MÉTODOS ESTADÍSTICOS PARA LA DETECCIÓN DE UMBRALES DE CONTAMINACIÓN EN ECOSISTEMAS ACUÁTICOS DE AGUA DULCE

STATISTICAL METHODS FOR THE DETECTION THRESHOLD OF POLLUTION IN FRESHWATER AQUATIC ECOSYSTEM

Autor:

Alexander Martínez Suárez [1]

ARTÍCULO DE INVESTIGACIÓN CIENTÍFICA Y TECNOLÓGICA

Cómo citar este artículo:
Cómo citar este artículo: Martínez Suárez, A. Métodos estadísticos para la detección de umbrales de contaminación en ecosistemas acuáticos de agua dulce. Innovaciencia facultad ciencias exactas fis. naturales. 2016; 4(1): 39 -48

Fecha de recepción:
Artículo recibido el 05 octubre de 2016 y aceptado para publicación el 25 de noviembre de 2016.

DOI:
http://dx.doi.org/10.15649/2346075X.402

Palabras clave
IC Intervalos de Confianza, Bootstrap, Devianza, LAD Least Absolute Deviation, umbral.

RESUMEN

Introducción: Este trabajo presenta la metodología y los resultados de la implementación de métodos estadísticos en la determinación de umbrales de contaminación en ecosistemas acuáticos. Los objetivos eran analizar algunos de los métodos con los cuales se pueden obtener umbrales o puntos de cambio en una variable de interés, desarrollar los respectivos programas estadísticos en el software R y aplicar estos métodos en la determinación de umbrales para contaminantes en reservorios de Puerto Rico.

Una de las herramientas estadísticas utilizadas en el desarrollo de cada uno de los métodos es la aplicación de probabilidad condicional. Para ello se necesitan dos variables Q y X, con la condición que a una de ellas se le conoce el umbral. Supóngase Q el umbral conocido y llámese X la variable de interés, variable a la que se le desea conocer el umbral. El proceso consiste en, dadas las dos variables, utilizar la probabilidad condicional $P(Q \mid X)$ y mediante los métodos, determinar el umbral de la variable de interés X.

[1] Candidato a Doctor en Educación, Universidad de la Salle, Costa Rica. Unidades Tecnológicas de Santander, Bucaramanga, Colombia. Correo electrónico: alexandermartinezsuarez@hotmail.com

ABSTRACT

Introduction: This paper presents the methodology and results of the implementation of statistical methods in determining thresholds of pollution in aquatic ecosystems. The objectives are to analyze some of the methods which can be obtained thresholds or points of change in a variable of interest, develop the respective statistical programs in the R software and apply these methods in determining thresholds for pollutants in reservoirs Puerto Rico.

One of the statistical tools used in the development of each of the methods is the application of conditional probability. For this, two variables Q and X are needed, provided that one of them is called the threshold. Q Suppose the known threshold and call it X variable interest variable that you want to know the threshold. The process involves, given the two variables, using the conditional probability $P(Q \mid X)$ and by the methods, determine the threshold of the variable of interest X.

Keywords
IC Confidence Intervals , Bootstrap, Deviance, and Least Absolute Deviation, threshold.

INTRODUCCIÓN

Dada la creciente importancia de la elaboración de normas de calidad del agua en los ecosistemas acuáticos, es necesario definir las condiciones numéricas de referencia para los diferentes nutrientes. Estos criterios deben ser geográficamente específicos, y reflejan la relación particular entre cada nutriente y una respuesta del medio ambiente (por ejemplo, la clorofila o turbidez). El nitrógeno y el fósforo son algunas de las causas más importantes de enriquecimiento excesivo de nutrientes en ríos y lagos. La Clorofila_a (*Chl_a*) se usa como un indicador de la condición del agua. Para los lagos, la Agencia de Protección Ambiental de los Estados Unidos USEPA [1] ha asociado en el 2009 el estado trófico de la Clorofila_a:

Estado Trófico	Chl_a
Oligotrófico	$< 2\ ug/L$
Mesotrófico	$2 - 7\ ug/L$
Eutrófico	$7 - 30\ ug/L$
Hipereutrófico	$> 30\ ug/L$

Se plantea determinar los umbrales para el Fósforo Total (TP) y el Nitrógeno Total (TN) en cada uno de los límites tróficos de la Clorofila_a. Además, comparar los resultados por los diferentes métodos para calcular umbrales, y proponer un método con mejores resultados que sea menos sensible a valores atípicos y con estimaciones más robustas.

En lo que respecta a datos, los embalses analizados en este estudio son Guajataca, Cerrillos y la Plata de Puerto Rico, un total de 165 datos recolectados del 2005 a 2010. Los métodos que se proponen son Intervalos de Confianza que no se traslapan, Ajuste de un modelo lineal, Modelo Jerárquico Bayesiano, Reducción de la Devianza y Metodología LAD, que tienen por objeto determinar umbrales para nutrientes en ecosistemas acuáticos.

MÉTODOS PARA DETERMINAR UMBRALES

Para determinar el umbral X_c de un nutriente X se debe conocer un umbral q_c de una variable respuesta Q. Una de las herramientas utilizadas en cada uno de los métodos es la probabilidad condicional $P(Q > q\_c \mid X > x\_i)$ para cada valor x_i. El valor x_c se determina el punto de cambio en la relación $P(Q > q\_c \mid X > x\_i)$ vs. x_i. Los métodos que se propuestos para determinar el umbral son:

- Intervalos de Confianza que no se Traslapan (Paul and McDonald, 2005) [2].
- Ajuste de un Modelo No Lineal (Paul and McDonald, 2005) [2].
- Ajuste de Modelo Jerárquico Bayesiano (Qian et al., 2003) [3].
- Reducción de la Devianza no paramétrica (Qian et al., 2003) [3].
- Metodología LAD (Martínez Suárez, 2010) [4].

IC que no se Traslapan

El método de los intervalos de confianza (IC) que no se traslapan determina el umbral x_c, como el primer valor para el cual el IC de $P(Q > q\_c \mid X > r\_c)$ no traslapa con el IC de $P(Q > q\_c)$ [2]. En casos donde el comportamiento no es estrictamente monótono el resultado no es razonable. Los intervalos de confianza que no se traslapan consisten en construir intervalos de confianza del 95% de la probabilidad no condicional y condicional mediante la técnica Bootstrap.

Ajuste de un modelo no lineal

El método del Modelo No Lineal consiste en definir un comportamiento o ajuste de un modelo no lineal a la curva de probabilidad condicional $P(Q > q_c \mid X > x\_i)$ vs. x_i . Un modelo que parte en dos el conjunto de los datos, dando lugar a dos parámetros que definirán el punto de cambio, a partir de un estimador de buen ajuste como la Devianza [2].

El comando nls() es la aplicación en **R** [5,8,10] que trabaja con este tipo de modelos. En **R**, la diferencia principal entre los modelos lineales y modelos no lineales es tener que especificar la naturaleza exacta de la ecuación como parte de la fórmula del modelo no lineal [6,9,11].

En nuestro caso $p_1, p_2, \ldots, p_n$ son las probabilidades condicionales, $x_1, x_2, \ldots, x_n$ los valores de la variable de interés y f una función no lineal que se ajusta a los datos. Las estimaciones de los parámetros β de f se determinan proporcionando el mejor ajuste que explique las observaciones p, obtenidas por el criterio de mínimos cuadrados con respecto a β:

$$RSS(\beta) = \sum_{i=1}^{n} (p_i - f(x_i, \beta))^2$$

Esta medida es útil para comparar diferentes modelos que se ajustan para el mismo conjunto de datos. Una forma para obtener el mínimo *RSS*(β) es utilizar la Devianza del modelo. El punto x_c (Umbral) es el valor que minimiza el método de mínimos cuadrados o el método numérico de ajuste.

Ajuste del modelo Jerárquico Bayesiano

El modelo bayesiano asume que las observaciones provienen de una mezcla de dos distribuciones normales, cuyos parámetros tienen distribuciones previas. El método supone que las probabilidades condicionales $P_1, P_2, \ldots, P_N$ provienen de una familia aleatoria de distribución normal,

$$\begin{aligned} P_1, P_2, \ldots, P_r &\sim \pi(P_i \mid \theta_1) \\ P_{r+1}, P_{r+2}, \ldots, P_N &\sim \pi(P_i \mid \theta_2) \end{aligned}$$

donde r ($1 \leq r \leq n$) representa el punto de cambio.

El punto r se presenta como un parámetro con una distribución de probabilidad dependiente de *P*. El objetivo es establecer su distribución y poder calcular el estimador de *r* [5] .

Reducción de la devianza no paramétrica

El procedimiento asume un modelo de mezcla con solo dos medias y un punto de cambio. El método de reducción de la devianza no paramétrica consiste en separar la variable respuesta en dos partes y utilizar la devianza no paramétrica para calcular el punto de cambio de la variable de interés. Análogo a los métodos anteriores, los valores de la variable respuesta son las probabilidades condicionales $p_1, p_2, \ldots, p_n$.

Si se asume un modelo con dos medias μ_p, para valores de $i \leq r$, y otra para valores de $i > r$, el proceso de la reducción de la devianza no paramétrica consiste en realizar una partición de la variable explicativa en un punto *r* (variable ya ordenada), con ello se forman dos sub muestras $S_{\leq r}$ y $S_{>r}$, una a la izquierda y otra a la derecha de r, respectivamente, luego determinar el punto donde

$$D_r = D(S_{\leq r}) + D(S_{>r})$$

Minimiza la suma de la devianza, teniendo en cuenta que

$$D(S) = \sum_{i \in s} (p_i - \bar{p}_s)^2$$

El objetivo es formar conjuntos más homogéneos, es decir, la partición que da mayor homogeneidad [3].

Metodología LAD (Least Absolute Deviatión)

Hasta donde se ha podido verificar, este método no ha sido usado previamente para determinar umbrales. El proceso es similar a la metodología de la reducción de la devianza no paramétrica, Pero en lugar de usar los mínimos cuadrados como criterio de homogeneidad, se utiliza *LAD* (Least Absolute Deviation) [4].

Se propone una alternativa de encontrar el umbral usando el criterio LAD (Least Absolute Deviation) [4]. Suponga que se tienen n pares de datos (x_i, p_i) donde $p_i = P(Q > q_c \mid x > x_i)$ y se define

$$LAD_{\leq r} = \sum_{i=1}^{r} |p_i - \bar{p}_{\leq r}| \quad y \quad LAD_{>r} = \sum_{i=r+1}^{n} |p_i - \bar{p}_{>r}|$$

El punto de cambio c que define el umbral x_c es

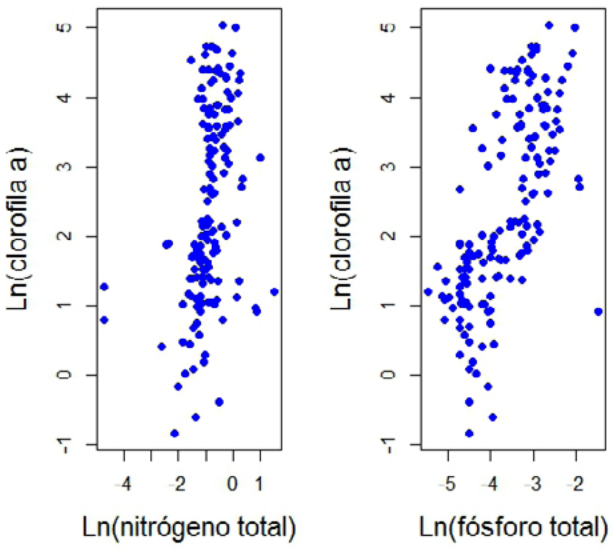
$$c = argmin_{r=1,2,..,n}(LAD_{\leq r} + LAD_{>r})$$

A pesar que los mínimos cuadrados son más clásicos y más ampliamente estudiados, sus estimaciones son muy sensibles a valores atípicos y su rendimiento puede verse comprometido cuando los errores son grandes y heterogéneos. A diferencia del método de los mínimos cuadrados, el método *LAD* no es tan sensible a valores atípicos y produce estimaciones más robustas.

APLICACIONES Y RESULTADOS

El propósito es aplicar los métodos para calcular umbrales en los nutrientes Nitrógeno Total (NT) y Fósforo Total (PT), basados en cada uno de los cuatro límites tróficos de la Clorofila_a (*ver tabla 1*). El proceso general es realizar 1000 Bootstrap y obtener el umbral como el promedio de los umbrales muéstrales. El IC se calcula usando el método de los percentiles en las muestras Bootstrap [7]. Cada uno de los resultados obtenidos y los respectivos algoritmos de programación están escritos en **R**.

Fig. 1. Diagrama de dispersión clorofila_a vs fósforo total y nitrógeno total



Dado el comportamiento que presenta la probabilidad condicional $pi = P(Q > q_c \mid X > x_i)$ vs. x_i, los dos primeros métodos (Intervalos de Confianza que no se traslapan y el Ajuste de un Modelo No Lineal) no pueden ser usados para calcular el umbral en los nutrientes. Las gráficas de probabilidad condicional en cada uno de los casos no presentan relaciones monótonas, lo que dificulta determinar un único punto donde el intervalo de confianza de la probabilidad no condicional no traslapa con los intervalos de confianza, para valores mayores de un punto de la variable de interés. Y para el modelo lineal debido al comportamiento de la probabilidad condicional en cada una de las variables, no es sencillo ajustar un modelo que represente estos comportamientos.

Los embalses de Puerto Rico analizados en este estudio son Guajataca, Cerrillos y la Plata. A partir de los niveles umbral de la Clorofila (*Ch_a*) se determina el umbral para las variables de interés Fósforo Total (*TP*) y Nitrógeno Total (*TN*) de un total de 165 datos recolectados del 2005 a 2010. La probabilidad condicional que se calcula, por ejemplo, para el límite oligotrófico en TP es $P(Chl\_a > 2 \mid TP > TP_i)$.

Las tablas muestran los resultados obtenidos al aplicar los métodos propuestos en cada uno de los límites y sus respectivos umbrales para el fósforo total y nitrógeno total. Los intervalos de confianza para cada umbral se determinaron mediante la técnica de Bootstrap a 1000 muestras. El umbral se estimó como la media de los puntos de cambio de la probabilidad condicional calculados en cada iteración Bootstrap y los intervalos de confianza por el método de los percentiles.

TABLA II
UMBRALES DE TP DETERMINADOS POR CADA UNO DE LOS MÉTODOS, EN CADA UNO DE LOS LÍMITES TRÓFICOS

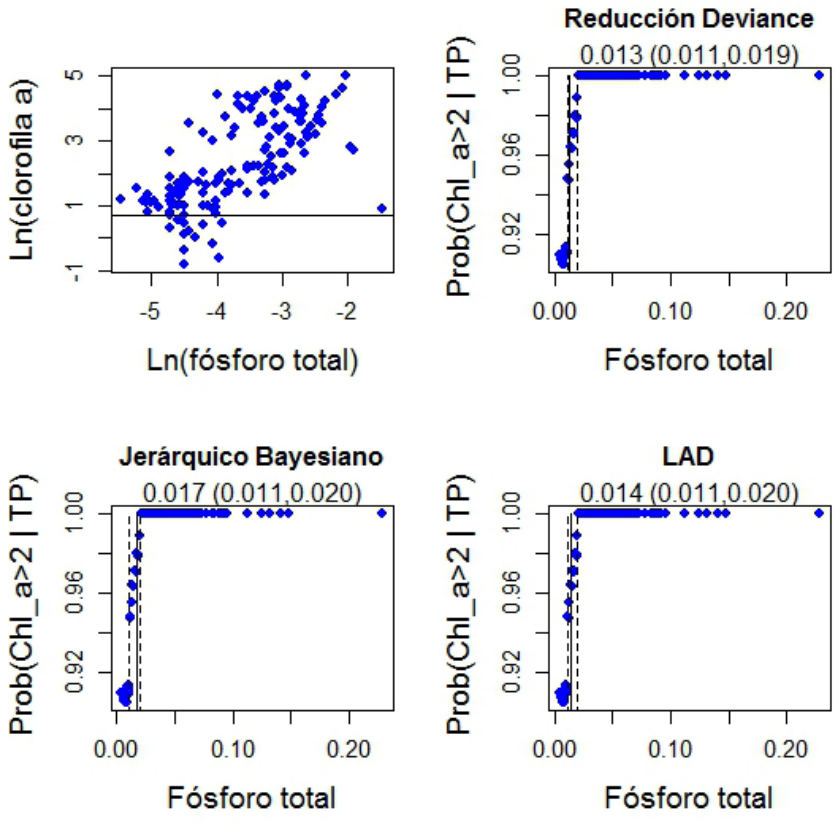
Clorofila_a	Método	Umbral TP (IC 95 %)
>2 mg/L	Bayesiano	0,017 (0,011; 0,020)
	Reducción Deviance	0,013 (0,011; 0,019)
	Metodología LAD	0,014 (0,011; 0,020)
>7 mg/L	Bayesiano	0,033 (0,010; 0,148)
	Reducción Deviance	0,036 (0,011; 0,148)
	Metodología LAD	0,016 (0,011; 0,021)
>30 mg/L	Bayesiano	0,037 (0,011; 0,131)
	Reducción Deviance	0,063 (0,015; 0,131)
	Metodología LAD	0,021 (0,011; 0,069)

TABLA III
UMBRALES DE TN DETERMINADOS POR CADA UNO DE LOS MÉTODOS, EN CADA UNO DE LOS LÍMITES TRÓFICOS

Clorofila_a	Método	Umbral TN (IC 95 %)
>2 mg/L	Bayesiano	0,369 (0,202; 0,590)
	Reducción Deviance	0,334 (0,159; 0,590)
	Metodología LAD	0,397 (0,202; 0,590)
>7 mg/L	Bayesiano	0,951 (0,252; 2,689)
	Reducción Deviance	1,116 (0,274; 2,689)
	Metodología LAD	0,535 (0,264; 1,349)
>30 mg/L	Bayesiano	0,865 (0,297; 1,291)
	Reducción Deviance	1,107 (0,410; 1,291)
	Metodología LAD	0,787 (0,293; 1,291)

Resultados: Si se comparan los valores obtenidos en cada uno de los niveles tróficos se puede observar que solo la metodología LAD mantiene la monotonicidad en los umbrales de los nutrientes, cuando el límite trófico de *Chl_a* aumenta, el umbral en cada uno de los nutrientes también aumenta. Esta propiedad es necesaria dada la relación monótona entre *Chl_a* y TN, TP, se observa que en los otros dos métodos no se cumple siempre. Los umbrales para cada uno de los nutrientes deben seguir un patrón de crecimiento en el aumento de los límites tróficos en *Chl_a*. Las gráficas siguientes presentan el punto de cambio, y su respectivo intervalo de confianza del 95% en cada uno los métodos y por cada uno de los niveles.

Fig. 2. Umbrales para fósforo total cuando chl_a> 2. Bajo las técnicas reducción de la deviance, jerárquico bayesiano, metodología lad.



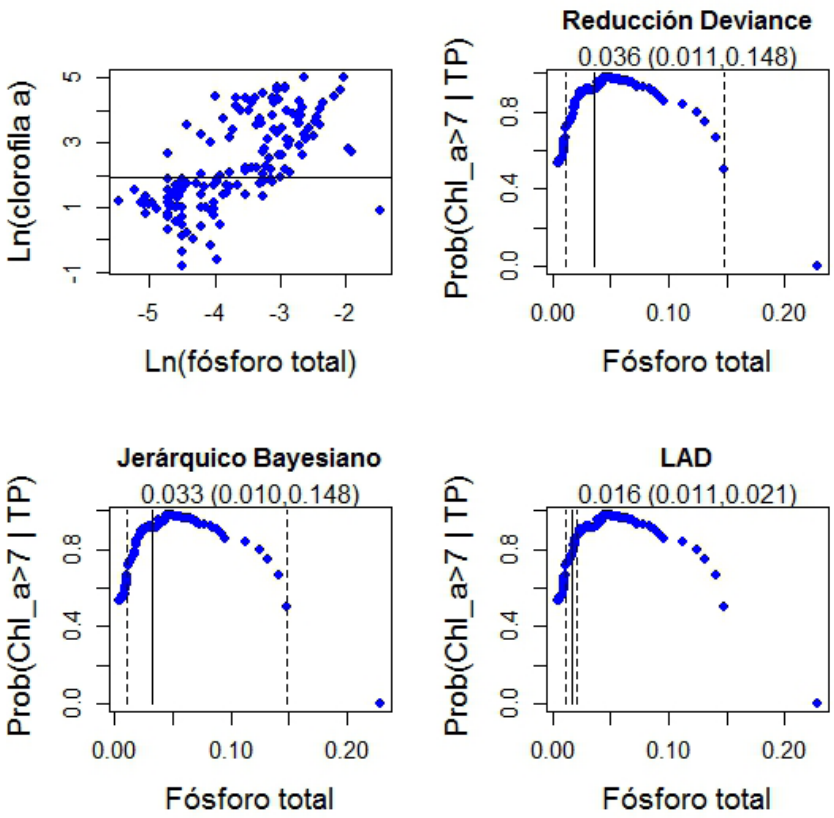


Fig. 3. Umbrales para fósforo total cuando chl_a > 7. Bajo las técnicas bajo las técnicas reducción de la deviance, jerárquico bayesiano, metodología lad.

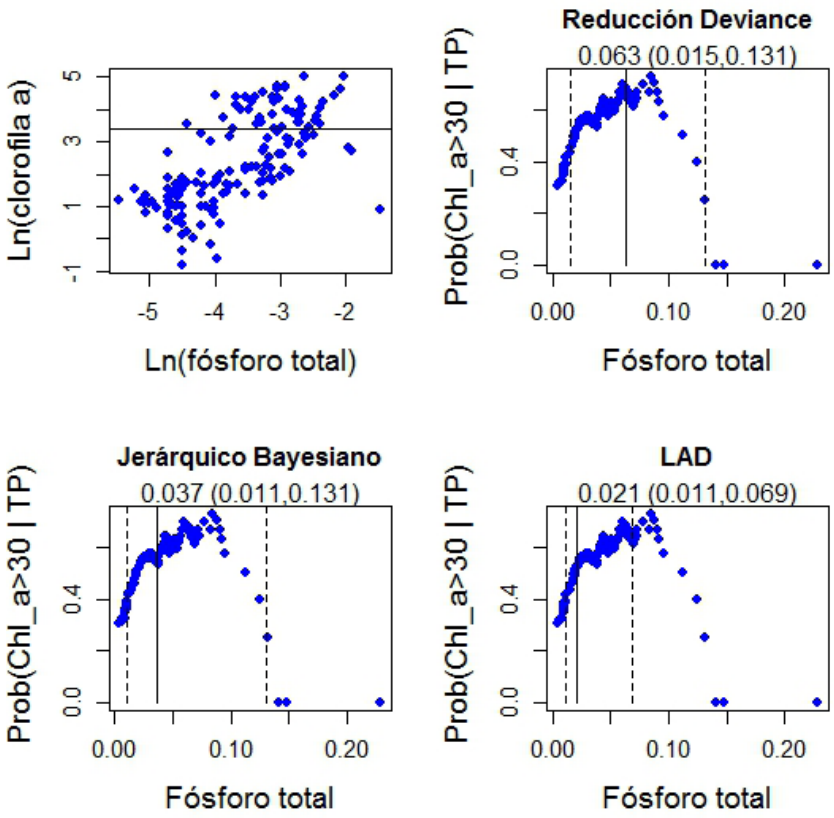


Fig. 4. Umbrales para fósforo total cuando chl_a > 30. Bajo las técnicas bajo las técnicas reducción de la deviance, jerárquico bayesiano, metodología lad.

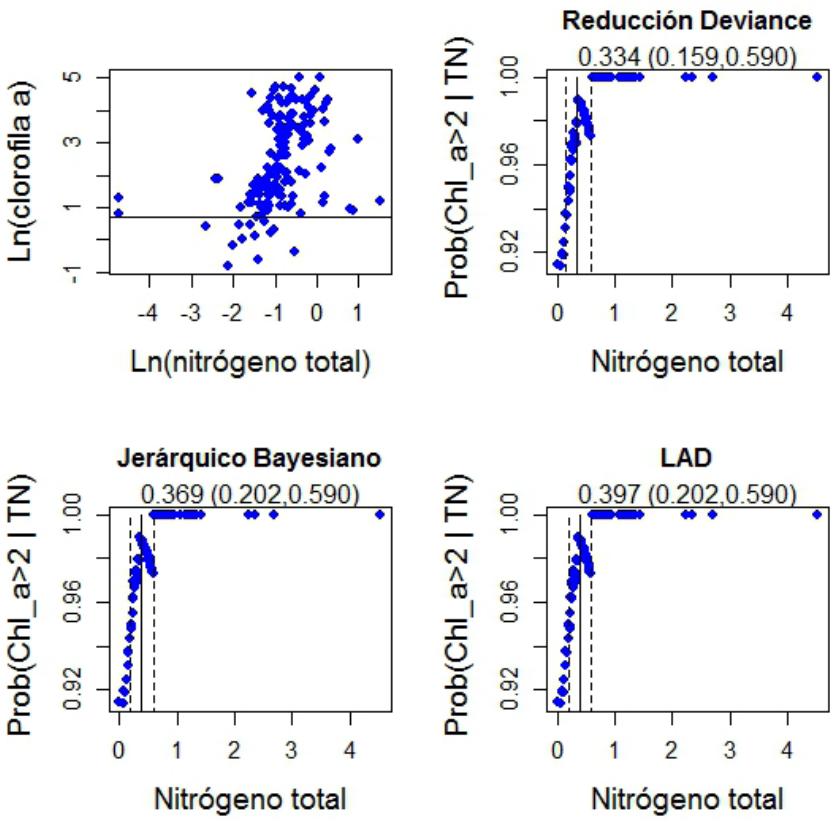


Fig. 5. Umbrales para nitrógeno total cuando chl_a > 2. Bajo las técnicas reducción de la deviance, jerárquico bayesiano, metodología lad.

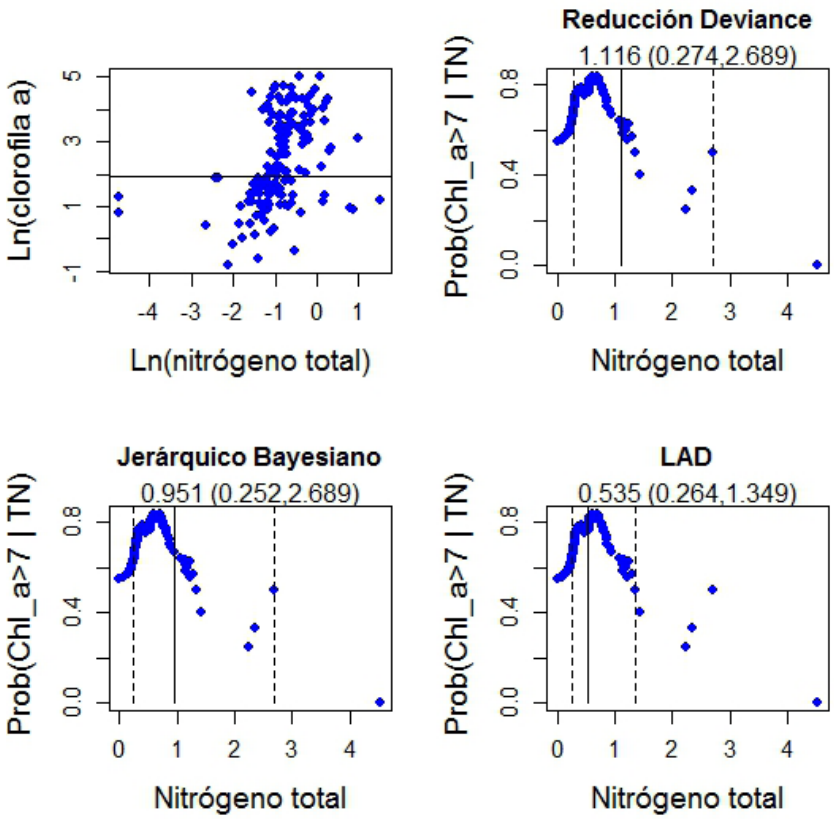


Fig. 6. Umbrales para nitrógeno total cuando chl_a > 7. Bajo las técnicas reducción de la deviance, jerárquico bayesiano, metodología lad.

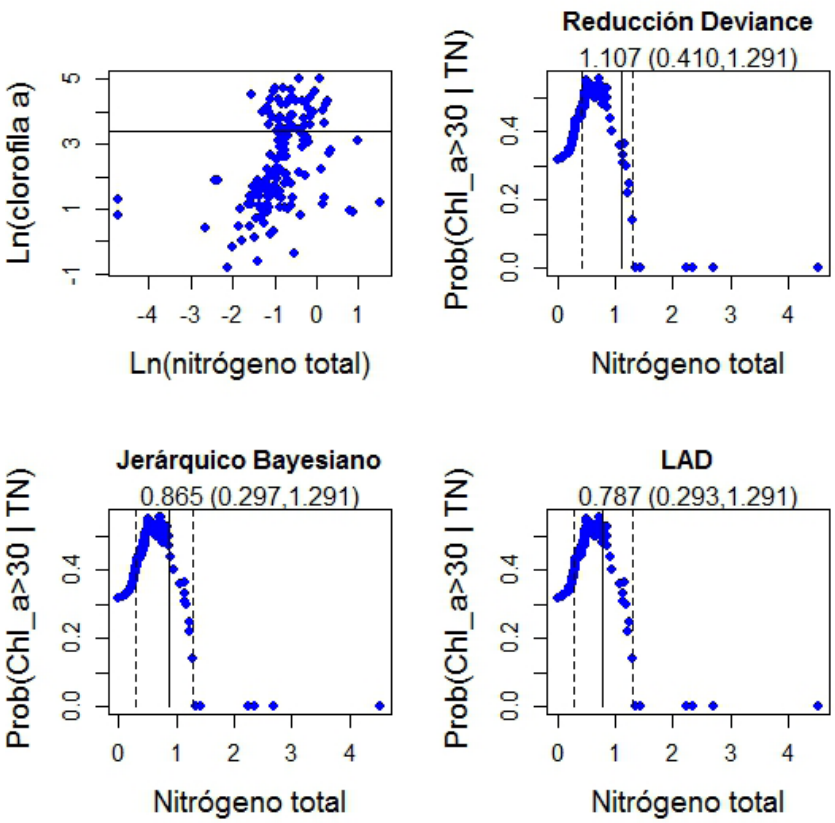


Fig. 7. Umbrales para nitrógeno total cuando chl_a > 30. Bajo las técnicas reducción de la deviance, jerárquico bayesiano, metodología lad.

Se puede observar en las figuras 2, 3, 4, 5, 6 y 7 cómo los métodos se aproximan en la determinación de los umbrales y en los intervalos de confianza dependiendo del comportamiento de la probabilidad condicional y de los valores atípicos. Para el caso de las gráficas de las (*figuras 2 y 5*) describen los comportamientos de la probabilidad condicional para *Chl_a >2*, estos umbrales para este límite trófico son muy similares entre los métodos y no muestran diferencias notables, en gran parte esto se debe a que las curvas de probabilidad condicional son bastante regulares.

En las gráficas (*figuras 3,4,6 y 7*) se presentan valores atípicos, estas observaciones causan diferencias en los resultados al determinar los umbrales en cada uno de los métodos, debido a que algunos de ellos son muy sensibles a la presencia de estos datos y por lo tanto, presentan más variación en sus resultados y mayor amplitud en sus intervalos de confianza.

Se puede observar que la metodología LAD es menos sensible a la presencia de estas observaciones, los umbrales y sus intervalos de confianza no son fuertemente afectados.

CONCLUSIONES

Se concluye que mediante los métodos intervalos de confianza que no se traslapan y el Ajuste de un modelo no lineal no han sido procesos muy útiles en la determinación de los umbrales para embalses en Puerto Rico. Algunas de las posibles causas, en la naturaleza no siempre se cumplen los comportamientos que plantea la literatura y además, los pocos datos hacen más frecuente este tipo de situaciones. Los métodos (Jerárquico Bayesiano, Reducción de la Deviance y Metodología LAD), los umbrales son comparables solo en los límites oligo- mesotrófico. En los límites meso-eutrófico y eu-hipereutrófico hay diferencia, y LAD termina siendo la metodología más robusta, menos sensible a datos atípicos.

Solo los umbrales determinados por LAD mantienen la monotonicidad para ambas variables. Propiedad necesaria dada la relación monótona

entre el umbral de *Chl_a* y TN, TP. El estimador Bootstrap resultó ser un estadístico de mejor aproximación a los umbrales esperados, que los determinados solo a partir de una muestra.

Por último, cabe destacar que la reducción de la devianza no paramétrica es el método que más se ve afectado por valores atípicos a pesar de ser una de las metodologías más utilizadas para la determinación de umbrales.

AGRADECIMIENTOS

A Dios por darme la oportunidad de trabajar en este proyecto, y por los resultados obtenidos. A mi familia por la paciencia y la confianza, a mi esposa por ser una motivación más en mis logros.

CONFLICTO DE INTERESES

El autor expresa que no existen conflictos de intereses y acepta todo el contenido.

REFERENCIAS BIBLIOGRÁFICAS

1. USEPA (United States Environmental Protection Agency). National Lakes Assessment: A Collaborative Survey of the Nation's Lakes. Washington, D.C. 2009; EPA 841-R-09-001.
2. Paul, J. F. y McDonald, M. E. Development of empirical, geographically specific water quality criteria: A conditional probability
3. Qian, S. S., King, R. S., y Richardson, C. J. Two statistical methods for the detection of environmental thresholds. Journal of Ecological Modelling. 2003; 166(1-2), 87-97.
4. Martínez Suárez, A. Métodos Estadísticos para la Detección de Umbrales de Contaminación en Ecosistemas Acuáticos de Agua Dulce. M.S. Thesis, University of Puerto Rico-Mayagüez. 2010. 10.
5. Ritz, C. y Streibig, J. C. Nonlinear Regression with R. New York: Springer. 2008.
6. Faraway, J. J. Extending the linear Model with R. New York: Chapman & Hall. 2006.
7. Efron, B. y Tibshirani, R. An Introduction to the Bootstrap. New York: Chapman & Hall. 1993.
8. Crawley, M. The R Book. Londres. Wiley. 2007.
9. Dobson, A. J. y Barnett, A. An Introduction to Generalized Linear Models. (3a. ed.) New York: Chapman & Hall. 2009.
10. Dalgaard, P. Introductory Statistics with R. (2a. ed.) New York: Springer. 2008.
11. Paul, J. F. y McDonald, M. E. Development of empirical, geographically specific water quality criteria: A conditional probability analysis approach. Journal of American Water Resources Association. 2005; 41(5), 1211-23.