{13} & r_{14} \\ r_{21} & 1 & r_{23} & r_{24} \\ r_{31} & r_{32} & 1 & r_{34} \\ r_{41} & r_{42} & r_{43} & 1 \end{pmatrix}$$

vector de correlaciones finales

$$B = \begin{pmatrix} r_{15} \\ r_{25} \\ r_{35} \\ r_{45} \end{pmatrix}$$

los coeficientes de sendero (b)

$$b = A^{-1}B$$

el residual E

$$E = \sqrt{1 - b'B}$$

y el coeficiente de determinación

$$R^2 = 1 - E^2$$

3. Resultados y discusión

En el Cuadro 1, se puede observar que al correlacionar los caracteres con número de frutos, es la cobertura la que nos indica la mayor asociación con un 99 porciento de confianza, esto indica que a medida que se incrementa la cobertura se incrementa el número de frutos, igualmente para peso de fruto tanto diámetro ecuatorial como diámetro polar tienen asociación significativa con peso de fruto, es decir, a medida que se incrementa el diámetro ecuatorial como diámetro polar, se incrementa el peso de fruto con un 99 porciento de confianza; en lo referente a la correlación entre número de frutos y peso de fruto contra rendimiento, esta fué de buena calidad, pero es número de frutos el que mayor sobresalió con respecto a rendimiento, con un 99 porciento de confianza.

Los efectos directos e indirectos se muestran en el Cuadro 2, se observa que el efecto directo de mayor ponderación es de 0.8938, debido a que ninguno excede la unidad se puede concluir que la multicolinealidad no ha producido coeficientes de sendero inflados. de los efectos directos sobre número de frutos; cobertura fue la más sobresaliente con un valor de 0.7273, de los efectos indirectos diámetro de ramas presentó el valor más alto a través de cobertura con un valor de 0.3338; el análisis de sendero para número de frutos no explicó en gran proporción la variación del número de frutos como se indica por el bajo valor del coeficiente de determinación $R^2 = 0.59$ y por el correspondiente efecto del residual ($\text{Res} = 0.6391$). Se observa que los componentes diámetro ecuatorial y diámetro polar influyeron en el peso de fruto directa e indirectamente, los efectos indirectos fueron sobresalientes, sin embargo, los efectos directos tuvieron mayor impacto sobre el peso de fruto y fueron casi de igual magnitud; el efecto del residual fue alto ($\text{Res} = 0.5581$) y el coeficiente de determinación fue bajo $R^2 = 0.69$ por lo que el diámetro ecuatorial y el diámetro polar no explicaron en gran proporción la variación en el peso de fruto. En el análisis de sendero para rendimiento de fruto, se observa que los efectos directos e indirectos de número de frutos y peso de fruto también fueron todos positivos, número de frutos presentó el efecto directo más alto (0.8938) por otra parte, los efectos indirectos fueron relativamente bajos comparados con los efectos

Carácter	Altura de árbol	Número de ramas	Diámetro de ramas	Cobertura	Número de frutos
a)					
Altura de árbol	1.000	-0.041	0.389*	0.356*	0.333
Número de ramas		1.000	-0.475**	0.320	0.298
Diámetro de ramas			1.000	0.459*	0.313
Cobertura				1.000	0.763**
Número de frutos					1.000
b)					
	Diámetro ecuatorial	Diámetro polar	Peso de fruto		
Diámetro ecuatorial	1.000	0.668**	0.748**		
Diámetro polar		1.000	0.767**		
Peso de fruto			1.000		
c)					
	Número fruto	Peso de fruto	Rendimiento de fruto		
Número de frutos	1.000	0.169	0.943**		
Peso de fruto		1.000	0.442**		
Rendimiento de fruto			1.000		

* Significativo al 5%

** Significativo al 1%

Tabla 1: Coeficientes de correlación entre varios caracteres relacionados con: a) número de frutos; b) peso de fruto; c) rendimiento de fruto, en árbol de Cirián

directos; el efecto del residual fue bajo ($\text{Res}=0.1689$) y el coeficiente de determinación fue alto $R^2 = 0.97$, por lo que número de frutos y peso de fruto explicaron el 97 porciento de la variación en el rendimiento de fruto.

4. Conclusiones

1. El análisis de coeficientes de sendero mostró que la cobertura fue un factor importante para determinar el número de frutos.
2. Los efectos directos de diámetro ecuatorial y diámetro polar sobre peso de fruto, manifestaron buena relación, y explicaron en un 69 porciento la variación en el peso de fruto.
3. El análisis de sendero para rendimiento de fruto, muestra que el incremento en el número de frutos es el factor más importante para mejorar el rendimiento de fruto por árbol. El coeficiente de determinación fue alto y muestra que el número de frutos y el peso de fruto explicaron en un 97 porciento la variación en el rendimiento de fruto.

Referencias

- Avila, R.A. (1999). Ecología y Evaluación del Fruto del Cirián (*Crescentia alata* H.B.K.) Como Recurso Forrajero en la Localidad el Llano, Municipio de Coahuayana, Michoacán, México. Tesis de Maestría, Universidad Autónoma Agraria Antonio Narro, Buenavista, Saltillo, Coahuila, México. p. 1-71.
- Dewey, D.R. y Lu, K.H. (1959). A Correlation and Path Coefficient Analysis of Components of Crested Wheatgrass Seed Production. *Agronomy Journal*, **51**, 515-518.
- Li, C.C. (1975). *Path Analysis: A Primer*. Boxwood Press, Pacific Grove, C.A. MATLAB; The Language of Technical Computing. Version 7.0.0. 19920 (R14). Copyright (1984-2004). The MathWorks. Inc.
- Wright, S. (1934). The Method of Path Coefficients. *Ann. Math. Stat.*, **5**, 161-215.

Carácter	Altura de árbol	Número de ramas	Diámetro de ramas	Cobertura	Número de frutos
a)					
Altura de árbol	0.0875	-0.0023	-0.0111	0.2589	0.333
Número de ramas	-0.0036	0.0552	0.0136	0.2327	0.298
Diámetro de ramas	0.0340	-0.0262	-0.0282	0.3338	0.313
Cobertura	0.0311	0.0177	-0.0131	0.7273	0.763**
Residual=0.6391					
$R^2 = 1 - (0.6391)^2 = 0.59$					
b)					
	Diámetro ecuatorial	Diámetro polar	Correlación con peso de fruto		
Diámetro ecuatorial	- 0.4255	0.3225	0.748**		
Diámetro polar	0.2842	0.4828	0.767**		
Residual=0.5581					
$R^2 = 1 - (0.5581)^2 = 0.69$					
c)					
	Número fruto	Peso de fruto	Correlación con rendimiento de fruto		
Número de frutos	0.8938	0.0492	0.943**		
Peso de fruto	0.1511	0.2909	0.442**		
Residual=0.1689					
$R^2 = 1 - (0.5581)^2 = 0.69$					

Tabla 2: Efectos directos (negreado) e indirectos del análisis de coeficientes de sendero para: a) número de frutos; b) peso de fruto; c) rendimiento de fruto, en árbol de Cirián

Comparación de poblaciones normales asimétricas

Paulino Pérez Rodríguez^a, José A. Villaseñor Alva^b
Colegio de Postgraduados

1. Introducción

Las distribuciones normales asimétricas constituyen una familia de distribuciones de tres parámetros: localidad, escala y forma, la cual contiene a la familia normal cuando el parámetro de forma es 0 y a la distribución media-normal cuando dicho parámetro tiende a infinito. Esta familia de distribuciones tiene algunas de las propiedades de la familia normal, lo que la hace atractiva desde el punto de vista de aplicaciones. Esta familia apareció de forma independiente varias veces en la literatura estadística (ver Roberts (1966), O'Hagan y Leonard (1976)); sin embargo, fue Azzalini (1985) quien estudió sus principales propiedades, propuso algunas generalizaciones y le dio el nombre con el cual se le conoce actualmente. Una revisión completa sobre esta distribución se encuentra en Azzalini (2005).

En este trabajo se presenta una solución al problema de comparación de dos poblaciones normales asimétricas, la cual utiliza una prueba de razón de verosimilitudes generalizada con respecto a mezclas de normales asimétricas.

^aperpdgo@colpos.mx

^bjvillasr@colpos.mx

2. La distribución normal asimétrica

Definición 2.1. Una v.a. Z tiene distribución normal asimétrica con parámetro de forma γ si su función de densidad es:

$$f_Z(z; \gamma) = 2\phi(z)\Phi(\gamma z)I_{(-\infty, \infty)}(z), \quad (1)$$

donde $\phi(\cdot)$ y $\Phi(\cdot)$ denotan la función de densidad y de distribución normal estándar, $\gamma \in \mathbb{R}$.

Si Z tiene la función de densidad (1) entonces usualmente se escribe $Z \sim SN(\gamma)$. Si $Y = \xi + \omega Z$ con $\xi \in \mathbb{R}$ y $\omega \in \mathbb{R}^+$, entonces $Y \sim SN(\xi, \omega, \gamma)$ y su función de densidad es:

$$f_Y(y; \xi, \omega, \gamma) = 2\frac{1}{\omega}\phi\left(\frac{y - \xi}{\omega}\right)\Phi\left[\gamma\left(\frac{y - \xi}{\omega}\right)\right]I_{(-\infty, \infty)}(y).$$

3. Comparación de poblaciones

Sea X_1, \dots, X_n una m. a. de $SN(\xi_1, \omega_1, \gamma)$ y Y_1, \dots, Y_m una m.a. de $SN(\xi_2, \omega_2, \gamma)$ y se supone que las muestras son independientes. Se desea saber si las observaciones vienen de una distribución normal asimétrica o de una mezcla de dos distribuciones normales asimétricas. Si los datos son de una mezcla, es de interés conocer los parámetros y la proporción de los componentes individuales que forman la misma. La función de densidad de la mezcla está dada por

$$f_W(w; \pi, \xi_1, \omega_1, \xi_2, \omega_2, \gamma) = \pi f_{Z_1}(w; \xi_1, \omega_1, \gamma) + (1 - \pi)f_{Z_2}(w; \xi_2, \omega_2, \gamma),$$

donde $Z_1 \sim SN(\xi_1, \omega_1, \gamma)$, $Z_2 \sim SN(\xi_2, \omega_2, \gamma)$, $\pi \in [0, 1]$.

Se plantea el siguiente juego de hipótesis:

$$H_0 : \xi_1 = \xi_2 = \xi, \omega_1 = \omega_2 = \omega > 0, \pi \in [0, 1], \gamma \in \mathbb{R} \text{ vs } H_1 : \xi_1 \neq \xi_2 \text{ ó } \omega_1 \neq \omega_2, \pi \in [0, 1], \gamma \in \mathbb{R}.$$

Para probar este juego de hipótesis se propone usar una prueba de razón de verosimilitudes generalizada.

Estimadores de máxima verosimilitud en todo el espacio de parámetros

En este caso es necesario calcular 6 parámetros, i.e. $\boldsymbol{\theta} = (\pi, \xi_1, \omega_1, \xi_2, \omega_2, \gamma)'$. Para estimar los parámetros se utiliza el algoritmo Esperanza-Mazimización Generalizado (GEM). Se parte del hecho de que se tiene una sola muestra Z_1, \dots, Z_{n+m} de una mezcla de dos normales asimétricas $f_W(w; \pi, \xi_1, \omega_1, \xi_2, \omega_2, \gamma)$. Sea $\delta_i = 1$ si la observación i -ésima viene del primer componente de la mezcla y $\delta_i = 0$ en caso contrario. Entonces la verosimilitud para los datos completos W_1, \dots, W_{n+m} con $W_i = (Z_i, \delta_i)$ está dada por:

$$f(w_1, \dots, w_{n+m}; \boldsymbol{\theta}) = \prod_{i=1}^{n+m} \pi^{\delta_i} (1 - \pi)^{1 - \delta_i} (f_{Z_1}(z_i; \xi_1, \omega_1, \gamma))^{\delta_i} (f_{Z_2}(z_i; \xi_2, \omega_2, \gamma))^{1 - \delta_i} \quad (2)$$

Etapa E

Al tomar el logaritmo de (2) se obtiene:

$$\begin{aligned} l(\boldsymbol{\theta}) &\propto \sum_{i=1}^{n+m} (1 - \delta_i) \log(1 - \pi) - \sum_{i=1}^{n+m} \delta_i \left(\log \omega_1 + \frac{1}{2} \left(\frac{z_i - \xi_1}{\omega_1} \right)^2 - \log \Phi \left(\gamma \frac{z_i - \xi_1}{\omega_1} \right) \right) \\ &\quad - \sum_{i=1}^{n+m} (1 - \delta_i) \left(\log \omega_2 + \frac{1}{2} \left(\frac{z_i - \xi_2}{\omega_2} \right)^2 - \log \Phi \left(\gamma \frac{z_i - \xi_2}{\omega_2} \right) \right) + \sum_{i=1}^{n+m} \delta_i \log \pi. \end{aligned}$$

Como el operador esperanza es lineal, la función

$$Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(j-1)}) = E \left\{ \log f(w_1, \dots, w_{n+m}; \boldsymbol{\theta}) | z_1, \dots, z_{n+m}, \boldsymbol{\theta}^{(j-1)} \right\}$$

es la log-verosimilitud de los datos completos, pero se sustituye δ_i por su valor esperado, es decir, $p_i = E\{\delta_i | z_1, \dots, z_{n+m}, \boldsymbol{\theta}^{(j-1)}\} = P(\delta_i = 1 | z_1, \dots, z_{n+m}, \boldsymbol{\theta}^{(j-1)})$. Luego por el teorema de Bayes:

$$p_i = \frac{P(\delta_i = 1)P(z_i, \boldsymbol{\theta}^{(j-1)} | \delta_i = 1)}{P(\delta_i = 0)P(z_i, \boldsymbol{\theta}^{(j-1)} | \delta_i = 0) + P(\delta_i = 1)P(z_i, \boldsymbol{\theta}^{(j-1)} | \delta_i = 1)}.$$

Etapa M

Hay que maximizar

$$\begin{aligned} Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(j-1)}) &\propto -\sum_{i=1}^{n+m} p_i \left(\log \omega_1 + \frac{1}{2} \left(\frac{z_i - \xi_1}{\omega_1} \right)^2 - \log \Phi \left(\gamma \frac{z_i - \xi_1}{\omega_1} \right) \right) \\ &\quad - \sum_{i=1}^{n+m} (1 - p_i) \left(\log \omega_2 + \frac{1}{2} \left(\frac{z_i - \xi_2}{\omega_2} \right)^2 - \log \Phi \left(\gamma \frac{z_i - \xi_2}{\omega_2} \right) \right) \\ &\quad + \sum_{i=1}^{n+m} p_i \log \pi + (1 - p_i) \log(1 - \pi). \end{aligned}$$

La etapa de maximización no se puede hacer de manera analítica, por lo cual no se podrá aplicar el algoritmo EM de forma directa y será necesario recurrir al algoritmo GEM. Es necesario ser muy cuidadosos al momento de calcular $\boldsymbol{\theta}$, hay que recordar que $\pi \in [0, 1]$, lo cual podría causar inestabilidades numéricas al momento de implementar el algoritmo GEM. El problema anterior puede evitarse reduciendo la dimensionalidad de la función a maximizar, si se considera que $\xi_1, \omega_1, \xi_2, \omega_2, \gamma$ son fijos y entonces $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(j-1)})$ es solo función de π , empleando la técnica de derivadas podemos conocer el valor de π que maximiza la función, es decir:

$$\frac{\partial Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(j-1)})}{\partial \pi} = \sum_{i=1}^{n+m} \frac{p_i}{\pi} - \frac{1 - p_i}{1 - \pi} = 0.$$

De donde se obtiene $\hat{\pi} = (n + m)^{-1} \sum_{i=1}^{n+m} p_i$. Una vez que conocemos el valor del parámetro π se buscan $\xi_1, \omega_1, \xi_2, \omega_2, \gamma$ tal que $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^*) \geq Q(\boldsymbol{\theta}^*; \boldsymbol{\theta}^*)$. Sea $H(\boldsymbol{\theta}; \boldsymbol{\theta}^{(j-1)}) = -Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(j-1)})$ entonces el problema de buscar $\boldsymbol{\theta}$ tal que $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^*) \geq Q(\boldsymbol{\theta}^*; \boldsymbol{\theta}^*)$ es equivalente a buscar $\boldsymbol{\theta}$ tal que $H(\boldsymbol{\theta}; \boldsymbol{\theta}^*) \leq H(\boldsymbol{\theta}^*; \boldsymbol{\theta}^*)$. Para resolver este problema se emplea el método del descenso más rápido. Las etapas de Esperanza-Maximización se realizan de manera alternada hasta que se alcanza la convergencia. El algoritmo GEM fue programado en el paquete R.

Estimadores de máxima verosimilitud bajo la hipótesis nula

El cálculo de los estimadores bajo la hipótesis nula se realiza de la manera usual.

Prueba de razón de verosimilitudes generalizada

Para probar la hipótesis de interés, se utiliza la estadística dada por el cociente de razón de verosimilitudes $\lambda(\mathbf{Z})$. Se rechaza la hipótesis nula al nivel de significancia α si y solo si $-2 \log \lambda(\mathbf{Z}) \geq \chi^2_{2,1-\alpha}$.

Potencia de la prueba

Con la finalidad de estudiar la potencia de la prueba descrita en la sección anterior, se consideran algunas alternativas. En la tabla 1 se presentan las potencias estimadas mediante simulación Monte Carlo con $B = 1,000$ réplicas de los tamaños indicados, $\alpha = 0.05$.

El algoritmo GEM resultó ser herramienta efectiva para la obtención de los estimadores de máxima verosimilitud requeridos para la solución del problema de comparación de poblaciones normales asimétricas. En la tabla 1, se observa que a medida que la distancia entre parámetros de localidad se hace más grande la potencia se incrementa, como era de esperarse.

Referencias

- Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, **12**, 171-178.
- Azzalini, A. (2005). The skew normal distribution and related multivariate families. *Scandinavian Journal of Statistics*, **32**, 159-188.
- O'Hagan, A. y Leonard, T. (1976). Bayes Estimation Subject to Uncertainty About Parameters Constraints. *Biometrika*, **63**, 201-203.
- Roberts, C. (1966). A correlation model useful in the study of twins. *Journal of the American Statistical Association*, **61**, 1184-1190.

Tamaño de muestra(n)	Diferencia entre parámetros de loc.		
	$\xi_2 - \xi_1 = 2$	$\xi_2 - \xi_1 = 3$	$\xi_2 - \xi_1 = 4$
$\pi = 0.5$			
50	0.210	0.633	0.781
100	0.265	0.921	0.952
150	0.371	0.981	0.988
$\pi = 0.6$			
50	0.258	0.657	0.912
100	0.403	0.739	0.984
150	0.439	0.988	0.988
$\pi = 0.7$			
50	0.240	0.711	0.937
100	0.394	0.977	0.995
150	0.484	0.997	0.998
$\pi = 0.8$			
50	0.162	0.399	0.772
100	0.239	0.716	0.985
150	0.281	0.894	0.986
$\pi = 0.9$			
50	0.102	0.162	0.614
100	0.119	0.267	0.923
150	0.119	0.313	0.970

Tabla 1: Potencia de la prueba de razón de verosimilitudes para mezclas de poblaciones SN, obtenida mediante simulación Monte Carlo con $B = 1,000$, $\gamma = 1$, $\omega_1 = \omega_2 = 1$

Análisis espectral aplicado al electroencefalograma

Verónica Saavedra Gastélum^a *Universidad Autónoma de Querétaro*

Thalía Fernández Harmony *Universidad Nacional Autónoma de México*

Eduardo Castaño Tostado *Universidad Autónoma de Querétaro*

Víctor Manuel Castaño Meneses *Universidad Nacional Autónoma de México*

1. Introducción

Una serie de tiempo puede analizarse en el dominio del tiempo o en el dominio de las frecuencias. El análisis espectral permite describir el comportamiento de la señal en el dominio de las frecuencias. El método más común para describir el dominio de las frecuencias es realizar el análisis espectral de la serie a través de la transformada de Fourier, (Evans, 1999) para obtener el espectro de potencias, el cual contribuye el diagnóstico neurológico o psiquiátrico de un sujeto. Sin embargo, la transformada de Fourier presenta problemas en su aplicación a datos reales en los que el supuesto de estacionariedad no se cumpla. El presente trabajo pretende describir este problema aplicándolo a un ElectroEncefaloGramma (EEG) y proponer una manera de análisis espectral vía Ondeletas.

2. Método

El EEG es una gráfica del voltaje en función del tiempo que representa la actividad eléctrica cerebral en diferentes regiones del cuero cabelludo. El registro del EEG se realizó de acuerdo a la norma internacional 10/20 propuesta por Henri Jasper en 1958 la cual incluye 19 derivaciones monopolares: Fp1, Fp2, F3, F4, C3, C4, P3, P4, O1, O2, F7, F8, T3, T4, T5, T6, Fz, Cz, y Pz. Además se utilizaron dos electrodos cortocircuitados, A1 y A2, en las

^averoclessg@yahoo.com.mx

orejas como referencia. Las derivaciones reciben su nombre de acuerdo a su localización, los números pares indican que el electrodo está localizado en el hemisferio derecho del cerebro y los números nones indican el hemisferio izquierdo. Fp1 y Fp2 corresponden a la región Prefrontal, F3, F4, F7, F8 y Fz corresponden a la región Frontal, C3 y C4 corresponden a la región Central, P3, P4 y Pz a la región Parietal, O1 y O2 a la región Occipital, T3, T4, T5 y T6 a la parte Temporal y Cz corresponde al Vertex o Rolándico.

2.1. Densidad espectral

La densidad espectral se obtiene aplicando la transformada de Fourier a la función de autocovarianza (ACF) de la serie que se desea analizar; siempre y cuando la serie sea estacionaria al menos de segundo orden. Se obtiene mediante:

$$\widehat{f}(\nu) = \sum_{h=-(n-1)}^{(n-1)} \widehat{\gamma}(h) \exp\{-2\pi i \nu h\} \quad (1)$$

$$\widehat{\gamma}(h) = n^{-1} \sum_{t=1}^{n-h} (x_{t+h} - \bar{x})(x_t - \bar{x}) \quad (2)$$

donde h se define como el retraso de la serie y ν es la frecuencia fundamental. Aplicando la fórmula de Euler, la densidad espectral se puede rescribir para cada frecuencia fundamental ν_k como sigue:

$$X_C(\nu_k) = n^{-1/2} \sum_{t=1}^n x_t \cos(2\pi\nu_k t), \text{ y} \quad (3)$$

$$X_S(\nu_k) = n^{-1/2} \sum_{t=1}^n x_t \sin(2\pi\nu_k t) \quad (4)$$

Con lo que es posible obtener el periodograma, el cual permite estimar la potencia en la frecuencia fundamental ν_k como:

$$I(\nu_k) = X_C^2(\nu_k) + X_S^2(\nu_k). \quad (5)$$

La aplicación de la transformada de Fourier, es válida siempre y cuando la serie sea estacionaria al menos de segundo orden y satisfaga

$$\sum_{h=-\infty}^{\infty} |\gamma(h)| < \infty \quad (6)$$

Por construcción, la transformada de Fourier no permite estudiar qué frecuencias están presentes a qué tiempos específicos.

2.2. Transformada de ondeleta

La transformada de ondeleta puede pensarse como una operación lineal que descompone en el tiempo una señal en bloques elementales que aparecen en diferentes escalas o resoluciones. Al descomponer una señal en escala-tiempo, permite determinar escalas de frecuencias dominantes a tiempos específicos. Los bloques elementales se analizan de manera individual.

La transformada de ondeleta utiliza una función base, también llamada ondeleta madre, la cual debe ser oscilatoria y decaer rápidamente a cero. La transformada de ondeleta de una señal continua se define como

$$CWT_{\psi}x(a,b) = W_x(a,b) = \int_{-\infty}^{\infty} x(t) \psi_{a,b}^*(t) dt \quad (7)$$

donde la función base se define como

$$\psi_{a,b}(t) = |a|^{-1/2} \psi\left(\frac{t-b}{a}\right) \quad (8)$$

Donde a representa un parámetro de escala (o resolución que corresponde al inverso de la frecuencia) y b un parámetro de corrimiento. Una vez que la función base es seleccionada, el procedimiento consiste en dilatar (o contraer) y trasladar la función base a partir de los parámetros a y b . El resultado de la transformada de ondeleta es un conjunto de señales a diferentes escalas o resoluciones. A esto se le conoce como Análisis de Multiresolución (MRA) (Walker, 1999).

Para el caso discreto, la translación y dilatación de la función base generan información redundante, para eliminarla es necesario redefinir los parámetros de escala y translación como: $a = 2^j$ (factor de dilatación) y $b = 2^j k$ (factor de localización). Lo que convierte a (8) en:

$$\psi_{j,k}(t) = 2^{-j/2} \psi\left(\frac{t-2^j k}{2^j}\right) \quad (9)$$

Con la MRA es posible reconstruir la señal a través de una suma de la señal promediada y señales de detalle como sigue:

$$x(t) = A^k + D^k + \dots + D^2 + D^1 \quad (10)$$

siempre y cuando la señal sea divisible k veces por 2. A^k es k -ésima señal promediada, y D^k, \dots, D^2, D^1 son las señales de detalle. A partir de (9):

$$x(t) \approx \sum_k s_{J,k} \phi_{J,k}(t) + \sum_k d_{J,k} \psi_{J,k}(t) + \dots + \sum_k d_{1,k} \psi_{1,k}(t) \quad (11)$$

donde J es el número de componentes en multiresolución (o escales), k toma valores entre 1 y el número de coeficientes en cada componente, $s_{J,k}$ es el k -ésimo coeficiente promediado y $d_{J,k}, \dots, d_{1,k}$ son los coeficientes de detalle.

Entre las ventajas que ofrece la transformada de ondeleta, es que permite trabajar con series no-estacionarias.

3. Aplicación al EEG transformada de ondeleta

Se realizaron dos EEG's en el cuero cabelludo de dos niños en reposo entre 6 y 11 años de edad, uno de ellos considerado como niño sano, es decir, sin trastorno de aprendizaje y el segundo, un niño con trastorno de aprendizaje. El EEG fue editado por expertos para remover todo artefacto derivado de cualquier actividad eléctrica cuyo origen no es cerebral, como son el parpadeo, la sudoración, etc. Los datos fueron registrados cada dos milisegundos.

En la toma de un EEG, se posicionan electrodos en lugares diferentes del cuero cabelludo conocidas como derivaciones. Se decidió analizar la derivación llamada Cz debido a que es la menos afectada por artefactos. De acuerdo con Kavale (1988), el cual refiere al trastorno de aprendizaje como una disfunción del cerebro, para efectos de simplicidad en este artículo llamaremos al niño que presenta trastorno de aprendizaje como sujeto con daño cerebral y sin daño cerebral al niño sin problemas de aprendizaje. La Figura 1 muestra un segmento del EEG libre de artefactos registrado por la derivación Cz, tanto para el niño con daño y sin daño cerebral. Como se puede observar en la Figura 1, el EEG del niño sin daño oscila más lento que el EEG del niño con daño cerebral.

La obtención del Periodograma se realiza para conocer los ritmos electroencefalográficos de una persona, definidos como las bandas de frecuencia delta [1,3.5]Hz, theta (3.5,7.5]Hz, alfa (7.5,12.5]Hz y beta (12.5,20]Hz. Existen normas extranjeras ya establecidas para determinar si un sujeto puede considerarse como normal o con daño cerebral, dependiendo de los valores obtenidos en las bandas de frecuencia de acuerdo con Thatcher (1998). Este método se

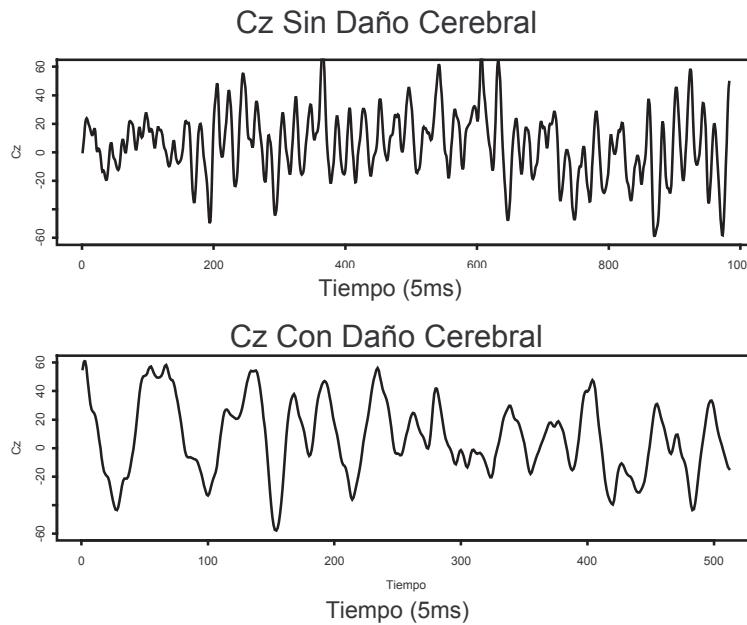


Figura 1: Representación gráfica del EEG registrado por Cz contra el tiempo, para el niño con daño y sin daño cerebral

realiza con base en la transformada Rápida de Fourier (FFT), por lo que es necesario primero verificar los supuestos inherentes al método.

Para corroborar el supuesto de estacionariedad de una serie de tiempo es necesario que su función de autocorrelación decaiga rápidamente a cero, sin embargo como se puede observar en la Figura 2, no sólo no decae rápido sino que muestra un comportamiento cíclico. Siendo más evidente en la persona con daño cerebral.

Debido a la falta de estacionariedad en las dos series, no tiene sentido aplicar FFT en el análisis espectral, por lo que se realizó dicho análisis aplicando la transformada de ondeleta, con la función base Daub4.

Para ambas series se calcularon (usando el paquete S plus) los coeficientes de niveles de resolución d1, d2, d3, d4, d5 y d6 obtenidos al aplicar la transformada ondeleta, donde d1 representa la escala más fina y d6 la escala más rugosa. Mientras que el coeficiente s6 representa al vector de coeficientes del comportamiento dominante suavizado de cada EEG. En la Figura 3 se puede observar que la persona con daño cerebral es mejor representada por coeficientes de escala rugosa, mientras que la persona sin daño cerebral, presenta coeficientes

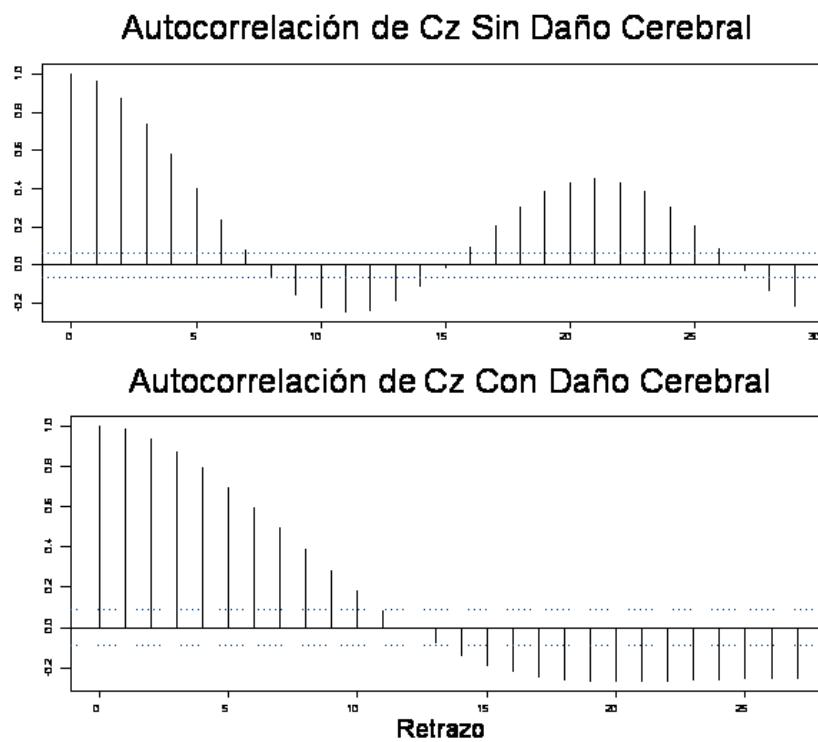


Figura 2: Función de autocorrelación del EEG registrado por Cz, para el niño con daño y sin daño cerebral

en escalas más finas. Además es posible observar en qué tiempo ocurren los detalles.

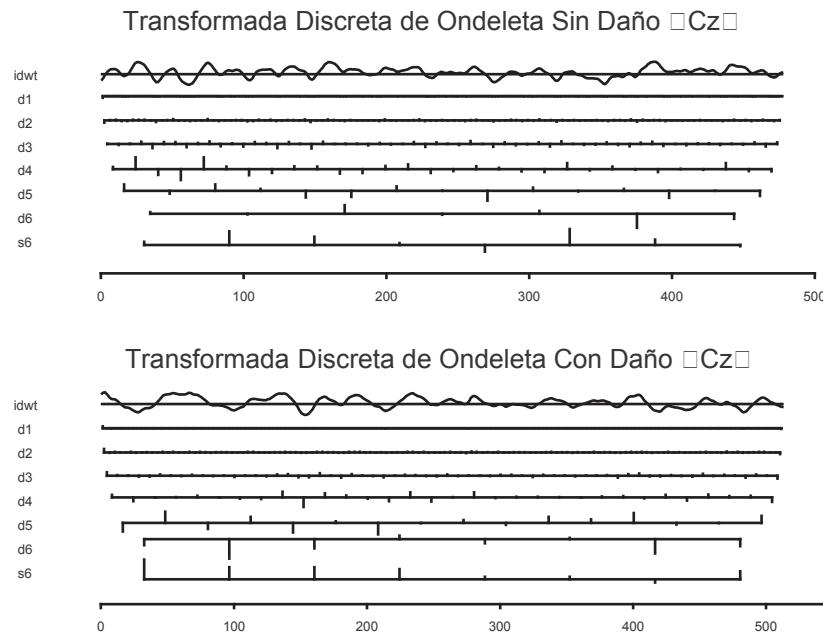


Figura 3: Transformada discreta de ondeleta del EEG registrado por Cz, para el niño sin daño y con daño cerebral

Al realizar un análisis de la energía capturada por los coeficientes de cada EEG, la cual nos proporciona la varianza de la señal, se puede notar que el EEG de la persona con daño queda mejor representada con el coeficiente de detalle d5, mientras que la persona sin daño, por el coeficiente de detalle d4; es decir, la ondeleta que mejor representa la actividad cerebral de los individuos con daño, es más rugosa que la ondeleta utilizada para representar la actividad cerebral de los individuos sin daño cerebral.

En la Figura 4, se puede observar el comportamiento de las ondeletas en los dos niveles de detalle. La ondeleta en el coeficiente de detalle d4 es más suave y más amplia, mientras que en el detalle d5, se vuelve más rugosa y más corta.

4. Conclusiones

El método utilizado para calcular la Potencia en las bandas de frecuencia definidas previamente presenta varias deficiencias en su aplicación. Por un lado, supone que la serie es

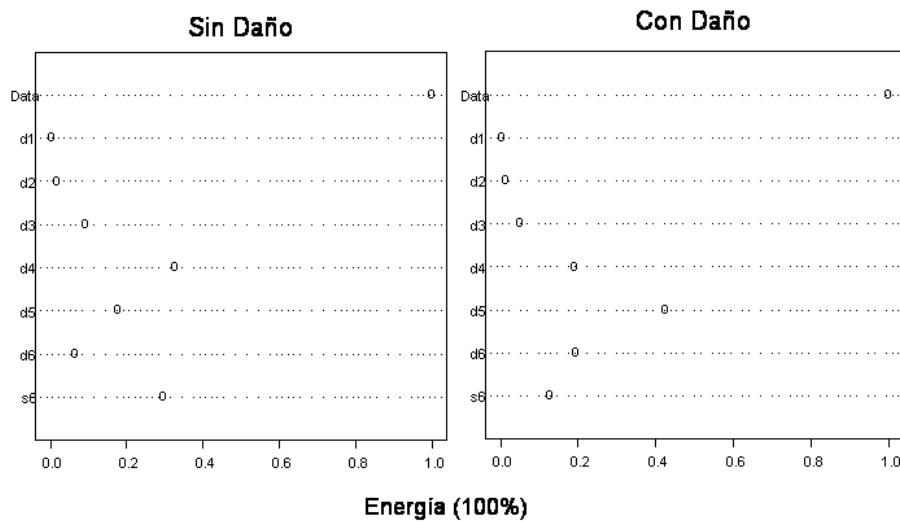


Figura 4: Porcentaje de energía de los coeficientes de ondeleta en “Cz”

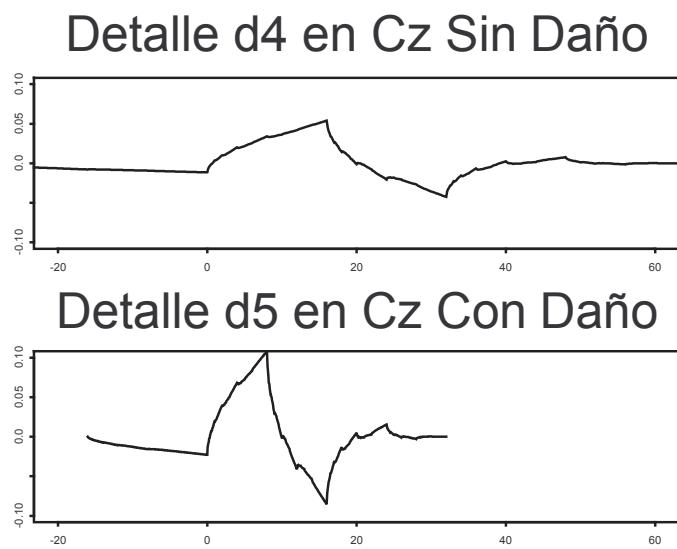


Figura 5: Ondeleta representativa de los detalles d4 y d5 del EEG registrado por Cz, para el niño sin daño y con daño cerebral

estacionaria, cuando en realidad existen segmentos en donde dicho supuesto no se cumple, como el que se mostró en la Figura 2. En la actualidad cualquier actividad cerebral que no se comporte de manera estacionaria es eliminada del análisis, lo que implica una pérdida de información que podría ser importante para la valoración de un sujeto.

La transformada de ondeleta provee mayor información y es posible utilizarlas con series no estacionarias. Además el tiempo no se pierde al realizar el Análisis Espectral y además es posible reconstruir la señal con un menor número de datos.

Cabe mencionar que diversos autores han utilizado las ondeletas en datos donde la no estacionariedad es evidente, como es el caso de pacientes epilépticos. Lo que se pretende es encontrar una manera nueva de establecer normas para México con base en la transformada de Ondeleta y así discriminar de una manera correcta a los sujetos.

Este estudio fue realizado sólo con dos sujetos, en un solo segmento y una sola derivación, por lo que se pretende ampliarlo a un grupo, analizar las derivaciones más importantes, con el mayor número de segmentos posibles libres de artefacto.

Agradecimientos

Nuestro más preciado agradecimiento a la Universidad Nacional Autónoma de México, por proporcionarnos los EEGs, así como a la Dra. Lourdes Díaz por la creación del programa que permite tener los registros en formato texto.

Referencias

Bruce, A. y Gao, H.Y. (1996). Applied Wavelet Analysis with S-Plus, pp. 11-62. New York: Springer.

Evans, J. R. (1999). Introduction to Quantitative EEG and Neurofeedback, pp. 3-23, USA: Academic Press.

Kavale, K. A. (1988). Learning Disability and Cultural-Economic Disadvantage: The Case for a Relationship. Learning Disability Quarterly. XI, 3, 195-210

- Saavedra-Gastélum, V., Fernández Harmony, T., Harmony-Baillet, T. y Castaño Meneses, V. M. (2006). Ondeletas en Ingeniería, Principios y Aplicaciones. Ingeniería Investigación y Tecnología, VII, 3, 185-190.
- Shumway, R. H. y Stoffer, D. S. (2000). Time series analysis and its applications, pp. 213-289. New York: Springer.
- Thatcher, R.W., (1998). Normative EEG Databases and EEG Biofeedback. Journal of Neurotherapy, II, 4, 8 – 39.
- Walker, J. S., (1999). A primer on wavelets and their scientific applications, pp. 2-49.USA: Chapman & Hall.

Software que trata las principales causas de la diabetes

Bárbara Emma Sánchez Rinza^a, Jessica Giovanna Huerta López*, Jazmin Jiménez Bedolla*, M. Bustillo Díaz, A. Rangel Huerta

Universidad Autónoma de Puebla – Facultad de Ciencias de la Computación

1. Introducción

La diabetes junto con enfermedades del corazón y cáncer son las tres principales causas de muerte en México, estas también forman parte de los problemas graves de índole pública, junto con la obesidad. Actualmente en nuestro país la enfermedad de la Diabetes es un asunto preocupante, ya que hasta hace 31 años las principales causas de muerte en nuestro país era la diarrea, neumonía, e infecciones respiratorias agudas. La diabetes ocupaba el séptimo lugar, mientras que ahora la diabetes mellitus es una de las principales causas de muerte en nuestro país. Para ser más exacto hasta la fecha ocupa el tercer lugar de mortalidad en nuestro país. México se encuentra ubicado a nivel mundial como uno de los países con el mayor número registrado de casos de diabetes. Y se espera un incremento a futuro de este número de casos de personas diabéticas. De acuerdo con organismos mundiales de la salud en 1995, México ocupaba el décimo lugar en casos de diabetes, pero se espera que para el año 2025, ocupará el séptimo con 10 millones de personas diabéticas. Esto se podría afirmar con el siguiente dato, en México por cada año mueren aproximadamente 40 000 personas a causa de la diabetes. En México, 11 % de la población entre 20 y 69 años padece diabetes, la cual en la última década se ha ubicado como la enfermedad crónico-degenerativa con mayor carga de mortalidad y discapacidad entre quienes la padecen esta enfermedad. Ahora se describirá

^abrinza@cs.buap.mx

*Estudiante de la BUAP

un poco acerca de la Diabetes Mellitus que es una enfermedad crónico-degenerativa esta surge por la falta de insulina, una hormona cuya función principal es permitir la entrada de glucosa a las células, así por falta de esta hormona es que se genera un incremento en los niveles de azúcar en la sangre (glucosa sanguínea), condición denominada Hiperglicemia (la Hiperglicemia sucede cuando el azúcar en la sangre alcanza un nivel de 180 mg/dl o más).

2. Desarrollo del trabajo

A continuación se mostrarán pantallas de una base de datos que fue diseñada para mostrar encuestas hechas a personas diabéticas, la mayoría de edad avanzada. En la imagen anterior

Tabla	Acción	Registros	Tipo	Cotejamiento	Tamaño
datos_generales		51	MyISAM	latin1_swedish_ci	3.3 KB
factores_herencia		50	MyISAM	latin1_swedish_ci	2.4 KB
factores_ingestion		51	MyISAM	latin1_swedish_ci	3.3 KB
factores_sintomaticos		51	MyISAM	latin1_swedish_ci	2.6 KB
factores_stress		51	MyISAM	latin1_swedish_ci	2.5 KB
seccion_mujer		30	MyISAM	latin1_swedish_ci	2.4 KB
tipo_personalidad		51	MyISAM	latin1_swedish_ci	3.0 KB
7 tabla(s)	Número de filas	335	InnoDB	latin1_swedish_ci	19.6 KB

Figura 1: Muestra las tablas de la encuesta

se muestran las 7 tablas que abarca nuestra tabla de datos. En datos generales se almacenan los datos personales de la población encuestada, como se muestra en la Figura 2. Como se mencionó anteriormente en esta tabla se almacenan los datos generales de la persona, como lo son su sexo, edad, estatura, y peso. Cada persona va a tener un id, este id servirá en las siguientes tablas para saber qué es lo que contestó la persona en cada pregunta. El software tiene una base de datos de personas enfermas, de esa base de datos se sacan algunas preguntas claves que fueron elaboradas por médicos especialistas en el cuidado de enfermos de diabetes. Una vez que el sistema tiene las preguntas cualquier persona puede resolver este cuestionario y el software le dará un porcentaje de padecer la enfermedad o no, pero es importante recalcar que la última palabra la tiene un médico y unos análisis hechos en un laboratorio clínico.

←↑→	id_datosGrales	sexo	peso	estatura	edad
█	1	femenino	76.5	1.5	46
█	2	femenino	74	1.49	55
█	3	femenino	62	1.48	26
█	4	femenino	64	1.54	48
█	5	femenino	53	1.53	72
█	6	femenino	57	1.63	54
█	7	femenino	79	1.56	69
█	8	femenino	58	1.57	59
█	9	femenino	63	1.42	48
█	10	femenino	63	1.48	56
█	11	femenino	68	1.5	42
█	12	femenino	66	1.58	70
█	13	femenino	57	1.5	79
█	14	femenino	72	1.5	40
█	15	femenino	53	1.47	55
█	16	femenino	66	1.55	75
█	17	femenino	55	1.48	76
█	18	femenino	58	1.49	70

Figura 2: Datos generales de las personas encuestadas

3. Secciones de la encuesta

- Tipo de personalidad
- Factores de stress
- Factores de ingestión
- Factores hereditarios
- Sección para contestar únicamente si es mujer

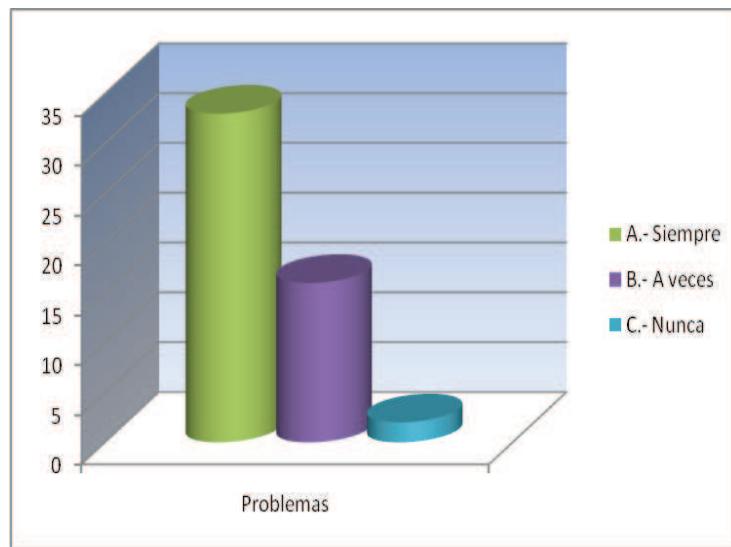
Para cada una de estas secciones se creó una tabla, esta tabla va a contener todas las preguntas hechas por sección. Por ejemplo:

En la Figura 3 se muestra que con el Id de la persona va a contestar las preguntas de la sección de tipo de personalidad, dentro de esta enfermedad es muy importante el tipo de personalidad ya que puede existir que varios hermanos pueden tener la misma tendencia genética y alimenticia, pero solo algunos la padecen por el tipo de personalidad que tienen. Como podemos apreciar en la grafica las personas A. Son las que tienden a aislarce cuando

	Campo	Tipo	Cotejamiento	Atributos	Nulo	Predeterminado	Extra	Acción
■	<u>id_personalidad</u>	int(3)			No		auto_increment	        
■	<u>problemas</u>	varchar(3)	latin1_swedish_ci		No			        
■	<u>dificultades</u>	varchar(3)	latin1_swedish_ci		No			        

Figura 3: Gráfica del consumo de carne en su dieta

tienen problemas, el caso B es a veces se aíslan cuando tienen problemas y el caso C es nunca. Y predomina más en los diabéticos el aislamiento.

**Figura 4:** Gráfica del tipo de personalidad

4. Factores de stress

Otro de los factores que se analiza es si las personas con diabetes han estado sometidas a factores de stress por periodos prolongados y podemos observar lo siguiente: A es durante un largo periodo de su vida, B durante un tiempo medio de su vida y C nunca, donde podemos apreciar que las personas con diabetes han estado expuestas a factores de stress por largo tiempo

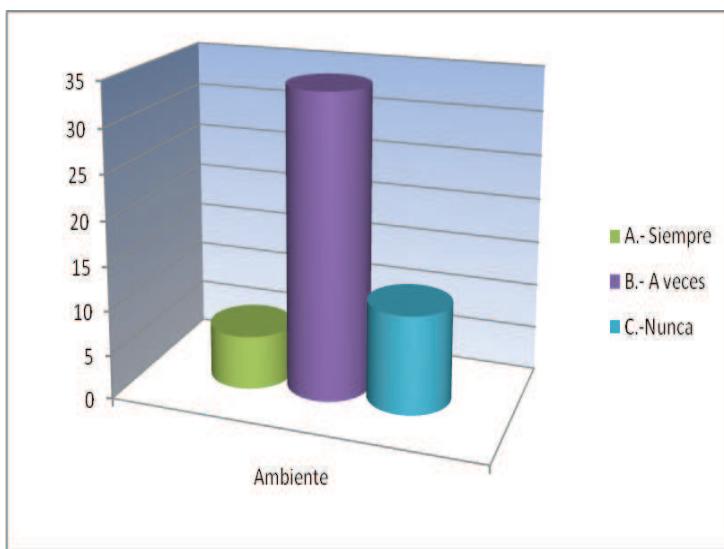


Figura 5: Gráfica de factores de stress

5. Factores de ingestión

Otro factor que se analiza es si son comedores compulsivos cuando tienen problemas, y se encuentra que A) si, B) no, en los enfermos de diabetes es que si son comedores compulsivos. Otro factor fue el consumo de carne en su alimentación A) mucho B) frecuentemente C) nunca, con mayor índice fue la B) regularmente. El Consumo de Alcohol entre los diabéticos es el siguiente: dio como resultado que antes de padecer la enfermedad la mayoría de los enfermos consumían poco alcohol. El sedentarismo fue otro de los factores que se analizó y se concluyó que las personas realizaron poco ejercicio físico durante su vida.

6. Factores de herencia

Los factores de herencia son muy importantes en este tipo de enfermedades. Se pudo observar en la gráfica que la enfermedad se transmite de padres a hijos, y también que existe un gran número de personas que no tuvieron factores de herencia lo cual indica que por otros factores se está incrementando el número de enfermos.

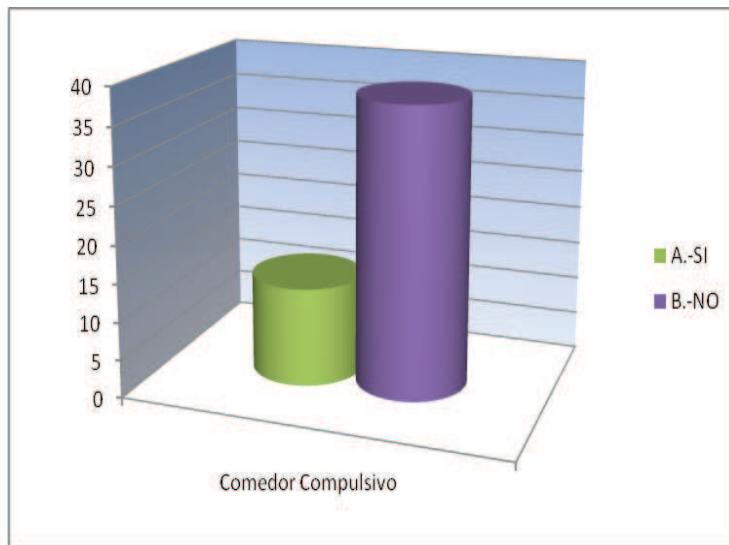


Figura 6: Si los enfermos son comedores compulsivos

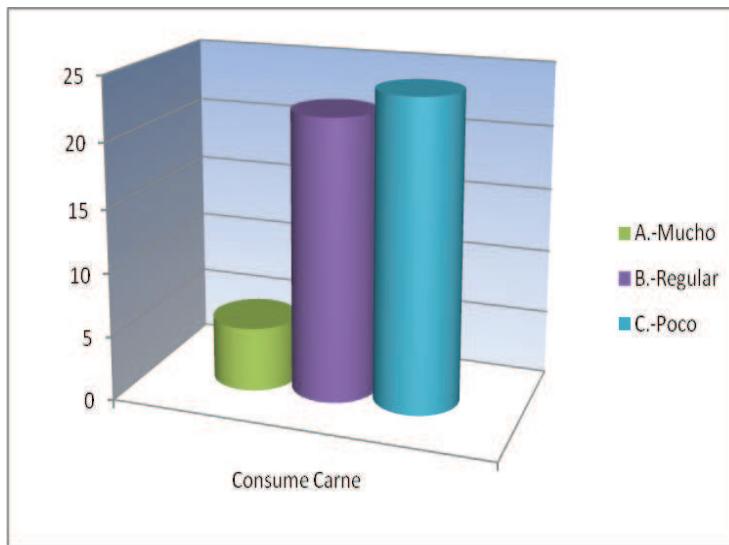


Figura 7: Gráfica del consumo de carne en su dieta

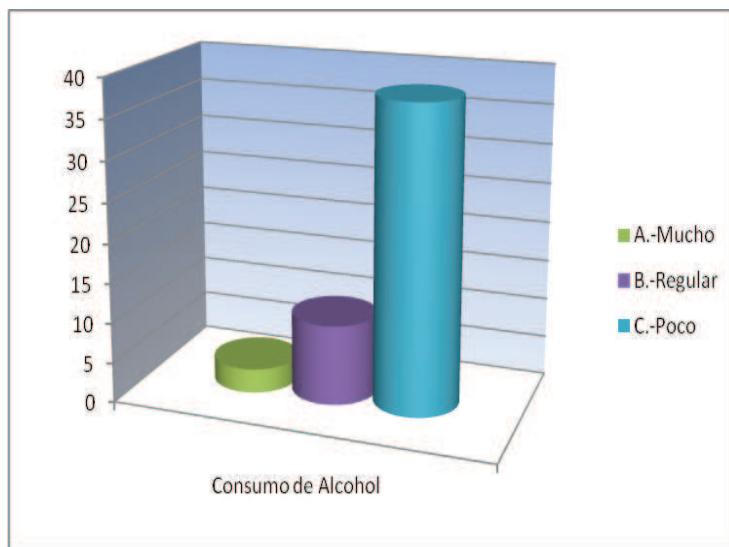


Figura 8: Gráfica del consumo de alcohol de los encuestados antes de padecer la enfermedad

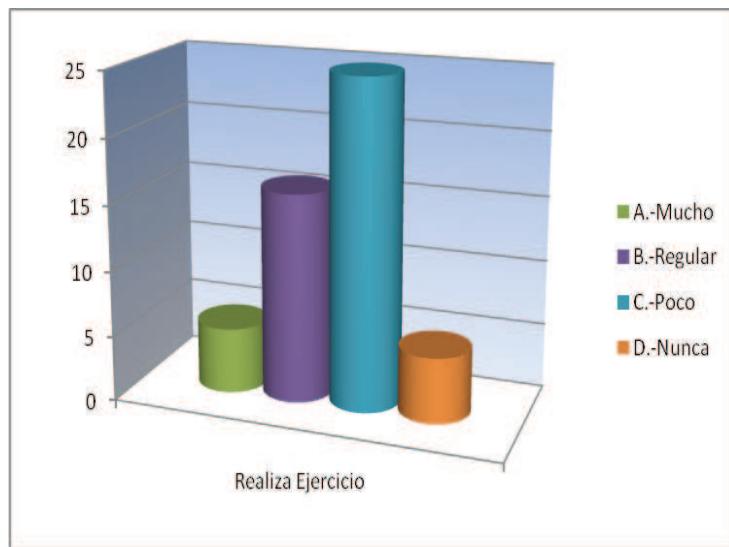


Figura 9: Gráfica de la actividad física que realizaban los enfermos de diabetes antes de padecer la enfermedad

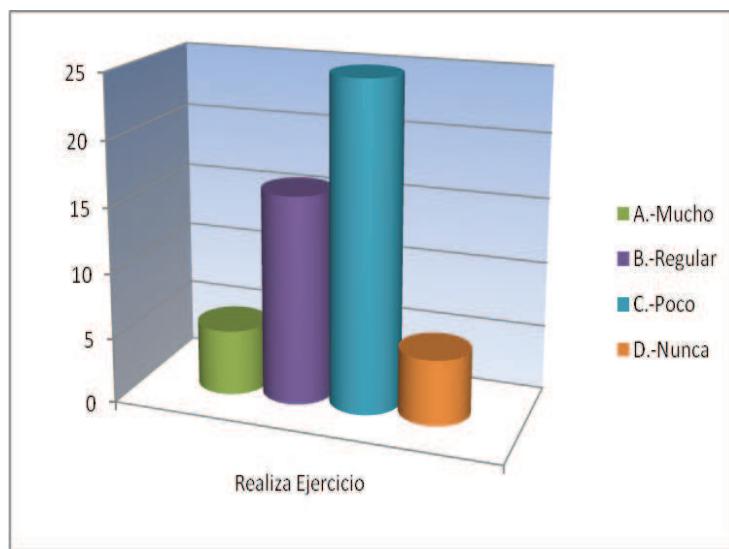


Figura 10: Gráfica de factores hereditarios relacionados con el parentesco

7. Conclusiones

Podemos apreciar mediante este estudio los principales factores de riesgos de los diabéticos y darles algunos tips en línea, e.g. cuidados con una dieta sana y apta para ellos, ejercicio físico adecuado, formas de manejo del stress, entre otras.

Referencias

Carretero Pérez, Jesús, García Carballeira, Félix, Pérez Lobato, José M., Pérez Menor, José M. Problemas Resueltos de Programación en Lenguaje Java. Paraninfo.

Ceballos Atienza, Rafael. Novedades En Diabetes: Atención Integral Y Tratamiento. Formación Alcalá, S.L.

Diabetes De La A A La Z. Todo Lo Que Necesita Saber Acerca De La Diabetes, Explicado Con Claridad Y SencillezAmerican Diabetes Association. Paidos.

Eckel, Bruce THINKING IN JAVA. Prentice Hall.

Milton, J. Susan y Arnold, Jesse C. Probabilidad Y Estadística Con Aplicaciones Para Ingeniería Y Ciencias Computacionales. Editorial Mcgraw-Hill.

Walker, Rose M. Diabetes. H. Blume.

Comparación de algunas pruebas estadísticas asintóticas de no-inferioridad para dos proporciones independientes

David Sotres Ramos^a *Colegio de Postgraduados*

Félix Almendra Arao^b *UPIITA del Instituto Politécnico Nacional*

1. Introducción

Las pruebas estadísticas asintóticas de no-inferioridad se utilizan muy frecuentemente en ensayos clínicos. Estas pruebas sirven para demostrar que una terapia nueva (con mínimos efectos secundarios o bajo costo) no es sustancialmente inferior en eficacia a la terapia estándar, ver Farrington-Manning (1990). El principal objetivo de este trabajo es comparar las pruebas asintóticas para no-inferioridad para dos proporciones independientes de Blackwelder, Farrington-Manning, Böhning-Viwatwongkasen, Hauck-Anderson, la prueba de razón de verosimilitudes y dos variantes de estas pruebas con base en sus niveles de significancia y en sus potencias reales y para los tamaños de muestra $25 \leq n \leq 100$.

2. Pruebas estadísticas consideradas

Sean X_1 y X_2 dos variables aleatorias independientes con distribución binomial y con parámetros (n_1, p_1) y (n_2, p_2) respectivamente, donde p_1 y p_2 representan las probabilidades de respuesta de los tratamientos estándar y nuevo, respectivamente. La hipótesis de interés (hipótesis de no-inferioridad) a ser probada es la alternativa (H_a) en el siguiente juego de hipótesis:

$$[H_0 : p_1 - p_2 \geq d_0] \quad \text{Vs.} \quad [H_a : p_1 - p_2 < d_0] \quad (1)$$

^asotres.davida@kendle.com

^bfalmendra@ipn.mx

donde d_0 es el llamado límite de no-inferioridad, el cual es una constante positiva y conocida. En el contexto de ensayos clínicos los valores usuales para d_0 son 0.10, 0.15 y 0.20.

Seis de las estadísticas de prueba consideradas son del tipo

$$T(X_1, X_2) = \frac{\hat{p}_1 - \hat{p}_2 - d_0}{\hat{\sigma}} \quad (2)$$

donde $X_1 = \sum X_{1j}$, $X_2 = \sum X_{2j}$, y $\hat{p}_i = \frac{X_i}{n_i}$ es el estimador de máxima verosimilitud de p_i para $i = 1, 2$ y $\hat{\sigma}$ es un estimador consistente de la desviación estándar de $\hat{d} = \hat{p}_1 - \hat{p}_2$; la séptima estadística es aquella correspondiente a la prueba de razón de verosimilitudes,

$$\lambda(X_1, X_2) = \frac{\sup_{\Theta_0} L(d|(X_1, X_2))}{\sup_{\Theta} L(d|(X_1, X_2))} \quad (3)$$

La diferencia entre las seis estadísticas del tipo (2) radica en la estimación que se elige para la desviación estándar de \hat{d} . Se consideran los siguientes seis estimadores

$$\begin{aligned} \hat{\sigma}_1 &= \left(\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2} \right)^{1/2}, & \hat{\sigma}_4 &= \left(\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1 - 1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2 - 1} \right)^{1/2}, \\ \hat{\sigma}_2 &= \left(\frac{\check{p}_1(1 - \check{p}_1)}{n_1} + \frac{\check{p}_2(1 - \check{p}_2)}{n_2} \right)^{1/2}, & \hat{\sigma}_5 &= \left(\frac{\check{p}_1(1 - \check{p}_1)}{n_1 - 1} + \frac{\check{p}_2(1 - \check{p}_2)}{n_2 - 1} \right)^{1/2}, \\ \hat{\sigma}_3 &= \left(\tilde{p}_1(1 - \tilde{p}_1) + \tilde{p}_2(1 - \tilde{p}_2) \right)^{1/2}, & \hat{\sigma}_6 &= \left(\frac{\tilde{p}_1(1 - \tilde{p}_1)}{n_1 - 1} + \frac{\tilde{p}_2(1 - \tilde{p}_2)}{n_2 - 1} \right)^{1/2}, \end{aligned}$$

donde $\hat{p}_i = \frac{X_i}{n_i}$ es el estimador de máxima verosimilitud de p_i , \check{p}_i es el estimador de máxima verosimilitud restringida bajo la hipótesis nula de p_i , ver Farrington y Manning(1990); además, $\tilde{p}_i = \frac{X_i+1}{n_i+2}$, ver Böhning y Viwatwongkasen(2005).

La estadística T en (2) tiene distribución asintótica normal estándar para cualquier estimador consistente $\hat{\sigma}$ de la desviación estándar de \hat{d} ; para la estadística λ , la distribución asintótica de $-2 \ln \lambda$ es $\frac{1}{2} + \frac{1}{2} F_{\chi_1^2}$ donde $F_{\chi_1^2}$ denota la función de distribución acumulada de una variable aleatoria ji-cuadrada con un grado de libertad. Las pruebas asintóticas correspondientes a las estadísticas del tipo (2), para un nivel de significancia nominal α , tienen región de rechazo de la forma

$$R_T(\alpha) = \{(x_1, x_2) \in \{0, \dots, n_1\} \times \{0, \dots, n_2\} : T(x_1, x_2) < -z_\alpha\}$$

donde z_α es el percentil superior α de la distribución normal estándar, es decir, $\Phi(z_\alpha) = 1 - \alpha$, donde Φ es la función de distribución acumulativa de una variable aleatoria normal estándar.

La región de rechazo para la prueba asintótica correspondiente a la estadística (3) es

$$R_T(\alpha) = \{(x_1, x_2) : -2 \ln \lambda(x_1, x_2) > \chi^2_{1-2\alpha}(1)\};$$

donde $\chi^2_{1-2\alpha}(1)$ es el percentil superior $1 - 2\alpha$ de la distribución ji-cuadrada con un grado de libertad, en otras palabras $P(\chi^2_1 > \chi^2_{1-2\alpha}(1)) = 1 - 2\alpha$.

Las correcciones por continuidad que se consideran aquí son $C_0 = 0$, $C_1 = \frac{1}{4 \min(n_1, n_2)}$, $C_2 = 2C_1$, $C_3 = \frac{1}{2n_1} + \frac{1}{2n_2}$, $C_4 = \frac{3}{2 \min(n_1, n_2)}$, $C_5 = \frac{2}{\min(n_1, n_2)}$. C_0, C_2 , y C_3 , son consideradas por Hauck y Anderson(1986) para el caso de las estadísticas T_1 y T_4 . El análisis se realizó para $n_1 = n_2 = n$. Así las estadísticas de prueba consideradas son para $i = 1, 2, 3, 4, 5, 6$ y $j = 0, 1, 2, 3, 4, 5$,

$$T_{iCj}(X_1, X_2) = \frac{\hat{p}_1 - \hat{p}_2 - d_0 + C_j}{\hat{\sigma}_i} \quad (4)$$

$$T_{7Cj}(X_1, X_2) = \lambda(X_1, X_2) + C_j = \frac{\sup_{\Theta_0} L(d | (X_1, X_2))}{\sup_{\Theta} L(d | (X_1, X_2))} + C_j \quad (5)$$

Las pruebas T_{iCj} fueron propuestas en los siguientes artículos: T_{1C_0} en Blackwelder(1982), T_{2C_0} en Farrington y Manning(1990), T_{3C_0} en Böhning y Viwatwongkasen(2005), T_{4C_0} en Hauck y Anderson(1986). T_{5C_0} es una combinación de T_{2C_0} y T_{4C_0} , mientras que T_{6C_0} es una combinación de T_{3C_0} y T_{4C_0} . Finalmente T_{7C_0} es la conocida estadística de razón de verosimilitudes, ver Casella y Berger(2002). El nivel de significancia nominal usado en todo este trabajo fue $\alpha = 0.05$. Las pruebas estadísticas serán simbolizadas de la misma forma que sus correspondientes estadísticas de prueba.

3. Cálculo del nivel de significancia real

En virtud de que X_i tiene distribución Binomial con parámetros (n_i, p_i) para $i = 1, 2$, se tiene que la función de verosimilitud conjunta es

$$L(p_1, p_2; x_1, x_2) = \binom{n_1}{x_1} p_1^{x_1} (1 - p_1)^{n_1 - x_1} \binom{n_2}{x_2} p_2^{x_2} (1 - p_2)^{n_2 - x_2}$$

y la función de potencia es

$$\beta_T(p_1, p_2) = \sum_{(x_1, x_2) \in R_T(\alpha)} L(p_1, p_2; x_1, x_2),$$

además, el espacio nulo es

$$\Theta_0 = \{(p_1, p_2) \in \Theta : p_1 - p_2 \geq d_0\}$$

y el nivel de significancia queda dado por

$$\sup_{(p_1, p_2) \in \Theta_0} \beta_T(p_1, p_2) \quad \text{y con} \quad \Theta = \{(p_1, p_2) : (p_1, p_2) \in [0, 1] \times [0, 1]\}.$$

Chan(1998) calculó el nivel de significancia para la prueba de Farrington-Manning (T_{2C_0}) tomando el supremo no en todo el espacio nulo (Θ_0) sino calculando el máximo únicamente en $\Theta_0^* = \{(p_1, p_2) \in \Theta : p_1 - p_2 = d_0\}$, el cual es solamente una parte de la frontera del espacio nulo. Computacionalmente ésto representa una inmensa ventaja, pues el tiempo de cómputo se reduce aproximadamente al 0.22 % del tiempo original. Sin embargo, el autor mencionado no justificó formalmente la validez de este argumento. Fue hasta 2005 cuando Röhmel(2005) presenta una prueba formal que justifica el procedimiento utilizado por Chan(1998). En este trabajo se siguió la misma estrategia de Chan(1998), para lo cual en lo que resta de esta sección se verifica la validez de la llamada condición de convexidad de Barnard y de la condición de simetría en la misma cola (ver definiciones abajo) para todas las pruebas.

Definición 3.1. *Se dice que una prueba estadística, para el problema en (1), con región de rechazo R_T cumple la condición de convexidad de Barnard (C) si satisface las dos propiedades:*

- a) $(x, y) \in R_T \implies (x - 1, y) \in R_T \quad \forall \quad 1 \leq x \leq n_1, 0 \leq y \leq n_2$
- b) $(x, y) \in R_T \implies (x, y + 1) \in R_T \quad \forall \quad 0 \leq x \leq n_1, 0 \leq y \leq n_2 - 1$

Definición 3.2. *Si $n_1 = n_2 = n$, se dice que una región de rechazo R , para el problema en (1), cumple la condición de simetría en la misma cola si $(x, y) \in R \implies (n - y, n - x) \in R$.*

Proposición 3.1. *Sean $n_1 = n_2 = n$ y $R(\alpha)$ una región crítica para el problema de prueba de hipótesis en (1), si $R(\alpha)$ cumple la condición de convexidad de Barnard y la condición de simetría en la misma cola, entonces el nivel de significancia exacto de la prueba $R(\alpha)$ está dado por*

$$\max_{\substack{p_2=p_1-d_0 \\ p_1 \in [d_0, \frac{1-d_0}{2}]}} \sum_{x_1=0}^{n_1} \sum_{x_2=0}^{n_2} \binom{n_1}{x_1} p_1^{x_1} (1-p_1)^{n_1-x_1} \binom{n_2}{x_2} p_2^{x_2} (1-p_2)^{n_2-x_2} I_{[(x_1, x_2) \in R(\alpha)]} \quad (6)$$

d₀	Prueba	Porcentaje	d₀	Prueba	Porcentaje
0.10	T _{2C1}	97.37	0.15	T _{2C1}	97.53
	T _{5C1}	93.42		T _{5C1}	91.36
	T _{7C4}	93.42	0.20	T _{2C1}	95.06
				T _{5C1}	95.06

Tabla 1: Porcentaje de niveles de significancia reales que caen dentro del intervalo [.045, .055], en base a los 76 tamaños de muestra en el rango $25 \leq n \leq 100$.

Demostración. *Se omite por razones de espacio.*

Proposición 3.2. *Todas las pruebas asintóticas T_{iCj} , $i = 1, 2, \dots, 7$; $j = 0, 1, \dots, 5$, con estadísticas de prueba definidas en (4) y (5) satisfacen la condición de convexidad de Barnard y la condición de simetría en la misma cola.*

Demostración. *Se omite por razones de espacio.*

Con base en las proposiciones 3.1 y 3.2, el cálculo del nivel de significancia de las pruebas consideradas se hizo aplicando la fórmula en (6) y particionando el intervalo $[d_o, (1 - d_0)/2]$ en subintervalos de longitud 0.001. Esto quiere decir que aproximamos el nivel de significancia exacto reemplazando en la fórmula (6) el intervalo continuo $[d_o, (1 - d_0)/2]$ por el conjunto de valores finito $p_1 \in [d_o](0.001)[(1 - d_0)/2]$, y a esta aproximación la llamamos el nivel de significancia real y la denotamos por α_R .

4. Resultados y conclusiones

Para c/u de las 42 pruebas estadísticas consideradas en este trabajo (T_{iCj} con $1 \leq i \leq 7$ y $0 \leq j \leq 5$) se evaluó la aproximación de su correspondiente nivel de significancia real (α_R) al nivel de significancia nominal (α), calculando para los 76 tamaños de muestra en el rango $25 \leq n \leq 100$, el porcentaje de niveles de significancia reales de la prueba que caen dentro del intervalo [0.045, 0.055]. Las pruebas con los mejores porcentajes se reportan en el Cuadro 1.

Adicionalmente se compararon las potencias de las pruebas del Cuadro 1 para aquellos tamaños de muestra donde la máxima diferencia entre los niveles de significancia reales re-

sultó menor o igual que 0.0001 y donde al menos una de las potencias a comparar fuera mayor o igual que 0.7. En más del 94 % de los puntos evaluados, las potencias de las pruebas resultaron idénticas y en aquellos puntos dónde las potencias resultaron distintas, las diferencias observadas fueron de centésimas. Conjugando este resultado sobre las potencias con los resultados sobre los niveles de significancia reales reportados en el Cuadro 1, se puede afirmar que la prueba de Farrington-Manning con el factor de corrección $C_1(T_{2C1})$ es la mejor de las pruebas.

Referencias

- Blackwelder, W. (1982). Proving the null hypothesis in clinical trials. *Controlled Clinical Trials*, **3**, 345-353
- Böhning, D. y Viwatwongkasen, C. (2005). Revisiting proportion estimators. *Statistical methods in medical research*, **14**, 1-23.
- Casella, G., y Berger, L. (2002). *Statistical Inference*. Duxbury, USA.
- Chan, I. (1998). Exact tests of equivalence and efficacy with a non zero lower bound for comparative studies. *Statistics in Medicine*, **17**, 1403-1413.
- Farrington, C. y Manning,G. (1990).Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk. *Statistics in Medicine*, **9**, 1447-1454.
- Hauck, W., y Anderson,S. (1986). A comparison of large-sample confidence interval methods for the difference of two binomial probabilities. *The American Statistician*, **40**, 318-322.
- Röhmel, J. (2005).Problems with existing procedures to calculate exact unconditional p-values for noninferiority and confidence intervals for two binomials and how to resolve them. *Biometrical Journal*, **47**, 37-47.

Procedimientos para analizar los datos no detectados en contaminación ambiental

Fidel Ulín Montejo^a *Matemáticas, Div. Acad. de Ciencias Básicas, Universidad Juárez Autónoma de Tabasco*
Humberto Vaquera Huerta *Estadística, Campus Montecillo, Colegio de Postgraduados*

1. Introducción

Los datos no detectados son pequeñas señales producidas por los contaminantes, que no pueden ser cuantificadas por los instrumentos. Tales observaciones son datos censurados por la izquierda, siendo el límite de detección (LD) el punto de referencia, (Helsel, 2005). En algunos estudios (Gilbert, 1987; Millard y Deverel, 1988; EPA, 1992), se comparan las poblaciones mediante sus medianas, donde los datos no-detectados son omitidos o substituidos por el LD; sin embargo, con esta técnica, los alcances y resultados son limitados. Los organismos de regulación ambiental requieren que los riesgos sean caracterizados en términos de la concentración media (El-Shaarawi y Viveros, 1997). En este sentido, este trabajo aborda el problema de comparación de concentraciones medias de poblaciones lognormales desde un enfoque paramétrico, reparametrizando el modelo lognormal con variables indicadoras. El criterio de comparación se establece mediante regiones de confianza aproximada. El algoritmo EM (Flury y Zoppé, 2001), la verosimilitud y el método de Wald (Meeker y Escobar, 1998) son empleados para la inferencia necesaria. El procedimiento fue implementado en R (2006) y se desarrolló un ejemplo con datos ambientales. El método resultó eficiente para dos poblaciones, sin embargo puede extenderse a más de dos poblaciones con covariables.

^afidel.ulin@basicas.ujat.mx

2. Metodología

El método de máxima verosimilitud (MV) provee herramientas versátiles para ajustar modelos y pueden ser aplicados a modelos paramétricos con datos censurados, el ajuste considera combinación de parámetros y de modelos para las cuales la probabilidad sea alta.

2.1. Función de verosimilitud

La función de verosimilitud es proporcional a la probabilidad conjunta de los datos. Para un conjunto de datos y y un modelo especificado $F(y;\theta)$, la verosimilitud es vista como una función de los parámetros desconocidos θ . Para una muestra censurada por la izquierda, de n observaciones independientes, la verosimilitud muestral se define como

$$L(\theta) = C \prod_{i=1}^n L_i(\theta; y_i) = C \prod_{i=1}^n [f(y_i; \theta)]^{\delta_i} [F(y_i; \theta)]^{1-\delta_i} \quad (1)$$

donde $L_i(\theta; y_i)$ es la probabilidad de los datos y_i para la i -ésima observación , $\delta_i = 1$ para una observación detectada y $\delta_i = 0$ para una no-detectada; C es una constante independiente de θ . El valor de θ que maximiza $L(\theta)$ provee un estimador de MV y se denota por $\hat{\theta}$.

2.2. El algoritmo EM

El algoritmo EM , es una herramienta poderosa para calcular estimadores de MV con datos incompletos (faltantes, censurados, etc.). Sean y los datos observados y x los datos desconocidos, θ el parámetro de interés y $\ell_c(\theta; y, x)$ la log-verosimilitud de los datos completos. Iniciando con $\theta^{(0)}$ el algoritmo EM repite los siguientes dos pasos hasta la convergencia.

- E: calcular $\ell^{(j)}(\theta) = E_{x|y, \theta^{(j-1)}} [\ell_c(\theta; Y, X)]$.
- M: encontrar $\theta^{(j)}$ que maximize $\ell^{(j)}(\theta)$.

2.3. Matriz de información vía el algoritmo EM

El algoritmo EM no genera estimadores para la matriz de covarianzas de los EMVs, por lo que se ha modificado para resolver este problema. Una modificación simple y muy útil fue

hecha por (Oakes, 1999), quien logró demostrar que, si $\log L(y; \theta)$ es la log-verosimilitud de la muestra, entonces la varianza aproximada de $\hat{\theta}$ se calcula con

$$Var(\hat{\theta}) \approx \left(\frac{\partial^2 \ell^{(j)}(\theta)}{\partial \theta^{(j-1)2}} + \frac{\partial^2 \ell^{(j)}(\theta)}{\partial \theta^{(j-1)} \partial \theta} \right)^{-1} \quad (2)$$

Con lo que es posible obtener regiones e intervalos de confianza aproximados.

2.4. Regiones e intervalos de confianza aproximados

La aproximación normal para la distribución de estimadores de MV puede ser usada para obtener regiones de confianza aproximadas para θ . Esto se conoce como el Método de Wald o metodo de aproximación normal. En particular, una región aproximada del $100(1-\alpha)\%$ de confianza para θ es el conjunto de los valores de θ en el elipsoide

$$W(\theta) = (\hat{\theta} - \theta)'(\hat{\Sigma}_{\hat{\theta}})^{-1}(\hat{\theta} - \theta) \leq \chi_{1-\alpha,r}^2, \quad (3)$$

donde r es la dimensión de $\theta = (\theta_1, \theta_2, \dots, \theta_r)$, $\hat{\Sigma}_{\hat{\theta}}$ es la matriz de varianzas y covarianzas, estimada por (2), $\chi_{(1-\alpha;r)}^2$ es el $1-\alpha$ cuantil de la variable aleatoria χ_r^2 . Entonces, un intervalo de confianza de aproximación-normal para θ_i es obtenido de la formula familiar

$$[\underline{\theta}_i, \bar{\theta}_i] = \hat{\theta}_i \pm z_{(1-\alpha/2)} \hat{s}e_{\hat{\theta}_i} \quad (4)$$

donde $\hat{s}e_{\hat{\theta}_i}$ es la raíz cuadrada de la ii -ésima entrada en (2), $z_{(1-\alpha/2)}$ es el $1 - \alpha/2$ cuantil de la distribución normal estándar. Este intervalo puede verse como una aproximación para la logverosimilitud marginal de θ_i en $\hat{\theta}_i$.

3. Procedimiento de comparación de medias

Empleando los métodos y la teoria descritos anteriormente, se desarrolló el procedimiento de comparacion para muestras de datos no detectados de poblaciones lognormales.

En estudios ambientales se ha reportado que las concentraciones de contaminantes en aire y suelo, y de metales en ríos, tienen distribución lognormal (Gilbert, 1987; Ott, 1995). Si y es una variable aleatoria lognormal, tiene función de densidad de probabilidad,

$$f(y; \mu, \sigma) = \frac{1}{y\sigma\sqrt{2\pi}} e^{-\frac{[\log(y)-\mu]^2}{2\sigma^2}}, \quad 0 < y < \infty, -\infty < \mu < \infty, \sigma > 0. \quad (5)$$

La mediana de Y , e^μ , depende solo de μ . En cambio, su media $e^{\mu+\frac{\sigma^2}{2}}$ depende de μ y σ , por lo que es necesario un análisis simultáneo para ambos parámetros al comparar medias.

La función logcuantil del modelo de regresión de log-localización y escala para la distribución lognormal, involucrando la variable indicadora x para comparar dos muestras es,

$$t_p(x) = \log[y_p(x)] = \mu(x) + \Phi^{-1}(p)\sigma = \beta_0 + \beta_1(x) + \Phi^{-1}(p)\sigma \quad (6)$$

Φ es la función de distribución de la normal estándar; $x = 0$ para una muestra y $x = 1$ para la otra.

3.1. Modelo bajo homogeneidad del parámetro σ

La función de verosimilitud para dos muestras lognormales independientes, de tamaño n_i , $i = 1, 2$; con observaciones exactas y censuradas por la izquierda tiene la forma

$$L(\beta_0, \beta_1, \sigma) = \prod_{i=1}^2 \prod_{j=1}^{n_i} \left[\frac{1}{y_{ij}\sigma} \phi\left(\frac{\log(y_{ij}) - \mu_i}{\sigma}\right) \right]^{\delta_{ij}} \left[\Phi\left(\frac{\log(y_{ij}) - \mu_i}{\sigma}\right) \right]^{1-\delta_{ij}}. \quad (7)$$

Donde ϕ es la función de densidad para la normal estándar, $\mu_i = \mu(x) = \beta_0 + \beta_1(x)$. Luego, para cada muestra, $\mu_1 = \mu(0) = \beta_0$ y $\mu_2 = \mu(1) = \beta_0 + \beta_1$, de donde $t_p(1) - t_p(0) = \mu(1) - \mu(0) = \beta_1$, que no depende de ningún cuantil. Entonces, si un intervalo aproximado del $(1 - \alpha)\%$ de confianza para β_1 contiene al cero, se concluye que no existe diferencia significativa entre las dos medias poblacionales.

3.2. Modelo bajo heterogeneidad del parámetro σ

La comparación se realiza agregando variables indicadoras en los modelos de regresión para ambos parámetros, reescribiendo (6) y optimizando (7) con $\mu_i = \beta_0 + \beta_1(x)$, $\log(\sigma_i) = \gamma_0 + \gamma_1(x)$. Entonces, $\mu_1 = \beta_0$, $\log(\sigma_1) = \gamma_0$ para la primera muestra y $\mu_2 = \beta_0 + \beta_1$, $\log(\sigma_2) = \gamma_0 + \gamma_1$ para la segunda. Una región de confianza permite, simultáneamente, analizar si β_1 y γ_1 son ceros, y aseverar respecto a la diferencia entre las medias poblacionales. Esto es, las poblaciones tienen concentraciones medias iguales si

Tabla 1: Datos de concentraciones de cobre

$$W(0) = \hat{\theta}'(\hat{\Sigma}_{\hat{\theta}})^{-1}\hat{\theta} \leq \chi^2_{1-\alpha}. \quad (8)$$

4. Ejemplo

En ocasiones, puede sospecharse que un grupo contiene concentraciones de contaminantes altas y para probar esta sospecha se compara con un grupo control. Otras veces, solo se desea saber si un grupo es mejor o peor que el otro. En todo caso, el interés es saber si los niveles de contaminantes son iguales o diferentes en ambos grupos.

Millard y Deverel (1988) reportan niveles de cobre en mantos freáticos, muestreados en dos zonas del valle San Joaquín en California, la Zona Alluvial Fan y la Zona Basin-Trough, presentados en el Cuadro 1. Cerca del 20 % de los datos son *no-detectados*, denotados por el signo <. Se desea comparar las concentraciones medias.

4.1. Homogeneidad del parámetro σ

Al optimizar (7) se encuentra que $\hat{\beta}_1 = -0.116$, $\hat{\beta}_0 = 1.050$, $\hat{\sigma} = 0.800$ y el intervalo del 95 % de confianza aproximado para β_1 es $(-0.413, 0.181)$. De la estimación de β_1 y de su intervalo se concluye que la diferencia en el nivel de cobre en las dos zonas es prácticamente nula. Así, a la luz de los datos, las concentraciones medias de cobre son iguales.

4.2. Heterogeneidad del parámetro σ

La comparación de medias se realizó a través de una región de confianza aproximada para β_1 y γ_1 . De los resultados, $\hat{\beta}_1 = 0.038$ y $\hat{se}_{\hat{\beta}_1} = 0.137$; mientras que para γ_1 , su EMV fue de -0.102 y su error estándar igual a 0.124 . Además, $\hat{\mu}_1 = 0.944$, $\hat{\mu}_2 = 0.906$, $\hat{\sigma}_1 = 0.817$ y $\hat{\sigma}_2 = 0.738$. Con lo anterior $W(0) = 1.816 \times 10^{-4} \leq 5.991 = \chi^2_{(0.95;2)}$; por lo tanto, al 95 % de confianza, las concentraciones medias de cobre no difieren significativamente.

5. Conclusiones

El método propuesto se basó en la función de máxima verosimilitud y en el algoritmo EM, el cual resultó versátil y simple para comparar poblaciones mediante modelos de regresión lognormal. En el ejemplo se compararon dos poblaciones bajo los supuestos de homogeneidad y heterogeneidad de σ , con una reparametrización para los parámetros del modelo. El criterio de comparación resultó eficiente al observar un intervalo de confianza aproximado y el estadístico de Wald. El algoritmo EM desempeñó un papel sustancial en la optimización de las funciones de verosimilitud. Debido a que la implementación del método no es difícil, éste puede extenderse a tres o más poblaciones.

Referencias

- El-Shaarawi, A. H., and Viveros, R. (1997). Inferences about the mean in log-regression with environmental applications. *Environmetrics*, **8**, 569–582.
- EPA (1992). *Statistical Training Course for Ground-Water Monitoring Data Analysis*. EPA530-R-93-003. Office of Solid Waste. U.S. Environmental Protection Agency, Washington, DC.
- Flury, B., and Zoppè, A. (2001). Exercise in EM. *The Am. Statist*, **54**, pp. 207 - 209.
- Gilbert, R.O. (1987). *Statistical Methods for Environmental Pollution Monitoring*. New York: Wiley
- Helsel, D. R. (2005). *Nondetects And Data Analysis*. New York: Wiley

- Meeker, W. Q., and Escobar, L. A. (1998). *Statistical Methods for Reliability Data*, New York: Wiley
- Millard, S. P., and Deverel, S. J. (1988). Nonparametric statistical methods for comparing two sites. *Water Resources Research*, **24**, 2087-2098.
- Oakes, D. (1999). Direct calculation of the information matrix via the EM algorithm. *Journal of the Royal Statistical Society. Series B*, **61** , 479 - 482.
- Ott, W. R. (1995). *Environmental Statistics and Data Analysis*, FL: CRC Press.
- R Development Core Team. (2006). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, www.R-project.org.

Evaluating cluster solutions with reference to data generation processes - a simulation study

Alexander von Eye^a *Michigan State University*

Patrick Mair *Wirtschaftsuniversität Wien*

1. The four steps of testing for absence of structure

In earlier papers, four steps were proposed for the process of determining whether an existing cluster solution contradicts hypotheses about the absence of a structure that can be detected using cluster analytic methods (von Eye, 2008, von Eye and Mair, 2007). These steps are

1. Clustering cases. Cluster methods are selected based on the decisions discussed by von Eye et al. (2004). For the purposes of the methods used and the simulations reported in this paper, we select clustering methods that create compact, that is, convex clusters.
2. Circumscribing clusters. In the present work, we use spheroids and ellipsoids to circumscribe the subregion that is defined by a cluster. The hull of an ellipsoid or spheroid is, in p -space, $x_d' R' V R x_d = 1$, where x_d is the $p \times 1$ vector of the differences of a point on the hull from the centroid of the hull, R is the $p \times p$ matrix of the orientation of the ellipse, and V is the inverse of the $p \times p$ matrix that contains the squared lengths of the semi-axes of the ellipsoid in its diagonal. If the semi-axes are equal in length, the hull circumscribes a spheroid, otherwise an ellipsoid. For an optimal description of clusters, we create Löwner ellipsoids which minimize both the volume of the region that is constituted by a cluster and the possible overlap between clusters.
3. Determining the expected number of cases in a cluster. Two data generation processes are considered. The first is a homogeneous Poisson process. For the second process, we assume that the data are multinormally distributed (for details, see von Eye, 2008, von Eye and Mair, 2007).

^aavoneye@msu.edu

4. Testing against hypotheses of lack of cluster structure. Exact tests, e.g., the binomial test, or asymptotic tests, e.g., the χ^2 -test, can be considered for testing hypotheses of lack of cluster structure.

2. Simulating data generation processes

We now ask whether clusters from hierarchical agglomerative methods still stand out when examined under the homogeneous Poisson and the multinormality models. We do this for both spherical and ellipsoid cluster hulls. To answer this question, we perform a simulation whose basic 2×2 design results from crossing the variables Model of Pattern Analysis and Shape of Cluster Hull.

The simulation proceeded as follows. Three to seven variables were created using the random number generators named below. The variables were generated one after the other, thus creating independent variables. These generators yielded either normally distributed random numbers or uniformly distributed random numbers. The uniformly distributed numbers were then transformed as described below so that they had the desired distributional characteristics. After these transformations, some of the variables were transformed again so that they were correlated to the degrees specified below. The result of these three steps were data sets with the characteristics needed for the simulation. The simulation was written in FORTRAN and executed under Windows XP. In addition, the following variables were used as factors of the simulation design.

Shape of distribution: The following five distribution shapes were realized (cf. von Eye et al., 2008):

1. Normal distribution: The generator GASDEV from the Numerical Recipes FORTRAN collection (Press et al., 1989) was used to create $N(0, 1)$ -distributed data.
2. Uniform distribution: The generator RANDOM, available in the Power Station's PortLib function pool, was used to create pseudo random numbers, z from the interval $0 \leq z < 1$.
3. Logarithmic distribution: Uniform variates x were subjected to the logarithmic transformation $\log(x)$. The resulting data were expected to exhibit some skewness and

elevated kurtosis.

4. Inverse Laplace-transformed: The Laplace probability distribution, also known as double exponential distribution, has a mean, a skewness, and a kurtosis of 0. A uniform distribution has no skew but exhibits increased kurtosis. Performing an inverse Laplace transformation on a uniform distribution should, therefore, result in a distribution with reduced kurtosis and possibly elevated skewness. Because the Laplace function has no inverse, the transformation introduced by von Eye et al. (2008) was performed. This transformation to the uniformly distributed random numbers results in a distribution with both a slightly elevated skewness and an elevated kurtosis. The kurtosis of the transformed uniform distribution has a positive sign. The kurtosis of the uniform distribution was negative. Thus, this transformation changed the distribution from being heavy-tailed to heavy around the belt line.
5. Cube root transformation: This transformation was used to create $y = 0.5x^{1/3}$ from the uniform x scores. Considering that the uniform scores that were cube root-transformed had no skewness and an only slightly elevated kurtosis, the resulting scores should have elevated skewness and elevated kurtosis.

Method of clustering: Six methods of clustering were used: Ward's method, complete linkage, average linkage, McQuitty's method, median linkage, and the centroid method.

The following variables were used as covariates:

- Sample size (N): The sample size in the simulation runs varied from 70 to 150 objects, in increments of 20.
- Number of variables (NVAR): the number of variables varied from 3 to 7, in increments of one.
- Cluster size (CLUSIZE): The size of each cluster was taken into account.
- Number of clusters (NCLUSTER): The number of clusters considered ranged from 2 to 9, increasing in steps of 1. For each of the created data sets, all eight hierarchical cluster solutions were analyzed.

Finally, it was counted whether a cluster contained more or fewer objects than expected under a probability model. The resulting variable, EGTK, was clearly data driven. However, it was assumed that clusters with fewer objects than expected might display different characteristics.

The resulting design was thus a 5 (TRANSFOR; type of distribution) \times 6 (CLUSMETH; method of clustering) \times 2 (POISSNOR; Poisson versus multinormaliy model) \times 2 (CIRCELLI; spherical versus ellipsoid cluster hull) \times 2 (EGTK; more versus fewer cases than expected for a cluster) design. The total number of observations considered was 466.664. This number is smaller than the number of 480.000 data sets that had been created in the simulation. However, clusters with fewer than 3 members were excluded from analysis because they occupy spaces with volumes of zero, and, therefore, the number of objects that is estimated based on volume would have led to an expected cluster size of zero. Note that clusters that contain three or more objects that are exactly aligned or have exactly the same coordinates also occupy spaces with zero volume, and would, therefore, have been excluded also. However, these cases did not occur. The probability of a cluster under the two probability models was used as dependent measure. The binomial test was used to calculate this probability.

The version of the binomial test used in the simulations can be described as follows. Let p be the probability and $q = 1 - p$. Let N be the sample size, n the observed frequency of an event, and e the expected frequency. Then, the probability that n or a larger number of cases was observed under p is

$$B(p) = \sum_{i=n}^N \binom{N}{i} p^i q^{N-i}. \quad (1)$$

3. Results

In Table 3, the results of an ANOVA of the design described above are summarized. This table was created using SAS. Overall, the model explained 18.45% of the variance of the dependent measure, and the mean of the dependent measure was 0.07.

As expected based on the large number of data sets, all effects are significant, with the POISSNOR*CIRCELL*EGTK interaction being the only exception. As could also be

Source	DF	Type III SS	MSQ	F Value	Pr < F
N	1	0.339	0.339	22.37	< .0001
NVAR	1	10.318	10.318	680.89	< .0001
NCLUSTER	1	2.301	2.301	151.81	< .0001
CLUSIZE	1	131.956	131.956	8707.27	< .0001
TRANSFOR	4	9.804	2.451	161.73	< .0001
CLUSMETH	5	0.888	0.178	11.72	< .0001
POISSNOR	1	0.982	0.982	64.78	< .0001
CIRCELLI	1	0.813	0.814	53.70	< .0001
EGTK	1	30.615	30.615	2020.20	< .0001
CLUSMETH*CIRCELLI	5	0.298	0.060	3.93	0.0015
CIRCELLI*EGTK	1	0.112	0.112	7.38	0.0066
POISSNOR*CIRCELLI	1	0.955	0.955	63.04	< .0001
TRANSFOR*CIRCELLI	4	2.487	0.622	41.03	< .0001
CLUSMETH*EGTK	5	0.464	0.093	6.13	< .0001
CLUSMETH*POISSNOR	5	1.109	0.222	14.64	< .0001
TRANSFOR*CLUSMETH	20	7.967	0.398	26.29	< .0001
POISSNOR*EGTK	1	3.311	3.311	218.47	< .0001
TRANSFOR*EGTK	4	7.610	1.902	125.53	< .0001
TRANSFOR*POISSNOR	4	9.131	2.283	150.64	< .0001
CLUSMET*CIRCELL*EGTK	5	0.261	0.052	3.45	0.0041
CLUSME*POISSNOR*CIRCEL	5	0.332	0.066	4.38	0.0005
TRANSF*CLUSME*CIRCEL	20	1.394	0.070	4.60	< .0001
POISSNOR*CIRCELL*EGTK	1	0.020	0.020	1.35	0.2452
TRANSFO*CIRCELL*EGTK	4	2.226	0.556	36.71	< .0001
TRANSF*POISSNOR*CIRCEL	4	1.823	0.456	30.07	< .0001
CLUSMET*POISSNOR*EGTK	5	4.191	0.838	55.31	< .0001
TRANSFO*CLUSMET*EGTK	20	4.217	0.210	13.91	< .0001
TRANSF*CLUSME*POISSNOR	20	3.071	0.154	10.13	< .0001
TRANSFO*POISSNOR*EGTK	4	2.177	0.544	35.91	< .0001
CLUS*POIS*CIRCE*EGTK	5	0.338	0.068	4.46	0.0005
TRAN*CLUS*CIRCE*EGTK	20	1.107	0.055	3.65	< .0001
TRAN*CLUS*POIS*CIRCE	20	1.937	0.097	6.39	< .0001
TRAN*POIS*CIRCE*EGTK	4	2.129	0.532	35.13	< .0001
TRAN*CLUS*POISSNOR*EGTK	20	4.598	0.230	15.17	< .0001
TRA*CLU*POI*CIR*EGTK	19	1.581	0.083	5.49	< .0001

Table 1: ANOVA of the simulation study

expected based on the nature of the random data which do not contain a pre-engineered cluster structure, the average probability for a cluster is above the significance threshold. However, the majority of the significance values was close to zero. Most of the effects sizes are very small. In fact, 26 of the 32 partial η^2 had zeros as their first three decimals. Therefore, in the following paragraphs, we illustrate a selection of the stronger effects graphically.

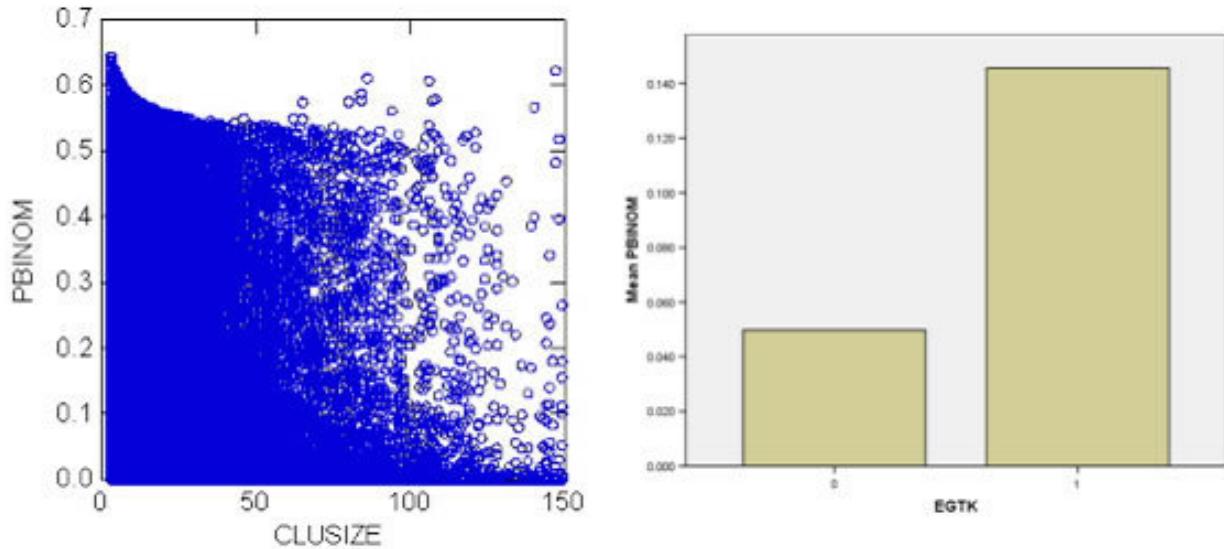


Figure 1: Effects of cluster size and EGTK

The strongest effect was that of cluster size. This is illustrated in Figure 1 (left panel). We see that the probability of a cluster decreases as its size increases. The number of clusters with high probabilities is much smaller for larger clusters than for smaller clusters. This result reflects the increased statistical power of larger cluster sizes. However, it also suggests that larger clusters are created in particular if they represent local density maxima. This result does not vary with shape of cluster hull and EGTK (not shown here).

The second largest effect was observed for the variable EGTK, that is, for the variable that distinguishes between cluster sizes above and cluster sizes below expectation. Figure 1 (right panel) displays this effect. Figure 2 shows that for those cases in which the observed cluster size is larger than expected, the average probability is 0.05. This covers 78.5% of the simulated cases. For those clusters that contain fewer objects than expected, the average probability is about 0.15. This covers 21.5% of the simulated cases.

The transformation the data were subjected to also had a strong effect. This is shown in the left panel of Figure 2. The right panel shows the interaction with EGTK.

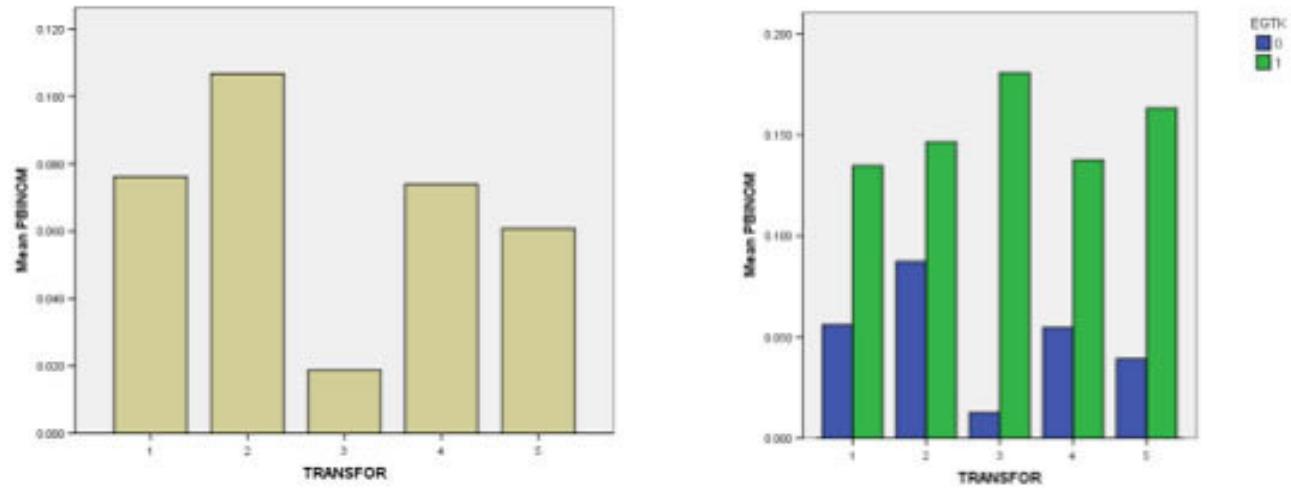


Figure 2: Bar chart of the effect of variable transformation (the transformations are 1 = normal distribution, 2 = uniform, 3 F= logarithmic, 4 = inverse Laplace, and 5 = cube root) without (left panel) and with (right panel) consideration of the effect of EGTK

The two panels in Figure 2 show that the transformations have effects on the probability of clusters. Specifically, on average (left panel of Figure 2), the average probabilities of the logarithmic transformation lead to data structures far away from the uniform and the multinormal distributions. We thus can conclude that the tests employed here are particularly sensitive when the underlying distributions are far from the uniform or the multinormal. Accordingly and as expected, the average probability of clusters is the highest for the multinormal and uniform distributions.

The interaction of the transformations with EGTK, displayed in the right panel of Figure 2, suggests that just the opposite is observed for those clusters that contain fewer cases than expected. Here, the clusters from the log-transformed data come with the highest probability, and the cluster from the multinormal data with the lowest. Overall, however, these probabilities are much higher than for the cases in which the clusters contain more cases than expected (see Figure 1).

The variable Clustering Method (CLUSMETH) had only minimal effects (not depicted here). The probabilities of the clusters from median linkage and the centroid method were

slightly below those for the other four methods. The probabilities for the clusters with fewer objects than expected were much higher than those for the clusters with more objects than expected. The rank orders of probabilities varied only minimally in the interactions with the transformation and the EGTK variables, as well as in the three-way interaction of these variables.

4. Discussion

The results of the present simulations are interesting in a number of respects. First, the six methods of hierarchical clustering used here seem to identify clusters even in uniform distributions. The methods proposed for testing whether clusters indeed contradict hypotheses that are derived from data generation models, therefore, fill an important gap in the arsenal of statistical methods. Second, the simulation shows that deviations from uniform distributions come with an increased probability that clusters will be identified. This is not a surprise and confirms well known earlier results. Third, and most importantly, the simulations show that some of the sectors that clustering methods identify as density centers are, in the light of hypotheses derived from data generation methods, just the opposite. They contain fewer cases than expected.

In a recent article by Bauer (2007), the issue is raised that clustering methods may present clusters even if the population does not contain any taxonic structure to be found. Therefore, the methods discussed here add tools to the user who, without these tools would have a hard time making decisions about the existence of clusters.

References

- Bauer, D. J. (2007). Observations on the use of growth mixture models in psychological research. *Multivariate Behavioral Research*, 42:757–786.
- Hand, D. J. and Bolton, R. J. (2004). Pattern discovery and detection: A unified statistical methodology. *Journal of Applied Statistics*, 31:885–924.

- Press, W. H., Flannery, B. P., Teukolsky, S. A., and Vetterling, W. T. (1989). *Numerical recipes. The art of scientific computing (FORTRAN version)*. Cambridge University Press, Cambridge.
- von Eye, A. (2008). Did you expect this cluster here? Distributional characteristics of clusters. Under editorial review.
- von Eye, A. and Mair, P. (2007). Examining distributional characteristics of clusters. In J. A. Dominguez Molina, A. V. Gonzalez Fragoso, and J. H. Sierra Cavazos, editors, *Memorias del XXI Foro Nacional de Estadistica*, pages 1–6. Instituto Nacional de Estadistica, Geographia e Informatica, Aguascalientes, Ags., Mexico.
- von Eye, A., Mun, E. Y. and Indurkhya, A. (2004). Classifying developmental trajectories: A decision making perspective. *Psychology Science*, 46:65–98.
- von Eye, A., von Eye, M. and Bogat, G. A. (2008). Multinormality and symmetry: A comparison of two statistical tests. *Psychology Science*, 48:419–435.

Lista de autores

- Almendra Arao, Félix <falmendra@ipn.mx>. *UPIITA del Instituto Politécnico Nacional*, 109
- Ariza Hernández, Francisco J. <arizahfj@colpos.mx>. *Colegio de Postgraduados*, 1
- Bustillo Díaz, M. *Universidad Autónoma de Puebla – Facultad de Ciencias de la Computación*, 99
- Castaño Meneses, Víctor Manuel. *Universidad Nacional Autónoma de México*, 89
- Castaño Tostado, Eduardo <ecastano@uaq.mx>. *Universidad Autónoma de Querétaro*, 25, 89
- Castillo Morales, Maribel. *Estudiante del Postgrado en Ciencias Ambientales, ICUAP Benemérita Universidad Autónoma de Puebla*, México, 55
- Dupuy, Jean François <dupuy@cict.fr>. *Université Paul Sabatier 3, Francia*, 33
- Escarela, Gabriel <ge@xanum.uam.mx>. *Universidad Autónoma Metropolitana – Iztapalapa*, 7, 33, 39, 61
- Fernández Harmony, Thalía. *Universidad Nacional Autónoma de México*, 89
- Godínez Jaimes, Flaviano <fgodinezj@gmail.com>. *Unidad Académica de Matemáticas, Universidad Autónoma de Guerrero*, 13
- González Estrada, Elizabeth <eliza_ge@yahoo.com.mx>. *Colegio de Postgraduados*, 19
- Guzmán Martínez, María <marnezmar@yahoo.com.mx>. *Universidad Autónoma de Querétaro*, 25
- Hernández Gallardo, Lorelie <heilerol@yahoo.com.mx>. *Universidad Autónoma Metropolitana – Iztapalapa*, 39
- Hernández Quintero, Angélica <angyka302@gmail.com>. *Universidad Autónoma Metropolitana – Iztapalapa*, 33
- Huerta López, Jessica Giovanna. *Universidad Autónoma de Puebla – Facultad de Ciencias de la Computación*, 99

- Jiménez Bedolla, Jazmin <akashajajibe@hotmail.com>. *Universidad Autónoma de Puebla – Facultad de Ciencias de la Computación*, 99
- Linares Fleites, Gladys <gladys.linares@icbuap.buap.mx>. *Departamento de Investigaciones en Ciencias Agrícolas. Instituto de Ciencias de la Benemérita Universidad Autónoma de Puebla*, 47, 55
- Mair, Patrick <pmair@stat.ucla.edu>. *Wirtschaftsuniversität Wien*, 123
- Méndez Ramírez, Ignacio <imendez@servidor.unam.mx>. *Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, UNAM*, 13, 75
- Moreno Zúñiga, Tania <tania_8304@hotmail.com>. *Universidad Autónoma Metropolitana – Iztapalapa*, 61
- Muñoz Urbina, Armando. *Universidad Autónoma Agraria Antonio Narro*, 75
- Olmedo, Leonardo <leonardo.olmedo@hotmail.com>. *Universidad Autónoma Metropolitana – Iztapalapa*, 69
- Padrón Corral, Emilio <epadron@mate.uadec.mx>. *Universidad Autónoma de Coahuila*, 75
- Pérez Rodríguez, Paulino <perpdgo@colpos.mx>. *Colegio de Postgraduados*, 83
- Rangel Huerta, A. *Universidad Autónoma de Puebla – Facultad de Ciencias de la Computación*, 99
- Rodríguez Yam, Gabriel A. <grodrigu@correo.chapingo.mx>. *Universidad Autónoma Chapingo*, 1
- Ruiz Suárez, Luis G. *Centro de Ciencias de la Atmósfera. Universidad Nacional Autónoma de México*, 47
- Saavedra Gastélum, Verónica <veroclessg@yahoo.com.mx>. *Universidad Autónoma de Querétaro*, 89
- Saldaña Munive, José Adrián. *Departamento de Investigaciones en Ciencias Agrícolas. Instituto de Ciencias de la Benemérita Universidad Autónoma de Puebla*, 47
- Sánchez Rinza, Bárbara Emma <brinza@cs.buap.mx>. *Universidad Autónoma de Puebla – Facultad de Ciencias de la Computación*, 99
- Sotres Ramos, David <sotres.davida@kendle.com>. *Colegio de Postgraduados*, 109
- Ulín Montejo, Fidel <fidel.ulin@basicas.ujat.mx>. *Matemáticas, Div. Acad. de Ciencias Básicas, Universidad Juárez Autónoma de Tabasco*, 115

Valera Pérez, Miguel Ángel. *Departamento de Investigaciones en Ciencias Agrícolas. Instituto de Ciencias de la Benemérita Universidad Autónoma de Puebla*, 55

Vaquera Huerta, Humberto <hvaquera@colpos.mx>. *Estadística, Campus Montecillo, Colegio de Postgraduados*, 115

Villaseñor Alva, José A. <jvillasr@colpos.mx>. *Colegio de Postgraduados*, 19, 83

von Eye, Alexander <voneye@msu.edu>. *Michigan State University*, 123

Esta publicación consta de 1 085 ejemplares y se terminó
de imprimir en julio de 2008 en los talleres gráficos del
Instituto Nacional de Estadística, Geografía e Informática
Av. Héroe de Nacozari Sur Núm. 2301, Puerta 11, Nivel Acceso
Fracc. Jardines del Parque, CP 20270
Aguascalientes, Ags.
México