

# MODELOS DE ÁRBOL DE REGRESIÓN BAYESIANO: UN ESTUDIO DE CASO.

O. Juárez<sup>\*1</sup> y E.Castells<sup>\*\*2</sup>

\*Facultad de Matemática. Universidad Autónoma de Guerrero. México.

\*\*Área Económico-Administrativa. Universidad Veracruzana, México

## RESUMEN

En este trabajo se explora el método Bayesiano de selección de árboles de regresión propuesto por Chipman et al. (1998a), así como otros resultados afines de los mismos autores. Las aplicaciones del referido método requieren que se generen modelos utilizando los métodos de Monte Carlo vía Cadenas de Markov, y para las mismas ha resultado conveniente un agrupamiento de dichos modelos, para lo cual se necesitan métricas. Aquí proponemos una nueva métrica para agrupar los modelos. También se hace un estudio por simulaciones con el objetivo de explorar la posibilidad de disminuir el número de modelos que son necesarios generar para determinar (con el enfoque de Chipman et al. (1998a)) un árbol que explique satisfactoriamente a un conjunto de datos. Se aplica la metodología a los datos que utilizaron Denison et al. (1998) en la ilustración de su método, lo que permite hacer comparaciones. Por último se presentan los resultados de la aplicación en el estudio socio-económico que precedió a un proyecto hidrológico.

## ABSTRACT

The Bayesian method for selecting regression models proposed by Chipman et al. (1998a) as well as some related results of the same authors are explored. In applications of this method, forming groups of the generated models has resulted a very useful tool (models are generated by Monte Carlo Markov Chains) so some metrics are required and here we propose a new one for grouping the models. The possibility of reducing the number of necessary models to be generated in order to determine (with Chipman et al. (1998a) approach) a tree which give a satisfactory explanation of the data is explored by means of a simulation study. We apply the method to the data used by Denison et al. (1998) and some comparisons are made. Lastly we analyze the results of the application to the data of some socio-economical study.

**KEY WORDS:** Monte Carlo Markov Chain, Regression tree, Metric on Models, Log. Likelihood, Integrated Likelihood.

MSC:62C12

## 1. INTRODUCCIÓN

Los modelos de Árbol de Regresión y Clasificación (C&RT, Classification & Regression Trees), fueron introducidos en la Estadística por Breiman et al. (1984). Los autores utilizan el término “modelos de árbol de regresión” cuando la variable respuesta es cuantitativa y el de “modelos de árbol de clasificación” cuando ésta es cualitativa. La idea fundamental de dicha metodología consiste en particionar el espacio de las variables independientes ( $X_1, X_2, \dots, X_p$ ) en forma tal que los valores de la variable de respuesta sean cada vez más homogéneos dentro de las clases de dicha partición. Los modelos C&RT son fáciles de aplicar e interpretar, por tal motivo se han hecho populares en diferentes áreas como Ensayos clínicos, Medicina, Meteorología, Finanzas (ver Andriyashin (2005), Camdeviren et al. (2007), Davis y Elder (2002), Teng et al. (2007), Lewis (2000) y Wada y Nakamura (2004)).

---

1 tavis53@yahoo.com.mx

2 ernestinaeg@yahoo.com

Chipman et al. (1998a) y Denison et al. (1998), proponen nuevas metodologías que utilizan los modelos C&RT en combinación con los métodos de Monte Carlo vía cadenas de Markov (MCMC, Markov Chain Monte Carlo). Ellos utilizan los métodos MCMC para la exploración de la distribución a posteriori. En el segundo trabajo, la distribución a priori del modelo se toma como una distribución Poisson truncada con parámetro  $k$  ( $k$  identifica el número de nodos terminales) y para la exploración de la distribución a posteriori se utiliza el método MCMC de saltos reversibles propuesto por Green (1995).

Chipman et al. (1998a), proponen un procedimiento algorítmico para definir en forma implícita la distribución a priori,  $p(T)$ , del árbol y especifican una distribución condicional de la variable de respuesta en los nodos terminales de éste. Para los parámetros de esta distribución se toma una distribución a priori conjugada a la cual nos referimos más adelante y con ella se deduce una expresión de la distribución a posteriori que posibilita su exploración mediante el algoritmo de Metropolis-Hastings.

Una característica que tiene el enfoque de Chipman et al. (1998a), es la generación de distintos modelos C&RT a partir de un mismo conjunto de datos. Chipman et al. (1998b y 2001) exploran un conjunto de métricas que permiten agrupar los árboles de acuerdo a ciertas similitudes. Este agrupamiento facilita la selección de alguno de ellos. Nosotros proponemos aquí una métrica definida en términos de la peor predicción que hace cada modelo.

En este trabajo se reportan los resultados que arrojó el uso del método de Chipman et al. (1998a) en el análisis de datos de un proyecto hidrológico en Guerrero, México. Con el objetivo de completar la ilustración y mostrar la utilidad de ese método, se presentan los resultados obtenidos utilizando un conjunto de datos publicados en Bruntz et al. (1974), dichos datos fueron utilizados por Denison, et al. (1998) para ilustrar su método, que aunque también es un método Bayesiano tiene otros supuestos teóricos. Además, se reportan los resultados de un estudio por simulaciones que se realizó para explorar la posibilidad de disminuir el número de corridas que son necesarias en las aplicaciones.

## 2. MODELOS DE ÁRBOL DE REGRESIÓN BAJO EL ENFOQUE BAYESIANO

### 2.1. Definiciones primarias

Para la construcción de un modelo de árbol binario se requiere de observaciones de una variable de respuesta  $Y$ , y de variables explicativas o predictoras  $X = (X_1, X_2, \dots, X_p)$ , las cuales pueden ser continuas o categóricas.

#### Definición 2.1 Árbol binario.

Un árbol es un conjunto de nodos definidos a través de los valores de ciertas variables, la figura No.5 da una ilustración gráfica (ver Breiman et al., 1984). Los nodos se clasifican en inicial o raíz, internos o de división y terminales. Si la división de los nodos ocurre de forma tal que de cada nodo que se divide resultan dos nuevos nodos, se dice que el árbol es binario. El nodo que se divide se llama nodo padre y a los dos resultantes se les llama nodos hijos izquierdo y derecho.

#### Definición 2.2. Modelo de árbol de regresión.

Un modelo de árbol de regresión es una descripción de la distribución condicional de  $Y$  dado  $X$ . Los dos componentes fundamentales del modelo son: un árbol binario con  $b$  nodos terminales y el vector de parámetros  $\Theta = (\theta_1, \theta_2, \dots, \theta_b)$ , donde el parámetro  $\theta_i$  está asociado al nodo terminal  $N_i$ . La densidad condicional de  $Y$  dado  $\theta_i$  se denota por  $f(y | \theta_i)$  (Chipman et al., 1998a).

#### Definición 2.3 Profundidad de un nodo.

La profundidad  $d$  de un nodo se define como el número de nodos de división que se encuentran por arriba de él en el árbol (Chipman et al., 1998a).

A continuación se explica la regla de división de los nodos.

### **Regla de división.**

Para dividir un nodo de un árbol binario se consideran dos pasos: 1.- se selecciona una variable  $X_i$  y 2.- Si la variable seleccionada resulta cuantitativa, se selecciona aleatoriamente un valor  $r$  y se asignan al nodo hijo izquierdo las observaciones que cumplan la condición  $X_i \leq r$ , las restantes son asignadas al nodo hijo derecho. Si la variable seleccionada resulta cualitativa, se selecciona un subconjunto  $A$  de las categorías de la variable  $X_i$  y las observaciones con valores pertenecientes al conjunto  $A$  se asignan al nodo hijo izquierdo, las restantes al nodo hijo derecho (Chipman et al., 1998a).

La aplicación sucesiva de la regla anterior, partiendo de un nodo raíz, genera un árbol binario.

## **2.2 Especificación de la distribución a priori del árbol y de los parámetros.**

### **2.2.1 Distribución a priori del árbol.**

La propuesta de Chipman et al. (1998a) para la especificación de la distribución a priori del árbol es un proceso estocástico árbol-generador que consiste en un algoritmo de tres pasos:

i) Se inicia con el árbol trivial de un nodo.

ii) Cada nodo Terminal se divide con probabilidad  $p = \alpha(1 + d)^{-\beta}$ , donde  $\alpha \in (0,1)$  y  $\beta \geq 0$  son parámetros seleccionados previamente que definen el tamaño y la forma del árbol.

iii) Cuando el nodo se puede dividir, se elige una variable predictora  $X_i$  utilizando un mecanismo uniforme discreto; de igual manera se selecciona el valor de  $r$  o el conjunto  $A$  necesarios para aplicar la regla de división. Por último, se asignan los valores de las variables a los nodos hijos izquierdo y derecho.

Se repiten los pasos ii) y iii). El algoritmo termina cuando no existen nodos que se puedan dividir. Los nodos que no son divididos se llaman nodos terminales.

Este algoritmo asocia a los nodos de división (de los árboles que genera) la probabilidad de dividirse  $p = \alpha(1 + d)^{-\beta}$  y a los nodos terminales el complemento  $q = 1 - p$ . Por lo tanto, a un árbol  $T$  se le asocia una probabilidad a priori ( $p(T)$ ) que es el producto de las probabilidades asociadas a cada nodo.

### **2.2.2 Distribución a priori de los parámetros.**

Los parámetros del modelo de árbol de regresión en cada uno de los  $b$  nodos terminales ( $i = 1, 2, \dots, b$ ) son la media y varianza de la variable de respuesta en dichos nodos. Sobre estos parámetros se especifica un modelo probabilístico.

### **2.2.3 Modelo de cambio de media y varianza.**

En este modelo se supone desigualdad entre las medias y entre las varianzas de los diferentes nodos terminales, esto es,  $\theta_i = (\mu_i, \sigma_i^2)$ . Se supone, además, que la variable  $Y$  sigue una distribución normal,

La distribución a priori conjugada de los parámetros del modelo es una normal-gamma inversa, esto es:

$$y_{i1}, y_{i2}, \dots, y_{in_i} | (\mu_i, \sigma_i^2) \sim N(\mu_i, \sigma_i^2) \quad \text{para } i=1, 2, \dots, b \quad (2.1)$$

$$\mu_i | \sigma_i^2 \sim N(\mu, \frac{\sigma_i^2}{a}) \quad \text{y} \quad \sigma_i^2 \sim IG(\frac{\nu}{2}, \frac{\nu\lambda}{2}) \quad (2.2)$$

(Chipman et al., 1998a).

### 2.3 Derivación de la distribución a posteriori del modelo.

Aplicando el Teorema de Bayes, se obtiene que la distribución a posteriori del modelo de árbol de regresión, salvo la constante de integración, se expresa como:

$$p(T|Y, X) \propto p(Y|T, X)p(T) \quad (2.3)$$

donde  $p(Y | T, X)$  es la verosimilitud integrada que se obtiene mediante la integral en (2.4):

$$p(Y|T, X) = \int_{\Theta} p(Y|\theta, T, X)p(\theta|T, X)d\theta \quad (2.4)$$

siendo  $p(Y | \theta, T, X)$  la verosimilitud de las observaciones de  $Y$  y  $p(\theta | T, X)$  la distribución de los parámetros del modelo. La verosimilitud integrada juega un papel importante en la obtención de una expresión analítica para la distribución a posteriori del modelo y debe permitir que ésta pueda ser evaluada en cualquier punto del espacio de los modelos de árbol de regresión.

Como la variable respuesta se distribuye según una normal (2.1) la verosimilitud de los datos está dada por la expresión en (2.5):

$$p(Y|\theta) = \left[ \prod_{i=1}^b (2\pi\sigma_i^2)^{-\frac{n_i}{2}} \right] \exp \left\{ \sum_{i=1}^b \left( -\frac{1}{2\sigma_i^2} \sum_{j=1}^{n_i} (y_{ij} - \mu_i)^2 \right) \right\} \quad (2.5)$$

La forma conjugada de la distribución a priori de los parámetros  $(\mu_i, \sigma_i^2)$  es la dada en (2.6):

$$p[(\mu_i, \sigma_i^2) | \bar{\mu}, a, \frac{\nu}{2}, \frac{\nu\lambda}{2}] = \frac{\sqrt{a}}{\sqrt{2\pi\sigma_i^2}} \exp\left\{-\frac{a}{2\sigma_i^2}(\mu_i - \bar{\mu})^2\right\} \times \frac{(\nu\lambda/2)^{\frac{\nu}{2}}}{\Gamma(\nu/2)} (\sigma_i^2)^{-(\frac{\nu}{2}+1)} \exp\left(-\frac{\nu\lambda}{2\sigma_i^2}\right) \quad (2.6)$$

Sustituyendo (2.5) y (2.6) en (2.4) y calculando las integrales involucradas se obtiene (2.7) (Ver Chipman et al., 1998a):

$$p[Y | X, T] = \prod_{i=1}^b (\pi)^{-\frac{n_i}{2}} (\nu\lambda)^{\frac{\nu}{2}} \frac{\sqrt{a}}{\sqrt{n_i + a}} \frac{\Gamma(\frac{n_i + \nu}{2})}{\Gamma(\nu/2)} (s_i + t_i + \nu\lambda)^{-\frac{n_i + \nu}{2}} \quad (2.7)$$

con  $s_i$  y  $t_i$  dados por:

$$s_i = \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2, \quad t_i = \frac{n_i a}{(n_i + a)} (\bar{y}_i - \bar{\mu})^2 \quad (2.8)$$

La expresión (2.7) representa la verosimilitud integrada del modelo de cambio de media y varianza.

### 3. EXPLORACIÓN DE LA DISTRIBUCIÓN A POSTERIORI.

#### 3.1. El algoritmo de Metropolis-Hastings.

Para su exploración Chipman et al. (1998a) simulan una muestra de la distribución a posteriori por medio de uno de los métodos de MCMC, el llamado algoritmo de Metropolis-Hastings. Este método tiene la ventaja de que la distribución no tiene que estar completamente definida y es muy eficiente aun cuando la dimensión del problema es muy grande. Tiene, además, la cualidad de simular muestras de distribuciones aun cuando el vector de parámetros tenga dimensión variable, y esto es justamente lo que sucede aquí.

Como lo explican Chib y Greenberg (1995) y Gamerman (1997), cuando se tiene una distribución  $\pi$  de la cual se quiere obtener una muestra; pero por su forma resulta complicado o costoso obtenerla, se puede generar una cadena de Markov vía el algoritmo de Metropolis-Hastings. Para implementar este algoritmo es necesario construir un kernel de transición  $p(T^i, T^*)$  de tal forma que la distribución de equilibrio de la cadena de Markov sea la distribución  $\pi$ , para que esto ocurra, el kernel de transición debe cumplir la condición de reversibilidad (3.1):

$$\pi(T^i) p(T^i, T^*) = \pi(T^*) \cdot p(T^*, T^i) \quad (3.1)$$

Expresando el kernel de transición como el producto en (3.2),

$$p(T^i, T^*) = q(T^i, T^*) \alpha(T^i, T^*) \quad (3.2)$$

donde  $q(T^i, T^*)$  es la densidad

arbitraria generadora de candidatos, que para el caso discreto debe satisfacer la condición

$\sum_j q(T^i, T^j) = 1$ ; y  $\alpha(T^i, T^*)$  es la probabilidad de aceptación. Hastings (1970), propone que esta probabilidad de aceptación sea

$$\alpha(T^i, T^*) = \min \left\{ \frac{\pi(T^*) q(T^*, T^i)}{\pi(T^i) q(T^i, T^*)}, 1 \right\} \quad (3.3)$$

Con esta definición de probabilidad de aceptación y la densidad generadora de candidatos, se logra un kernel de transición que satisface la condición de reversibilidad (3.1) y una cadena de Markov que tiene distribución de equilibrio  $\pi$ .

#### 3.2 Kernel de transición.

El algoritmo de Metropolis-Hastings se emplea, en el caso que nos ocupa, para simular una cadena de Markov de árboles  $(T^0, T^1, T^2, \dots)$  a partir de la cual se puede obtener una muestra simulada de la distribución a posteriori. El algoritmo se inicia con un árbol  $T^0$ , y a partir de éste se generan los demás elementos de la cadena de Markov, esto se hace de la forma siguiente: a partir del  $i$ -ésimo elemento  $T^i$  de la cadena se genera un árbol-candidato  $T^*$ , y se decide si  $T^*$  será el elemento  $T^{i+1}$  de la cadena. Esta decisión se toma con probabilidad  $\alpha(T^i, T^*)$  dada en (3.4):

$$\alpha(T^i, T^*) = \min \left\{ \frac{q(T^*, T^i) p(Y | X, T^*) p(T^*)}{q(T^i, T^*) p(Y | X, T^i) p(T^i)}, 1 \right\} \quad (3.4)$$

en caso contrario se considera  $T^{i+1} = T^i$ .

Para obtener el árbol candidato  $T^*$  a través de una modificación de  $T^i$ , se selecciona en forma aleatoria uno de los cuatro pasos siguientes, propuestos por Chipman et al. (1998a):

- i) Crecimiento: utilizando una distribución uniforme discreta se selecciona un nodo terminal, que se dividirá o no de acuerdo a las reglas de selección de variables y valores de división.
- ii) Podado: utilizando una distribución uniforme discreta se selecciona un nodo interno que tenga sus nodos hijos, izquierdo y derecho, como nodos terminales, se eliminan estos nodos hijos.
- iii) Cambio de la regla de división de un nodo interno: con probabilidades dadas por una distribución uniforme discreta se selecciona un nodo interno para cambiar la regla de división.
- iv) Intercambio de las reglas de división de dos nodos internos: utilizando una distribución uniforme discreta se selecciona un par de nodos internos, padre e hijos, y se intercambian sus reglas de división.

La tasa de aceptación es el porcentaje del número de veces que la cadena acepta un candidato como elemento de la cadena. Este número es indicativo del comportamiento de la cadena (Chib y Greenberg, 1995).

Para implementar el algoritmo se requiere una densidad generadora de candidatos en la cual se incorporen los cuatro movimientos anteriores (las propiedades de esta densidad se pueden encontrar en Chipman et al., 1998a). El kernel de transición depende no sólo del número de variables, sino también del número de valores disponibles de la variable involucrada y del número de nodos donde sea posible realizar el movimiento.

### **3.3 Estrategia de búsqueda.**

La distribución de probabilidad a posteriori  $p(T | X, Y)$  es multimodal y esto tiene repercusiones en la estrategia de búsqueda ya que el algoritmo busca los puntos modales, una vez que los localiza, la probabilidad de pasar a otro punto modal es pequeña, empleando en esto mucho tiempo de cómputo. La cadena se "atasca" alrededor de un punto modal porque los pasos definidos por el kernel de transición son pequeños y también por el propio algoritmo de Metropolis-Hastings (Besag y Green, 1993). Para evitar que el algoritmo se quede atascado durante un tiempo largo, Chipman et al. (1998a) proponen reiniciar el algoritmo varias veces para una búsqueda rápida y una mejor exploración de la distribución a posteriori, en contraposición a las corridas largas. En la Sección 7 se presentan los resultados de un estudio por simulación que se efectuó para analizar si es posible disminuir el número de modelos a generar.

### **3.4 Criterios de selección de los modelos.**

El criterio para seleccionar un buen modelo, en este caso un buen árbol de regresión, depende de los objetivos con los que se utilizará el mismo. Un criterio es seleccionar aquel árbol con mayor probabilidad a posteriori (2.3), otro pudiera ser la mayor verosimilitud integrada del modelo  $p(Y | X, T)$  (éste para evitar la influencia de la distribución a priori del árbol  $p(T)$ ), o como en el caso de la regresión lineal, la suma de los cuadrados de los residuales.

## **4. MEDIDAS PARA LA AGRUPACIÓN DE MODELOS**

En el contexto de la selección de modelos de árbol de regresión, las métricas se utilizan con el objetivo de agrupar los modelos que sean semejantes según alguna característica. De esta forma se reduce el número de modelos para una comparación final, tratando de seleccionar aquel modelo que mejor explique el comportamiento de los datos y no entre en contradicción con las características del fenómeno en estudio.

Existen diversas propuestas de definición de distancia entre los modelos de árbol de regresión. Dichas propuestas se definen en base a los diferentes elementos que caracterizan un modelo de árbol, como pueden ser la suma de los cuadrados de la diferencia entre el valor observado y el valor ajustado, las reglas de división o sobre la partición del espacio de regresores, entre otros.

En Chipman et al. (1998b y 2001), se exponen las métricas de ajuste, de partición y de árbol. En la métrica de ajuste se calculan los valores ajustados de las observaciones, la distancia entre los modelos se define como la suma de los cuadrados de las diferencias entre los valores ajustados. Para la métrica de partición, primero se cuenta el número de parejas de observaciones que se encuentran en el mismo nodo terminal, esto se hace para cada modelo; la distancia entre los árboles es la diferencia entre estos conteos, escalada por el total de parejas posibles. Para el caso de la métrica de árbol, se comparan las reglas de división existentes en el desarrollo de los modelos de árbol de regresión.

#### 4.1 Métrica general.

En Chipman et al. (1998b), en la sección de discusión, se plantea la posibilidad de construir otras métricas, una de ellas consiste en construir para cada árbol un vector resumen que contenga características tales como el tamaño, la frondosidad, grado de uso de cada variable, entre otras. Entonces un modelo de árbol de regresión se puede representar como un punto en un espacio de dimensión mayor que uno. La dimensión vendría determinada por el número de características del árbol que se utilicen. Las características del árbol que se podrían incluir en la definición de un vector son:

1) El tamaño: Es el número de nodos terminales que tiene el árbol.

2) Índice de frondosidad: Si  $d$  representa la profundidad máxima del árbol, el índice se define como:

$I_f = \frac{\text{No.de nodos terminales}}{2^d} \times 100$ , cuando el árbol está completamente saturado en la profundidad  $d$ , el valor del índice es 100% .

3) Variable de la primera división: Se registra la variable independiente utilizada en la división del nodo raíz.

4) Variable de la segunda división izquierda: Se registra la variable independiente utilizada en la división del nodo hijo izquierdo del nodo raíz.

5) Variable de la segunda división derecha: Se registra la variable independiente utilizada para la división del nodo hijo derecho del nodo raíz. Cada variable se identifica con un número natural.

6) Grado de uso de las distintas variables independientes: Se define el grado de uso para cada variable independiente (en el modelo de árbol de regresión) como:

$\text{Grado de uso de } x_i = \frac{\text{No.de nodos donde se usa } x_i}{\text{No.total de nodos del arbol}} \times 100$  .

7) Suma de los cuadrados residuales (SCR): El valor ajustado para cada una de las observaciones de la variable de respuesta en los nodos terminales del árbol, se estima mediante el promedio de sus observaciones en el nodo terminal correspondiente, con lo que la SCR se calcula mediante:  $\sum_{ij} e_{ij}^2 = \sum_i^b \sum_j^{n_i} (y_{ij} - \bar{y}_i)^2$ ,

donde  $b$  es el número de nodos terminales.

8) Logaritmo de la verosimilitud integrada: tal como se definió en (2.4).

Con estas definiciones, cada árbol está representado por un punto en un espacio  $\mathfrak{R}^m$ , esto es, el árbol  $T_k$  expresado en términos de sus coordenadas es  $T_k = (T_{k1}, T_{k2}, \dots, T_{km})$ , por lo que la distancia entre los árboles se define como:

$$d(T_k, T_l) = \left[ \sum_{j=1}^m (T_{lj} - T_{kj})^2 \right]^{1/2} \quad (4.1)$$

## 4.2 Métrica basada en la peor predicción.

Para comparar el modelo cuando se quiere utilizar éste con fines predictivos, proponemos una métrica sobre el conjunto de los modelos de árbol de regresión. Esta métrica asocia a cada modelo el máximo de la diferencia entre cada observación y su valor predicho.

$$\text{Sea } m(T_k) = \text{Max} |y_{ijk} - \hat{y}_{ijk}| \text{ con } i = 1, 2, \dots, n; \quad j = 1, 2, \dots, n_i$$

y se define la distancia entre los modelos de árbol  $T_l$  y  $T_k$  como:

$$d(T_l, T_k) = |m(T_l) - m(T_k)|$$

Esta asociación entre los modelos de árbol de regresión y los números reales positivos tiene ventajas, pues posibilita ordenar los modelos en el rango de valores de  $m(T_k)$  o en el rango de valores de las distancias  $d(T_l, T_k)$ . También se pueden agrupar con sólo generar una partición en cualquiera de los anteriores rangos.

La distancia propuesta apunta hacia cuán diferentes son los modelos en términos de la peor de la predicción.

## 5. EJEMPLO 1: UN EJEMPLO CON DATOS PUBLICADOS POR BRUNTZ ET AL. (1974).

### 5.1 Descripción de los datos y las cadenas de Markov obtenidas.

En Bruntz et al. (1974) aparece un conjunto de datos que contiene 111 observaciones de la cantidad de ozono ( $Y$ ) en distintos lugares de la zona urbana de Nueva York. Esta variable se trata de explicar en función de otras dos características meteorológicas: la temperatura ( $X_1$ ) y la velocidad del viento ( $X_2$ ). Los datos están disponibles en el paquete estadístico S-Plus, y se les aplicó la misma transformación que aplicaron Denison et al. (1998). El valor mínimo de  $X_1$  es 57 y su máximo 97; los correspondientes valores para  $X_2$  son 2.3 y 20.7, respectivamente.

En cuanto a la variable  $Y$ , su mínimo es 1.11 y su máximo 5.42, su promedio es 3.248 y la varianza 0.738.

A los hiperparámetros  $\alpha$  y  $\beta$ , se les asignaron los valores de  $\alpha = 0.95$  y  $\beta = 1.0$ . Se determinó que los valores de los hiperparámetros de la distribución a priori conjugada, normal-gamma inversa, convenientes son:  $\nu = 16$ ,  $\lambda = 0.3$ ,  $\bar{\mu} = 3.248$ ,  $a = 1$ .

Para la realización de los ejemplos se programaron diversas rutinas en el código del paquete estadístico S-Plus. En una de estas rutinas se implementó un algoritmo para generar las cadenas de Markov, en cada ejecución se realizan 10,000 iteraciones, reiniciándose cada 1,000. El programa se ejecutó 10 veces, obteniéndose 100 modelos, cada uno de éstos es el árbol final de 1,000 simulaciones. Los modelos generados se numeran del 1 al 100 para su identificación.

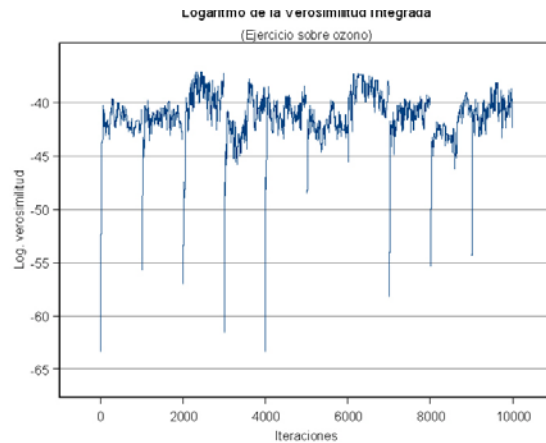


Figura No. 1



En la Figura No. 1, se muestra una ejecución del programa con los reinicios cada 1000 iteraciones. Se observa en cada cadena un periodo de calentamiento, después los valores del logaritmo de la verosimilitud de sus elementos se estabilizan dentro una banda de valores.

Con el criterio del mayor valor del logaritmo de la verosimilitud, el mejor modelo de árbol correspondiente al de la Figura No.1, que tiene un valor del logaritmo de la verosimilitud igual a  $-37.25$ ; además, tiene 10 nodos terminales, una frondosidad de 31.25%, entre otras características (Tabla No. 1).

## 5.2 Métrica general.

Puesto que se tienen 100 árboles de regresión, al calcular las correspondientes distancias se obtiene una matriz de  $100 \times 100$  con la diagonal principal nula. Para lograr una representación gráfica de las distancias en el plano  $\mathcal{R}^2$  (Figura No. 2), se aplica la técnica del escalamiento multidimensional (una excelente presentación de esta técnica se puede encontrar en Mardia et al., 1979) que está disponible en el paquete estadístico S-Plus.

Con esta métrica la distancia máxima existente entre dos árboles es de 89.36, esto es,  $d(89, 99) = 89.36$  y la distancia mínima es  $d(3, 9) = 0.64$ , esto posibilita agrupar los 100 modelos en tres subconjuntos.

En el Grupo 1 se tienen 26 elementos, el Grupo 2 tiene 40 y el Grupo 3 agrupa a 34 modelos de árbol.

En el primer grupo el mejor modelo es el marcado con el número 33, tiene un valor del logaritmo de verosimilitud igual a  $-37.25$ ; le siguen en orden de importancia los árboles 66 y 13; el número 33 es el mejor del conjunto de los 100 modelos. En el segundo grupo, el mejor es el número 15, con un valor del logaritmo de verosimilitud igual a  $-38.04$ ; le siguen el 90 y el 71, en orden de importancia; una característica de este grupo es que su SCR promedio es la menor. En el tercer grupo; el modelo más importante, utilizando el mismo criterio, es el número 92, con un valor del logaritmo de la verosimilitud igual a  $-38.62$ ; le siguen en importancia, los modelos números 89 y 74; una característica distintiva del grupo es que tiene el menor valor promedio del logaritmo de la verosimilitud y el menor número promedio de nodos terminales de los tres grupos.

## 5.4 Métrica basada en la peor predicción.

La gráfica en  $\mathcal{R}^2$  (Figura No. 3) correspondiente a las distancias de los modelos bajo esta métrica, muestra la forma de herradura que es frecuente encontrar al aplicar el escalamiento multidimensional a cualquier conjunto de números. Para este caso la distancia máxima entre dos modelos es igual a 1, esto ocurre en los modelos representados por los puntos que se encuentran en los extremos de la "herradura", que corresponden a los modelos 85 y 63, y la distancia mínima es cero y ocurre en 163 parejas de modelos. Con base en la gráfica resultante del escalamiento multidimensional y considerando la representación de los modelos en el plano, esto es, por parejas

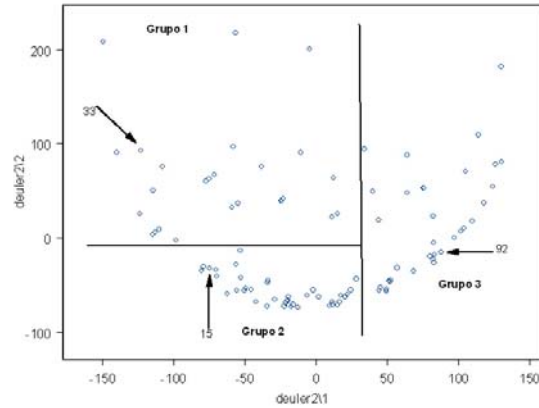


Figura No 2.

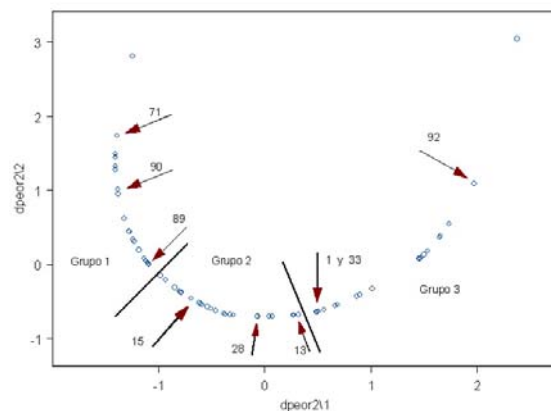


Figura No. 3

$(x, y) \in \mathfrak{R}^2$ , se realiza un agrupamiento.

El Grupo 1 lo constituyen los modelos cuya primera coordenada se encuentra en el intervalo  $x \in (-2, -1)$ , en él se encuentran 25 elementos; siendo el mejor modelo el marcado con el número 90, con un valor del logaritmo de verosimilitud igual a - 38.48 y con 10 nodos terminales, le siguen los modelos con los números 71 y 89, con un valor del logaritmo de verosimilitud igual a - 38.60 y -39.12, respectivamente.

El Grupo 2 está formado por aquellos que satisfacen  $x \in (-1, 0.4)$ , lo integran 38 modelos, el mejor es el número 15 con un valor del logaritmo de verosimilitud integrada de -38.04 y con 13 nodos terminales, le siguen en orden de importancia los modelos 28 y 13.

El Grupo 3 está constituido por 37 modelos de árbol de regresión y son aquellos representados en el plano para los cuales  $x \in (0.4, 2.4)$ ; el mejor es el número 33, con un valor del logaritmo de la verosimilitud igual a - 37.25; el segundo lugar lo ocupa el modelo marcado con el número 92, con un valor del logaritmo de verosimilitud igual a -38.62; y el tercero es el número 1, con un valor del logaritmo de verosimilitud igual a - 38.72.

Se observa que los dos mejores modelos del Grupo 1, en su primera división utilizan la variable  $X_2$  y en el caso del Grupo 2, los dos mejores modelos utilizan en la primera división la variable  $X_1$ .

**Tabla No. 1. Ejemplo 1.** Características del mejor modelo de árbol de regresión de cada ejecución del programa con los datos de Bruntz.

Ejecución	Logaritmo de verosimilitud	No. nodos terminales	Frondosidad (%)	Var. 1 <sup>a</sup> División	SCR	Tasa de aceptación (%)
1	-38.72	11	17.19	1	20.54	33.9
2	-38.04	13	20.31	1	17.02	29.0
3	-38.73	12	18.75	1	19.23	31.3
4	-37.25	10	31.25	1	19.48	30.0
5	-40.43	13	40.63	1	18.74	26.5
6	-39.35	12	18.75	1	19.84	34.6
7	-38.76	12	37.5	1	19.54	31.8
8	-38.60	15	11.72	1	14.60	30.6
9	-38.48	10	15.63	1	19.37	33.2
10	-38.62	10	15.63	1	21.44	24.5

Contrastando los resultados obtenidos con las dos distancias, se observa que cuando el criterio de comparación es el logaritmo de la verosimilitud, resulta seleccionado el modelo número 33 como el mejor en ambos casos; con un logaritmo de la verosimilitud igual a -37.25. Más aun, cuando se seleccionan los tres mejores modelos, estos coinciden y son los modelos 33, 15 y 90. Cuando el criterio de comparación es la suma de cuadrados residuales (SCR), los dos mejores modelos, en ambas métricas, son 71 y 15, con una SCR igual a 14.60 y 17.02, respectivamente.

## 6. EJEMPLO 2. APLICACIÓN: LA PAROTA.

La metodología expuesta se aplicó a datos provenientes del Censo Socioeconómico del Proyecto Hidrológico "La Parota" (Universidad Autónoma de Guerrero (2004)). Éste proyecto es un estudio de la zona donde se construirá una presa (sobre el cauce del río Papagayo) para la generación de energía eléctrica. La presa estará ubicada en el poblado denominado "La Parota", en el municipio de Acapulco, estado de Guerrero, México. Los datos corresponden a las viviendas ubicadas en la zona de inundación, una vez que se construya la presa. El número aproximado de viviendas afectadas sería aproximadamente de 700, sólo se tiene información acerca de 503 de ellas, la falta de información de las restantes viviendas es debido a problemas circunstanciales de la zona de estudio. Se utilizó toda la información disponible.

Las variables utilizadas fueron: Metros cuadrados de construcción de la vivienda ( $X_1$ ); Tipo de material del piso de la vivienda (dos categorías) ( $X_2$ ); Superficie de las tierras de cultivo pertenecientes a los miembros de la vivienda (has.) ( $X_3$ ); Índice de los bienes de la vivienda, este índice va de 0 a 100 (estufa de gas, refrigerador, aparato de sonido, etc.) ( $X_4$ ); Ocupación del jefe de familia (ocho categorías) ( $X_5$ ); Número de emigrantes de la vivienda (0,1,2 y 3) ( $X_6$ ).

La variable de respuesta ( $Y$ ) es el ingreso monetario total de los miembros de la familia que habitan en la vivienda. Esta variable no fue incluida en los censos, por lo que fue necesario realizar un muestreo de la población formada por las 503 viviendas censadas, obteniéndose una muestra de tamaño 126.

## 6. 1 Generación de las cadenas de Markov.

La determinación de los valores de los hiperparámetros arrojó:

$$\nu = 14, \lambda = 2.3, \mu = 3.225, a = 1.$$

Al igual que en el ejemplo anterior, se tienen 100 modelos como resultado de la realización de 10 ejecuciones del programa, cada ejecución consistió de 10,000 iteraciones con reinicios cada 1,000. En este caso se eligió el modelo con la mayor verosimilitud obtenida durante las 1,000 iteraciones, a diferencia de los casos presentados anteriormente donde se tomó el árbol final de cada reinicio. La razón del cambio en el criterio es la inestabilidad de las cadenas de Markov, como se observa en la gráfica de la Figura No. 4. De cada ejecución se toma el mejor modelo, formándose el conjunto de los 10 mejores. Los valores del logaritmo de la verosimilitud están comprendidos entre -118.082 y -103.764; para el caso de los 10 mejores modelos están entre -107.207 y -103.764. La tasa de aceptación promedio general fue del 21.26%, si se consideran sólo los 10 mejores modelos este valor corresponde al 21.59%. El número de nodos terminales para cada árbol de la cadena, en todo el ejemplo, varía entre 1 y 21; en el caso de los 10 mejores modelos este número varía entre 10 y 17. Considerando el conjunto de los 100 mejores modelos, 46 de ellos utilizaron la variable  $X_4$  en la primera división; 21 utilizaron la variable  $X_1$ ; cada una de las variables  $X_2$ ,  $X_3$  y  $X_5$  fue utilizada por 8 modelos y la variable  $X_6$  por 9. En el caso de los 10 mejores modelos también se observa que la mayor frecuencia de uso corresponde a la variable  $X_4$ .

Bajo el criterio del logaritmo de la verosimilitud integrada el mejor modelo de árbol de regresión, con un valor del criterio igual a -103.764, es el marcado con el número 43. Además, este modelo tiene dos características deseables: primera, que la cantidad de nodos terminales es 10, la menor cantidad dentro de los 10 mejores árboles; segunda, el porcentaje de frondosidad 62.5, que es el más alto entre los 10 mejores modelos.

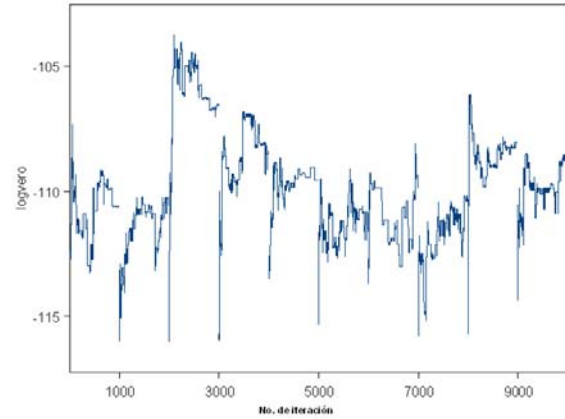
Para facilitar el agrupamiento de los modelos se eliminaron los 26 peores, aquellos cuyo valor del logaritmo de verosimilitud integrada es menor que -108. Los 74 modelos restantes se agruparon en tres subconjuntos (grupos 1, 2 y 3) con 24, 25 y 25 modelos, respectivamente.

Una característica relevante del mejor modelo en los Grupos 1 y 3 es que la variable de división en el nodo raíz es la  $X_4$ ; además, dos de los tres mejores modelos de estos grupos utilizan esta variable de división en el nodo raíz (Tabla No. 2)

En el Grupo 1, el mejor modelo (según el logaritmo de la verosimilitud) es el número 43, al cual ya se hizo referencia anteriormente. Este grupo se caracteriza por dos elementos: primero, tiene los valores del logaritmo de verosimilitud más grandes; segundo, tiene los valores de la SCR más pequeños. Los tres mejores modelos del Grupo 3, se caracterizan por: primero, sus árboles tienen los números de nodos terminales más pequeños y segundo, las cantidades correspondientes a la SCR son las más grandes.

Con el objetivo de hacer comparaciones, se eligió el mejor modelo de los grupos formados con esta métrica.

El mejor modelo del Grupo 3, que es el número 8, se descarta puesto que después de la división del nodo raíz, el árbol se desarrolla solamente sobre el nodo de la segunda división. Al comparar el mejor modelo del Grupo 1 (43) con el mejor modelo del Grupo 2 (17), se prefiere el 43 puesto que en la primera división utiliza la variable  $X_4$  y en la segunda división izquierda y derecha utiliza la variable  $X_1$ , siendo estas las variables más relevantes.



**Figura No. 4.**

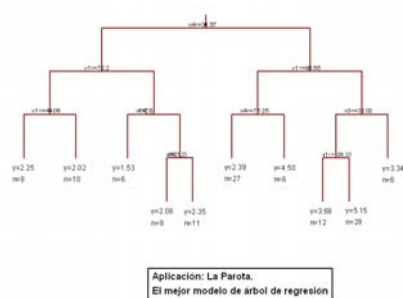
### 6. 3 Métrica basada en la peor predicción.

En la gráfica de las distancias entre los modelos según la métrica basada en la peor predicción, cada modelo puede ser representado por un punto  $(X, Y)$ . En este caso, el agrupamiento de los 100 modelos generados se realiza particionando el rango de los valores de la coordenada  $X$  en cuatro subconjuntos con igual cantidad de elementos. El mejor modelo, tanto del grupo 1 como del 3, utiliza la variable  $X_1$  en el nodo raíz y los de los grupos 2 y 4 utilizan la variable  $X_4$ . Considerando el conjunto de los 12 mejores modelos, 3 de cada grupo, 7 de ellos utilizan la variable  $X_4$  en el nodo raíz, lo que hace pensar que ésta es una variable importante. Para realizar la comparación entre los grupos, se considera el promedio del logaritmo de la verosimilitud intragrupo. Bajo este criterio el mejor grupo es el número 4, con un valor promedio del logaritmo de verosimilitud de -104.290 y el peor grupo es el número 3 con un valor del indicador de -105.589. Considerando el valor promedio por grupo de la SCR, el mejor grupo es el número 2 con un valor de 231.625 y el peor es el Grupo 3 con un valor del indicador de 255.920.

Bajo el criterio del logaritmo de la verosimilitud, los mejores modelos por grupo son los marcados con los números: 93, 33, 13 y 43, respectivamente. Ahora considerando la menor SCR, los mejores modelos por grupo son: 93, 2, 13 y 4, respectivamente. En la intersección de estos dos conjuntos se encuentran los modelos 93 y 13; que además, coinciden en el uso de la variable  $X_1$  en el nodo raíz y en el número de nodos terminales, 17.

Los modelos marcados con los números 17, 33, 43, 70 y 85 son los modelos que se encuentran en la intersección de los conjuntos formados por los mejores modelos de los grupos

correspondientes a ambas métricas (Tablas No.2 y No.3). Los modelos que en su primer nodo de división utilizan la variable  $X_4$  son 17, 33, 43 y 85. En opinión de los especialistas en el tema, el nivel de ingreso de las familias se refleja de manera importante en los bienes existentes en la vivienda, entre ellos: estufa de gas, refrigerador, licuadora, etc. En la presente aplicación, estos elementos son considerados en la variable  $X_4$ . Ahora, comparando los modelos 17, 33 y 43 (los modelos 70 y 85 están en grupos representados ya por uno de estos 3 modelos) se tiene que: Con el criterio del logaritmo de la verosimilitud, el mejor modelo es el número 43. Utilizando como criterio la menor SCR, el mejor modelo es el número 33 con un valor de 225.94. Considerando la menor cantidad de nodos terminales, los mejores modelos son 17 y 43 con 10 nodos terminales cada uno; finalmente, si el criterio es la mayor frondosidad, el mejor modelo es el 43.



**Figura No. 5**

El modelo marcado con el número 43, como se observa, tiene distintas propiedades deseables. Una cualidad adicional que tiene, es la utilización de la segunda variable explicativa más importante ( $X_1$ ) después de  $X_4$ , en la división de los nodos hijos izquierdo y derecho del nodo raíz (Figura No. 5). Esta cualidad no la tienen los modelos 17 y 33.

**Tabla No. 2. Ejemplo de aplicación.** Características de los tres mejores modelos por grupo: métrica general.

Grupo	No. de árbol	Logaritmo de verosimilitud	No. nodos terminales	Frondosidad (%)	Variable 1a. división	SCR
1	43	-103.76	10	62.50	4	265.93
1	33	-104.18	14	43.75	4	225.99
1	70	-104.41	11	17.19	1	255.98
2	17	-105.67	10	31.25	1	282.03
2	83	-105.94	12	37.50	4	268.34
2	85	-105.96	14	10.94	2	257.23
3	8	-104.66	6	37.50	4	301.18
3	38	-105.76	5	62.50	4	315.09
3	66	-105.99	8	50.0	1	305.87

Resulta interesante una evaluación del modelo 43 en términos de la SCR. La columna de SCR de la Tabla No. 2, muestra que el modelo número 43 es la mediana de la SCR de los nueve modelos presentes en dicha tabla, lo que refleja que hay cuatro modelos peores y cuatro modelos mejores que éste, por tanto, su selección sería una solución de compromiso entre el valor del logaritmo de la verosimilitud y la suma de los cuadrados residuales.

Considerando el mejor modelo de los 100 que se tienen y el hecho de que está presente en los mejores modelos de los grupos de ambas métricas (el número 43), se realizó un ejercicio de predicción de los ingresos monetarios de los habitantes de las viviendas. Por datos faltantes en alguna de las variables explicativas, el número de registros se redujo de los 504 a 353.

La distribución de las viviendas según su ingreso promedio es: 63.7% tienen ingresos menores que 2,500 pesos, 18.7% tienen ingresos mayores a 2,500 y menores a 4,500 pesos y en el 17.6% sus ingreso son superiores a 4,500 pesos mensuales. Cuando los resultados fueron discutidos con los especialistas, estos

hicieron valoraciones comparativas y estuvieron de acuerdo con los resultados y manifestaron que el análisis les había sido útil, además, por la exploración que se realizó.

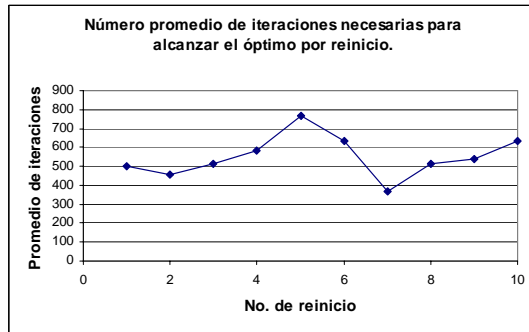
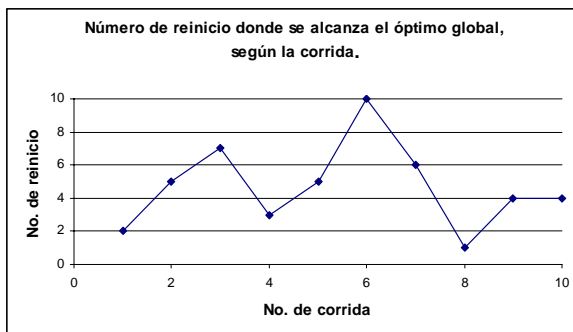


Figura No. 6

**Tabla No. 3 Ejemplo de aplicación.** Los tres mejores modelos por grupo, según métrica basada en la peor predicción.

Grupo	No. de árbol	Logaritmo de verosimilitud	No. nodos terminales	Frondosidad (%)	Variable 1a. división	SCR
1	93	-104.89	17	13.28	1	237.64
1	34	-105.90	15	11.72	4	238.74
1	10	-106.32	14	21.88	4	253.04
2	33	-104.18	14	43.75	4	225.99
2	70	-104.41	11	17.19	1	255.98
2	2	-104.71	19	14.84	4	212.89
3	13	-105.13	17	13.28	1	228.49
3	17	-105.67	10	31.25	4	282.03
3	85	-105.96	14	10.94	4	257.23
4	43	-103.76	10	62.50	4	265.93
4	4	-104.52	16	12.50	2	221.48
4	86	-104.58	11	8.59	3	243.50

## 7. RESULTADOS DE UN ESTUDIO POR SIMULACIONES PARA ANALIZAR SI ES POSIBLE REDUCIR EL NÚMERO DE MODELOS A GENERAR.

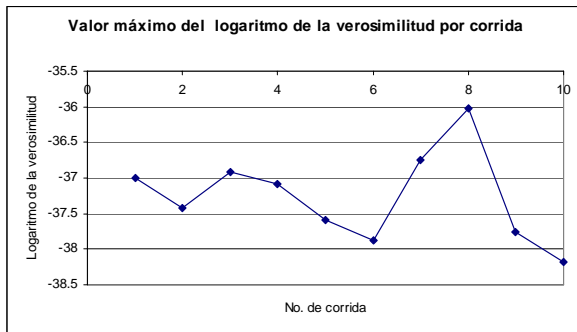


Un modelo será mejor en el sentido del logaritmo de la verosimilitud mientras más grande sea éste, por tanto, el objetivo aquí es tratar de estudiar el comportamiento del máximo del logaritmo de la verosimilitud, es decir, cuán temprano o tarde se alcanza este valor en cada corrida. Dicho de otra manera, la pregunta sería: ¿si se hacen menos corridas los valores del logaritmo de la verosimilitud obtenidos estarán muy por debajo de los que se hubieran obtenido haciendo corridas más largas?

Figura No. 7

Para realizar el estudio se tomó un conjunto de datos donde se tiene una variable dependiente cuantitativa (Y) y dos variables explicativas (X1 y X2). Se efectuaron 10 corridas del programa de 10,000 iteraciones cada una, haciendo un reinicio cada 1000 (en total 100 000 modelos generados). Se registraron, los valores máximos del logaritmo de la verosimilitud que se alcanzaron dentro del conjunto de iteraciones de cada reinicio, en cuál iteración se obtuvo este valor y a qué reinicio y corrida pertenecía.

En la Figura No. 6 se encuentra el gráfico del número del reinicio (eje vertical) donde se obtuvo el máximo



**Figura No. 8**

global para cada corrida (eje horizontal). Como se observa el máximo del logaritmo de la verosimilitud en casi todas las corridas, excepto en la corrida número 8, se alcanzó antes del reinicio 7 y en el 50% de las corridas se alcanzó en el reinicio 4 o antes. Este análisis conduce a pensar que se puede reducir el número de modelos a generar; pero para hacer una conclusión definitiva habría que aumentar el número de corridas (100 000 simulaciones no parecen suficientes).

Por otra parte, analizando el gráfico que se presenta en la Figura No.7 (ver además la Tabla No.4) se

concluye que el promedio del número de iteraciones necesarias para alcanzar un máximo, por reinicio, no exhibe un comportamiento muy errático y en todos los

reinicio, excepto en el número 5, son necesarias menos de 650 iteraciones.

El gráfico en la Figura No. 8 muestra el valor máximo del logaritmo de la verosimilitud para cada corrida. El máximo de estos máximos es -36.02, en el reinicio uno de la corrida ocho, además, para alcanzar este valor se requirieron 325 iteraciones. El mínimo de los máximos es -38.18 y pertenece al reinicio 4 de la corrida diez, para alcanzarlo se requirieron 965 iteraciones. La diferencia de 2.16 es informativa, en cierto sentido, del riesgo que se corre al disminuir el número de modelos.

## 8. CONCLUSIONES.

**Tabla No. 4** Número promedio de iteraciones necesarias para alcanzar el óptimo por reinicio.

No. de reinicio	Promedio de las iteraciones	No. de reinicio	Promedio de las iteraciones
1	498.9	6	631.3
2	458.3	7	367.7
3	512.6	8	515.6
4	581.6	9	538.5
5	765.5	10	632.1

Los trabajos de Chipman et al. (1998a, 1998b y 2001) proponen, de acuerdo a nuestra experiencia, herramientas muy eficientes para la selección de modelos de árbol de regresión. En el presente trabajo se ha intentado presentar con ejemplos y de forma unificada los procesos de generación de los árboles de regresión y la selección de los mejores modelos utilizando distintas propuestas de distancia entre modelos y se propone una nueva. Se aplicó el método para analizar los datos provenientes de un estudio socio-económico en el marco del proyecto hidráulico “La Parota”, obteniéndose resultados importantes que fueron respaldados por los especialistas. El empleo de los datos de Bruntz et al. (1974), empleados ya por Denison et al. (1998), permitió hacer comparaciones entre los resultados de los métodos y puso de manifiesto la importancia de que no es adecuado restringirse a un método ni a un único modelo, sino que una exploración por varios métodos y seleccionar un conjunto de modelos es importante (se pueden encontrar conclusiones similares en otros trabajos, ver por ejemplo Castells y Juárez, 2008).

Para la implementación práctica se programaron el algoritmo y todas las distancias empleadas sobre S-PLUS.

**RECEIVED JANUARY 2008**

**REVISED MARCH 2009**

## REFERENCIAS

- [1] ANDRIYASHIN, A. (2005): **Financial Applications of Classification and Regression Trees**. Tesis de Maestría. CASE - Center of Applied Statistics and Economics. Humboldt University, Berlin.
- [2] BESAG, J. AND GREEN, P.J. (1993): Spatial Statistics and Bayesian Computation. **Journal Royal Statistical Society B** 55, 25-37.
- [3] BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A. Y STONE, CH. J. (1984): **Classification and Regression Trees**. Wadsworth, Belmont
- [3] BRUNTZ, S. M., CLEVELAND, W. S. Y WARNER, J. L. (1974): The dependence of ambient ozone on solar radiation, temperature and mixing height. **In Symposium on Atmospheric Diffusion and Air Pollution, MA**: American Meteorological Society. Boston.
- [4] CAMDEVIREN, H. A.; YAZICI, A. C.; AKKUS, Z.; RESUL BUGDAYCI, R. AND SUNGU, M. A. (2007): **Comparison of logistic regression model and classification tree: An application to postpartum depression data**. Pergamon Press, Inc. Tarrytown, NYork.
- [5] CASTELLS, E. Y JUÁREZ, O (2008): Una estrategia en dos direcciones para la selección de modelos de regresión para el caso de una variable explicativa cualitativa. Aprobado para aparecer en la revista **Investigación Operacional**.
- [6] CHIB, S AND GREENBERG, E. (1995): Understanding the Metropolis-Hastings algorithm. **The American Statistician**, 49, 327-335.
- [7] CHIPMAN, H. A., GEORGE, E. I. AND MCCULLOCH, R. E. (1998a): Bayesian CART model search (with discussion). **Journal of the American Statistical Association**, 93, 935-948.
- [8] CHIPMAN, H. A., GEORGE, E. I. AND MCCULLOCH, R. E. (1998b): Extracting representative tree models from forest. **Working Paper 98-07, Department of Statistics and Actuarial Science, University of Waterloo, 1998**.
- [9] CHIPMAN, H. A., GEORGE, E. I. AND MCCULLOCH, R. E. (2001): Managing multiple models. **Artificial Intelligence and Statistics 2001**, (eds. Tommi Jaakkola, Thomas Richardson). 11-18. Morgan



Kaufman, San Francisco.

[10] DAVIS, R.E. AND ELDER, K. (2002): Application of Classification and Regression Trees: Selection of Avalanche Activity Indices at Mammoth Mountain. Disponible en [www.Avalanche.org](http://www.Avalanche.org). Leído el 15 de enero de 2008.

[11] DENISON, D. G. T., MALLICK, B. K. AND SMITH, A. F.M. (1998): A Bayesian CART Algorithm. **Biometrika** 85, 363-377.

[12] GAMERMAN, D.(1997): **Markov chain Monte Carlo. Stochastic simulation for Bayesian inference.** Chapman & Hall. London

[13] GREEN, P.J. (1995): Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. **Biometrika** 82, 711-732.

[14] HASTINGS, W.K. (1970): Monte Carlo sampling methods using Markov chain and their applications. **Biometrika** 57, 97-109.

[15] LEWIS, R. J. (2000): An Introduction to Classification and Regression Tree (CART) Analysis. **2000 Annual Meeting of the Society for Academic Emergency Medicine.** San Francisco.

[16] MARDIA, K.V., KENT, J.T. & BIBBY, J.M. (1979): **Multivariate analysis.** Academic Press. London.

[17] MATHSOFT (1999): **S-PLUS 2000.** Mathsoft Inc., Seattle, Washington.

[18] TENG, J.H., LIN K.C., HO, B.S. (2007): Application of classification tree and logistic regression for the management and health intervention plans in a community-based study. **Journal of Evaluation in Clinical Practice** 13 (5), 741-748. Resumen leído el 18 de enero 2008 en [www.ingentaconnect.com](http://www.ingentaconnect.com).

[19] WADA, T. AND NAKAMURA, T. (2004): Regression Trees and Its Application to Classification Problems: **IEIC Technical Report VOL.104; No..291**, 33-40.