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

ALEXANDER MARTÍNEZ SUÁREZ

TESIS SOMETIDA EN CUMPLIMIENTO PARCIAL DE LOS REQUISITOS PARA EL GRADO DE

MAESTRÍA EN CIENCIAS

EN MATEMÁTICAS (ESTADÍSTICA)

UNIVERSIDAD DE PUERTO RICO

RECINTO UNIVERSITARIO DE MAYAGÜEZ

DICIEMBRE 2010

Aprobado por:

_____	_____
Macchiavelli, Raúl, PhD	Fecha
Presidente, Comité Graduado	
_____	_____
Acuña Fernández, Edgar, PhD	Fecha
Miembro, Comité Graduado	
_____	_____
Quintana Díaz, Julio, PhD	Fecha
Miembro, Comité Graduado	
_____	_____
Alston, Dallas, PhD	Fecha
Representante de Estudios Graduados	
_____	_____
Colón, Silvestre, MS	Fecha
Director del Departamento Ciencias Matemáticas	

ABSTRACT

This paper presents the methodology and results from applying statistical methods to determine the thresholds of contamination of aquatic ecosystems. The objectives are to analyze methods to determine thresholds or turning points for a given variable by using the statistical software program **R** and to use the methods to determine pollutant thresholds in reservoirs in Puerto Rico.

The first part presents these methods, including their theoretical foundations, computational mathematics, and application to streams in the Mid-Atlantic states (USA) during 1993 and 1994. The objective is to compare the predicted results with the published results. The second part applies the methods to real data from Puerto Rico.

One of the statistical tools used to develop each of the methods is conditional probability, which requires two variables, Q and X. One variable, Q, represents the known threshold; the other variable X represents the unknown threshold of interest. Given the two variables, the process uses the conditional probability $P(q|x)$ and one of the methods to determine the threshold of interest.

In the Mid-Atlantic data, the benthic macroinvertebrate communities (EPT taxa richness) data were used to determine stream impact (the known variable threshold). The percentage of fine substrate (fraction silt/clay$<$ 0.06mm) was used to measure stream sedimentation as the variable of interest. With these two variables, five methods were applied to determine the fine substrate threshold percentage for these streams.

With respect to the data in Puerto Rico, chlorophyll a is the variable for which the threshold is known. The variables of interest for which the thresholds were determined were nutrients: total phosphorus and total nitrogen.

RESUMEN

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 son: 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.

En la primera parte del trabajo se desarrolla la presentación de estos métodos, su fundamentación teórica, matemática y computacional, y la aplicación en arroyos de la región de los estados Mid-Atlantic (Estados Unidos) durante 1993 y 1994. El objetivo es comparar los resultados predichos con los publicados. La segunda parte del trabajo consiste en aplicar los métodos a datos reales 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, a una de ellas se le conoce el umbral. Supóngase Q y llámese X la variable de interés, variable a que se le desea conocer el umbral. El proceso consiste en, dadas las dos variables, utilizar la probabilidad condicional $P(Q|X)$ y mediante los métodos, determinar el umbral de la variable de interés X.

En los datos del Mid Atlantic en los Estados Unidos, se utilizaron las comunidades de macroinvertebrados bénticos (riqueza de taxones EPT) como medida de referencia de si existe impacto en el arroyo (umbral de la variable conocida); y se usó como variable de interés X, el porcentaje de sustrato fino (fracción limo/arcilla$<$ 0.06mm) como un indicador de la sedimentación en los arroyos. A estas dos variables se aplican cinco métodos para determinar el umbral del porcentaje de sustrato fino en estos arroyos.

En lo que respecta a los datos en Puerto Rico, la clorofila a es la variable de la

cual se conoce el umbral, y se analizan como variables de interés los nutrientes fósforo total y nitrógeno total, para determinar el umbral en cada uno de ellos.

A mi señor Dios que siempre me llena de triunfos.

A mi familia en Colombia que ha sido fundamental

en el ánimo la fuerza y las ganas de poder

terminar mi maestría, en especial a mis padres por su entrega

fuerte y la confianza que depositan en mi. A mis hermanas

y a mi bella y hermosa Zareth Camila. A mi novia por su dedicación y amor.

Gracias Dios por todo.

AGRADECIMIENTOS

Primero que todo a Dios, siempre me ha dado lo mejor y me ha mostrado de mil maneras las gracias que me concede y la familia tan bella que puso a mi vida.

En segunda instancia, al Doctor Raúl Macchiavelli por su ejemplo de trabajo, por su colaboración por todo lo que pude aprender de él y por la gran persona que lo caracteriza. Un fuerte agradecimiento al Doctor Raúl.

Por último a todos mis compañeros y los docentes del departamento de Matemáticas que de diversas maneras me enseñaron a ser mejor persona y mejor profesional. Gracias.

Índice general

Índice de figuras

Índice de cuadros

Índice de algoritmos

Capítulo 1

Introducción

El estudio de métodos para determinar umbrales de contaminación en fuentes acuáticas, es una propuesta que trae consigo diversos beneficios en temas como la salud humana, la investigación estadística, el mejoramiento del ecosistema y la protección de la vida animal y vegetal en los arroyos, debido a que estos métodos promueven la formulación de criterios para garantizar la protección de la vida acuática. En Estados Unidos, existen agencias de protección ambiental como la USEPA (*United States Environmental Protection Agency*), que es la responsable de la implementación de leyes en la calidad del agua y en proveer estándares de calidad. Otra agencia es la Junta de Calidad Ambiental de Minnesota EQB (*Environmental Quality Board*), encargada de desarrollar las políticas, crear planes a largo alcance y revisar los proyectos propuestos que hayan podido influir en el medio ambiente de Minnesota. En Puerto Rico, se encuentra la Junta de Calidad Ambiental JCA que es un organismo dedicado al control de la contaminación, la degradación ambiental y la protección del aire, las aguas y los suelos.

Para el desarrollo de cada uno de estos métodos, se hace necesario el conocimiento de dos variables que estén correlacionadas en el contexto acuático. Si se denota a estas

dos variables como Q y X, una debe ser una variable de la cual se conozca el umbral y la otra, la variable de interés (la variable de la cual se desea determinar el umbral). Sea Q la variable de la cual se conoce el umbral y X la variable de interés. Un primer paso consiste en definir una variable Z, a partir de la variable Q. Si q_c es el umbral de Q, la variable Z se define como: $z = 1$ si existe impacto (por ejemplo $Q \leq q_c$) y $z = 0$ si no hay impacto ($Q > q_c$). Por lo tanto, Z es dicotómica.

Un segundo paso consiste en estudiar la probabilidad condicional, que es el instrumento en el análisis y la comprensión del proceso al determinar el umbral en cada uno de los métodos. En Casella y Berger (2002), se define la probabilidad condicional como

Definición 1.1. Sean A y B dos eventos en un espacio S, y $P(B) > 0$, la probabilidad condicional de A dado B, se escribe como $P(A \mid B)$, y es

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \tag{1.0.1}$$

donde $P(A)$, es la probabilidad de que ocurra el evento A.

Sea (x_1,q_1), (x_2,q_2),...,(x_n,q_n) una muestra de (X,Q). En cada uno de los métodos propuestos, se determina la probabilidad condicional de Q para cada uno de los valores mayores a x_i. La probabilidad condicional que se calcula es $p_i = P(Z = 1 \mid X > x_i)$, que significa la probabilidad que exista impacto $Z = 1$, dado que X ha excedido el valor x_i. La desigualdad en la probabilidad condicional depende del tipo de relación que se establece en la dos variables, lo que se explicará con más detalle en el capítulo 3.

Luego de calcular la probabilidad condicional, que en cada uno de los métodos es la variable respuesta, se analizan cada uno de los métodos: intervalos de confianza que no se traslapan, el ajuste de un modelo no lineal a la curva empírica de la probabilidad condicional (ajuste de una regresión no lineal por mínimos cuadrados), la técnica de la

reducción de la *deviance* no paramétrica, el ajuste de un modelo jerárquico Bayesiano y por último, como aporte de este estudio, la metodología LAD (desviación absoluta mínima). A continuación se presenta una breve descripción de los métodos.

El método de los intervalos de confianza que no se traslapan consiste en construir intervalos de confianza de la probabilidad no condicional y condicional mediante la técnica *Bootstrap*, y luego observar en qué punto dejan de traslaparse los intervalos de la probabilidad condicional con el de la probabilidad incondicional.

El método de ajuste de un modelo no lineal es un método que consiste en hacer una partición de la variable de interés y a cada partición asignarle un modelo no lineal a la probabilidad condicional que difiera en ciertos parámetros. Uno de estos parámetros es el punto de cambio de la curva de probabilidad condicional, que sería el umbral.

El método de la reducción de la *deviance* no paramétrica consiste en partir la curva de probabilidad condicional en dos grupos. El punto de división que minimiza la *deviance* no paramétrica determina el umbral.

El ajuste de un modelo jerárquico bayesiano, consiste en asumir que cada valor de la variable respuesta (probabilidad condicional) es una variable aleatoria que pertenece a una misma familia de distribución con parámetro θ, y suponer un punto de cambio en la variable respuesta, es decir, dos distribuciones de probabilidad que solo difieren en sus parámetros.

Por último, la metodología LAD basa su procedimiento en particiones de la variable de interés, sólo que el método utilizado para la bondad de ajuste es la desviación absoluta mínima , en vez de la *deviance* no paramétrica.

En algunos de estos métodos se hace indispensable aplicar técnicas como *Bootstrap* o *Gibbs sampler* para obtener el umbral. La técnica del *Bootstrap* aplicado a una

muestra $x_1, x_2, .., x_n$ es llamado remuestreo, debido que es una técnica estadística basada en una distribución discreta que da una probabilidad de $1/n$ a todos los puntos x_i y 0 a cualquier otro valor. El procedimiento *Bootstrap* utiliza esta distribución como sustituto de la verdadera distribución para construir, mediante muchos *Bootstrap*, estimaciones de varianza e intervalos de confianza.

La técnica del *Gibbs sampler* es una clase de algoritmo de la cadena de *Markov Monte Carlo* (MCMC), y es utilizado para generar una secuencia de muestras de la distribución de probabilidad conjunta de dos o más variables aleatorias, cuyo propósito es la aproximación a la distribución conjunta, o la aproximación a la distribución marginal de una de las variables, o algún subconjunto de las variables (por ejemplo, parámetros desconocidos o variables latentes), o el calcular una integral que podría ser el valor esperado de una de las variables.

Cada uno de estos algoritmos o procedimientos son explicados detalladamente en el capítulo 3. El programa estadístico utilizado en este trabajo es **R** y los algoritmos se presentan en el lenguaje del programa (Dalgaard, 2008). Estas técnicas se aplican a datos reales en embalses de Puerto Rico, mostrando su utilidad para diversas variables de interés.

Capítulo 2

Revisión de Literatura

Las técnicas utilizadas en este trabajo se han implementado en investigaciones de ecosistemas acuáticos y más comúnmente la reducción de la *deviance* no paramétrica.

2.1. Punto de Cambio (Umbral)

El análisis de punto de cambio es un método para identificar los umbrales en la relación entre dos variables. Más concretamente, es un método analítico que busca determinar un punto a lo largo de una distribución de valores donde las características de los valores antes y después del punto son diferentes. El análisis del punto de cambio se puede utilizar para identificar ese punto a lo largo del eje X donde las características a lo largo del eje Y cambian, como por ejemplo un cambio en el promedio de variación o cambio en la pendiente.

El punto de cambio es un método estadístico para identificar los umbrales y es esencial para el desarrollo de criterios de nutrientes. En el caso donde la variable respuesta

es la concentración de clorofila a y la predictora el fósforo total (TP), el análisis del punto de cambio puede ser utilizado para identificar la concentración promedio TP en que la clorofila a cambia de manera significativa antes y después del umbral de TP.

Se pueden seleccionar una variedad de enfoques para la determinación del punto de cambio como: Intervalos de Confianza que no se traslapan, el ajuste de un modelo no lineal (si los datos se ajustan), la reducción de la *deviance* no paramétrica, estimación bayesiana, entre otras. Se escogen los umbrales que dan lugar a una reducción sustancial en el porcentaje de error, así que la mayoría de las técnicas deben incluir alguna medida de precisión.

2.2. Ventajas y Desventajas de los Métodos

Es interesante conocer las ventajas y desventajas que se tienen a la hora de trabajar con métodos paramétricos y no paramétricos. Como la mayor parte del trabajo se desarrolla con métodos no paramétricos: Intervalos de confianza que no se traslapan, reducción de la *deviance* no paramétrica y la metodología LAD, la tabla 2.1 muestra algunas ventajas y desventajas de estos métodos, lo que permite afianzar ciertos resultados y explicar ciertas situaciones, como por ejemplo, no todos los métodos funcionan de manera muy útil, existen ciertos limitantes que en su mayor parte dependen del comportamiento de los datos de estudio.

Ventajas	Desventajas
Identificación clara de los umbrales	Cómputo intensivo
Libres de distribución y supuestos	No hay una fórmula que se produce
Obtención de los intervalos de confianza	

Tabla 2.1: Ventajas y desventajas de métodos no paramétricos.

2.3. Intervalos de Confianza que no se Traslapan

Existen varios trabajos realizados en lo que respecta a este método y a su aplicación, por ejemplo, se estudia la relación entre los Intervalos de Confianza (IC) y las Pruebas de Hipótesis, indicando: Si dos IC al 95 % se superponen, las medias pueden ser significativamente diferentes entre sí, en el nivel del 0.05 y concluye "*two means may be significantly different from one another, despite the two confidence intervals abutting or having a modest degree of overlap*" (Austin y Hux, 2002).

En Cumming (2009) se exponen los debates ya realizados en torno a IC que se traslapan y se amplía la investigación a p valores distintos de 0.05 y 0.01. Además, se estudian los IC del 95 % en 2 proporciones y en 2 correlaciones, encontrando propiedades similares que se aplican a la superposición de estos IC asimétricos.

Otro documento y soporte de este trabajo es el realizado en Paul y McDonald (2005) en el se determina el umbral de contaminación para el porcentaje de sustrato fino mediante este método. En el capítulo 3 se estudia con más detalle este documento y el respectivo método.

2.4. Ajuste de un Modelo No lineal

El método del ajuste de un modelo no lineal para determinar umbrales, consiste en ajustar un modelo no lineal a la curva empírica creada por la probabilidad condicional, donde uno de los parámetros es el punto de cambio de la curva, es decir, el umbral a determinar. Los parámetros se determinan por mínimos cuadrados no lineales. En Venables and Ripley (2002) se presenta un capítulo completo a la forma de estructurar este tipo de modelos, al análisis y a la comparación de los parámetros.

En Paul y McDonald (2005) se ajusta un modelo no lineal a la probabilidad condicional para determinar el umbral de impacto. Uno de los parámetros del modelo no lineal es un punto x_0 (umbral) que parte la curva de probabilidad en dos y establece parámetros diferentes en la partición. El modelo se especifica y se analiza con más detalle en el capítulo siguiente.

Para un conocimiento más profundo del manejo de este tipo de modelos, en Ritz y Streibig (2008) se realiza un análisis bien detallado de las diversas formas de modelos no lineales, la interpretación de los parámetros, los comandos a utilizar de acuerdo al modelo y a los objetivos que se pretenden.

2.5. Reducción de la *Deviance* no paramétrica

La *deviance* se define como una medida de homogeneidad y dependiendo del tipo de variable, existe una forma diferente de calcularla. Para una variable continua $D = \sum_{i=1}^{n}(y_k - \mu)^2$, donde D es la *deviance* no paramétrica, n el tamaño de la muestra, μ es la media de las n observaciones y_k. Si la variable es categórica, la *deviance* no paramétrica está definida como $D = -2\sum_{k=1}^{g} n_k log(p_k)$, donde g es el número de clases, p_k es la proporción y n es el número de observaciones en la clase k, respectivamente. La idea de usar la reducción de la *deviance* no paramétrica como método para determinar un umbral emerge del modelo de árbol de regresión. Segal (2008) da una explicación detallada de este tipo de regresión, fundamentada en los esfuerzos realizados en la elaboración de técnicas de regresión libres de algunos supuestos clásicos restrictivos.

El método de la reducción de la *deviance* no paramétrica consiste en dividir la variable respuesta (probabilidad condicional) en dos grupos, las submuestras formadas de esta manera se llaman nodos. Se escoge un criterio de división de buen ajuste, el cual se utiliza para evaluar el criterio de homogeneidad dentro de los nodos, que para este caso

sería la *deviance* no paramétrica. El punto donde se produzca la menor *deviance* es el punto de cambio que determina el umbral.

En Paul y McDonald (2005) se aplica este procedimiento a la probabilidad condicional para determinar el umbral de contaminación. Quian et al. (2003) estudian el método de la reducción de la *deviance* no paramétrica, un análisis más detallado de su procedimiento y del porqué este método no paramétrico puede usarse para calcular el umbral de una variable.

Otro estudio que hace uso de la reducción de la *deviance* no paramétrica es Peterson et al. (1999) quienes a partir de datos del interior de la cuenca del Río Columbia, en Estados Unidos, comparan cuatro métodos para predecir la distribución de siete taxones de salmónidos utilizando la información del paisaje. Una de las técnicas es el árbol de regresión, y para ello utilizan el método de la reducción de la *deviance* no paramétrica para datos categóricos. El documento expone que existen diversas medidas para la homogeneidad dentro de las particiones y para el estudió se utilizó la reducción de la *deviance* no paramétrica. La finalidad del método, es determinar la partición que produce la mayor reducción de la *deviance* no paramétrica.

2.6. Ajuste del Modelo Jerárquico Bayesiano

La formulación de un modelo jerárquico Bayesiano consiste en asumir que los valores de la variable respuesta $p_1, ..., p_n$, a lo largo de un gradiente de interés X, son muestras aleatorias de una secuencia de variables aleatorias $P_1, ..., P_n$. Para cada variable se define una distribución y se asume que todas pertenecen a una misma familia de distribuciones con parámetros θ. El proceso consiste en, dadas n variables aleatorias $P_1, ..., P_n$ existe un punto de cambio r, $(1 \leq r \leq n)$, donde r es el punto donde la

distribución de los P cambia en sus parámetros. Este punto de cambio r en la variable predictora x_r, es el umbral.

En Quian et al. (2003) uno de los métodos que se proponen para la detección de un umbral de contaminación es el método Jerárquico Bayesiano, el cual supone para la variable respuesta una distribución de probabilidad, cuyos parámetros tienen a su vez distribuciones de probabilidad. El objetivo en este método es determinar la distribución de probabilidad que representa el punto de cambio. La aplicación en este documento se realiza en macroinvertebrados como medida de referencia para la contaminación, debido a sus cambios corporales cuando son expuestos a contaminantes. El estudio se desarrolla en humedales del Everglades, U.S.A, donde el fósforo es un nutriente limitante.

En Perreault et al. (2000) se ofrece un enfoque bayesiano para caracterizar cuándo y en qué medida un único cambio se ha producido en una secuencia de variables aleatorias hidrometeorológicas. Las inferencias están basadas en el análisis de las distribuciones posteriores.

Para un análisis más detallado y profundo del método Bayesiano, en Casella y Berger (2002) se suponen distribuciones previas de los parámetros desconocidos del modelo. Se estudian con detalle supuestos de normalidad y de Poisson.

2.7. Estimación del Umbral

La metodología que se utiliza para determinar la estimación del umbral y su intervalo de confianza en la variable de interés, depende del método utilizado para calcular el umbral y de la técnica *Bootstrap*. El procedimiento del *Bootstrap* es una técnica propuesta para hallar intervalos de confianza en situaciones donde es imposible determinar analíticamente la distribución muestral del estimador. Es una técnica de remuestreo, de

cómputo intensivo. El *Bootstrap* consiste fundamentalmente en tratar la muestra como si fuera la población y aplicar un muestreo con reposición para generar una estimación empírica de la distribución muestral del estadístico. Al ser una técnica no paramétrica, el *Bootstrap* tiene la ventaja de que no precisa conocer la función de distribución teórica de los datos.

En Efron y Tibshirani (1993) se describen los pasos básicos en la estimación *Bootstrap*:

1. Se construye una distribución de probabilidad $\hat{P}(x)$ empírica a partir de la muestra disponible, asignando una probabilidad de $1/n$ a cada punto $p_1, ..., p_n$. Esta es la función de distribución empírica (FDE) de P, que constituye el estimador no paramétrico de máxima verosimilitud de la función de distribución de la población, $P(x)$.

2. Partiendo de $\hat{P}(x)$ se extrae una muestra aleatoria simple con reemplazo de tamaño n.

3. A partir de la muestra obtenida en el paso 2, se calcula el estadístico de interés $\hat{\theta}$, dando $\hat{\theta}_b^*$.

4. Se repite B veces los pasos 2 y 3. La magnitud de B depende de las pruebas que se van a aplicar a los datos. En general, B varía entre 50 a 200 para estimar el error estándar de $\hat{\theta}$ y es mayor que 1000 para estimar intervalos de confianza alrededor de $\hat{\theta}$ o si el parámetro es un percentil extremo de la distribución.

5. Se construye una distribución de probabilidad $\hat{\theta}_b^*$ a partir de los B *Bootstrap*, asignando una probabilidad de $1/B$ a cada punto $\hat{\theta}_1^*, \hat{\theta}_2^*, ..., \hat{\theta}_b^*$. Esta distribución es la estimación *Bootstrap* de la distribución muestral de $\hat{\theta}$ y puede usarse para hacer inferencias sobre θ. El estimador *Bootstrap* del parámetro se define como la media de

los valores del estadístico calculados en las B remuestreos *Bootstrap* y su expresión es:

$$\hat{\theta}^*_{(.)} = \frac{\sum_{b=1}^{B} \hat{\theta}^*_b}{B}$$

Por ejemplo, si $\hat{\theta}$ es la mediana muestral, entonces $\hat{\theta}^*_{(b)}$ es la mediana de cada muestra *Bootstrap*.

Este procedimiento se utiliza para estimar el umbral y los intervalos de confianza del 95 %. El umbral se determina como el estimador *Bootstrap,* es decir como la media de los estimadores *Bootstrap*, y los intervalos de confianza del 95 % como los percentiles correspondientes de la secuencia ordenada de estimadores *Bootstrap*.

El algoritmo 2.1 calcula el umbral para la variable de interés con el método de la reducción de la *deviance* no paramétrica, este algoritmo calcula el estimador *Bootstrap* del umbral y su intervalo de confianza del 95 %. El comando *sample* en el programa **R** realiza el remuestreo de los datos y *Deviance* es la función (ver algoritmo 3.4) que calcula el umbral mediante el método de reducción de la *deviance* no paramétrica. El algoritmo funciona incluyendo los datos de estudio, el umbral de la variable que se conoce y el número de remuestreos que se van a utilizar. Este mismo procedimiento se aplica para cada uno de los otros métodos.

Algoritmo 2.1 Estimación *Bootstrap*.

```
Inter_Dev=function(datos, umbral, n_resam)
{ data=datos
w=length(data[,1])
resultado=matrix(0,1,n_resam)
m=1
while(m<=n_resam) {
# Se calcula el umbral por la reducción de la deviance no paramétrica para cada remuestreo
x=Deviance(data,umbral)
# Se almacenan los resultados de los umbrales
resultado[1,m]=x
# Se realizan los remuestreos a los datos
data=datos[sample(w,w,replace=T),]
m=m+1 }
y=resultado[1,]
# Se calcula la media de los umbrales calculados en cada remuestreo
medi=mean(y,na.rm=TRUE)
y1=sort(y)
nobs=length(y1)
nboot=n_resam
level=95
alpha=1-.01*level
kperc=floor((nboot+1)*alpha*.5)
# Los intervalos de confianza
l.ic=y1[kperc]
u.ic=y1[nboot+1-kperc]
cat("El intervalo de confianza para el umbral", resultado[1,1]," del",level," % con una media ",medi," es:(",l.ic,",",u.ic,")\n") }
```

Capítulo 3

Métodos Para Determinar Umbrales

En este capítulo se realiza un análisis detallado de cada una de los métodos para determinar el umbral de una variable de interés. El análisis consiste en el desarrollo matemático y estadístico de cada método, en su algoritmo de programación y de un ejemplo aplicado a datos recolectados en la región de los estados Mid Atlantic durante 1993 y 1994 (ver anexo 1). Antes de realizar el estudio a los métodos, se dedica una sección a la explicación de las variables y su relación en el medio acuático.

3.1. Datos del Mid Atlantic

Para lograr una interpretación adecuada de los resultados que se obtienen en estadística, es importante conocer la naturaleza de los datos. Para la aplicación de los métodos de detección de umbrales es necesaria una observación detallada del comportamiento de las variables y del tipo de variables que se están analizando.

Los datos utilizados en este capítulo están disponibles en el sitio web USEPA (2005). Estos datos se recolectaron de arroyos en la región de los estados Mid Atlantic

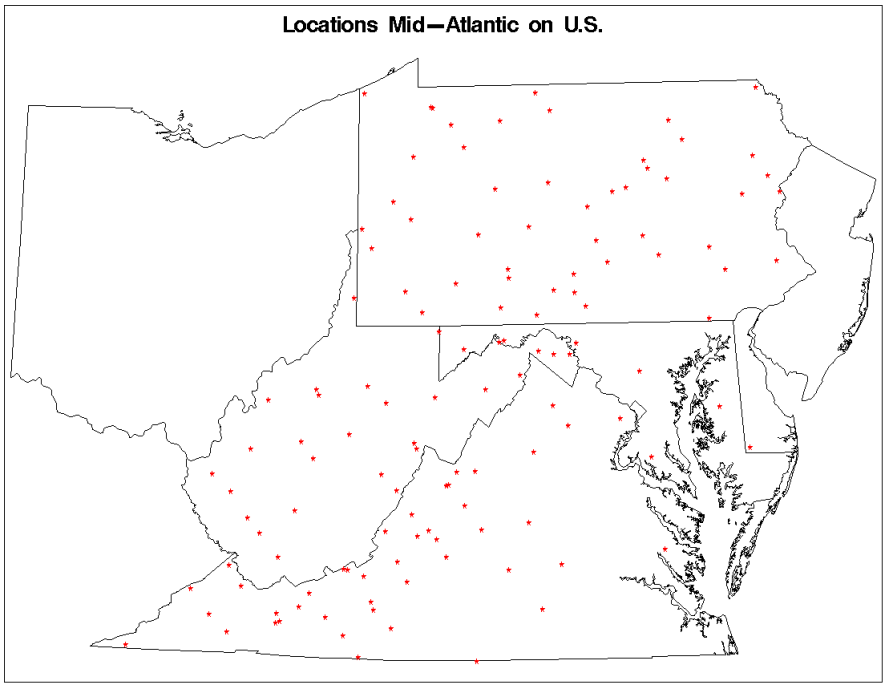

Figura 3.1.1: Los estados Mid Atlantic de USA. Sitios de las muestras usadas en el estudio.

en Estados Unidos durante 1993 y 1994 (ver figura 3.1.1). Son 99 segmentos de arroyos seleccionados aleatoriamente. El muestreo se realizó durante dos meses cada año, de abril a mediados de junio. Los datos se restringen a segmentos de corriente con piscinas, a sitios visitados una sola vez y a segmentos de primer a tercer orden. Las dos variables analizadas en este estudio son la riqueza en los taxones EPT (EPT_RICH) y el porcentaje de sustrato fino (PCT_FN).

Las EPT son comunidades bénticas de macroinvertebrados. Estas comunidades son una medida robusta de las condiciones del arroyo y están compuestas por los órdenes Ephemeroptera, conocidas como Efímeras, Plecoptera, conocidas comúnmente como *stoneflies* y Tripchoptera comúnmente tricópteros, que colectivamente se denominan EPT. Estos invertebrados reaccionan sensiblemente a los cambios en los niveles de sedimentación, exhibiendo un decrecimiento en su riqueza con el aumento de la sedimentación en el agua. El umbral en la riqueza de taxones EPT utilizado para identificar impacto

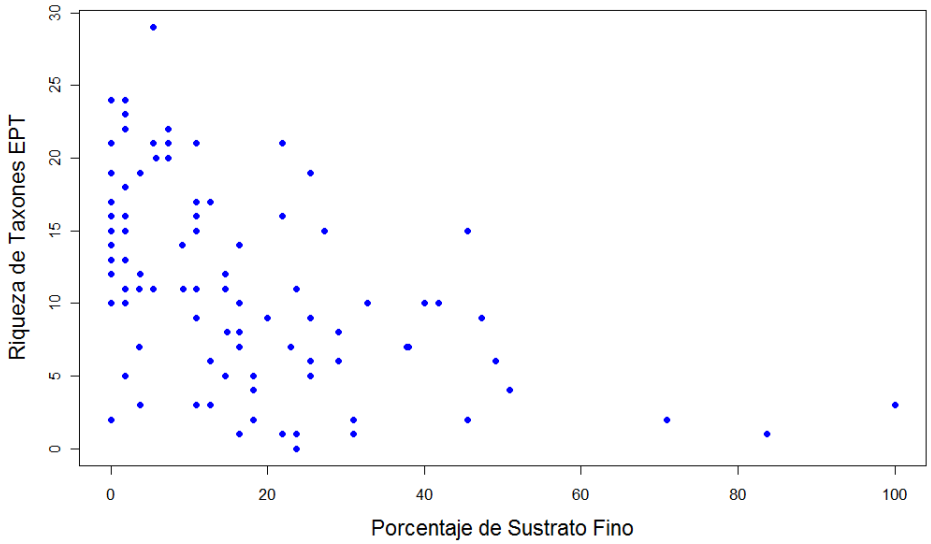

Figura 3.1.2: Riqueza de taxones EPT vs. Porcentaje de sustrato fino. Diagrama de dispersión de las muestras en los estados Mid Atlantic. Si la riqueza de taxones EPT<9, el arroyo está impactado.

en los segmentos de arroyo en Mid Atlantic es 9. Cuando la riqueza de taxones es menor que ese valor, el segmento de arroyo se considera como **impactado**.

El porcentaje de sustrato fino es usado como un indicador sustituto de la sedimentación. Las consecuencias con el aumento del porcentaje de sustrato fino son: turbidez, la disminución en la penetración de la luz, la reducción de los espacios que podrían ser usados para la reproducción, la alimentación, la cubierta para invertebrados y la disminución de peces. Todo esto es un indicador de un ambiente en no muy buenas condiciones.

La figura 3.1.2 muestra el comportamiento de las variables riqueza de taxones EPT y porcentaje de sustrato fino de los datos del Mid Atlantic. Se observa que a medida que aumenta el porcentaje de sustrato fino, disminuye la riqueza de EPT. Es decir, las dos variables están correlacionadas negativamente.

3.2. Procedimiento General: Probabilidad Condicional

Antes de dar a conocer los procedimientos de cada método, es necesario exponer el proceso general que se desarrolla en cada uno de ellos.

La probabilidad condicional relaciona las dos variables de estudio, es el procedimiento general que se desarrolla en cada uno de los métodos y es considerada como la variable respuesta de análisis. La notación usual a la probabilidad de observar un evento x está dada por $P(x)$.

La probabilidad condicional se define como la probabilidad de que un evento y ocurra, dado que un evento x ha ocurrido, y se denota por $P(y|x)$. Para efectos del análisis de las variables, por ejemplo para los datos del Mid Atlantic, Paul y McDonald (2005) analizan la probabilidad condicional $P(z = 1|x > x_i)$, que significa la probabilidad de que la riqueza de taxones EPT sea menor que 9 ($z = 1$), dado que el porcentaje de sustrato fino ha superado el valor x_i (ver figura 3.2.1). La probabilidad condicional se asume como la variable respuesta $p_i = P(z = 1|x > x_i)$. La dirección de la desigualdad de la probabilidad condicional está sujeta al comportamiento de la variable X y en su respuesta al impacto biológico. Para este caso la probabilidad condicional se expresa con la desigualdad planteada, debido a que la variable X aumenta y la posibilidad de impacto en las condiciones biológicas aumenta (ver figura 3.1.2). Si la variable de interés se comporta de forma distinta, es decir, si cuando ésta aumenta y la posibilidad de contaminación disminuye, la dirección de la desigualdad en la probabilidad condicional se invierte, y se analiza $P(z = 1|x < x_c)$.

La inclusión de la probabilidad condicional en la aplicación de los métodos es interesante, primero por que provee un único valor para cada punto específico de X (es común encontrar varias muestras en un mismo sitio o en sitios diferentes, con la misma observación de la variable de interés). Segundo, porque al determinar el punto de cambio

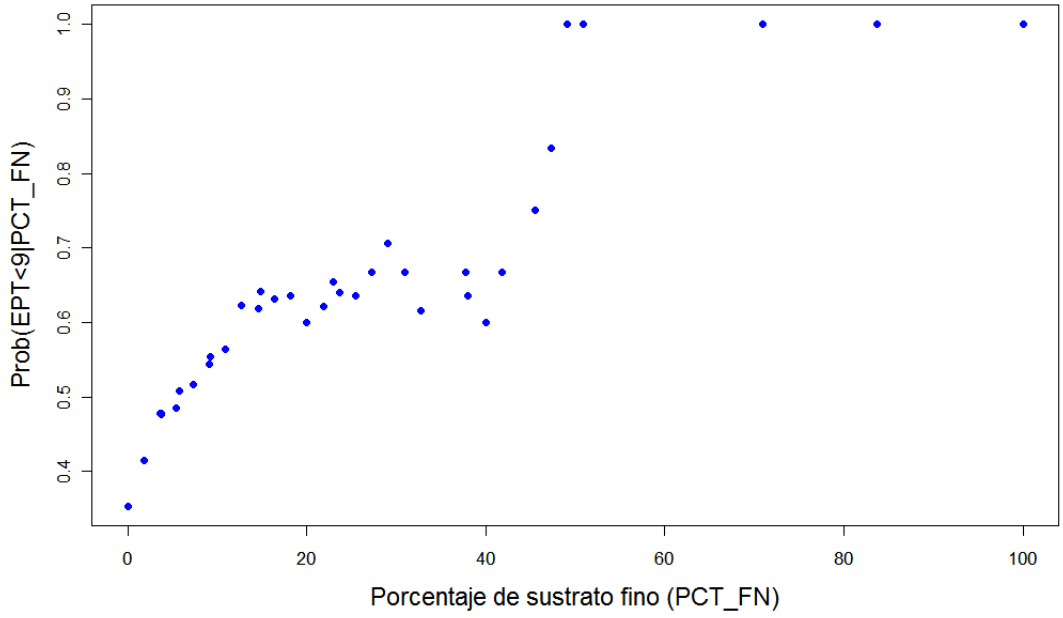

Figura 3.2.1: Probabilidad Condicional $P(\text{EPT}< 9|x > x_i)$ vs. Porcentaje de sustrato fino. Probabilidad que un arroyo este impactado, dado el porcentaje de sustrato fino. Correlación positiva entre el impacto y porcentaje de sustrato fino.

en la probabilidad condicional, se está hallando un punto que divide la probabilidad condicional en dos grupos que no son homogéneos por el nivel de impacto. El punto de cambio para la probabilidad condicional en estos datos, significa que si el porcentaje de sustrato fino es mayor para ese punto, el impacto en la riqueza de taxones EPT es diferente que para valores menores a ese punto. Las probabilidades obtenidas para los datos del Mid Atlantic se muestran en la figura 3.2.1. Sus correspondientes valores están en el anexo 2.

3.3. Intervalos de Confianza que no se Traslapan

Uno de los propósitos es analizar la relación entre los IC y las pruebas de hipótesis, además ilustrar IC que se traslapan y medias que son significativamente diferentes.

Antes de proceder al método como tal, se deja claro que para dos medias de dos grupos diferentes donde sus intervalos de confianza se traslapan no es generalmente cierta la afirmación que las medias no sean significativamente diferentes. El siguiente análisis muestra dos medias donde sus intervalos de confianza se traslapan y las medias resultan ser estadísticamente diferentes.

Supóngase dos muestras independientes, cada una de tamaño n, donde $\overline{x}_1$ y $\overline{x}_2$ son las medias de la primera y segunda muestra, respectivamente. Si se asume varianzas iguales (para simplificar el caso), $\overline{x}_1 < \overline{x}_2$ y p la proporción en que los intervalos se superponen.

El ancho del IC del 95 % es $2 \cdot 1.96 \cdot \sigma/\sqrt{n}$, entonces

$$\overline{x}_1 + 1.96 \cdot \sigma/\sqrt{n} = \overline{x}_2 - 1.96 \cdot \sigma/\sqrt{n} + p \cdot 2 \cdot 1.96 \cdot \sigma/\sqrt{n} \tag{3.3.1}$$

Reorganizando la diferencia de medias

$$\overline{x}_2 - \overline{x}_1 = 2 \cdot 1.96 \cdot \sigma/\sqrt{n} - p \cdot 2 \cdot 1.96 \cdot \sigma/\sqrt{n} \tag{3.3.2}$$

Ahora, para probar la hipótesis de que las medias son iguales en los dos grupos, se calcula la prueba z de dos muestras independientes, con varianzas iguales y conocidas. La prueba estadística z es:

$$z_{test} = \frac{\overline{x}_2 - \overline{x}_1}{\sigma \cdot \sqrt{\frac{1}{n} + \frac{1}{n}}} \tag{3.3.3}$$

al sustituir $\overline{x}_2 - \overline{x}_1$

$$z_{test} = \sqrt{2} \cdot 1.96 \cdot (1 - p) \tag{3.3.4}$$

Se rechaza la hipótesis de la igualdad de las dos medias, si $z_{test} > 1.96$, esto se cumple si $p < 0.29$. Por lo tanto, siempre y cuando los dos intervalos de confianza del 95 % se superponen en menos de un 29 %, se rechazará la hipótesis nula de la igualdad de las dos medias con un valor de $p < 0.05$ para muestras de igual tamaño.

El método para determinar umbrales utiliza los intervalos de confianza del 95 % que no se traslapan, a la probabilidad condicional como variable respuesta y a la variable de interés como predictora. El proceso consiste en determinar el intervalo de confianza de la probabilidad incondicional y condicional. Para los datos del Mid Atlantic sería: El IC de la probabilidad que la riqueza de taxones EPT sea menor que 9, $P(z = 1)$, y los IC para la probabilidad condicional $p_i = P(z = 1|x > x_i)$, determinados mediante la técnica *Bootstrap* . El umbral $x_r$se determina apartir del primer IC del 95 % para p_i que no traslapa al IC del 95 % para $P(z = 1)$.

En la mayoría de los casos, este método es muy sensible a la variabilidad de los datos. En la figura 3.3.1 se presentan los resultados de los IC obtenidos de la probabilidad condicional para cada valor de X. Los datos del Mid Atlantic muestran la siguiente situación: el intervalo de confianza para la probabilidad incondicional es $(0.27, 0.44)$, el primer punto donde el intervalo de confianza no traslapa con el IC de la probabilidad incondicional es $x = 12.727$ con IC $(0.5, 0.76)$, pero debido al comportamiento de los datos vuelve a traslaparlo, y para valores mayores sus intervalos de confianza dejan de superponerse con el de la probabilidad incondicional. Para este tipo de situaciones donde el comportamiento de las variables no es monótono, el método se ve muy afectado y por lo tanto, los IC que no se traslapan no son una estrategia muy útil para determinar el umbral.

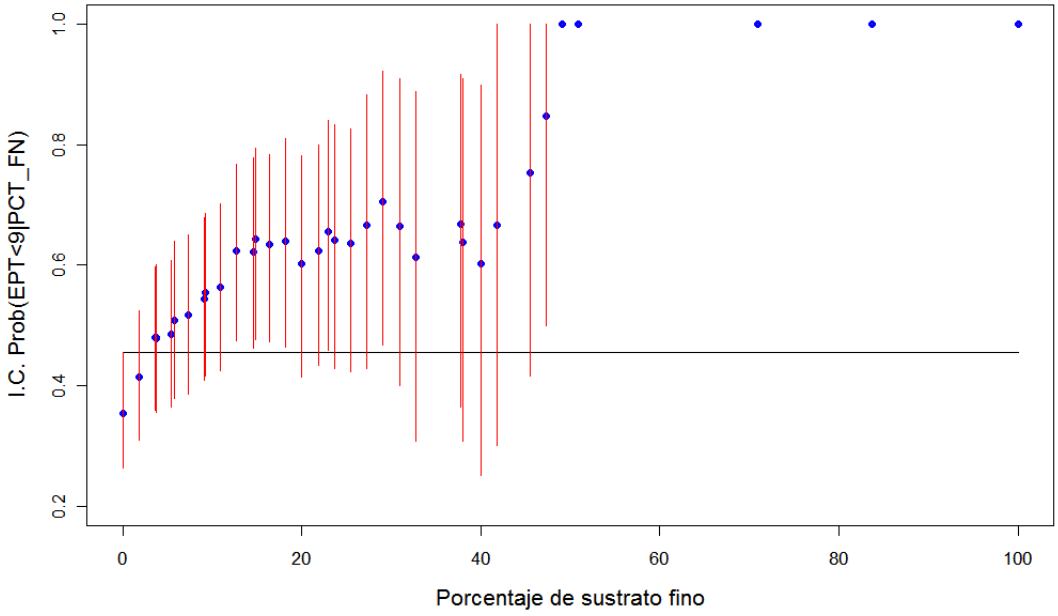

Figura 3.3.1: IC de las Probabilidades Condicionales. La línea horizontal marca el extremo superior del IC de la probabilidad incondicional: Indica el punto donde el IC de la probabilidad condicional no se traslapa.

El algoritmo 3.1 calcula la probabilidad condicional, las muestras *Bootstrap*, el estimador del umbral y los intervalos de confianza para cada valor de la probabilidad condicional. Aplicado a los datos del Mid Atlantic donde las variables son EPT_RICH y PCT_FN con un 1000 muestras *Bootstrap*,

```
datos=read.table("H:\\Tesis\\datosept.txt",header=T)
I.C.(datos,9,1000)
```

Algunos de los resultados obtenidos al aplicar el método de los intervalos de confianza

X	u.ic	l.ic
0.000	0.2525253	0.4444444
1.818	0.3037975	0.5189873
3.636	0.3589744	0.5945946
3.704	0.3538462	0.5967742
3.774	0.3559322	0.6000000
5.455	0.3606557	0.6031746
5.714	0.3787879	0.6250000
7.273	0.3870968	0.6349206
9.091	0.4107143	0.6666667
9.259	0.4166667	0.6730769
10.909	0.4259259	0.6896552

donde X representa la variable de interés, y u.ic, l.ic son los extremos del intervalo de confianza del 95 % para las probabilidades condicionales. La figura 3.3.1 expone estos resultados.

Algoritmo 3.1 IC que no se Traslapan

```
I.C.=function(datos,umbral, N_resam)
{# Se eliminan los datos faltantes
datos=na.omit(datos) # Ordenamos las variables por su gradiente de interés
datos=datos[order(datos[,1]),]
w=length(datos[,1]) # Se crea una matriz de almacenamiento
datos=matrix(0,w,2)
datos[,1]=datos[,1]
datos[,2]=datos[,2]
y=table(datos[,1])
c=length(y)
z=as.numeric(names(y)) # Se crea la matriz para almanecenar las probabilidades
probabilidad=matrix(0,c,N_resam+1) # La primera columna gradiente de interés
probabilidad[,1]=z
m=1 # Un while para Bootstrap, N_resam de veces
while(m<=N_resam)
{muestro=DATOS # Remuestreamos las dos variables por filas
if(m>=2) muestro=DATOS[sample(w,w,replace=T),]
k=1 ; i=0 ; r=0 ;p=0
# Se calcula la probabilidad condicional y se almanecena en probabilidad
while(k<=c) { i=i+1
if (muestro[i,1]>=z[k]) r=r+1
if (muestro[i,1]>=z[k] && muestro[i,2]<umbral) p=p+1
if(i==w) { i=0
if(r!=0) probabilidad[k,m+1]=p/r
if(r==0) probabilidad[k,m+1]=NA
k=k+1; i=0 ; p=0
r=0 }
}
m=m+1 } # Una matriz de almacenamiento de los datos
s=matrix(0,36,2)
for(i in 1:c)
{prob=probabilidad[i,]
prob=prob[-1]
medi=mean(prob,na.rm=TRUE)
y=sort(prob)
nobs=length(y)
nboot=1000
level=95
alpha=1-.01*level
kperc=floor((nboot+1)*alpha*.5)
s[i,1]=y[kperc]
s[i,2]=y[nboot+1-kperc] }
u.ic=s[,1] ; l.ic=s[,2]
l=data.frame(z,u.ic,l.ic)
print(l) }
```

3.4. Ajuste de un Modelo No Lineal

Este método consiste en definir un comportamiento o ajuste de un modelo no lineal a la curva de probabilidad condicional. Utiliza la variable respuesta y la variable predictiva como continuas. La manera más fácil de saber cuándo el análisis de regresión es adecuado, es observar un diagrama de dispersión y analizar el comportamiento de las variables y así determinar un modelo adecuado.

Existen diversos tipos de regresión, como la regresión lineal (el más simple, y que se utiliza con más frecuencia), de regresión polinomial (a menudo usado para probar la no-linealidad en una relación), la regresión logística (de amplio uso debido a su aplicación en datos binarios), la regresión robusta (modelos que son menos sensibles a los valores atípicos), la regresión múltiple (donde hay varias variables explicativas) y la que se utiliza, la regresión no lineal.

Es común querer tener un modelo que explique la relación entre Y y X, de tal forma que pueda estimar los parámetros y los errores estándares de los parámetros en una determinada ecuación no lineal a partir de los datos. Algunos modelos no lineales utilizados con frecuencia se muestran en la Tabla 3.1. Decir no lineal no significa que la relación sea curva, sino que la relación con los parámetros no puede ser linealmente expresada.

El comando nls() es la aplicación en **R** 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.

En nuestro caso $p_1, p_2, ..., p_n$ son las probabilidades condicionales, $x_1, x_2, ..., x_n$ los valores de la variable de interés y f una función no lineal que se ajusta a los datos. Las

Nombre	Ecuación
Funciones con Asíntota	
Michaelis–Menten	$y = \frac{ax}{1+bx}$
Exponencial de dos parámetros con asíntota	$y = a(1 - e^{-bx})$
Exponencial de tres parámetros con asíntota	$y = a - be^{-cx}$
Funciones en Forma de S	
Logística de dos parámetros	$y = \frac{exp(a+bx)}{1+exp(a+bx)}$
Logística de tres parámetros	$y = \frac{a}{1+b\,exp(-cx)}$
Logística de cuatro parámetros	$y = a + \frac{b-a}{1+exp((c-x)/d)}$
Weibull	$y = a - b\,exp(-cx^d)$
Gompertz	$y = a\,exp(-b\,exp(-cx))$
Curvas Unimodales	
Curva de Ricker	$y = axe^{-bx}$
Un compartimiento	$y = k\,(exp(-e^{ax}) - exp(-e^{bx}))$
Forma de Campana	$y = a\,exp(-\lvert bx\rvert^2)$
Biexponencial	$y = ae^{bx} - ce^{-dx}$

Tabla 3.1: Algunos modelos no lineales.

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:

$$RSS(\beta) = \sum_{i=1}^{n} (p_i - f(x_i, \beta))^2 \tag{3.4.1}$$

La solución de la minimización da como resultados las estimaciones de β, denotadas como $\hat{\beta}$. El algoritmo más común para la estimación en la regresión no lineal es el método de Gauss-Newton, que se basa en aproximaciones lineales a la función no lineal en cada paso. Es importante dar a conocer el valor $RSS(\beta)$ en el modelo propuesto, o un estimador equivalente de error residual estándar o varianza, como una medida del ajuste del modelo. Esta medida es útil para comparar diferentes modelos que se ajustan para el mismo conjunto de datos. Una medida para obtener el mínimo $RSS(\beta)$ es utilizar la *deviance* del modelo.

El procedimiento del método consiste en hacer un gráfico de dispersión de los datos, observar la curva que presentan los datos y definir un modelo que explique su

comportamiento, ver Tabla 3.1.

Para los datos del Mid Atlantic el gráfico de dispersión de las variables a modelar, la probabilidad condicional y el porcentaje de sustrato fino, está en la figura 3.2.1. El modelo que se ajusta a los datos es

$$P(z=1|x>x_c) = \begin{cases} 1+\frac{D_0-1}{1+e^{B_1(x_c-x_0)}}, & para\ x_c \leq x_0 \\ 1+\frac{D_0-1}{1+e^{B_0(x_c-x_0)}}, & para\ x_c > x_0 \end{cases} \tag{3.4.2}$$

donde D_0 es la probabilidad incondicional $p(z=1)$ y x_0 el umbral de interés. Los parámetros del modelo x_0, B_0 y B_1 se estiman por mínimos cuadrados no lineales.

El algoritmo 3.2 calcula la probabilidad condicional y ajusta el modelo de la ecuación (3.4.2) a la probabilidad condicional. Al aplicar a los datos del Mid Atlantic,

```
datos=read.table("H:\\Tesis\\datosept.txt",header=T)
Fitequation(datos,9)
```

Los resultados obtenidos al aplicar la funcion del modelo no lineal son:

```
El Umbral de Interés es: 44.66806
Los Estimados de los parámetros, Bo= 0.5678853 B1= 0.02495639
La deviance del modelo es: 0.1002749.
```

El umbral para el porcentaje de sustrato fino es 44.66, lo que significa que para valores de porcentaje de sustrato fino mayores a 44.66 existe un cambio en la probabilidad de impacto de contaminación respecto a valores menores a 44.66.

La desventaja de este método es que no siempre puede ser utilizado para calcular umbrales de un gradiente de interés, debido a que no siempre se tiene un modelo no lineal que explique el comportamiento de la probabilidad condicional.

Algoritmo 3.2 Ajuste del Modelo No lineal

```
Fitequation=function(datos,umbral)
{datos=na.omit(datos)
datos=datos[order(datos[,1]),]
w=length(datos[,1])
y=table(datos[,1])
c=length(y)
z=as.numeric(names(y))
probabilidad=matrix(0,c,2)
probabilidad[,1]=z
k=1 ; i=0 ; r=0 ; p=0
# Calculo de la probabilidad condicional
while(k<=c)
{i=i+1
if(datos[i,1]>=z[k]) r=r+1
if(datos[i,1]>=z[k] && datos[i,2]<umbral) p=p+1
if(i==w)
{probabilidad[k,2]=p/r
k=k+1; i=0 ; p=0 ; r=0
} }
probabilidad=data.frame(probabilidad)
require(graphics)
x=probabilidad[,1]
prob=probabilidad[,2]
# Ajuste del modelo planteado para los datos
Modelo<-nls(prob ~ I(x>p)*(1+ (0.3535354-1)/(1+exp(Bo*(x-p)))) + I(x<=p)*(1+ (
0.3535354-1)/(1+exp(BI*(x- p)))),data=probabilidad,start=list(Bo=0.04,BI=0.5, p=8))
plot(x,prob,xlab="Percent Fines in Substrate", ylab="Probability Conditional")
#lines(Probabilidad$PCT_FN, fitted(Modelo))
q=as.numeric(coef(Modelo))
cat("El Umbral de Interés es: ",q[3],"\n")
cat("Los Estimados de los parámetros, Bo=",q[1],"B1=",q[2],"\n")
dev=deviance(Modelo)
cat("La deviance del modelo es:",dev,"\n")
}
```

3.5. Ajuste de Modelo Jerárquico Bayesiano

El método consiste en asumir que los valores de la variable respuesta $p_1, p_2, ..., p_n$ (probabilidades condicionales) son muestras aleatorias de una secuencia de variables aleatorias $P_1, P_2, ..., P_n$. En otras palabras, las probabilidades condicionales se asumen como variables aleatorias que pertenecen a una misma familia de distribuciones con parámetro θ.

En situaciones como la descrita anteriormente, es natural construir una distribución a priori de una manera jerárquica. En este tipo de modelo, las observaciones provienen de las distribuciones condicionales dados los parámetros, y los parámetros a su vez tienen distribuciones que dependen de hiperparámetros. En concreto, se inicia especificando una distribución a los datos

$$P \sim f(P|\theta)$$

al vector θ se le asigna una distribución previa con hiperparámetros λ desconocidos.

$$\theta \sim g_1(\theta|\lambda)$$

El vector de hiperparámetros λ a su vez, se le asigna una distribución

$$\lambda \sim g_2(\lambda)$$

El ajuste del modelo jerárquico bayesiano, consiste al igual que en los métodos anteriores, en determinar un punto de cambio de la curva de probabilidad. Entonces, las variables

aleatorias $P_1, P_2, ..., P_N$ se definen como

$$\begin{aligned} P_1, P_2, ..., P_r &\sim \pi(P_i|\theta_1) \\ P_{r+1}, P_{r+2}, ..., P_n &\sim \pi(P_i|\theta_2) \end{aligned} \tag{3.5.1}$$

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 lograr determinar su distribución y poder calcular el estimador de r.

Ejemplo 1. Si se supone que las variables aleatorias $P_1, P_2, .., P_n$ son de una familia de distribución normal, el problema del punto de cambio se define como

$$P_i \sim \begin{cases} N(\mu_1, \sigma_1^2), & i = 1, ..., r \\ N(\mu_2, \sigma_2^2), & i = r+1, ..., n \end{cases} \tag{3.5.2}$$

Sea $\lambda_1 = 1/\sigma_1^2$ y $\lambda_2 = 1/\sigma_2^2$. El vector de parámetros es $\theta = (\mu_1, \lambda_1, \mu_2, \lambda_2)$.

Definición. Para cada punto de muestreo x, sea $\hat{\theta}(x)$ el valor del parámetro en el que la verosimilitud $L(\theta|x)$ alcanza su máximo en la función de θ, con x fijo. El estimador máximo verosímil (MLE) del parámetro θ basado en una muestra X es $\hat{\theta}(X)$.

Para los parámetros μ_1 y μ_2, se asumirá sus estimados (MLE) a partir de una distribución normal, $\bar{P}_1$ y $\bar{P}_2$, respectivamente. Para los parámetros λ_1 y λ_2 las distribuciones previas pertenecen a la familia de distribución gamma, es decir $\lambda_1 \tilde{} \Gamma(\alpha_1', \beta_1')$ y $\lambda_2 \tilde{} \Gamma(\alpha_2', \beta_2')$. En la práctica, los valores de los parámetros (α_1', β_1') y (α_2', β_2'), se eligen para hacer las distribuciones a priori casi planas, por lo tanto se pueden tomar valores de 0.001 para todos los parámetros. Las distribuciones a priori λ_1 y λ_2 aseguran una

adecuada distribución posterior para r. Por lo tanto, la distribución a priori del vector de parámetros θ está dada por

$$\pi(\theta, r) \propto \pi(\lambda_1)\pi(\lambda_2)$$

donde,

$$\pi(\lambda_1) = \frac{1}{\Gamma(\alpha_1')}\beta_1'^{\alpha_1'} e^{-(\lambda_1\beta_1')}\lambda_1^{(\alpha_1'-1)} \quad y \quad \pi(\lambda_2) = \frac{1}{\Gamma(\alpha_2')}\beta_2'^{\alpha_2'} e^{-(\lambda_2\beta_2')}\lambda_2^{(\alpha_2'-1)}$$

entonces,

$$\pi(\theta, r) \propto \frac{1}{\Gamma(\alpha_1')\Gamma(\alpha_2')}\beta_1'^{\alpha_1'}\beta_2'^{\alpha_2'}\lambda_1^{\alpha_1'-1}e^{-\lambda_1\beta_1'}\lambda_2^{\alpha_2'-1}e^{-\lambda_2'\beta_2'}$$

dado que P se distribuye como una normal con parámetros θ y r

$$\pi(P|\theta, r) = \prod_{i=1}^{r}\frac{1}{\sqrt{2\pi}}\lambda_1^{1/2}exp(-\frac{(p_i-\mu_1)^2\lambda_1}{2})\prod_{i=r+1}^{n}\frac{1}{\sqrt{2\pi}}\lambda_2^{1/2}exp(-\frac{(p_i-\mu_2)^2\lambda_2}{2})$$

al operar,

$$\pi(P|\theta, r) = \frac{1}{(2\pi)^{\frac{r}{2}}}\lambda_1^{\frac{r}{2}}exp(-\sum_{i=1}^{r}\frac{(p_i-\mu_1)^2\lambda_1}{2})\frac{1}{(2\pi)^{\frac{n-r}{2}}}\lambda_2^{\frac{n-r}{2}}exp(-\sum_{i=r+1}^{n}\frac{(p_i-\mu_2)^2\lambda_2}{2})$$

Por el teorema de Bayes, $\pi(r, \theta|P) = \pi(\theta, r)\prod_{i=1}^{n}(p_i|\theta, r)$, entonces, al sustituir

$$\begin{aligned}\pi(r, \theta|P) \quad = \quad & \frac{1}{\Gamma(\alpha_1')\Gamma(\alpha_2')(2\pi)^{\frac{n}{2}}}\beta_1'^{\alpha_1'}\beta_2'^{\alpha_2'}\lambda_1^{\frac{r}{2}+\alpha_1'-1}exp(-\tfrac{1}{2}r\lambda_1(\mu-\bar{P}_1)) \\ & exp(-\lambda_1\delta_1)\lambda_2^{-\frac{n-r}{2}+\alpha_2'-1}exp(-\tfrac{1}{2}(n-r)\lambda_2(\mu_2-\bar{P}_2)^2)exp(-\lambda_2\delta_2).\end{aligned}$$

donde δ_1 y δ_2 se encuentran al final del ejemplo.

Para determinar la distribución marginal de r, se debe integrar $\pi(r, \theta|P)$ con respecto a

θ, es decir, con respecto a μ_1, μ_2, λ_1 y λ_2.

$$\pi(r|P) = \int_0^\infty \int_{-\infty}^\infty \int_0^\infty \int_{-\infty}^\infty \pi(r,\theta|P) d\mu_1 d\lambda_1 d\mu_2 d\lambda_2 \tag{3.5.3}$$

realizando sustituciones y asumiendo $r \neq n$

$$\pi(r|P) = \frac{1}{\Gamma(\alpha_1')\Gamma(\alpha_2')(2\pi)^{\frac{n}{2}}} \beta_1'^{\alpha_1'} \beta_2'^{\alpha_2'} (\sqrt{2\pi})^2 \frac{1}{r^{1/2}} \frac{1}{(n-r)^{1/2}} \frac{\Gamma(\gamma_1)}{\delta_1^1} \frac{\Gamma(\gamma_2)}{\delta_2^{\gamma_2}} \tag{3.5.4}$$

Si $r = n$, entonces la ecuación sería $\pi(r,\theta|P) = \pi(\theta,r)\prod_{i=1}^n (p_i|\theta,r)$, lo que traduce a un solo par de parámetros en la normal μ y λ. Por lo tanto

$$\pi(\theta,r|P) = \tfrac{1}{\Gamma(\alpha)}\beta^\alpha e^{-\lambda\beta}\lambda^{\alpha-1}\prod_{i=1}^n \tfrac{1}{(2\pi)^{1/2}}\lambda^{1/2} exp(-\tfrac{(p_i-\mu)^2\lambda_2}{2})$$

al mutiplicar y despejar el cuadrado

$$\pi(\theta,r|P) = \tfrac{1}{\Gamma(\alpha)}\beta^\alpha e^{-\lambda\beta}\lambda^{\alpha-1}\tfrac{1}{(2\pi)^{\frac{n}{2}}}\lambda^{\frac{n}{2}} exp(-\sum_{i=1}^n \tfrac{(\mu-\bar{P}_i)^2\lambda}{2}) exp(-\tfrac{\lambda\sum_{i=1}^n P_i^2 - n\bar{P}}{2}),$$

al integrar con respecto a μ y λ, se obtiene

$$\pi(r|P) = \frac{\beta(2\pi)^{1/2}}{\Gamma(\alpha)(2\pi)^{\frac{n}{2}} n^{1/2}} \frac{\Gamma(\gamma)}{\delta_n^{\gamma_n}} \frac{\Gamma(\alpha_2')}{\beta_2^{\alpha_2'}} \tag{3.5.5}$$

para los casos de r. Por (3.5.4) y (3.5.5) la distribución marginal de r está dada por (Quian et al. 2003)

$$p(r|P) \propto \begin{cases} \frac{1}{r^{1/2}} \frac{1}{(n-r)^{1/2}} \frac{\Gamma(\gamma_1)\Gamma(\gamma_2)}{\delta_1^1 \delta_2^{\gamma_2}} & si\ r < n \\ \frac{1}{n^{1/2}} \frac{\Gamma(\gamma)\Gamma(\alpha_2')}{\delta_n^{\gamma_n} \beta_2^{\alpha_2'}} & si\ r = n \end{cases} \tag{3.5.6}$$

donde,

$$\begin{aligned}
\bar{P}_1 &= \tfrac{1}{r}\textstyle\sum_{i=1}^{r} P_i & \bar{P}_2 &= \tfrac{1}{n-r}\textstyle\sum_{i=r+1}^{n} P_i \\
\gamma_1 &= \tfrac{r-1}{2} + \alpha_1' & \gamma_2 &= \tfrac{n-r-1}{2} + \alpha_2' \\
\delta_1 &= \tfrac{1}{2}\left[\textstyle\sum_{i=1}^{r} P_i^2 - r\bar{P}_1^2\right] + \beta_1' & \delta_1 &= \tfrac{1}{2}\left[\textstyle\sum_{i=r+1}^{n} P_i^2 - (n-r)\bar{P}_2^2\right] + \beta_2' \\
\gamma_n &= \tfrac{n-1}{2} + \alpha_1' & \delta_n &= \tfrac{1}{2}\left[\textstyle\sum_{i=1}^{n} P_i^2 - n\bar{P}_1^2\right] + \beta_1'
\end{aligned}$$

y $\Gamma(\boldsymbol{.})$ representa la función gamma.

La distribución marginal de r es una distribución discreta. Además, el orden de la variable de interés es el mismo de la distribución de r, es decir, la probabilidad que x_i sea el umbral está definida por $p(r|Y)$ en $r = i$. Un estimado del punto de cambio del valor correspondiente a la variable del gradiente ambiental, es seleccionar la moda o bien la esperanza de esta distribución marginal.

Ejemplo 2. Si las variables $P_1, P_2, .., P_n$ son de una familia de distribución binomial, el problema del punto de cambio se define como el valor r, tal que

$$P_i \sim \begin{cases} Bin(\theta_1, N_i) & i = 1, ..., r \\ Bin(\theta_2, N_i) & i = r+1, ..., n \end{cases} \tag{3.5.7}$$

donde θ_1 y θ_2 son las probabilidades de éxito antes y después del cambio, respectivamente. Se asume las distribuciones a priori de los parámetros como uniformes $(0, 1)$. Por lo tanto

$$\pi(\theta_1, \theta_2, r|P) \propto \prod_{i=1}^{n} \pi(P|\theta_1, \theta_2, r) \tag{3.5.8}$$

por ser $\pi(\theta_1, \theta_2, r)$ una constante igual a 1, se tiene:

$$\begin{aligned}
\pi(\theta_1, \theta_2, r|P) &\propto \theta_1^{\sum_{i=1}^{r} P_i}(1-\theta_1)^{\sum_{i=1}^{r}(N_i - P_i)}\theta_2^{\sum_{i=r+1}^{n} P_i}(1-\theta_2)^{\sum_{i=r+1}^{n}(N_i - P_i)} \\
&= \theta_1^{S_{11}}(1-\theta_1)^{S_{12}}\theta_2^{S_{21}}(1-\theta_2)^{S_{22}}
\end{aligned} \tag{3.5.9}$$

donde $S_{11} = \sum_{i=1}^{r} P_i$, $S_{12} = \sum_{i=1}^{r}(N_i - P_i)$, $S_{21} = \sum_{i=r+1}^{n} P_i$ y $S_{22} = \sum_{i=r+1}^{n}(N_i - P_i)$. Integrando con respecto a θ_1 y θ_2, se obtiene la distribución marginal de la distribución de r

$$\pi(r|P) \propto \frac{\Gamma(S_{11}+1)\Gamma(S_{12}+1)}{\Gamma(S_{11}+S_{12}+1)} \frac{\Gamma(S_{21}+1)\Gamma(S_{22}+1)}{\Gamma(S_{21}+S_{22}+2)}. \tag{3.5.10}$$

Al igual que en el ejemplo para distribuciones normales, el punto de cambio se halla calculando la moda. La inferencia sobre si existe un punto de cambio se puede hacer por la probabilidad de que no existe punto de cambio, es decir $P(n|P)$.

Para las inferencias sobre las distribuciones posteriores para r, μ_1, μ_2, λ_1 y λ_2 del modelo normal, y r, θ_1 y θ_2 del modelo binomial, se utilizan técnicas como *Gibbs Sampler* o cadenas de *Markov Monte Carlo*.

Para los datos del Mid Atlantic se asume la probabilidad condicional de las variables pertenecientes a una familia de distribución normal. El algoritmo 3.3 al igual que los anteriores, funciona introduciendo las variables en su forma natural, el mismo elimina los valores faltantes, ordena los datos con respecto a la variable de interés, calcula la probabilidad condicional, ajusta el modelo jerárquico bayesiano y determina el valor del umbral.

Los resultados al entrar los datos

```
datos=read.table("H:\\Tesis\\datosept.txt",header=T)
Bayes(datos,9)
```

son los siguientes

```
El umbral de la variable de interés es 45.455.
```

Este resultado es muy similar al hallado en el método de ajuste de un modelo no lineal.

Algoritmo 3.3 Algoritmo Modelo Jerárquico Bayesiano

```
Bayes=function(datos,umbral)
{datos=na.omit(datos)
datos=datos[order(datos[,1]),]
w=length(datos[,1])
y=table(datos[,1])
s=length(y)
z=as.numeric(names(y))
probabilidad=matrix(0,s,2)
probabilidad[,1]=z
attach(datos)
k=1 ; i=0 ; r=0 ; p=0
while(k<=s)
{ i=i+1
if (datos[i,1]>=z[k]) r=r+1
if (datos[i,1]>=z[k] && datos[i,2]<umbral) p=p+1
if(i==w)
{probabilidad[k,2]=p/r
k=k+1 ; i=0 ; p=0 ; r=0
} }
datos=(probabilidad[,2])
par(mfrow=c(2,2))
hist(datos,col="50",breaks=5)
qqnorm(datos,datax=FALSE,col='51',plot=TRUE)
prob_change=matrix(0,s,1) for(r in 1:s)
{ alpha1=0.001 ; alpha2=0.001 ; beta1=0.001 ; beta2=0.001
n=length(datos)
c=1:r
y1=(r-1)/2+alpha1
y2=(n-r-1)/2+alpha2
yn=(n-1)/2+alpha1
sigma1=1/2*(sum(datos[c]^2)-r*(mean(datos[c]))^2)+beta1
if(r<n) sigma2=1/2*(sum(datos[-c]^2)-(n-r)*(mean(datos[-c]))^2)+beta2
else sigma2=0
sigman=1/2*(sum(datos^2)-n*(mean(datos[c]))^2)+beta1
if(r<n){prob_change[r,1]=-1/2*log(r)-1/2*log(n-r)-y1*log(sigma1)-
y2*log(sigma2)+lgamma(y1)+lgamma(y2)}    if(r==n){prob_change[r,1]=-1/2*log(n)-
yn*log(sigman)-alpha2*(log(beta2))+lgamma(yn)+lgamma(alpha2)}
}
# El siguiente proceso calcula la ubicación del umbral de la variable de interés
m=1:s
change=max(prob_change[,1])
r=m[prob_change[,1]==change]
cat("El umbral de la variable de interés es",z[r],"\n")
}
```

3.6. Reducción de la *Deviance* no paramétrica

El método de reducción de la *deviance* no paramétrica consiste en separar la variable respuesta en dos partes y utilizar la *deviance* 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 $p_1, p_2, ...p_n$.

La *deviance* es una medida de ajuste más general que la suma cuadrada de los residuales (RSS). Este valor permite medir qué tan bien se ajusta en un sentido absoluto el modelo a los datos, pero no, qué tan bien se ajusta en un sentido relativo. La *deviance* tiene una distribución aproximada χ^2, chi cuadrado, con $(n-1)$ grados de libertad, si el modelo es correcto. Por lo tanto, se puede utilizar para probar si un modelo es o no adecuado a los datos.

Dependiendo del modelo la *deviance* cambia su forma, por ejemplo, para modelos lineales generalizados, se tiene:

MGL	*Deviance*
Gaussiano	$\sum_i (y_i - \hat{\mu}_i)^2$
Poisson	$2\sum_i [y_i log(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i)]$
Binomial	$2\sum_i [y_i log(y_i/\hat{\mu}_i) + (m - y_i) log((m - y_i)/(m - \hat{\mu}_i))]$
Gamma	$2\sum_i [-log(y/\hat{\mu}) + (y - \hat{\mu})/\hat{\mu}]$

Los residuales representan la diferencia entre los datos y el modelo, por lo tanto son esenciales para explorar el ajuste de un modelo. La *deviance* está definida de forma análoga a los residuales Pearson. El residual Pearson r_p es $\sum r_p^2 = X^2$, para el conjunto de los residuales de la *deviance* es r_D, tal que *deviance* $D = \sum r_D^2 = \sum d_i$. Por lo tanto

$$r_{Di} = signo(y_i - \hat{\mu}_i)\sqrt{d_i}$$

donde d_i representa el i-ésimo término la *deviance* del modelo.

Si se asume un modelo con dos medias μ_p, una para valores de $i \leq r$, otras para valores de $i > r$, el proceso de la reducción de la *deviance* no paramétrica consiste en:

1. Realizar una partición de la variable explicativa en un punto r (variable ya ordenada), con ello se forman dos submuestras t_1 y t_2, una a la izquierda y otra a la derecha de r, respectivamente.

2. Escoger las muestras perteneciente a cada conjunto, es decir los $\{(x_i, p_i)\}$. Si $N(s)$ es el número de observaciones en la muestra s, entonces $\bar{y}(s) = \frac{1}{N(s)} \sum_{p_i \epsilon s} p_i$. La suma de cuadrados o la *deviance* no paramétrica, que para este caso es igual, está dada por

$$D(s) = \sum_{p_i \epsilon s} (p_i - \bar{p}(s))^2$$

3. Calcular el criterio de mínimos cuadrados para las submuestras, es decir

$$D_r = D(t_1) + D(t_2)$$

4. El punto de cambio de la variable de interés:

$$r = argmin_{r \epsilon \Omega} D_r$$

   donde Ω=$\{1, 2, ..., n\}$.

El objetivo es formar conjunto más homogéneos, es decir, la partición que da mayor homogeneidad.

El algoritmo 3.4 calcula las probabilidades condicionales, ordena las observaciones en términos de la variable de interés, elimina los valores faltantes, calcula la *deviance* no paramétrica para cada partición y determina el umbral.

Para los datos del Mid Atlantic,

```
datos=read.table("H:\\Tesis\\datosept.txt",header=T)
Deviance(datos,9)
```

El resultado obtenido con la aplicación del método de la reducción de la *deviance* no paramétrica es:

```
El umbral de la variable de interés es: 41.818.
```

Algoritmo 3.4 Algoritmo Reducción de la *Deviance* no paramétrica

```
Deviance=function(datos, umbral)
{# Para omitir los datos faltantes
datos=na.omit(datos)
# Ordena los datos por la primera columna
datos=datos[order(datos[,1]),]
w=length(datos[,1])
# Reconocer datos distintos
y=table(datos[,1])
# Valor importante, longitud de la tabla
c=length(y) # Convierte a vector
z=as.numeric(names(y)) # Crea una matrix para incluir los valores de la Deviance no paramétrica
D=matrix(0,c,1) # Este comando es para poder trabajar los datos por variables
attach(datos) # Matriz donde se incluira la probabilidad y la variable de interés
probabilidad=matrix(0,c,2) # Variable de interés
probabilidad[,1]=z
k=1; i=0 ; r=0 ;p=0
# Recorrido para obtener las probabilidades
#de cada punto de la variable de interés, que en total son c,
# se almancenan en Probabilidad[k,2]
while(k<=c) { i=i+1
if (datos[i,1]>=z[k]) r=r+1
if (datos[i,1]>=z[k] && datos[i,2]<umbral) p=p+1
if(i==w) { probabilidad[k,2]=p/r; k=k+1; i=0; p=0; r=0 } }
# funcion de la Deviance
Dev=function(x) { sum((x-mean(x))^2) }
# Calcula la deviance para cada partición, solo se trabaja con la probabilidad.
for(i in 1:c) {x=probabilidad[,2]
p=1:i ; y1=x[p] ; y2=x[-p]
D[i,1]=Dev(x)-(Dev(y1)+Dev(y2))
cat("La Deviance para la partición en",z[i]," es",D[i,1],"\n") }
# Calcula el umbral
m=1:c ; change=max(D[,1]) ; r=m[D[,1]==change]
cat("El umbral de la variable de interés es:", z[r],"\n")
}
```

3.7. Metodología LAD (*Least Absolute Deviation*)

Este método no ha sido usado previamente, hasta donde se ha podido verificar, para determinar umbrales. El proceso es similar a la metodología de la reducción de la *deviance* no paramétrica, pero en lugar de usar los mínimos cuadrados como criterio de homogeneidad, se utiliza LAD (*Least Absolute Deviation*).

El LAD es ampliamente conocido en el análisis estadístico de modelos de regresión lineal. En lugar de utilizar el criterio de los mínimos cuadrados, se minimiza la suma de los valores absolutos de los errores. No ha tenido tanta aplicación como los mínimos cuadrados, por la falta de procedimientos adecuados de inferencia general. Por ejemplo, es bastante complicado realizar el equivalente al método del análisis de varianza, que es una herramienta estadística basada en los mínimos cuadrados para probar hipótesis.

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 sensible a valores atípicos y produce estimaciones robustas.

La metodología del criterio de ajuste LAD, consiste en:

1. Dividir la variable de interés en un punto r, se forman dos submuestras t_1 y t_2 a la izquierda y derecha de r.

2. Calcular la desviación absoluta mínima $LAD(s) = \sum_{y \epsilon s} |p - \bar{p}(s)|$, aplicarlas a t_1 y t_2

$$LAD_r = LAD(t_1) + LAD(t_2)$$

3. Determinar el punto de cambio de la variable de interés, dado por

$$r = argmin_{r \epsilon \Omega} LAD_r$$

donde Ω=$\{1, 2, ..., n\}$.

El umbral seria x_r.

El algoritmo 3.5 realiza este procedimiento. Aplicado a los datos del Mid Atlantic, se tiene

```
datos=read.table("H:\\Tesis\\datosept.txt",header=T)
LAD(datos,9)
```

Los resultados de la aplicación del algoritmo 3.5 en los datos del Mid Atlantic, dan como resultado

```
La desviación absoluta mínima es 2.491580.
El Umbral de la variable de interés es: 45.455
```

El umbral se aproxima a los obtenidos en los métodos anteriores, aunque se ve una diferencia notable con el método de la reducción de la *deviance* no paramétrica. En el capítulo 4 se analizará con más detalle su comportamiento con otros datos.

Algoritmo 3.5 Algoritmo LAD

```
LAD=function(datos, umbral)
{# Omite los datos faltantes
datos=na.omit(datos)
datos=datos[order(datos[,1]),]# Ordena los datos por la primera columna
w=length(datos[,1])# Longitud del vector datos
y=table(datos[,1]) # Para reconocer los datos distintos
c=length(y)
z=as.numeric(names(y)) # Convierte vector
D=matrix(0,c,1) # Crea una matriz para incluir los valores de la LAD
attach(datos) # Este comando es para poder trabajar las variables separadas
probabilidad=matrix(0,c,2) # Matriz donde se incluirá la probabilidad y la variable de interés
probabilidad[,1]=z # Variable de interés
k=1; i=0; r=0; p=0
# Cada punto de la variable de interés, que en total son c, se almacenan en Probabilidad[k,2]
# Función calcula la probabilidad condicional
while(k<=c) { i=i+1;
if (datos[i,1]>=z[k]) r=r+1
if (datos[i,1]>=z[k] && datos[i,2]>umbral) p=p+1
if(i==w) { probabilidad[k,2]=p/r
k=k+1; i=0; p=0; r=0 }
}
LAD=function(x) { sum(abs(x-mean(x))) }# Función que calcula la LAD
for(i in 1:c) {x=probabilidad[,2]
p=1:i; y1=x[p]; y2=x[-p];
D[i,1]=(LAD(y1)+LAD(y2))
}
cat("La desviación absoluta mínima es ", min(D),"\n")
m=1:c
change=min(D[,1])
r=m[D[,1]==change]
cat("El Umbral de la variable de interés es:", z[r],"\n")
}
```

Capítulo 4

Aplicaciones en Embalses de Puerto Rico

Este capítulo aplica los métodos propuestos a un problema real. El objetivo fundamental es obtener umbrales para nutrientes y sus respectivos intervalos de confianza con la técnica del *Bootstrap* en embalses de Puerto Rico.

Para la determinación de los umbrales en los nutrientes, se incluyen: el umbral de una variable a partir de valores en la literatura científica (umbrales de la clorofila a), el análisis del punto de cambio utilizando las probabilidades condicionales y los métodos para determinar umbrales de impacto.

4.1. Datos de Puerto Rico

Los embalses de Puerto Rico analizados en este estudio son Guajataca, Cerrillos y La Plata. Se analizan tres variables: Como variable que se conoce el umbral la clorofila

a (Chl_a) y como variables de interés el fósforo total (TP) y el nitrógeno total (TN). Los datos han sido recolectados durante los años 2005 a 2010, un total 165 datos para cada variable de interés, en el mismo sitio a lo largo del año. (Datos provistos por el Dr. David Sotomayor, Departamento de Ciencias Agroambientales, Universidad de Puerto Rico, Mayagüez).

La clorofila es el pigmento de color verde presente en plantas y algas, es un elemento muy importante para la fotosíntesis. Las clorofilas son una familia de pigmentos que se encuentran en las cianobacterias y en todos aquellos organismos que contienen plastos en sus células, lo que incluye a las plantas y a los diversos grupos de protistas que son llamados algas. La clorofila puede detectarse fácilmente gracias a su comportamiento frente a la luz. Medir ópticamente la concentración de clorofila en una muestra de agua da poco trabajo y permite una estimación eficiente de la concentración de fitoplancton (algas microscópicas) e, indirectamente, de la actividad biológica; de esta manera, la medición de la clorofila es un instrumento importante para la detección en los procesos de **eutrofización**, es decir del enriquecimiento en nutrientes de un ecosistema, lo que conlleva a un impacto negativo en todo el lago, desde la diversidad de especies hasta su estética (USEPA, 2009).

El aumento de la biomasa está limitado por la escasez de nitrógeno y fósforo, llamados factores limitantes, los cuales son necesarios para el desarrollo de la producción orgánica. La contaminación puntual del agua por efluentes urbanos o por contaminación agraria o atmosférica, pueden aportar cantidades importantes de elementos limitantes. El resultado es un aumento de la producción orgánica con importantes consecuencias sobre la composición, estructura y dinámica del ecosistema (USEPA, 2009). Por lo tanto, la eutrofización produce de manera general un aumento de la biomasa y un empobrecimiento de la diversidad, siguiendo un patrón de crecimiento: A mayor clorofila mayor cantidad de nutrientes.

Estado Trófico	**Clorofila A**
Oligotrófico	< 2 ug/L
Mesotrófico	$2 - 7$ ug/L
Eutrófico	$7 - 30$ ug/L
Hipereutrófico	> 30 ug/L

Tabla 4.1: Límites tróficos para clorofila a.

Para la clorofila se fijaron estándares tróficos dados por USEPA (2009). El NLA "*National Lakes Assessment*" evalúa los embalses con respecto a su estado trófico, los clasifica de acuerdo a su producción primaria. El estado trófico representa la productividad biológica en los embalses, en especial la productividad primaria. Para el NLA, el estado trófico de los embalses se caracteriza a nivel nacional mediante las concentraciones de clorofila a, ver tabla 4.1.

Cada una de estas clasificaciones define el estado del lago. Un lago oligotrófico se define como cuerpo de agua con baja productividad primaria como resultado de bajos contenidos de nutrientes. Estos embalses tienen baja producción de algas, y consecuentemente, poseen aguas sumamente claras, con alta calidad de agua potable. Un lago mesotrófico es un cuerpo de agua con un nivel intermedio de productividad primaria, mayor que el de un lago oligotrófico, pero menor que el de un lago eutrófico. Estos embalses tienen comúnmente aguas claras y mantienen lechos de plantas acuáticas sumergidas, y niveles medios de nutrientes. Un lago eutrófico o hipereutrófico es un cuerpo de agua con gran aumento de fitoplancton, uno de los efectos ambientales negativos en estos embalses incluyen la anoxia, o falta de oxígeno en el agua con severas reducciones en los peces y las poblaciones de otros animales.

Los umbrales de las variables TP y TN se desarrollan basados en la probabilidad que Chl_a exceda los límites oligotróficos, mesotróficos y eutróficos. El análisis da lugar a criterios de valoración basados en estas condiciones.

4.2. Variables de Interés y sus Comportamientos

Las figuras 4.2.1 y 4.2.2 muestran la relación de las variables fósforo total (TP), nitrógeno total (TN) y clorofila a (Chl_a). La correlación entre las variables de interés y la clorofila a es positiva. Para TP y Chl_a la correlación es 0.559 y para TN y Chl_a la correlación es 0.539, una buena correlación entre los nutrientes y la clorofila. Los puntos atípicos de la probabilidad condicional que se observan en las gráficas son valores que pueden influir en los resultados de los umbrales, debido a que la presencia de estos puntos es una de las desventajas del métodos de la *deviance* no paramétrica.

En este trabajo se aplican los métodos propuestos para la determinación de los umbrales para TP y TN. El método de los intervalos de confianza que no se traslapan no resultó útil para estos datos, debido a que 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. El método del ajuste de un modelo no lineal tampoco es una técnica útil en estos datos debido al comportamiento de la probabilidad condicional en cada una de la variables, no es sencillo ajustar un modelo que represente estos comportamientos (ver figuras 4.2.1 y 4.2.2).

El comportamiento de cada una de las probabilidades condicionales en cada uno límites tróficos de los nutrientes se puede observar en las figuras 4.2.1 y 4.2.2. La probabilidad condicional que se calcula, por ejemplo, para el límite oligotrófico en TP es $P(\text{Chl\_a}> 2|\ \text{TP}>\text{TP}_i)$.

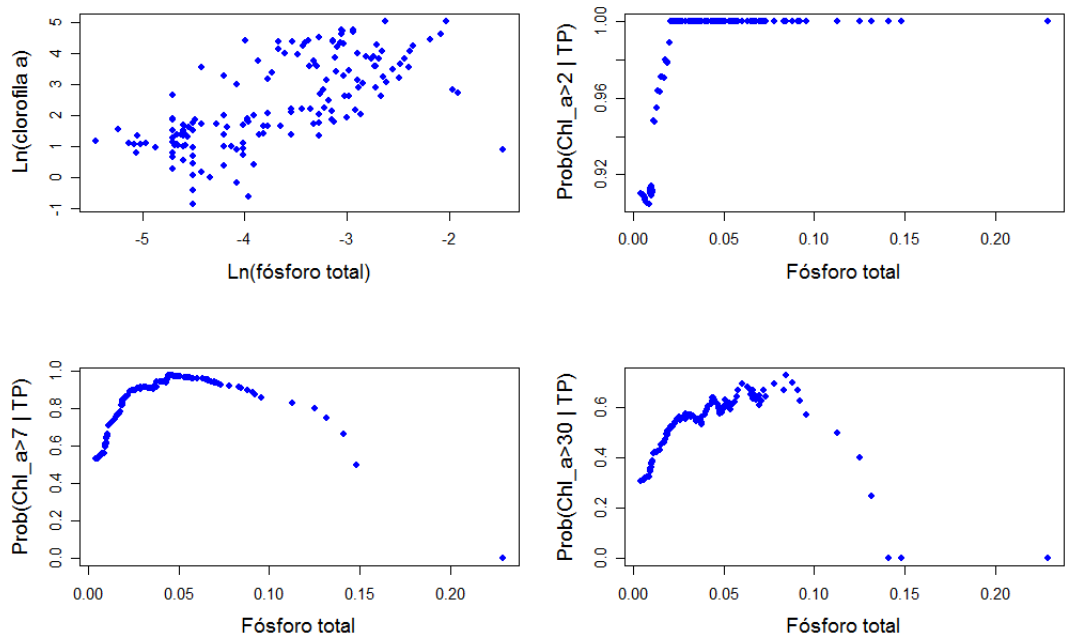

Figura 4.2.1: Probabilidad condicional vs. TP, para cada límite trófico. Se observan valores atípicos en Prob(Chl_a>7|TP) y Prob(Chl_a>30|TP).

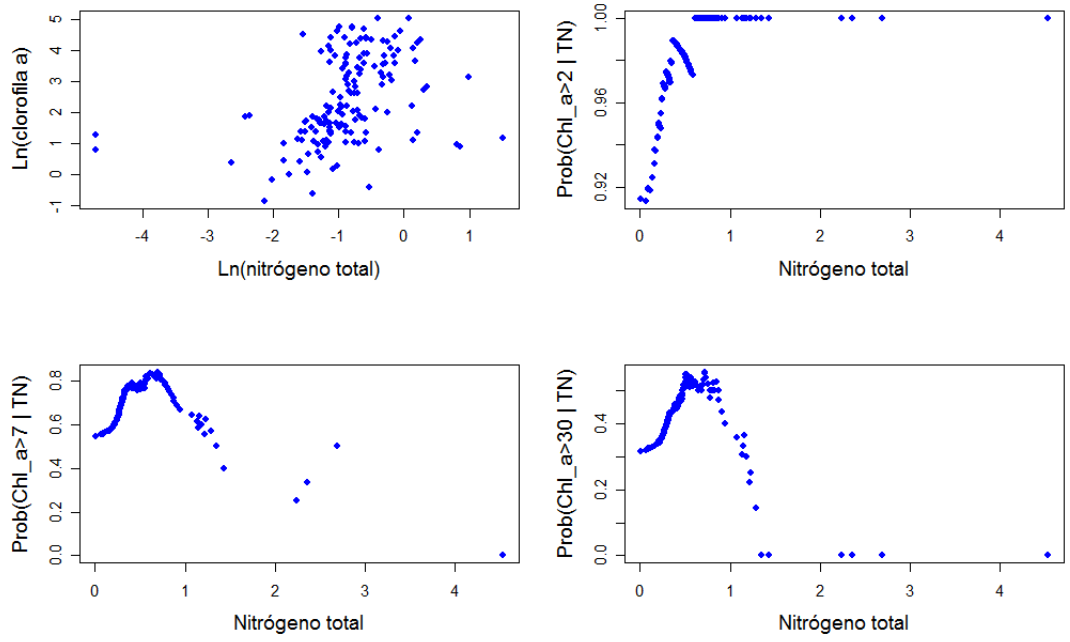

Figura 4.2.2: Probabilidad condicional vs. TN, para cada límite trófico. Se observan valores atípicos en Prob(Chl_a>7|TN) y Prob(Chl_a>30|TN).

4.3. Determinación de los Umbrales

Para cada variable de interés se determina el umbral en cada uno de los límites tróficos de Chl_a. Las tablas 4.1, 4.2 y 4.3 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 *Bootstrap*. 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, ver sección 2.7.

Método	TN	TP
Red. de la *Deviance* no P. (umbral, IC)	0.334 (0.159,0.590)	0.013 (0.011,0.019)
Modelo Jérarquico Bayesiano (umbral, IC)	0.369 (0.202,0.590)	0.017 (0.011,0.020)
Desviación Absoluta Mínima (umbral, IC)	0.397 (0.202,0.590)	0.014 (0.011,0.020)

Tabla 4.2: Umbrales para TN y TP si Chl_a > 2, en cada uno de los métodos.

Método	TN	TP
Red. de la *Deviance* no P. (umbral, IC)	1.116 (0.274,2.689)	0.036 (0.011,0.148)
Modelo Jérarquico Bayesiano (umbral, IC)	0.951 (0.252,2.689)	0.033 (0.010,0.148)
Desviación Absoluta Mínima (umbral, IC)	0.535 (0.264,1.349)	0.016 (0.011,0.021)

Tabla 4.3: Umbrales para TN y TP si Chl_a > 7, en cada uno de los métodos.

Método	TN	TP
Red. de la *Deviance* no P. (umbral, IC)	1.107 (0.410,1.291)	0.063 (0.015,0.131)
Modelo Jérarquico Bayesiano (umbral, IC)	0.865 (0.297,1.291)	0.037 (0.011,0.131)
Desviación Absoluta Mínima (umbral, IC)	0.787 (0.293,1.291)	0.021 (0.011,0.069)

Tabla 4.4: Umbrales para TN y TP si Chl_a > 30, en cada uno de los métodos.

Se pueden observar resultados similares para los umbrales de nutrientes en cada uno de métodos, para cada uno de los límites tróficos. Es importante destacar que estos umbrales son muy dependientes de las observaciones en los embalses, es decir, no es viable hacer una generalización de los resultados.

El método propuesto LAD tiene ciertas diferencias con los otros métodos en el umbral, pero se observa que los intervalos de confianza son bastante cercanos a los de los otros métodos.

Si se comparan los resultados 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. Esta propiedad es necesaria dada la relación monótona entre Chl_a y TN,TP. Los umbrales para cada uno de los nutrientes deben seguir un patrón de crecimiento en el aumento de los límites tróficos en Chl_a.

En cada una de las clasificaciones del estado trófico de un embalse se determinaron los umbrales para cada una de las nutrientes TP y TN. Con los umbrales de estos nutrientes se puede deducir el estado trófico de un embalse. En otras palabras, si la cantidad de nutriente en un lago supera el umbral correspondiente, el lago no se puede clasificar dentro de ese estado. Por ejemplo, para el límite oligotrófico-mesotrófico, el umbral para la variable fósforo total (TP) se puede asumir como 0.36 mg/L (ver tabla 4.2) lo que significa que si el embalse presenta un cantidad de TP superior a 0.36 mg/L, el embalse ya no se clasifica como oligotrófico y por lo tanto, presenta características diferentes a ese estado.

Las figuras 4.3.1 y 4.3.2 presentan los umbrales y los respectivos intervalos de confianza para cada nutriente, y por cada uno de los métodos.

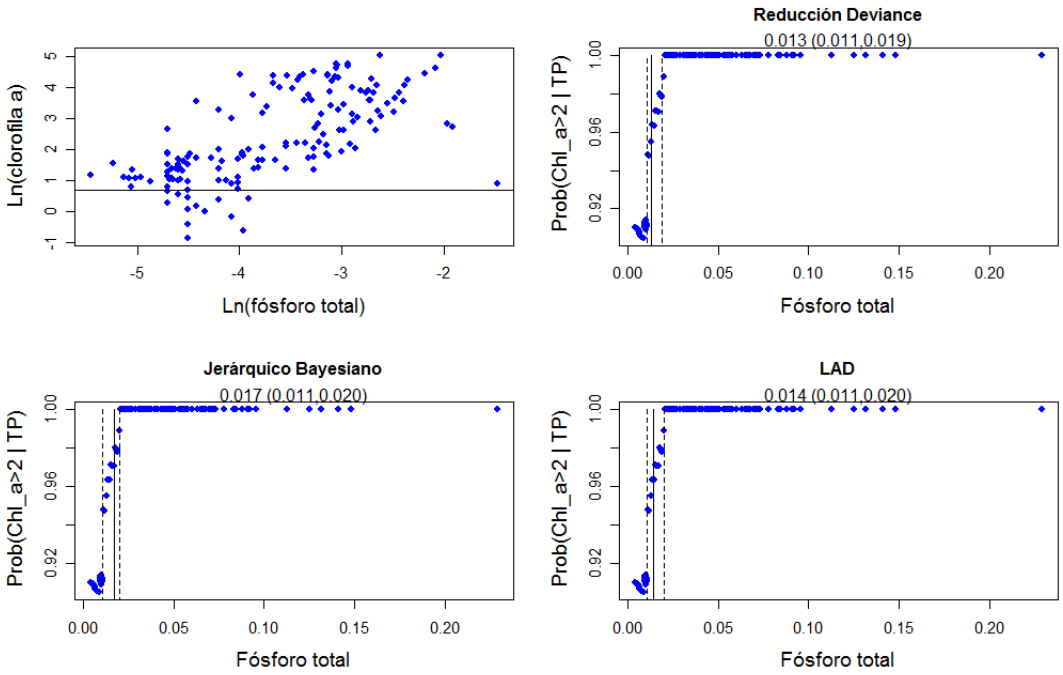

Figura 4.3.1: Umbrales para fósforo total cuando Chl_a > 2. Se observa que los umbrales y sus IC determinados por cada uno de los métodos, son cercanos.

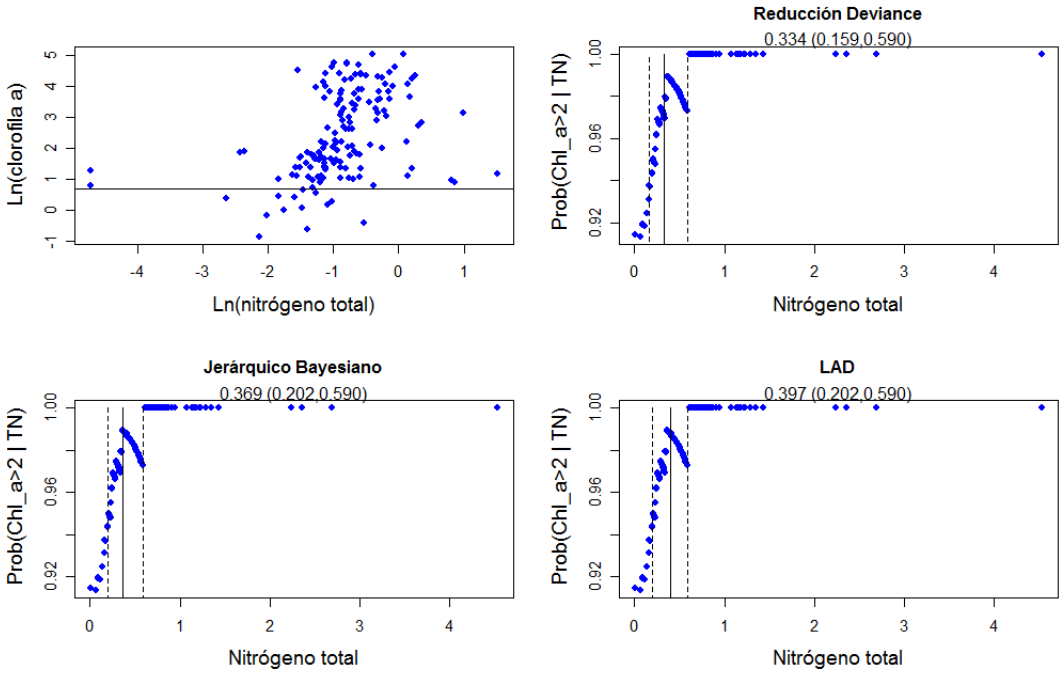

Figura 4.3.2: Umbrales para nitrógeno total cuando Chl_a > 2. Se observa que los umbrales y sus IC determinados por cada uno de los métodos, son cercanos.

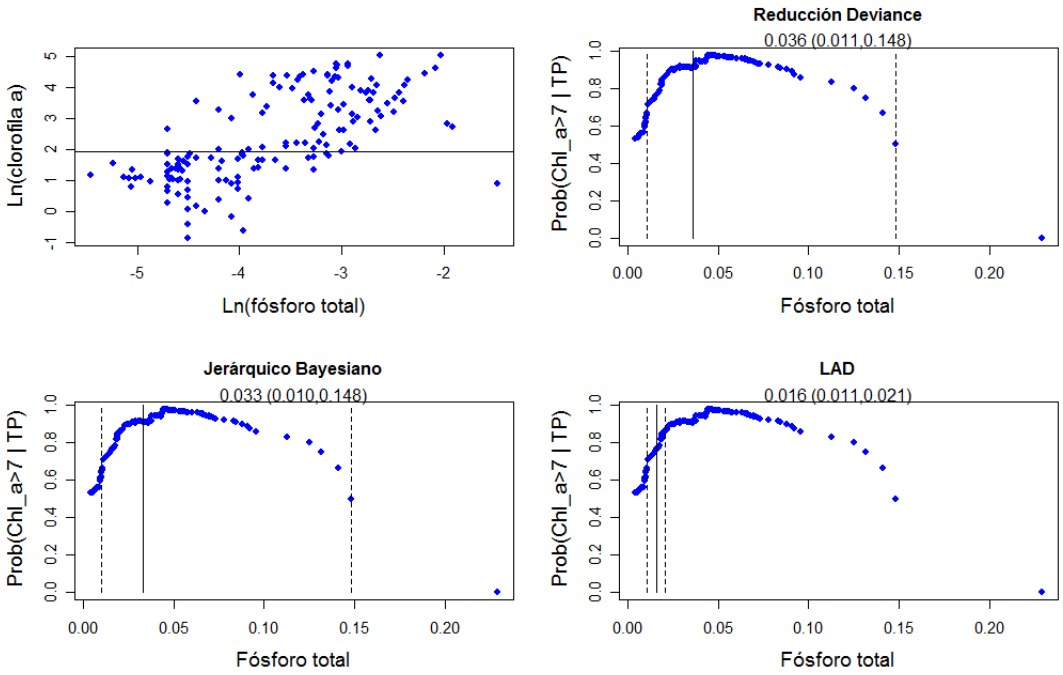

Figura 4.3.3: Umbrales para fósforo total cuando Chl_a > 7. Los umbrales y sus IC determinados por la reducción de la *deviance* no paramétrica y el método jerárquico bayesiano son más afectados por los valores atípicos de la probabilidad condicional, que los de la metodología LAD.

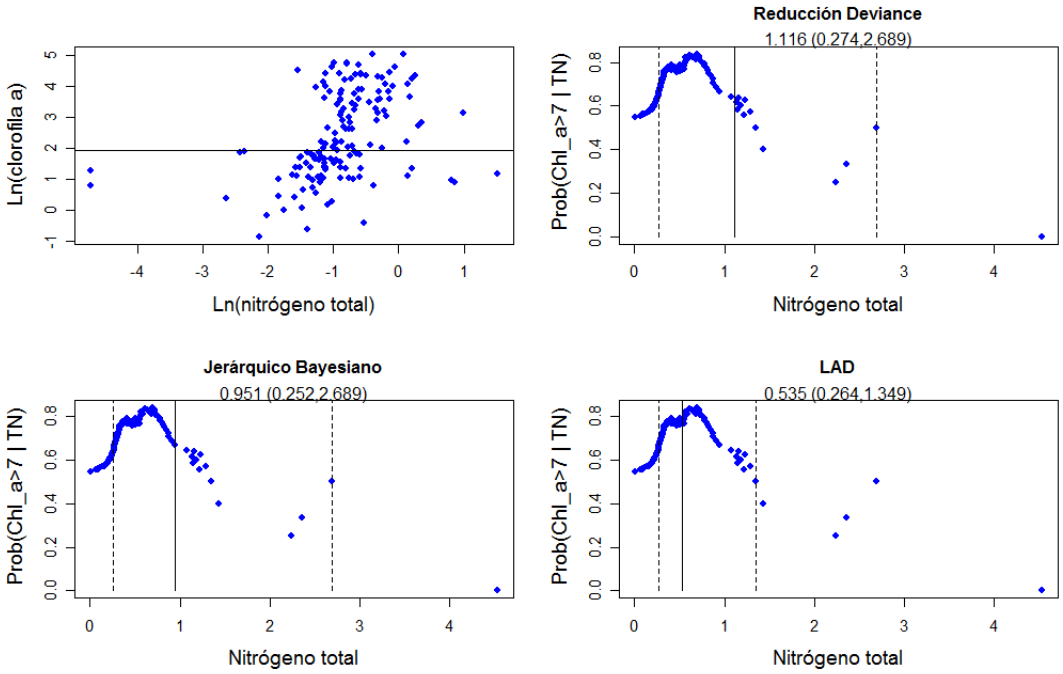

Figura 4.3.4: Umbrales para nitrógeno total cuando Chl_a > 7. Los umbrales y sus IC determinados por la reducción de la *deviance* no paramétrica y el método jerárquico bayesiano son más afectados por los valores atípicos de la probabilidad condicional, que los de la metodología LAD.

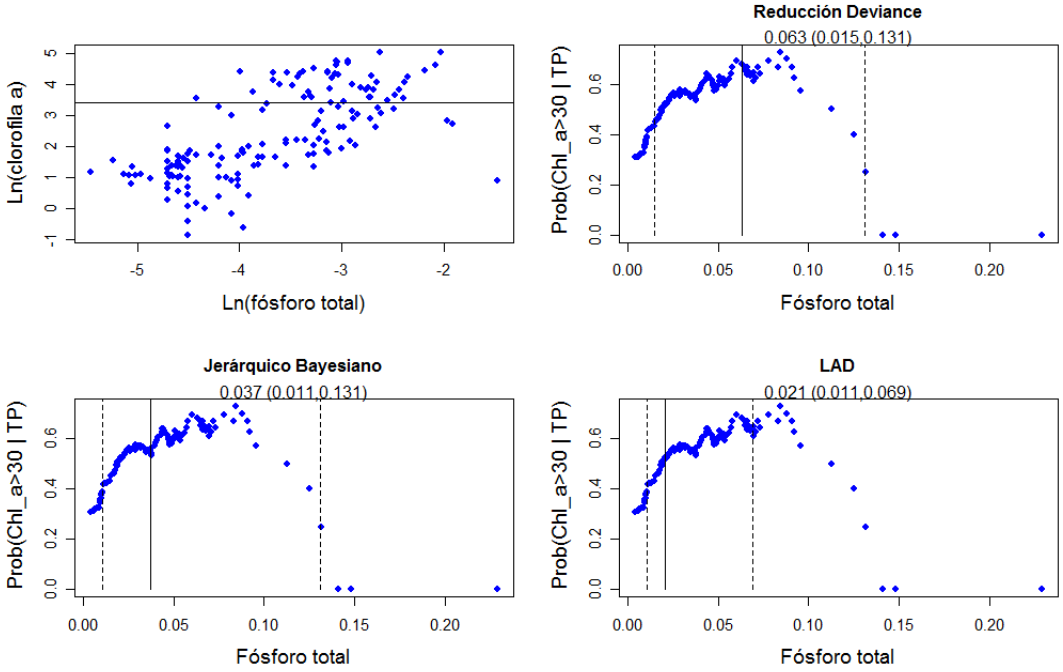

Figura 4.3.5: Umbrales para fósforo total cuando Chl_a $>$ 30. Los umbrales y sus IC determinados por la reducción de la *deviance* no paramétrica y el método jerárquico bayesiano se ven más afectados por los valores atípicos de la probabilidad condicional, que los de la metodología LAD.

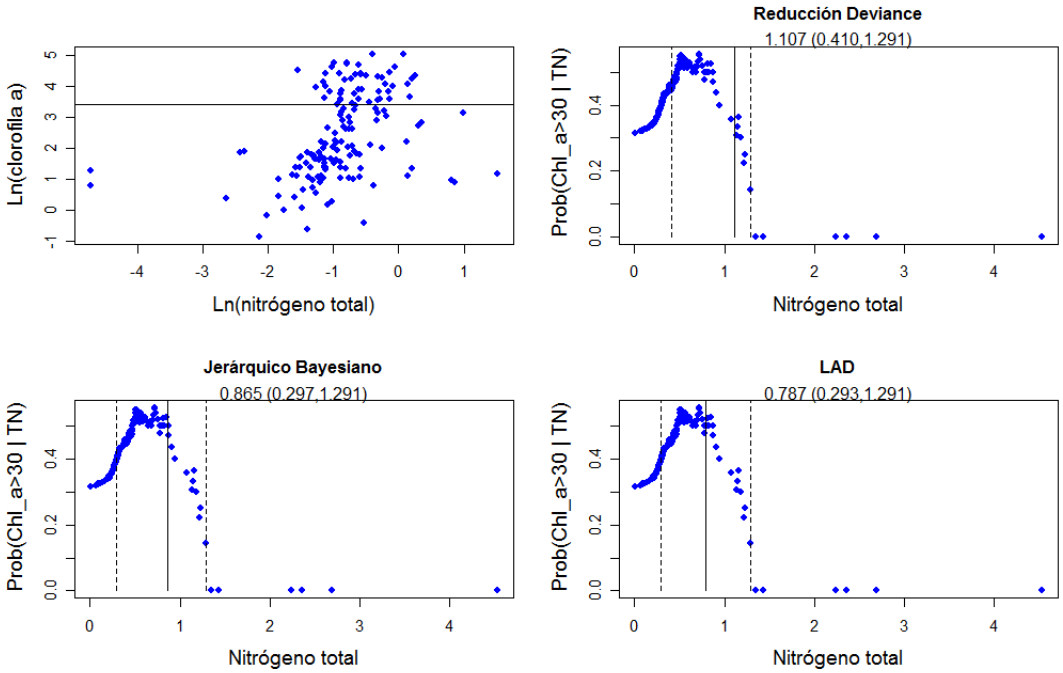

Figura 4.3.6: Umbrales para nitrógeno total cuando Chl_a $>$ 30. Para este caso, la reducción de la *deviance* no paramétrica es el método que más se afecta por los valores atípicos de la probabilidad condicional.

Se puede observar en las figuras anteriores 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. Las figuras 4.3.1 y 4.3.2 describen los comportamientos de la probabilidad condicional para Chl_a mayor que 2. Los umbrales para este límite trófico son muy similares y no muestran diferencias notables, en gran parte esto se debe a que para este valor las curvas de probabilidad condicional son bastante regulares.

En las figuras 4.3.3 y 4.3.4 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 valores y por lo tanto, presentan más variación en sus resultados y mayor amplitud en los 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.

En las figuras 4.3.3 y 4.3.4 se presenta mayor cantidad de valores atípicos lo que da como resultado mayor variación en la determinación de los umbrales por cada uno de los métodos y mayor amplitud en sus intervalos de confianza. Al igual que en las figuras 4.3.3 y 4.3.4 la metodología LAD es el método menos sensible a la presencia de estas observaciones.

Capítulo 5

Conclusiones

Algunos de los aspectos que más se pueden reflejar en el estudio realizado son los siguientes:

1. Los intervalos de confianza que no se traslapan y el ajuste de un modelo no lineal no han sido muy útiles en la determinación de los umbrales para embalses en Puerto Rico. Unas de las posibles causas son: En la naturaleza es difícil encontrar el comportamiento teórico que plantea la literatura, y los pocos datos hacen más frecuente este tipo de situaciones. Estas situaciones pueden lograr que ciertos métodos no sean muy útiles en la determinación de umbrales.

2. En los tres métodos, los umbrales son comparables solo en los límites oligo-mesotrófico. En los límites meso-eutrófico y eu-hipereutrófico hay diferencia, y LAD estima los umbrales más bajos.

3. Solo los umbrales determinados por LAD mantienen la monotonicidad para ambas variables. Esta propiedad es necesaria dada la relación monótona entre el umbral de Chl a y TN, TP.

4. El estimador *Bootstrap* resultó ser un estadístico de mejor aproximación a los umbrales esperados según la literatura, que los determinados solo a partir de una muestra.

5. La reducción de la *deviance* 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.

Capítulo 6

Trabajos Futuros

Aunque se utilizaron diversos métodos en la determinación de umbrales, aun quedan ciertos aspectos por aclarar y mejorar. Esto puede ser motivo de investigaciones futuras:

1. Si bien, se obtuvieron umbrales bastante similares por cada uno de los métodos, no está claro cuál de ellos da una mejor información en la determinación del umbral. Es necesario comparar los métodos mediante inferencia formal y simulaciones.

2. Se podría estudiar métodos alternativos de calcular los intervalos de confianza mediante la técnica del *Bootstrap* u otro método que pueda ser útil.

3. Se podría observar mediante simulaciones si se pueden lograr modificaciones en los métodos que puedan llevar a mejores resultados en la determinación de los umbrales.

Bibliografía

[1] Austin, P. C. y Hux, J. E. (2002). A Brief note on overlapping confidence intervals. *Journal of Vascular Surgery*, 36(1), 194-95.

[2] Casella, G. y Berger, R. L. (2002). *Statistical Inference*. (2a. ed.) Belmont, CA: Duxbury Press.

[3] Crawley, M. J. (2007). *The **R** Book*. Londres: Wiley.

[4] Cumming, G. (2009). Inference by eye: Reading the overlap of independent confidence intervals. *Statistics in Medicine*, 28(2) 205-220.

[5] Dobson, A. J. y Barnett, A. G. (2008). *An Introduction to Generalized Linear Models*. (3a. ed.) New York: Chapman & Hall.

[6] Dalgaard, P. (2008). *Introductory Statistics with **R***. (2a. ed.) New York: Springer.

[7] Efron, B. y Tibshirani, R. (1993). *An Introduction to the Bootstrap*. New York: Chapman & Hall.

[8] Faraway, J. J. (2006). *Extending the linear Model with **R***. New York: Chapman & Hall.

[9] Paul, J. F. y McDonald, M. E. (2005). Development of empirical, geographically specific water quality criteria: A conditional probability analysis approach. *Journal of American Water Resources Association*, 41(5), 1211-23.

[10] Perreault, L., Bernier, J., Bobée, J. y Parent, E. (2000). Bayesian change-point analysis in hydrometeorological time series. Part 1. The normal model revisited. *Journal of Hydrology*, 235(3-4), 242-63.

[11] Peterson, J. T., Haas, T. C. y Lee, D. C. (1999). *An Evaluation of Parametric and Nonparametric Models of Fish Population Response* (Informe No. 92-032-00). Portland OR: BPA Report DOE/BP-25866-8.

[12] Qian, S. S., King, R. S., y Richardson, C. J. (2003). Two statistical methods for the detection of environmental thresholds. *Journal of Ecological Modelling*, 166(1-2), 87-97.

[13] Ritz, C. y Streibig, J. C. (2008). *Nonlinear Regression with* ***R***. New York: Springer.

[14] Robert, C. P. y Casella, G. (2009). *Introducing Monte Carlo Methods with* ***R***. New York: Springer.

[15] Segal, M. R. (2008). Regression trees for censored data. *Journal of the International Biometric Society,* 44(1), 35-47.

[16] USEPA (U.S. Environmental Protection Agency). (2005). [base de datos]. Mid Atlantic Streams 1993-1996. Disponible en: http://www.epa.gov/emap/remap/html/three/data/index.html [2005, agosto].

[17] USEPA (United States Environmental Protection Agency). (2009). *National Lakes Assessment: A Collaborative Survey of the Nation's Lakes*. Washington, D.C. EPA 841-R-09-001.

[18] Venables, W. N. y Ripley, B. D. (2002). *Modern Applied Statistic With S.* (4a. ed.) New York: Springer.

Anexo 1: Datos del Mid Atlantic

YEAR	SAMPLED	VISIT_NO	ORDER	PCT_POOLS	EPT_RICH	PCT_FN
1994	Yes	1	2	90	1	83.636
1993	Yes	1	1	12.6667	23	1.818
1994	Yes	1	1	34.3137	1	30.909
1994	Yes	1	3	18	2	18.182
1994	Yes	1	1	30.6667	11	5.455
1994	Yes	1	1	10	12	14.545
1993	Yes	1	1	10	2	45.455
1993	Yes	1	3	11	24	1.818
1993	Yes	1	1	45.3333	11	1.818
1993	Yes	1	1	5.8824	15	0
1993	Yes	1	3	5	10	41.818
1993	Yes	1	2	11.4286	19	3.774
1993	Yes	1	1	37.8641	5	25.455
1993	Yes	1	2	6	12	3.704
1993	Yes	1	1	32.6531	4	18.182
1993	Yes	1	1	6.1644	9	20
1993	Yes	1	3	7	14	16.364
1993	Yes	1	3	8	17	12.727
1993	Yes	1	1	93.3333	10	40
1993	Yes	1	1	5	11	10.909
1994	Yes	1	1	31.9444	13	0
1993	Yes	1	2	19	2	30.909
1993	Yes	1	2	45	3	10.909
1993	Yes	1	3	0	3	10.909
1993	Yes	1	2	4	21	0
1993	Yes	1	3	5.814	7	22.917
1993	Yes	1	2	7.0707	1	21.818
1993	Yes	1	1	15.6463	9	10.909
1994	Yes	1	2	0	16	1.818
1994	Yes	1	3	11	22	1.818
1994	Yes	1	2	0	6	12.727
1994	Yes	1	1	6	19	25.455
1994	Yes	1	2	12.8713	10	32.727
1994	Yes	1	1	13	17	10.909
1994	Yes	1	2	38	6	25.455
1994	Yes	1	1	2	10	16.364
1994	Yes	1	1	0	20	5.714
1994	Yes	1	1	14	8	14.815
1994	Yes	1	2	36	11	3.636

```
1994 Yes 1 1 14 22 7.273
1994 Yes 1 2 0 21 5.455
1994 Yes 1 2 19.0476 19 0
1994 Yes 1 1 1 13 1.818
1994 Yes 1 1 32.8859 2 70.909
1994 Yes 1 2 45 11 23.636
1993 Yes 1 1 14 17 0
1993 Yes 1 1 6.7308 2 0
1993 Yes 1 2 16 16 0
1993 Yes 1 3 0 7 3.636
1993 Yes 1 1 15.4639 5 14.545
1993 Yes 1 1 8 8 29.091
1993 Yes 1 1 15 8 16.364
1993 Yes 1 3 2 10 0
1993 Yes 1 2 1 24 0
1993 Yes 1 2 22 21 10.909
1993 Yes 1 1 6.0606 7 16.364
1993 Yes 1 3 17 15 1.818
1993 Yes 1 1 23.7624 11 9.259
1993 Yes 1 2 13 6 29.091
1994 Yes 1 3 34 1 16.364
1994 Yes 1 3 11 9 47.273
1994 Yes 1 1 98.0769 1 23.636
1994 Yes 1 3 14 21 7.273
1994 Yes 1 2 5 16 21.818
1994 Yes 1 1 10 14 0
1994 Yes 1 1 5.3333 17 10.909
1994 Yes 1 3 18 11 14.545
1994 Yes 1 1 22.7586 4 50.909
1994 Yes 1 2 0 3 10.909
1994 Yes 1 1 23.3333 5 1.818
1994 Yes 1 3 3 15 27.273
1994 Yes 1 1 15 5 18.182
1994 Yes 1 3 64.3564 9 25.455
1994 Yes 1 3 2 21 21.818
1994 Yes 1 2 23 14 0
1994 Yes 1 3 3 3 3.704
1994 Yes 1 1 0.7752 7 38
1993 Yes 1 2 16.0377 7 37.736
1993 Yes 1 3 0 3 100
1993 Yes 1 1 0.9091 15 45.455
1993 Yes 1 1 38 10 1.818
1993 Yes 1 3 43.8776 3 12.727
```

```
1993 Yes 1 1 0.6803 15 10.909
1993 Yes 1 1 21.3333 12 0
1993 Yes 1 1 0 29 5.455
1993 Yes 1 1 0 15 1.818
1993 Yes 1 2 19 11 1.818
1993 Yes 1 3 0 16 0
1993 Yes 1 3 0 0 23.636
1994 Yes 1 2 3 13 0
1994 Yes 1 3 9 13 0
1994 Yes 1 2 12 10 0
1994 Yes 1 1 8 23 1.818
1994 Yes 1 2 0 20 7.273
1994 Yes 1 2 0 18 1.818
1994 Yes 1 1 9 24 0
1994 Yes 1 3 0 14 9.091
1994 Yes 1 3 9 16 10.909
1994 Yes 1 1 16.1074 6 49.091
```

Anexo 2: Probabilidad Condicional Datos Mid Atlantic

PCT_FN	Prob. Condicional
0	0.3535354
1.818	0.4146341
3.636	0.4782609
3.704	0.4776119
3.774	0.4769231
5.455	0.484375
5.714	0.5081967
7.273	0.5166667
9.091	0.5438596
9.259	0.5535714
10.909	0.5636364
12.727	0.6222222
14.545	0.6190476
14.815	0.6410256
16.364	0.6315789
18.182	0.6363636
20	0.6
21.818	0.6206897
22.917	0.6538462
23.636	0.64
25.455	0.6363636
27.273	0.6666667
29.091	0.7058824
30.909	0.6666667
32.727	0.6153846
37.736	0.6666667
38	0.6363636
40	0.6
41.818	0.6666667
45.455	0.75
47.273	0.8333333
49.091	1
50.909	1
70.909	1
83.636	1
100	1