UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Comunicaciones en Estadística
Junio 2017, Vol. 10, No. 1, pp. 9–40

Estrategia de imputación con la media bajo el uso de árboles de regresión

Imputation strategy with media using regression trees

Victor Márquez Pérez[a]
victore.marquezp@gmail.com

Lelly Useche[b]
usechel@unesur.edu.ve

Dulce Mesa[c]
dmesa@ivad.com

Ana Idés Chacón[d]
anaidesch@ula.ve

Resumen

Se presenta un diseño de imputación en el que conjuga la clasificación y la imputación con el fin de mejorar la calidad del dato imputado. La imputación se lleva a cabo para datos cuantitativos bajo pérdida completamente aleatoria con el uso de árboles de regresión, comparando la técnica de imputación con la media teórica y empíricamente con el uso de los árboles de regresión, con ello se busca desarrollar una estrategia integral de clasificación e imputación.

Se obtuvieron estimadores insesgados desarrollando el valor esperado del estimador, se evaluaron las propiedades de los estimadores mediante el desarrollo de sus varianzas y sesgos, en el que se observó insesgamiento. En cuanto a la varianza del estimador insesgado de la media, no se probó suficiencia para el estimador de la media.

***Palabras clave*:** datos faltantes, imputación, CART, árboles de regresión, estimadores insesgados, simulación.

Abstract

An imputation design is presented to combine classification and imputation in order to improve the quality of imputed datum. Imputation is done with completely randomized missing quantitative data and using regression trees. Media imputation techniques is compared, theoretical and empirically, using regression trees, in order to develop an integral classification and imputation strategy.

[a]Profesor, Escuela Superior Politecnica de Chimborazo, Ecuador.
[b]Profesora, Universidad Nacional Experimental Sur del Lago, Venezuela.
[c]Profesora, Universidad Central de Venezuela, Venezuela.
[d]Profesora asistente, Universidad de los Andes, Venezuela.

Unbiased estimators were obtained developing the expected value of the estimator. estimators proprieties were evaluated trough their variance and bias development, which showed non bias. as for the unbiased estimator variance of the media, sufficiency was not proved for the media estimator.

***Keywords*:** missing data, imputation, CART, regression trees, unbiased estimators, simulation.

1. Introducción

En la mayoría de los estudios muestrales y/o censales, principalmente en la medición de unidades, nos encontramos con múltiples obstáculos, como perder una medición, lo cual genera espacios vacíos que pueden producir problemas en el análisis posterior. Desde hace ya varias décadas, se viene estudiando la forma de "llenar" estos espacios vacíos (problema de imputación de datos faltantes), con el fin de obtener un conjunto de datos completos para ser analizados por la vía de los métodos estadísticos tradicionales. Sin embargo, esta situación se complica cuando las pérdidas se presentan en una matriz de datos formada por diversas variables sobre la cual se realizarán estudios multivariantes, lo que exige la generación de metodologías que se adapten a particularidades (Useche & Mesa 2006).

Recientemente, se han desarrollado nuevas formas de tratar los datos faltantes en matrices multivariantes, de esta manera se obtiene una variedad de técnicas basadas en diferentes enfoques según las características del dato. Aun así, todavía son muchas las deficiencias que enfrentan las técnicas actuales ante la falta de respuesta a temas como: los sesgos en las estimaciones, la alteración de la relación entre las variables y los cambios en las varianzas. La ausencia de respuesta, plantea un gran problema que enfrenta el analista. En ocasiones, por ejemplo, con grandes conjuntos de datos y con poca proporción de pérdida, pudiéramos ignorar la ausencia de datos, pero esto no es conveniente ni cuando se trata de pocos datos o de altas proporciones de pérdida.

Cuando no se deben ignorar los datos faltantes, la manera más adecuada de tratarlos es llenar esos espacios faltantes con valores plausibles, a este procedimiento es lo que denominamos "imputación". Actualmente, las técnicas de imputación de datos son utilizadas en la fase de análisis exploratorio de datos como parte del preprocesamiento de datos (López 2001).

Lohr (2009), refiriéndose a la importancia de los procedimientos de imputación, señala que esta no radica sólo en reducir el posible sesgo por la ausencia de respuestas, sino también para producir un conjunto de datos reticulares y "limpios" sin valores faltantes, para llevar a cabo los análisis estadísticos, ya que todos estos análisis usan bases de datos completas.

Son muchas las técnicas de imputación que han surgido, sobre todo desde la década de los sesenta, empleando enfoques univariantes y multivariantes. Se han empleado enfoques basados en modelos como: funciones de verosimilitud, regresión (Buck

1960) y descomposición de matrices en valores singulares (Dempster et al. 1977) y (Krzanowski 1988). A pesar de estos avances, no se ha encontrado una metodología capaz de reproducir el dato original o que pueda resolver el tratamiento de los datos faltantes en forma satisfactoria, ya que no todas las metodologías son adecuadas para todos los problemas relacionados con alteraciones de la distribución de los datos (una vez imputados), con la alteración en la relación de las variables o el sesgo en las estimaciones e inflación de la varianza (Little & Rubin 2014). Por esta razón, se sigue en la búsqueda de mejorar la calidad del dato imputado.

Las técnicas actuales para imputación se diferencian en que son más automáticas y son modelos más exploratorios, tal como lo expresa Piela et al. (2001). Uno de los enfoques más recientes utiliza modelos de árboles de decisión, lo que permite obtener grupos cada vez más homogéneos y en consecuencia mejores estimaciones de los datos perdidos dentro de cada subgrupo, ya que se supone que mientras más parecido sea el donante al receptor, mejor será la imputación. Estas investigaciones se han centrado básicamente en la ausencia de respuesta cualitativa (Mesa 2004), pero para variables cuantitativas no se han encontrado muchos aportes, es por ello que se buscará en esta investigación, optimizar el uso de las técnicas de imputación con la ayuda de técnicas de árboles de regresión CART para datos faltantes cuantitativos.

Con base en lo anterior, esta investigación está fundamentada en la presentación de una propuesta de imputación de datos cuantitativos que integre el uso de árboles de regresión CART en el proceso propio de imputación, como una estrategia integral de clasificación e imputación de datos. De forma más especifica se planteará el uso de árboles de regresión para optimizar la técnica de imputación con la media para datos cuantitativos faltantes.

A partir de la segunda sección, se describen los fundamentos generales de la falta de respuesta, desde los tipos de errores presentes en una investigación científica, tipos de falta de respuesta, patrón de pérdida e imputación. En la tercera sección se detalla la técnica de imputación con la media, en la cuarta sección se desarrollan los conceptos de árboles de regresión CART. En la quinta sección se lleva a cabo el desarrollo algebraico de la imputación bajo el uso de árboles de regresión hallando el estimador insesgado de la media del total y de la varianza. En la sexta sección se hace una evaluación empírica de la técnica de imputación con la media incluyendo el uso de árboles de regresión diseñando una estrategia integral de imputación para datos faltantes con la aplicación de la técnica propuesta. La cual se implementa usando los datos del VI Censo Agrícola 1998, con patrones inducidos de pérdida entre el 5 % y 30 %. En la sétima sección se muestran los resultados obtenidos, los cuales fueron bastantes satisfactorios, ya que al introducir los CART en el método de imputación se logra robustez en la técnica, pues así se mantienen los parámetros distribucionales independientemente del porcentaje de pérdida.

2. Fundamentos generales

2.1. Tipos de errores presentes en una investigación estadística

En una investigación estadística de encuestas por muestreo, generalmente se producen dos tipos de errores: errores muestrales y errores no muestrales. Los errores muestrales son los errores producidos al observar una muestra de la población y no la totalidad de ella. Este tipo de error está compuesto por la variabilidad del estimador ante todas las muestras posibles y su sesgo, llamado sesgo técnico (Mesa 2004).

Los errores no muestrales son aquellos errores presentes en una investigación, no atribuibles al hecho de observar una muestra. Pueden ser aleatorios o sistemáticos (Mesa 2004). Una forma de clasificar los errores no muestrales es según su fuente, los cuales pueden ser (Little & Rubin 2014): Errores de cobertura, errores de respuesta y errores de no respuesta, estos últimos surgen cuando las unidades de observación seleccionadas para la encuesta no proporcionan todos los datos que debería recogerse, produciéndose espacios vacíos. Esta investigación se centrará solo en los errores producidos por la no respuesta.

2.2. Tipos de falta de respuesta

La ausencia de respuesta, aun cuando no se quiera, estará presente en toda investigación que involucre medición, de la cual, muchas veces no se obtendrán registros completos por diversas causas ajenas al investigador. Según Little & Rubin (2014) la falta de respuesta, puede presentarse de dos maneras: la falta de respuesta total, es cuando hace falta todo el registro de una base de datos, puesto que no hay una unidad para ser medida (persona, vivienda, empresa, etc.) o por impedimento de efectuar un conjunto total de mediciones de variables en un determinado momento específico; es decir, no se recoge ningún dato de la unidad de la muestra. Por ejemplo, cuando se lleva a cabo la aplicación de encuesta de hogares y en algunas viviendas seleccionadas no se encuentran, al momento de aplicar el instrumento, los miembros del hogar, generándose una pérdida total de las respuestas del cuestionario correspondiente a ese hogar.

La falta de respuesta parcial, se presenta cuando hay ausencia de uno o más valores para un registro, sin llegar a la ausencia completa de un registro, ejemplo; un individuo que va a encuestado está presente en el momento de la entrevista, pero no responde algunas preguntas del cuestionario o a una unidad no se le efectuaron algunas mediciones por fallas en los equipos de medición; sin embargo, otras mediciones sí se llevaron a cabo. Esta investigación se basará en el estudio de la falta de respuesta parcial.

2.3. Patrón de pérdida de respuesta

Como lo expresa Little & Rubin (2014) y Schafer (1997), uno de los puntos para tener en cuenta en la ausencia de respuesta, es el patrón de pérdida de los datos faltantes, ya que estos pueden influir en la manera como se lleva a cabo la imputación. Dichos patrones de pérdida pueden ser ignorables o no ignorables.

Los patrones de pérdida son ignorables si los valores faltantes no dependen de la variable que está perdida, ocurren de manera completamente aleatoria (MCAR, *Missing Completely At Random*) o de manera aleatoria (MAR, *Missing At Random*). El primer caso (MCAR) ocurre cuando la ausencia de información no depende de la variable con datos faltantes ni de otras variables presentes en la matriz de datos. Para el segundo caso (MAR) ocurre cuando la ausencia de los datos no depende de la variable con datos faltantes, pero sí de alguna variable presente en la matriz de datos. Esta investigación se basa en pérdidas completamente aleatorias (MCAR).

2.4. Imputación

Imputación es llenar los espacios vacíos de la base de datos incompleta con valores plausibles y obtener así un archivo completo, con el fin de analizarlo. Es usado como tratamiento de la falta de respuesta parcial.

2.5. Descripción del modelo de imputación

Bajo el enfoque basado en el modelo (en el que se basa nuestro estudio), x_{ik} y y_i son consideradas variables aleatorias con distribucion $f(X, Y|\theta)$ indexadas por el parámetros (conjunto de parámetros) θ. La respuesta R es también incluida como una variable aleatoria con distribución $f(R|X, Y, \theta)$ dado que X está completamente observada y Y está sujeta a no respuesta, la forma completa de la distribución puede ser escrita como $f(X, Y, R|\theta, \phi)$ indexada por los parámetros θ y ϕ con R como indicador de respuesta.

La distribución conjunta de X, Y y $f(X, Y, R|\theta, \phi)$ puede ser descompuesta como el producto de la distribución de probabilidad de X y Y, indexada por el parámetro θ y la distribución condicional de R dado X y Y (la distribución para el mecanismo de data perdida), indexada por el parámetro ϕ, esto es:

$$f(X, Y, R|\theta, \phi) = f(X, Y|\theta) f(R|X, Y, \phi) \tag{1}$$

Si Y está sujeta a no respuesta, podemos escribir $Y = (Y_{\text{obs}}^T, Y_{\text{per}}^T)^T$, donde Y_{obs} es el vector de tamaño $m \times 1$, que representa los valores observados de Y, mientras que Y_{per} es el vector de tamaño $(N - m) \times 1$ que representa los valores faltantes de Y.

Además, la distribucion $f(X, Y, R|\theta, \phi)$ puede ser escrita como $f(X, Y_{\text{obs}}, Y_{\text{per}}, R|\theta, \phi)$ y la ecuación (1) puede ser escrita como:

$$f(X, Y_{\text{obs}}, Y_{\text{per}}, R|\theta, \phi) = f(X, Y_{\text{obs}}, Y_{\text{per}}|\theta) f(R|X, Y_{\text{obs}}, Y_{\text{per}}, \phi)$$

La distribución de la data observada se puede obtener integrando Y_{per} fuera de la distribución conjunta de X, Y y R, es decir,

$$f(X, Y_{\text{obs}}, R) = \int f(X, Y, R) dY_{\text{per}}$$

Más específicamente,

$$f(X, Y_{\text{obs}}, R|\theta, \phi) = \int f(X, Y_{\text{obs}}, Y_{\text{per}}|\theta) f(R|X, Y_{\text{obs}}, Y_{\text{per}}, \phi) dY_{\text{per}}$$

Los supuestos acerca del modelo son normalmente hechos con el objeto de obtener una estimación válida. Uno de los supuestos más comunes es que los valores faltantes son "faltantes aleatoriamente", MAR. Es decir,

$$f(R|X, Y_{\text{obs}}, Y_{\text{per}}, \phi) = f(R|X, Y_{\text{obs}}, \phi)$$

Entonces, asumiendo MAR y dada la data actual observada (X, Y_{obs}, R) donde ahora tenemos:

$$f(X, Y_{\text{obs}}, R|\theta, \phi) = f(X, Y_{\text{obs}}|\theta) f(R|X, Y_{\text{obs}}, \phi)$$

En este caso, el procedimiento común usado para la data observada completa es el de máxima verosimilitud para estimar el parámetro θ requerido cuando la data está incompleta (data con valores faltantes). Es decir, θ puede ser estimado por maximización $f(X, Y_{\text{obs}}|\theta)$ de la data observada. Así, el mecanismo de data perdida es ignorable; es decir, la segunda parte del lado derecho de la última ecuación puede ser ignorado en la estimación de θ.

Si MAR se mantiene, la inferencia de θ está basada en la función de verosimilitud $L(\theta|X, Y_{\text{obs}})$, el cual es función DE θ proporcional a $f(X, Y_{\text{obs}}|\theta)$ (Little & Rubin 2014).

3. Técnicas de imputación

Varios estudios, como Goicoechea (2002) y Service (1996), indican que las técnicas de imputación se pueden clasificar de la siguiente manera:

- Técnicas fundamentadas en información externa: cuando son basadas en variables relacionadas con una encuesta perteneciente a otras bases de datos externas o reglas previas. Entre estas se encuentran: los métodos deductivos y las tablas Look-up.

- Técnicas determinísticas: cuando, al imputar diferentes unidades bajo las mismas condiciones, se producen las mismas respuestas, se clasifican en: Imputación con la media o modo, imputación con la media de clases, imputación por regresión, imputación por emparejamiento media, imputación por el vecino más cercano, imputación por el algoritmo EM (*Expectation Maximization*), imputación mediante redes neuronales, e imputación mediante modelos de series de tiempo.

- Técnicas aleatorias o estocásticas: son aquellas que cuando se repite el método de imputación bajo las mismas condiciones para una unidad, pueden producir resultados diferentes. Entre ellas tenemos; imputación de un caso seleccionado aleatoriamente, imputación de un caso seleccionado aleatoriamente entre clases, imputación Hot-Deck secuencial, imputación jerárquica Hot-Deck, imputación por regresión aleatoria e imputación por regresión logística.

3.1. Imputación con la media

Se seleccionó esta técnica por ser la más conocida y comúnmente usada para diversas situaciones. Es fácil ejecutarla, es por ello es aplicada generalmente por usuarios que desconocen otras técnicas de imputación, además, por esta misma razón, es una de las opciones que ofrecen la mayoría de los software estadísticos de análisis de datos.

En el método de imputación con la media, se estima la media absoluta de los registros presentes en la base de datos completa para la variable a imputar, por lo tanto el valor resultante (media absoluta) será el valor donante para los registros con datos faltantes de esta variable. De esta misma forma se aplica para cada una de las variables que presenten al menos un registro ausente.

Según (Little & Rubin 2014) se expresa de la siguiente manera: sea y_{ij} el valor de Y la unidad i en la variable j, $i = 1, 2, \ldots, N$, $j = 1, \ldots, J$. La media imputada sustituye la media $\bar{y}_{jR}$ de la m_j unidad respondiente en la celda j para unidades que son muestreadas, pero que no responden. Para diseños igualmente ponderados, la media poblacional $\bar{Y}$ podría ser estimada por la media de las unidades observadas e imputadas, a saber:

$$\frac{\sum_{j=1}^{J} n_j \hat{\bar{y}}_j}{\sum_{j=1}^{J} n_j}$$

donde $\hat{\bar{y}}_j$ es la media de las unidades observadas e imputadas en la celda j. Ahora,

$$\hat{\bar{y}}_j = \frac{m_j \bar{y}_{jR} + n_j m_j \bar{y}_{jR}}{n_j}$$

4. Árboles de regresión

4.1. Importancia de la clasificación en la imputación

La eficiencia del uso de una técnica de imputación, en una base de datos, muchas veces se ve afectada por la heterogeneidad de los datos, que hace que el valor donante pertenezca a un registro de una característica muy distinta a la del registro a imputar, pudiéndose obtener datos inconsistentes y/o sesgos muy grandes. Por ejemplo, para imputar usando un registro aleatorio proveniente de la misma base, se obtendría mejores resultados si se clasificaran los datos formando grupos homogéneos y luego se seleccionara el donante dentro del grupo al que pertenece el receptor, el cual tendría características más similares a este receptor.

4.2. Árboles de decisión

En (Borgoni & Berrington 1990), definen a los árboles de decisión como una secuencia de divisiones de un conjunto de datos a través de reglas de decisión formuladas con respecto a valores de variables explicativas. Esta secuencia de divisiones se mantiene hasta obtener grupos homogéneos con respecto a una(s) variable(s) en particular (variable a clasificar).

Breiman et al. (1984) manifiestan que un modelo basado en árboles de decisión es un grupo de reglas de clasificación que particiona el conjunto total de datos en subgrupos mutuamente exhaustivos y excluyentes. Las reglas son definidas en términos de valores de un grupo de variables explicativas. El modelo es construido particionando progresivamente el conjunto de datos en subgrupos más pequeños que son crecientemente más homogéneos, con respecto a la variable respuesta. El proceso de partición continúa hasta que el criterio de detención se cumpla.

Cabe aclarar que dentro de los métodos basados en árboles se pueden distinguir dos tipos, dependiendo del tipo de variable que se vaya a clasificar; árboles de clasificación y árboles de regresión. Los primeros son empleados cuando las variables que van a ser clasificadas son categóricas, tanto nominales como ordinales, mientras que los árboles de regresión son los que se emplean para variables continuas que vayan a ser clasificadas.

4.3. Elementos de un árbol de decisión:

Según Goicoechea (2002), los elementos de un árbol son:

- Nodo raíz: es el grupo inicial formado por todos los elementos para segmentar.
- Nodos intermedios: son aquellos que generan dos o más segmentos descendientes inmediatos llamados nodos hijos.
- Nodo terminal: cuando un grupo no se divide más.
- Rama de un nodo: consta del grupo de todos los nodos dependientes de un nodo intermedio.
- Árbol de decisión completo: es aquel conjunto de nodos (raíz, intermedios y terminales) obtenido cuando cada nodo terminal no se puede ramificar.
- Subárbol: se obtiene de la poda de una o más ramas del árbol completo.

4.4. Árboles de clasificación y regresión (CART)

El árbol propuesto para esta investigación es CART, descrito como un algoritmo de segmentación desarrollado por (Breiman et al. 1984). Este algoritmo es conocido como una partición binaria recursiva que representa sus resultados en forma de árbol de decisión. Es binaria porque los nodos padres son siempre partidos en dos subgrupos (hijos) y recursivo porque cada hijo puede también ser tratado como un padre y por lo tanto puede ser también particionado.

Bajo el supuesto basado en el modelo de, Breiman et al. (1984), podemos expresarlo escribiendo la función de probabilidad de, la cual está sujeto a no respuesta, dado el nodo terminal definido por X_t como $f(Y|\bar{x}_i \in X_t)$, es decir, la función de probabilidad de Y dado un conjunto de los valores de las variables explicatorias, identificando el nodo terminal (grupo de clasificación). Para simplificar la notación, escribimos:

$$f(Y|\bar{x}_i \in X_t) = f(Y|t)$$

4.5. Ventajas y desventajas de los Árboles de Decisión CART

Varios autores están de acuerdo con el uso de árboles de decisión como método de clasificación, debido a tener ventajas tales como: "su flexibilidad, escasez de supuestos, resistencia a datos atípicos" (Bárcena & Tusell 1999). Goicoechea (2002), expone también como ventajas que: las reglas de asignación son claras y, por lo tanto, la interpretación de los resultados es directa y sencilla de observar, son válidas sea cual sea la naturaleza de las variables explicativas: continuas, nominales u

ordinales y que es una técnica no paramétrica que tiene en cuenta las interacciones que pueden existir entre los datos.

Como desventaja de los árboles de decisión CART para imputar es que se requiere de grandes masas de datos para asegurarse de que la cantidad de observaciones de los nodos finales sea significativa, a menos que el proceso se detenga a una altura tal que garantice grandes grupos para poder conseguir donantes para llevar a cabo el proceso de imputación. También cabe destacar que esta metodología no es rápida, debido a que se necesita del diseño del árbol, de la clasificación de los datos para luego proceder a la imputación. No es recomendable usar el árbol diseñado para imputar una variable distinta para la que fue creado.

5. Imputación bajo árboles de regresion

5.1. Desarrollo algebraico

Hansen et al. (1953) reconocen que tratar los valores imputados como valores observados pueden llevar a desestimaciones de la varianza de los estimadores de la media, varianza y proporción si son usadas las fórmulas estándar. Tal como lo plantea (Little & Rubin 2014), sea U una población finita de N unidades $U = \{U_i; i = 1, \ldots, N\}$ y sea $Y = (y_i)$ un vector de tamaño N de la variable respuesta y_i que representa el i-ésimo elemento. Sea $X = (x_{ik})$ una matriz de tamaño $N \times K$ de variables auxiliares, con x_{ik} como la realización de la k-ésima variable para el i-ésimo elemento. X puede ser representada como $X = (x_1, \ldots, x_k, \ldots, x_K)$ donde $x_k = (x_{1k}, \ldots, x_{Nk})^T$ es un vector de N valores.

Asumimos que la variable y_i está sujeta a no respuesta y x_{ik} es completamente observada. Definimos $R = (r_i)$ como un vector de indicadoras de tamaño N para y_i, el cual identifica si y_i está o no perdida. Esto es:

$$r_i = \begin{cases} 1 & \text{si } y_i \text{ es observada} \\ 0 & \text{si } y_i \text{ no es observada} \end{cases}$$

La población puede ser representada como sigue:

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1k} & \cdots & x_{1K} \\ x_{21} & \cdots & x_{2k} & \cdots & x_{2K} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mk} & \cdots & x_{mK} \\ x_{m+1,1} & \cdots & x_{m+1,k} & \cdots & x_{m+1,K} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{N1} & \cdots & x_{Nk} & \cdots & x_{NK} \end{bmatrix} \qquad Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad R = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Donde m es el número de registros, el cual Y es observada (medida) y los ceros representan los valores perdidos. Tomando, sin pérdida de generalidad que $r_1 = \cdots = r_m = 1$ y $r_{m+1} = \cdots = r_N = 0$. Este caso corresponde al caso univariante, en el cual sólo una variables está sujeta a no respuesta.

Si el modelo tratado r_i como variables aleatorias, consideremos un modelo donde y_i, x_{ik} y r_i son unidades aleatorias. Entonces, escribimos la distribución conjunta como $f(Y, X, R)$.

Cuando R es independiente de Y, que es

$$f(y_i|r_i = 1) = f(y_i|r_i = 0)$$

la data es llamada pérdida completamente aleatoria, MCAR, donde $f(Y|R)$ denota la función de probabilidad de Y dado R.

Uno de los objetivos de esta investigación consiste en hallar los estimadores insesgados para la media, totales y varianzas de cada una de las técnicas de imputación propuestas (con la media, Hot Deck secuencial y selección aleatoria) haciendo uso previo de CART, esto permitirá evaluar desde el punto de vista teórico las técnicas, así como la comparación entre ellas.

La evaluación teórica (otro objetivo que persigue esta investigación) de los estimadores insesgados encontrados se hará observando las propiedades de los estimadores hallando su sesgo y varianza para probar insesgamiento, consistencia y eficiencia.

5.2. Técnica de imputación con la media

Sea $\mathbf{m}_q$ el número de registros observados y $\mathbf{N}_q - \mathbf{m}_q$ el número de registros no observados, la cual será el donante de los registros ausentes. Sea $\bar{Y}_{\text{obs}}$ la media obtenida con los registros observados. Suponiendo que $Y \sim N(\mu, \sigma^2)$ y $R \sim F_R(\varphi, \sigma_r)$, asumiendo independencia entre $\mathbf{r_i}$ y $\mathbf{y}_i$ ya que se supone MCAR.

Estimador de la media:

$$\hat{\bar{Y}} = \frac{1}{Q} \sum_{q=1}^{Q} \left[\frac{1}{N_q} \left[\sum_{i=1}^{N_q} y_i r_i + \bar{Y}_{\text{obs}}(N_q - m_q) \right] \right] \tag{2}$$

Hallando el estimador insesgado para la media, se calcula la esperanza del estimador (2)

$$
\begin{aligned}
E(\hat{\bar{Y}}) &= E\left[\frac{1}{Q}\sum_{q=1}^{Q}\left[\frac{1}{N_q}\left[\sum_{i=1}^{N_q} y_i r_i + \bar{Y}_{\text{obs}}(N_q - m_q)\right]\right]\right] \\
&= \frac{1}{Q}\sum_{q=1}^{Q}\left[\frac{1}{N_q}\left[E\left(\sum_{i=1}^{N_q} y_i r_i\right) + E\left(\bar{Y}_{\text{obs}}(N_q - m_q)\right)\right]\right] \\
&= \frac{1}{Q}\sum_{q=1}^{Q}\left[\frac{1}{N_q}\sum_{i=1}^{N_q} E(y_i r_i) + \frac{1}{N_q}(N_q - m_q)E\left(\bar{Y}_{\text{obs}}\right)\right] \\
&= \frac{1}{Q}\sum_{q=1}^{Q}\left[\frac{1}{N_q}\sum_{i=1}^{N_q} E(y_i)E(r_i) + \frac{1}{N_q}(N_q - m_q)E\left(\bar{Y}_{\text{obs}}\right)\right]
\end{aligned}
$$

Luego,

$$
\begin{aligned}
E(\hat{\bar{Y}}) &= \frac{1}{Q}\sum_{q=1}^{Q}\left[\mu\varphi + \frac{(N_q - m_q)}{N_q}E\left(\frac{1}{m_q}\sum_{i=1}^{N_q} y_i r_i\right)\right] \\
&= \frac{1}{Q}\sum_{q=1}^{Q}\left[\mu\varphi + \frac{(N_q - m_q)}{N_q}\frac{1}{m_q}\sum_{i=1}^{N_q} E(y_i)E(r_i)\right] \\
&= \frac{1}{Q}\sum_{q=1}^{Q}\left[\mu\varphi + \frac{(N_q - m_q)}{N_q}\frac{N_q}{m_q}\mu\varphi\right] \\
&= \mu\varphi + \frac{(N_q - m_q)}{m_q}\mu\varphi \\
&= \left[1 + \frac{(N_q - m_q)}{m_q}\right]\mu\varphi \\
&= \frac{N_q}{m_q}\varphi\mu
\end{aligned}
$$

Ahora bien, obteniendo un estimador insesgado,

$$\begin{aligned}\hat{\bar{Y}}' &= \frac{m_q}{N_q\varphi}\hat{\bar{Y}} \\ &= \frac{m_q}{N_q\varphi}\frac{1}{N_q}\left[\sum_{i=1}^{N_q} y_i r_i + \bar{Y}_{\text{obs}}(N_q - m_q)\right] \\ &= \frac{m_q}{N_q^2\varphi}\left[\sum_{i=1}^{N_q} y_i r_i + \frac{(N_q - m_q)m_q}{m_q N_q^2\varphi}\sum_{i=1}^{N_q} y_i r_i\right]\end{aligned}$$

Este será el estimador insesgado de la media por la técnica de imputación con la media. Ahora, se desea calcular la varianza de este estimador para hacer una comparación con las demás técnicas y así evaluar cual es el estimador más eficiente.

$$\begin{aligned}Var\left(\hat{\bar{Y}}'\right) &= Var\left[\frac{m_q}{N_q^2\varphi}\sum_{i=1}^{N_q} y_i r_i + \frac{(N_q - m_q)m_q}{m_q N_q^2\varphi}\sum_{i=1}^{N_q} y_i r_i\right] \\ &= \frac{m_q^2}{N_q^4\varphi^2}Var\left[\sum_{i=1}^{N_q} y_i r_i + \frac{(N_q - m_q)}{m_q}\sum_{i=1}^{N_q} y_i r_i\right] \\ &= \frac{m_q^2}{N_q^4\varphi^2}\underbrace{E\left[\left(\sum_{i=1}^{N_q} y_i r_i + \frac{(N_q - m_q)}{m_q}\sum_{i=1}^{N_q} y_i r_i\right)^2\right]}_{(Eq.\ 1)} \\ &\quad - \frac{m_q^2}{N_q^4\varphi^2}\underbrace{\left[E\left(\sum_{i=1}^{N_q} y_i r_i + \frac{(N_q - m_q)}{m_q}\sum_{i=1}^{N_q} y_i r_i\right)\right]^2}_{(Eq.\ 2)}\end{aligned}$$

Ahora, para hacer más sencillos los cálculos, se dividió la expresión en dos partes. Para la primera parte, $(Eq.\ 1)$, se tiene la siguiente expresión:

$$
\begin{aligned}
&E\left[\left(\sum_{i=1}^{N_q} y_i r_i + \frac{(N_q - m_q)}{m_q}\sum_{i=1}^{N_q} y_i r_i\right)^2\right] \\
&= \left(1 + \frac{(N_q - m_q)}{m_q}\right)^2 E\left[\left(\sum_{i=1}^{N_q} y_i r_i\right)^2\right] \\
&= \left(1 + \frac{(N_q - m_q)}{m_q}\right)^2 \left[Var\left(\sum_{i=1}^{N_q} y_i r_i\right) + \left[E\left(\sum_{i=1}^{N_q} y_i r_i\right)\right]^2\right] \\
&= \left(1 + \frac{(N_q - m_q)}{m_q}\right)^2 \left[\sum_{i=1}^{N_q} Var(y_i r_i) + \left[\sum_{i=1}^{N_q} E(y_i)E(r_i)\right]^2\right] \\
&= \left(1 + \frac{(N_q - m_q)}{m_q}\right)^2 \left[\sum_{i=1}^{N_q} \left(\mu\sigma_r^2 + \varphi\sigma^2 + \sigma_r^2\sigma^2\right) + \left[\sum_{i=1}^{N_q} \mu\varphi\right]^2\right]
\end{aligned}
$$

Nótese que,

$$
\begin{aligned}
&E\left[\left(\sum_{i=1}^{N_q} y_i r_i + \frac{(N_q - m_q)}{m_q}\sum_{i=1}^{N_q} y_i r_i\right)^2\right] \\
&= \left(1 + \frac{(N_q - m_q)}{m_q}\right)^2 \left[\sum_{i=1}^{N_q} \left(\mu\sigma_r^2 + \varphi\sigma^2 + \sigma_r^2\sigma^2\right) + \left[\sum_{i=1}^{N_q} \mu\varphi\right]^2\right] \\
&= \left(1 + \frac{(N_q - m_q)}{m_q}\right)^2 \left[N_q\left(\mu\sigma_r^2 + \varphi\sigma^2 + \sigma_r^2\sigma^2\right) + [N_q\mu\varphi]^2\right] \\
&= \left(1 + \frac{(N_q - m_q)}{m_q}\right)^2 \left[N_q\left(\mu\sigma_r^2 + \varphi\sigma^2 + \sigma_r^2\sigma^2\right) + N_q^2\mu^2\varphi^2\right] \\
&= \frac{N_q^2}{m_q^2}\left[N_q\left(\mu\sigma_r^2 + \varphi\sigma^2 + \sigma_r^2\sigma^2\right) + N_q^2\mu^2\varphi^2\right]
\end{aligned}
$$

Ahora, desarrollando la segunda parte (Eq. 2), se tiene

$$\begin{aligned}
&\left[E\left(\sum_{i=1}^{N_q} y_i r_i + \frac{(N_q - m_q)}{m_q}\sum_{i=1}^{N_q} y_i r_i\right)\right]^2 \\
&= \left(1 + \frac{(N_q - m_q)}{m_q}\right)^2 \left[\sum_{i=1}^{N_q} \mu\varphi\right]^2 \\
&= \frac{N_q^2}{m_q^2}\left[N_q \mu\varphi\right]^2 \\
&= \frac{N_q^4 \mu^2 \varphi^2}{m_q^2}
\end{aligned}$$

Por lo tanto,

$$\begin{aligned}
Var\left(\hat{\bar{Y}}'\right) &= \frac{m_q^2}{N_q^4\varphi^2}\frac{N_q^2}{m_q^2}\left[N_q\left(\mu\sigma_r^2 + \varphi\sigma^2 + \sigma_r^2\sigma^2\right) + N_q^2\mu^2\varphi^2\right] - \frac{m_q^2}{N_q^4\varphi^2}\frac{N_q^4\mu^2\varphi^2}{m_q^2} \\
&= \frac{1}{N_q^2\varphi^2}\left[N_q\left(\mu\sigma_r^2 + \varphi\sigma^2 + \sigma_r^2\sigma^2\right) + N_q^2\mu^2\varphi^2\right] - \mu^2
\end{aligned}$$

Esta es la varianza del estimador de la media por el método imputación con la media. Para evaluar la consistencia del estimador, como el estimador es insesgado, basta con evaluar si

$$\lim_{n\to\infty} Var\left(\hat{\bar{Y}}'\right) = 0$$

Estimador del total: la expresión para la estimación de los totales es igual a la expresión del estimador de la media, pero sin dividir entre N_q (número total de registros), por lo tanto, se tiene la siguiente expresión:

$$\hat{T} = \sum_{i=1}^{N_q} y_i r_i + \bar{Y}_{\text{obs}}(N_q - m_q) \tag{3}$$

Con el fin de encontrar un estimador insesgado para el total, se procede a calcular el valor esperado de $\hat{T}$ de la siguiente manera:

$$
\begin{aligned}
E(\hat{T}) &= E\left(\sum_{i=1}^{N_q} y_i r_i + \bar{Y}_{\text{obs}}(N_q - m_q)\right) \\
&= N_q E(y_i)E(r_i) + (N_q - m_q)E(y_{obs}) \\
&= N_q \mu\varphi + (N_q - m_q)E\left(\frac{1}{m_q}\sum_{i=1}^{N_q} y_i r_i\right) \\
&= N_q \mu\varphi + (N_q - m_q)\left(\frac{1}{m_q}\right)\mu\varphi \\
&= \left[N_q + \frac{(N_q - m_q)}{m_q}\right]\mu\varphi \\
&= \left[\frac{m_q N_q + (N_q - m_q)}{m_q}\right]\mu\varphi \\
&= \left[\frac{m_q(N_q - 1) + N_q}{m_q}\right]\mu\varphi
\end{aligned}
$$

Ahora bien, obteniendo el estimador insesgado de (3):

$$
\begin{aligned}
\hat{T}' &= \frac{m_q}{(m_q(N_q - 1) + N_q)}\varphi \\
&= \frac{m_q}{(m_q(N_q - 1) + N_q)\varphi}\left[\sum_{i=1}^{N_q} y_i r_i + \bar{Y}_{obs}(N_q - m_q)\right] \\
&= \frac{m_q}{(m_q(N_q - 1) + N_q)}\varphi\left[\sum_{i=1}^{N_q} y_i r_i + \frac{(N_q - m_q)}{m_q}\sum_{i=1}^{N_q} y_i r_i\right]
\end{aligned}
$$

Este será el estimador insesgado del total por la técnica de imputación con la media.

Varianza del estimador del total:

$$
\begin{aligned}
V(\hat{T}') &= V\left[\frac{m_q}{(m_q(N_q - 1) + N_q)\varphi}\left[\sum_{i=1}^{N_q} y_i r_i + \frac{(N_q - m_q)}{m_q}\sum_{i=1}^{N_q} y_i r_i\right]\right] \\
&= \frac{m_q^2}{[(m_q(N_q - 1) + N_q)\varphi]^2} V\left[\sum_{i=1}^{N_1} y_i r_i + \frac{(N_q - m_q)}{m_q}\sum_{i=1}^{N_q} y_i r_i\right]
\end{aligned}
$$

Como lo único que varia es la constante con respecto al estimador de la media, tenemos:

$$V(\hat{T}') = \frac{m_q^2}{[(m_q(N_q-1)+N_q)\varphi]^2}\left(1+\frac{2(N_q-m_q)}{m_q}+\frac{(N_q-m_q)^2}{m_q^2}\right) \\ (\mu\sigma_r^2+\varphi\sigma^2+\sigma_r^2\sigma^2+N_q^2\mu^2\varphi^2-\mu^2)$$

De está manera se obtiene la varianza del estimador del total por el método de imputación con la media. No se calcularon los sesgos para la media y los totales pues en ambos casos se obtuvieron estimadores insesgados.

Estimador de la varianza: Una de las fórmulas generales para la estimación de la varianza es:

$$S^2 = \frac{\sum_{i=1}^{N} y_i^2 - N\bar{Y}^2}{N-1} \tag{4}$$

Para el método de imputación con la media se tiene:

$$\sum_{i=1}^{N} y_i^2 = \sum_{i=1}^{N} y_i^2 r_i + (N-m_q)\bar{Y}_{obs}^2$$

$$N_1\bar{Y}^2 = N_q\left[\frac{1}{N_q}\left(\sum_{i=1}^{N_q} y_i r_i + (N_q-m_q)\bar{Y}_{obs}\right)\right]^2$$

$$= N_q\frac{1}{N_q^2}\left(\left(\sum_{i=1}^{N_q} y_i r_i\right)^2 + 2\left(\sum_{i=1}^{N_q} y_i r_i\right)(N_q-m_q)\bar{Y}_{obs} + (N_q-m_q)^2\bar{Y}_{obs}^2\right)$$

como

$$\bar{Y}_{obs} = \frac{\sum_{i=2}^{N_q} y_i r_i}{m_q}$$

Entonces

$$\sum_{i=1}^{N_q} y_i r_i = m_q\bar{Y}_{obs}$$

Sustituyendo:

$$N_q\bar{Y}^2 = N_q\frac{1}{N_q^2}\left(m_q^2\bar{Y}_{obs}^2\left(\sum_{i=1}^{N_q} y_i r_i\right)^2 + 2m_q(N_q-m_q)\bar{Y}_{obs}^2 + (N_q-m_q)^2\bar{Y}_{obs}^2\right)$$

$$= \frac{1}{N_q}(m_q^2+2m_qN_q-2m_q^2+N_q^2+2m_qN_q+m_q^2)\bar{Y}_{obs}^2$$

$$= \frac{1}{N_q}N_q^2\bar{Y}_{obs}^2$$

$$N_q\bar{Y}^2 = N_q\bar{Y}_{obs}^2$$

Sustituyendo en

$$S^2 = \frac{\sum_{i=1}^{N_q} y_i^2 r_i + (N_q - m_q)\bar{Y}_{obs}^2 - N_q\bar{Y}_{obs}^2}{N_q - 1} \tag{5}$$

$$S^2 = \frac{\sum_{i=1}^{N_q} y_i^2 r_i + (N_q - m_q - N_q)\bar{Y}_{obs}^2}{N_q - 1} \tag{6}$$

$$S^2 = \frac{\sum_{i=1}^{N_q} y_i^2 r_i + m_q\bar{Y}_{obs}^2}{N_q - 1} \tag{7}$$

De está manera se obtiene el estimador de la varianza de la variable para la técnica imputación con la media.

Esperanza del estimador de la varianza:

$$\begin{aligned} E(S^2) &= E\left(\frac{\sum_{i=!}^{N_q} y_i^2 r_i + m_q\bar{Y}_{obs}^2}{N_q - 1}\right) \\ &= \frac{1}{N_q - 1}\left[E\left(\sum_{i=1}^{N_q} y_i^2 r_i\right) - m_q E(\bar{Y}_{obs}^2)\right] \\ &= \frac{1}{N_q - 1}\left[\sum_{i=1}^{N_q} E(y_i^2)E(r_i) - \frac{m_q}{m_q^2}E\left(\sum_{i=1}^{N_q} y_i^2 r_i\right)\right] \\ &= \frac{1}{N_q - 1}[N_q(V(y_i) + (E(y_i))^2)E(r_i)] \end{aligned}$$

Varianza del estimador de la varianza: para está parte solo se deja planteada para futuras investigaciones

$$\begin{aligned} V(S^2) &= V\left(\frac{\sum_{i=1}^{N_q} y_i^2 r_i + m_q\bar{Y}_{obs}^2}{N_q - 1}\right) \\ &= \frac{1}{(N_q - 1)^2}V\left(\sum_{i=1}^{N_q} y_i^2 r_i + m_q\bar{Y}_{obs}^2\right) \end{aligned}$$

Dado que,

$$V(x) = E(x^2) - E^2(x)$$

Se tiene:

$$
\begin{aligned}
=&\frac{1}{(N_q-1)^2}\left[E\left(\sum_{i=1}^{N_q}y_i^2r_i+m_q\bar{Y}_{obs}^2\right)^2-\left[E\left(\sum_{i=1}^{N_q}y_i^2r_i+m_q\bar{Y}_{obs}^2\right)\right]^2\right]\\
=&\frac{1}{(N_q-1)^2}E\left[\left(\sum_{i=1}^{N_q}y_i^2r_i\right)^2+2\left(\sum_{i=1}^{N_q}y_i^2r_i\right)m_q\bar{Y}_{obs}^2+m_q^2\bar{Y}_{obs}^4\right]-\\
&\frac{1}{(N_q-1)^2}\left[E\left(\sum_{i=1}^{N_q}y_i^2r_i\right)-m_qE[\bar{Y}_{obs}^2]\right]^2\\
=&\frac{1}{(N_q-1)^2}E\left[\left(\sum_{i=1}^{N_q}y_i^2r_i\right)^2+2E\left[m_q\left(\sum_{i=1}^{N_q}y_i^2r_i\right)\bar{Y}_{obs}^2\right]+E(m_q\bar{Y}_{obs}^4)\right]-\\
&\frac{1}{(N_q-1)^2}\left[E\left(\sum_{i=1}^{N_q}y_i^2r_i\right)-m_qE[\bar{Y}_{obs}^2]\right]^2\\
=&\frac{1}{(N_q-1)^2}E\left[\left(\sum_{i=1}^{N_q}y_i^2r_i\right)^2+2E\left[m_q\left(\sum_{i=1}^{N_q}y_i^2r_i\right)\bar{Y}_{obs}^2\right]+m_q^2E(\bar{Y}_{obs}^4)\right]-\\
&\frac{1}{(N_q-1)^2}\left[\left(\sum_{i=1}^{N_q}E(y_i^2)E(r_i)\right)-m_qE[\bar{Y}_{obs}^2]\right]^2
\end{aligned}
$$

Debido a la dificultas de desarrollar ecuaciones de cuarto orden se dejará expresado de está manera.

6. Bases de datos reales

6.1. Descripción de la base de datos original

En esta sección se hace una breve descripción de la base de datos que va a ser utilizada con el fin de llevar a cabo un EDA (Análisis Exploratorio de Datos). De esta manera se pueden conocer las principales características de la base y poder saber cuáles que técnicas se pueden utilizar para el análisis comparativo.

Dicho proceso se llevará a cabo con el uso de la base de datos proveniente del VI Censo agrícola (1997-1998) de Venezuela para tratar los datos faltantes en este tipo de censo, en particular.

6.2. Descripción de las variables de estudio

Para el desarrollo empírico de esta investigación se hace uso del VI Censo agrícola de Venezuela de 1998.

6.3. Patrón de pérdida

En esta base de datos se establecieron pérdidas aleatorias artificiales, compuestas por ciertos patrones de pérdida con una, dos, tres, cuatro o cinco variables perdidas simultáneamente.

Una vez obtenidos los patrones de datos y el porcentaje que representan cada uno de ellos en la base de datos original, se elimina de la base y se selecciona, en función a este, nuevos porcentajes de registros de la base de datos completa restante (BDC). Estas pérdidas artificiales obtenidas se imputan mediante las técnicas que se deseen compararse.

Por lo tanto, según el patrón de pérdida obtenido, del número de unidades para eliminar en la base de datos completa se obtiene:

Tabla 1: *Número de unidades para eliminar en la base de datos completa según el patrón de pérdida obtenido. V16: vacas paridas; V17: vacas en ordeño; V19: novillos; V27: gallinas reproductoras; V43: babas. Fuente: elaboración propia.*

	Para 5 %	**Para 10 %**	**Para 20 %**	**Para 30 %**
V16	2	4	8	11
V27	2	4	6	8
V16+ V17	10	18	32	41
V19+ V43	4	7	13	17
V16+ V17+ V19	8	13	23	32
V17+ V27+ V43	3	4	8	11
V19+ V27+ V43	9	18	32	41
V16+ V17+ V19+V27	5	9	16	21
V16+ V19+ V27+V43	5	9	16	21
V16+ V17+ V19+V27+V43	1	2	3	4

6.4. Construcción de los árboles de regresión (CART)

Una vez conocida la base de datos original y aplicado el patrón de pérdida de los datos se obtiene la base de datos completa y se procede a la construcción de los árboles de regresión para poder obtener grupos más similares, los cuales serán aquellos que se encuentren dentro de cada uno de los nodos. Para llevar a cabo el diseño de un árbol es necesario tener en cuenta ciertas características:

- Porcentaje de los datos para la muestra de aprendizaje o para la construcción del árbol.
- Porcentaje de los datos para la muestra de validación. Una vez construido el árbol, se evalúa con esta muestra para saber si el modelo fue bien construido el modelo.

Para la construcción de los arboles se utilizó el software `SPAD`, el cual cuenta con un manejo simple e intuitivo, y permite realizar diversos procedimientos al mismo tiempo. Se puede montar una cadena de análisis estadísticos que enlace un análisis de correspondencias con un análisis de clasificación en un momento. Además, por defecto aparece la composición de cada uno de los grupos que se han creado, esto representa una gran ventaja para el software `SPAD`.

El porcentaje de poda no es extraído del 100 % sino de los datos restantes a la muestra de validación, es decir, si tengo 100 datos y tomo el 40 % de muestra de validación y 20 % de muestra de poda, ese 20 % de muestra de poda es del 60 % restante de los datos, una vez extraído el porcentaje de validación. Se usaron ciertos índices para evaluar el árbol, es decir, para comparar modelos de árboles. Estos índices fueron el AIC (*Akaike Information Criterion*) y BIC (*Bayesian Information Criterion*), tanto para la muestra de validación como para la muestra de aprendizaje.

Tabla 2: *Valores de AIC y BIC para diferentes porcentajes de validación y poda. Fuente: elaboración propia.*

M. Valid.	M. Poda	Und segme.	Divi sion	Akaike Learn	Akaike test	BIC Learn	BIC Test	Error Red.Lear	Error Red.Test	Mean Err. Learn	Mean Err. Test	N Learrn	N Test
33 %	33 %	5	10	***105.712***	157.527	***108.705***	161.272	***0.0065***	***0.3357***	***0.000***	5.516.562	418	309
75 %	13 %	30	10	143.054	161.680	143.873	162.004	0.3449	0.8305	***0.000***	5.479.471	202	701
15 %	10 %	30	10	150.901	***148.964***	151.094	***149.591***	0.3331	0.3691	***0.000***	***1.894.567***	713	141
45 %	15 %	30	10	147.980	161.216	148.167	161.408	0.4776	0.6012	***0.000***	3.553.838	436	421
33 %	33 %	30	10	***130.201***	***157.004***	***131.263***	***158.333***	***0.0826***	***0.3626***	***0.000***	***5.580.531***	418	309
50 %	0 %	30	10	136.730	161.876	138.209	163.297	0.1564	0.6339	***0.000***	***686.478***	443	467

Los valores que están en negrilla y cursiva en (Tabla 2) son los que ofrecen menor valor de AIC y BIC que es lo que se busca. Mientras menor sean estos valores, mejor será el modelo de árbol. Estos modelos de árbol fueron el primero y el sexto modelo de árbol. Aquí se plantea el sexto, es decir, 67 % de los registros para la construcción del árbol (muestra de aprendizaje), un 33 % de la muestra de prueba y un 33 % de la muestra de aprendizaje para la muestra de poda.

El diseño del árbol que se seleccionó fue: árbol (33,30,10,33) que indica lo siguiente: 33 % de los registros componen la muestra de aprendizaje, 30 son el número de elementos mínimos para segmentar, los árboles se pueden crear hasta 10 niveles y 33 % de los registros conforman la muestra de prueba (validación). El tiempo en el que el computador arrojó los resultados fue entre 20 minutos y 5 horas, esto se debe posiblemente a la presencia de variables categóricas, las cuales toman muchos valores posibles, lo que origina tiempos mayores de corrida. (Breiman et al. 1984).

En la figura 1 se muestra uno de los árboles de regresión creados para imputar los datos faltantes de la variable “vacas paridas”, detallando un poco, con las

variables auxiliares y bajo las condiciones mencionadas anteriormente se crea el árbol. Posteriormente, con los registros que tienen para esta variable de ausencias se aplica la clasificación, usando este árbol de regresión y según en el nodo final donde quede cada registro se imputará según la media de los registros presentes dentro de cada nodo.

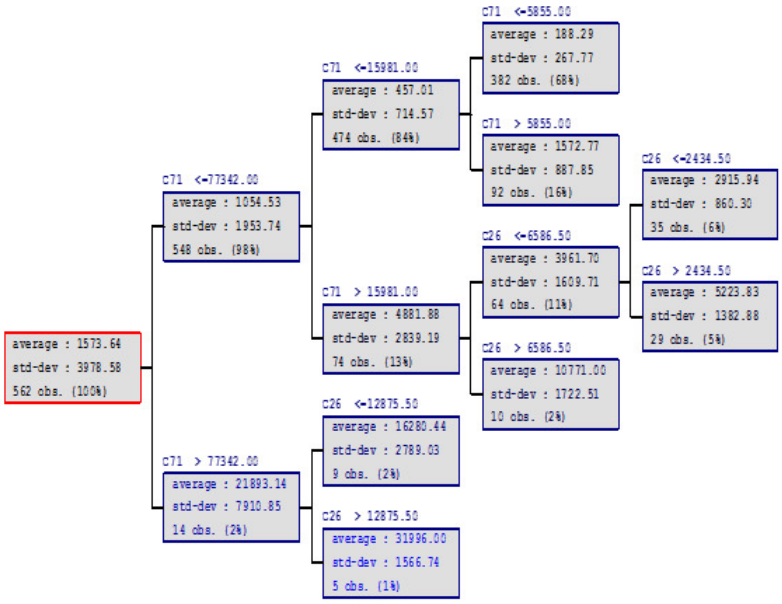

Figura 1: *Árbol de regresión para la variable "vacas paridas" para un 10 % de pérdida. Fuente: elaboración propia.*

7. Resultados y discusión

7.1. Comprobación de los supuestos fundamentales de análisis de la varianza

Con la finalidad de comparar las medias de las variables imputadas a diferente porcentaje de pérdida con la base de datos original, se realizó un análisis de la varianza, el cual permite corroborar si la técnica propuesta de imputación mantiene el promedio de las variables de estudio a diferentes protocolos o tratamientos de pérdida. Antes de la aplicación del análisis de la varianza se deben probar los supuestos de dicha técnica; es decir que los errores sean independientes, normalmente distribuidos con media cero y varianza constante. Para probar la normalidad se llevó a cabo un histograma de frecuencias y gráfico Q-Q, y se determinó que las

variables originales no lo cumplían. Por lo tanto, se recurrió a la transformación de las variables. Los resultados se pueden apreciar en la figura 2, teniendo en cuenta la transformación logarítmica más adecuada, y realizando de nuevo los histogramas y el gráfico Q-Q así cumple con el supuesto de normalidad. Para evaluar el supuesto de independencia se elaboraron los gráficos de residuos contra la secuencia del tiempo, en el cual no se observó ninguna correlación entre ellos, por lo tanto, se puede inferir que cumple así con el supuesto. Finalmente para el supuesto de aleatoriedad se llevó a cabo el gráfico de residuos contra el valor ajustado de los datos, donde no se aprecia ningún patrón inusual, por ello se puede decir que mantienen una varianza constante, cumpliendo así con la suposición de homogeneidad.

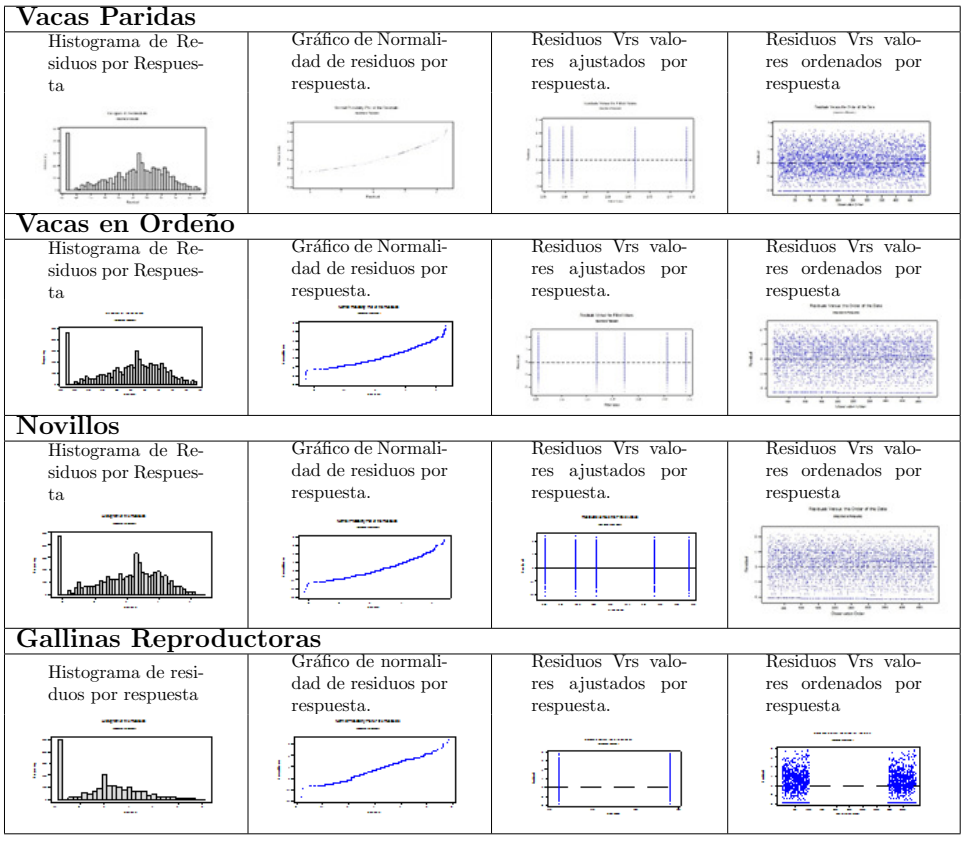

Figura 2: *Gráficos de los supuestos de aleatoriedad, normalidad e independencia de los ANOVAS aplicados. Datos transformados utilizando la función logaritmo. Fuente: elaboración propia.*

7.2. ANOVAS de comparaciones de los datos transformados

Según porcentaje de pérdida. Se desea probar si las medias entre los porcentajes de pérdida son significativamente iguales. Es decir, se realizará una comparación de medias, de la base de datos original (BDO) con la base de datos imputada a distintos porcentajes de pérdida.

$$H_0 : \mu_{\text{EDO}} = \mu_5 = \mu_{10} = \mu_{20} = \mu_{30}$$
$$H_1 : \text{al menos uno de los promedios difiere del resto}$$

Tabla 3: *P-valores de ANOVAS de una vía entre porcentajes de pérdida. Imputación con la media. Fuente: elaboración propia.*

Variable	P-Valor	Resultado
Vacas paridas	0,603	No hay diferencias significativas
Vacas en ordeno	0,812	No hay diferencias significativas
Novillos	0,364	No hay diferencias significativas
Gallinas reproductoras	0,675	No hay diferencias significativas

Como se muestra en la tabla 3 las imputaciones de la data perdida no son significativamente diferentes para los distintos porcentajes de pérdida. El estadístico F del análisis de la varianza que permite establecer la igualdad o no del efecto de los porcentajes de pérdida sobre las imputaciones, arrojó valores estadísticamente no significativos ($p \geq 0,05$). Por lo tanto, los datos confirman la hipótesis de igualdad de varianzas para los cuatro porcentajes de pérdida (5 %, 10 %, 20 %, 30 %) y la base de datos original.

Los residuos muestran un comportamiento normal, luego de que los datos fueron transformados. Los residuos no muestran relación aparente con los valores estimados para cada observación, tampoco manifiestan un aumento o disminución de la magnitud de las respuestas estimadas. Igualmente, no presentan autocorrelación, tampoco se evidencia una tendencia definida de los residuos en función del orden de observación, esto puede deberse a la experiencia del empadronador, eficiencia del organismo que recolectó los datos en el municipio, o al mal uso del software al imputar, etc.. Por lo tanto, se obtienen resultados válidos y en consecuencia conclusiones confiables.

Comparación de los sesgos según pérdida. Imputación con la media. Para ello se tomó la diferencia entre el valor real y el valor imputado y se compara sus promedios.

$$H_0 : \mu_{\text{sesgo5 \%}} = \mu_{\text{sesgo10 \%}} = \mu_{\text{sesgo20 \%}} = \mu_{\text{sesgo30 \%}}$$
$$H_1 : \text{al menos uno de los promedios difiere del resto}$$

Tabla 4: *Comparación de sesgos. P-valores de ANOVAS de una vía entre porcentajes de pérdida. Imputación con la media. Fuente: elaboración propia.*

Variables	**P-Valor**	**Resultado**
Vacas paridas	0,583	No hay diferencias significativas entre los sesgos de los diferentes porcentajes de pérdida para la vaca parida
Vacas en ordeño	0,216	No hay diferencias significativas entre los sesgos de los diferentes porcentajes de pérdida para la vaca en ordeño
Novillos	0,006	Se rechaza Ho, es decir, existen diferencias significativas entre 20 % y 30 % los sesgos aumentan.

7.3. Comparación de las relaciones de las variables mediante matrices de varianza y covarianza

Para poder evaluar si las técnicas de imputación mantienen la variabilidad de los datos y las relaciones entre las variables en forma simultánea (multivariante) se llevó a cabo una prueba de comparación de matrices de varianzas y covarianzas entre las variables que tenían al menos un dato perdido (Rencher, 2002). Las matrices para comparar fueron aquellas formadas por las varianzas y covarianzas de los datos de las variables completas; es decir, sin tomar en cuenta los valores para imputarlos.

Un supuesto para las pruebas MANOVA y T^2 al comparar dos o más vectores de media es que las matrices de varianzas poblacionales son iguales a $\Sigma_1 = \Sigma_2$. Bajo este supuesto, las matrices de covarianzas muestrales S_1, S_2 se asume una población común Σ. Si $\Sigma_1 = \Sigma_2$ no es verdadero, muy posiblemente se rechazaría $\mu_1 = \mu_2$. Por tanto se desea, probar la igualdad de matrices de varianzas y covarianzas.

7.4. Pruebas multivariantes para igualdad de matrices de covarianza

Según Rencher (2002) para k poblaciones multivariantes, la hipótesis de igualdad de matrices de covarianza es:

$$H_0 : \Sigma_1 = \Sigma_2 = \ldots = \Sigma_k$$
$$H_0 : \Sigma_1 = \Sigma_2$$

Para $k = 2$. No es exactamente una prueba de $H_0 : \Sigma_1 = \Sigma_2$, como en el coso univariante. Se asume muestras independientes de tamaño $n_1, n_2, \ldots, n_k$ de distribuciones normal multivariante.

Para hacer esta prueba se calcula:

$$M = \frac{|S_1|^{v_1/2}|S_2|^{v_2/2}\ldots|S_k|^{v_k/2}}{|S_{pl}|^{\sum_i v_i/2}} = \left(\frac{|S_1|^{v_1/2}}{|S_{pl}|}\right)\left(\frac{|S_2|^{v_2/2}}{|S_{pl}|}\right)\ldots\left(\frac{|S_k|^{v_k/2}}{|S_{pl}|}\right)$$

Para $S_1 = S_2 = S_pl$, entonces $M = 1$ y para una disparidad entre $S_1, S_2, \ldots, S_k$ se incrementa y M se acerca a cero.

Si $v = n_i - 1$, es una matriz de covarianzas de la i-ésima muestra, y S_{pl} es la muestra conjunta de la matriz de covarianzas.

$$S_{pl} = \frac{\sum_{i=1}^{k} v_i S_i}{\sum_{i=1}^{k} v_i} = \frac{E}{v_E}$$

Donde E viene dado por

$$E = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)(y_{ij} - \bar{y}_{i.})' = \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}y'_{ij} - \sum_{i=1}^{k}\frac{1}{n_i}y_i y'_i$$

y $v_E = \sum_{i=1}^{k} v_i = \sum_i n_i - k$. Está claro que se tiene $v_u > p$. Por otra parte si $|S_i| = 0$ para algún i, M podría ser cero.

Mayor a los puntos $-2lnM = v(kln|S_{pl}| - \sum_i ln|S_i|)$, para el caso $v_1 = v_2$, se hace uso de la tabla propuesta por (Lee et al. 1975), sin embargo, (BOX 1949) propone que dado X^2 se hace una aproximación F para la distribución M.

La aproximación X^2 se calcula:

$$C_1 = \left[\sum_{i=1}^{k}\frac{1}{v_i} - \frac{1}{\sum_{i=1}^{k} v_i}\right]\left[\frac{2p^2 + 3p - 1}{6(p+1)(k-1)}\right]$$

Entonces: $\mu = -2(1-C_1)lnM$ es aproximadamanete $X^2\left[\frac{1}{2}(k-1)p(p+1)\right]$; donde

$$lnM = \frac{1}{2}\sum_{i=1}^{k} v_1 ln|S_i| - \frac{1}{2}\left(\sum_{i=1}^{k} v_i\right) ln|S_{pl}|$$

Se rechaza H_0 sí $\mu > X^2_\alpha$. Si $v_1 = v_2 = \ldots = v_k = v$ donde C_1 se simplifica a

$$C_1 = \frac{(k+1)(2p^2+3p-1)}{6kv(p+1)}$$

Para justificar los grados de libertad de la aproximación x^2, note que el número total de parámetros estimados bajo H_1 es $k\left[\frac{1}{2}p(p+1)\right]$. Bajo H_0 estimamos sólo Σ, el cual tiene $p + \binom{p}{2} = \frac{1}{2}p(p+1)$ parámetros. La cantidad $k\left[\frac{1}{2}p(p+1)\right]$ surge del supuesto de todas Σ_i, $i = 1, 2, \ldots, k$, son diferentes. Técnicamente, H_1 puede estar como $\Sigma_i \neq \Sigma_j$ para algún $i \neq j$. Sin embargo, el caso más general para todo Σ_i sea diferente.

Comparando con un valor teórico para $k = 5$, $p = 5$ variables y $u = 268$ se obtiene los resultados de la tabla 5.

Tabla 5: *Valores T obtenidos y valores teóricos. Técnica de imputación con la media. Fuente: elaboración propia.*

Valor obtenido		Valor teórico	
T1	0.51	MT1	0.51
T2	9.55	MT2	9.51
T3	18.49	MT3	18.40
T4	15.91	MT4	15.83

Buscando una aproximación a:

$$X^2\left(\frac{1}{2}(k-1)p(p+1)\right)$$

Se rechaza si

$$MT \geq X^2_\alpha$$

8. Comparación de la base de datos imputadas con y sin el uso de CART

Finalmente, para saber, desde el punto de vista empírico, la influencia que tiene el uso de los CART sobre la estimación de los datos faltantes se compararon las bases de datos completas con la base de datos imputada, usando la técnica de imputación con la media, en presencia y ausencia de los CART. Se usó un

Tabla 6: *Medias y desviaciones estándar de las bases de datos imputadas con y sin CART. Fuente: elaboración propia.*

Variable	Vacas Paridas		Vacas en Ordeño		Novillos	
Parámetro	**Media**	**D.E**	**Media**	**D.E**	**Media**	**D.E**
BDC	1152,87	3239,3	1686,31	4207,21	926,91	2186,28
Media sin CART	1015,42	2704,26	1718,05	3431,87	966,53	1710,56
Media CART	1052	2389,11	1682,95	3795,38	887,92	1763,78

porcentaje de pérdida del 30 % que sería el peor de los casos, es decir, cuando hay mayor pérdida de información. Tabla 6.

Además se llevó a cabo una ANOVA en el que se compararon las bases de datos; la completa (con los datos reales), las imputadas según técnica de la media, con y sin el uso de CART. Se desea probar si la media entre la base de datos completa y la imputación con el uso de CART y la imputación sin el uso del CART son iguales. (Tabla 7)

$$H_0 : \mu_{\text{BDC}} = \mu_{\text{mediasinCART}} = \mu_{\text{mediaconCART}}$$
$$H_1 : \text{al menos uno de los promedios difiere del resto}$$

Tabla 7: *P-valores de ANOVAS de una vía entre bases de datos imputadas con y sin el uso de CART y la base de datos completa. Fuente: elaboración propia.*

Variable	P-Valor	Resultado
Vacas Paridas	0	Difiere BD Media sin CART
Vacas en Ordeño	0,001	Difiere BD Media sin CART
Novillos	0	Difiere BD Media sin CART

9. Conclusiones

La metodología de imputación diseñada inicia con el diseño de un árbol de clasificación y regresión CART, con los valores de Akaike y BIC más bajos, con un 33 % de muestra de aprendizaje (para crear el árbol) y un 33 % de muestra de prueba o validación. Como la finalidad es obtener en cada nodo final donantes para imputar, se segmenta el árbol si posee 30 elementos mínimos en el nodo y 10 niveles máximos por árbol. Este CART fue elaborado con el uso de todas las variables auxiliares de la base de datos (se construyó un CART por cada variable pérdida). Una vez obtenido el CART, dentro de cada nodo donde se encontrara un dato perdido, se aplica mediante una técnica clásica de imputación: un donante del mismo nodo para imputar al ausente. Para este caso se decidió probar con las técnicas clásicas de imputación con la media.

Se obtuvieron los estimadores para la media y los totales bajo el uso de los CART y se evaluaron las propiedades de los estimadores mediante el desarrollo de sus

varianzas y sesgos, observando insesgamiento. En cuanto a la varianza del estimador insesgado de la media se probó que es suficiente (varianza nula cuando N tiende a infinito), a diferencia del estimador de la media y los totales. Por lo tanto, esta técnica debe tomarse con cierta precaución. De esta manera, si el objetivo es la estimación de los totales, la técnica de imputación con la media no es recomendada, debido a la tendencia que tiene esta técnica a la subestimación de la varianza, concentrándose las frecuencias más altas hacia los valores centrales, lo cual no afecta la media pero sí a sus totales. Esta condición se mantiene aun realizando previamente la clasificación mediante CART, no resolviendo esta desventaja.

Como resultado de la investigación, esta metodología tiene ciertas limitaciones, como que si hay variables para ser imputadas que no mantienen correlación con alguna otra variable de la base de datos, el CART no encontrará variable auxiliar que clasifique a esta, por lo tanto, no pueden formar árbol como lo que ocurrió con las variables "babas" y "gallinas reproductoras" en el desarrollo empírico de la propuesta. Es necesario que la base de datos cuente con un conjunto de variables auxiliares clasificadoras, que guarden correlación con la variable a imputar, además, la metodología probada en este estudio implica un procedimiento más extenso y conlleva más tiempo y recursos.

Como ventajas de la metodología propuesta se tiene que: CART permite apreciar el registro "atípico" sin necesidad de otro estudio previo, el cual será aquel que quede sólo en un nodo desde el primer nivel del árbol y si es un registro que tiene un dato ausente, no se podrá imputar porque no lo acompaña ningún posible donante; se comprueba una vez más la resistencia que este método tiene frente datos atípicos y la escasez de supuestos, pues lo supuestos que se probaron fueron para hacer las comparaciones mediante ANOVAS y MANOVAS; las técnicas tradicionales como la media puede ofrecer buenos resultados si se combinan con técnicas de clasificación como los CART; la metodología es sencilla y fácil de interpretar; se conservan de los valores agregados como la media, varianzas y covarianzas; se mantiene la distribución de los datos en cuanto al apuntamiento de la curva de distribución (kurtosis) y la asimetría y la técnica se conserva robusta ante el aumento del porcentaje de pérdida, al menos hasta un 30 %.

Al llevarse a cabo la comprobación y evaluación de la metodología propuesta haciendo uso de los datos del VI Censo agrícola de 1998 con pérdidas artificiales entre el 5 % y 30 %, con patrones de pérdida de una, dos, tres, cuatro y cinco variables ausentes simultáneamente y aplicando la técnica de imputación; con la media, asumiendo una pérdida completamente aleatoria, se obtuvieron los siguientes resultados bajo el uso de CART: las técnicas conservan la relación entre las variables, tanto en comparaciones bivariantes como multivariantes; no se encontraron diferencias significativas entre los distintos porcentajes de pérdida, es decir, CART mantiene robustez en la metodología propuesta sin importar el tamaño de pérdida de la base de datos original; en cuanto a las distribuciones de las variables,

se observaron cambios en sus kurtosis, sin embargo, empiezan a verse diferencias cuando el porcentaje de pérdida es del 30 % , que aumenta la kurtosis; es decir, haciéndose más puntiaguda. En cuanto a la comparación del uso de la metodología CART antes de la aplicación de la técnica de imputación con respecto a no usar CART se aprecia la influencia positiva en la mejora de las estimaciones con el uso de CART para la técnica de imputación con la Media; Al llevar a cabo la imputación de manera conjunta (esto es imputando todas las pérdidas de un registro con el mismo donante) se aprecia que de esta manera se obtengan estimaciones similares a las llevadas a cabo de manera simple, debido a la poca correlación entre las variables no afecta que se use un donante para imputar por ítem o un donante que impute todos los ítems faltantes de un mismo registro de manera simultánea.

Recibido: 17 de enero de 2016
Aceptado: 6 de octubre de 2016

Referencias

Bárcena, M. J. & Tusell, F. (1999), 'Enlace de encuestas: una propuesta metodológica y aplicación a la encuesta de presupuestos de tiempo'.

Borgoni, R. & Berrington, A. (1990), 'A sequential tree-based procedure for multivariate imputation of complex missing data structure', *Journal of the American Statistical Association* **85**(410), 376–386.

BOX, G. E. P. (1949), 'A general distribution theory for a class of likelihood criteria', *Biometrika* **36**.

Breiman, L., Freidman, J., Olshen, R. & Stone, C. (1984), *Classification and Regression Tree*, 1 edn, Wadsworth.

Buck, S. F. (1960), 'A method of estimation of missing values in multivariate data suitable for use with an electronic computer', *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 302–306.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977), 'Maximum likelihood from incomplete data via the em algorithm', *Journal of the royal statistical society. Series B (methodological)* pp. 1–38.

Goicoechea, A. P. (2002), 'Imputación basada en árboles de clasificación', *Eustat. Available in: http://www.eustat.es/documentos/datos/ct* **4**.

Hansen, M., Hurwits, W. & Madow, W. (1953), *Sample survey Methods and Theory*, 1 edn, Wiley & Sons.

Krzanowski, W. (1988), 'Missing value imputation in multivariate data using the singular value decomposition of a matrix', *Biometrical letters* **25**(1-2), 31–39.

Lee, J., Chang, T. & Krishnaiah, P. (1975), 'Approximations to the Distributions of the likelihood Ratio Statistics for testing certain structures on the Covariance Matrices of Real Multivariate Normal Populations', *in Multivariate Analysis* pp. 105–118.

Little, R. J. & Rubin, D. B. (2014), *Statistical analysis with missing data*, John Wiley & Sons.

Lohr, S. (2009), *Sampling: design and analysis*, Nelson Education.

López, T. (2001), *Estudio de técnicas de análisis de datos para selección de variables, detección de valores atípicos y estimación de valores faltantes en entradas al sistema NEUROMASTER.*

Mesa, D. (2004), 'Imputación y árboles de decisión', *Guía práctica. Postgrado en Estadística, Universidad Central de Venezuela, Venezuela* .

Piela, P., Laaksonen, S. & Finland, S. (2001), Automatic interaction detection for imputation œ tests with the waid software package, *in* 'Contributed Paper for the Federal Committee on Statistical Methodology Research Conference, Washington, DC Area', Citeseer.

Rencher, A. C. (2002), *Methods of multivariate analysis*, Wiley series in probability and mathematical statistics, 2nd ed edn, J. Wiley.

Schafer, J. L. (1997), *Analysis of incomplete multivariate data*, CRC press.

Service, G. S. (1996), Report of the task force on imputation, *in* 'GSS Methodology Serie'.

Useche, L. & Mesa, D. (2006), 'Una introducción a la imputación de valores perdidos', *Revista Terra* **22**(31), 127–151.