MATEMÁTICA
ESPOL - FCNM JOURNAL
2022, Vol. 20, No. 2 Julio - Diciembre

Árboles de regresión para optimizar la imputación Hot Deck

Regression trees to optimize Hot Deck imputation

Lelly María Useche Castro, Wilmer Eduardo Barrera Yayes, Jean Carlos Pérez Parra y Victor Márquez Pérez.

Recepción: 14/03/2022 **Aceptación:** 19/07/2022 **Publicación:** 31/07/2022

Abstract Faced with the persistent need to have a complete database and the search to improve the classic techniques for estimating missing data, the research presents a proposed methodology for the imputation of the loss or lack of data, by combining a Segmentation analysis, specifically a Regression Tree and the classic sequential Hot Deck imputation technique. The estimators for the mean, total and variance are calculated, as well as the empirical validation of the proposal, in which unbiased estimators were obtained. The technique is considered to improve the robustness between losses of 5 and 30 % of the data, the variability of the data and the relationships between the variables are maintained. The new technique improves estimates over Hot Deck without the use of segmentation.
Keywords Imputation, absence of data, regression trees, Hot Deck.

Resumen Ante la persistente necesidad de tener una base de datos completa y la búsqueda de mejorar las técnicas clásicas de estimación de los datos ausentes, la in-

Lelly María Useche Castro, Ph.D.
Docente, Universidad Técnica de Manabí, UTM, Instituto de Ciencias Básicas, Departamento de Matemática y Estadística, Portoviejo, Ecuador, e-mail: lelly.useche@utm.edu.ec,
https://orcid.org/0000-0002-4294-9009

Wilmer Eduardo Barrera Yayes, M.Sc.
Docente, Universidad Técnica de Manabí, UTM, Instituto de Ciencias Básicas, Departamento de Matemática y Estadística, Portoviejo, Ecuador, e-mail: wilmer.barrera@utm.edu.ec,
https://orcid.org/0000-0002-5736-1865

Jean Carlos Pérez Parra, Ph.D.
Docente, Universidad Técnica de Manabí, UTM, Instituto de Ciencias Básicas, Departamento de Química, Portoviejo, Ecuador, e-mail: jean.perez@utm.edu.ec,
https://orcid.org/0000-0002-7971-1782

Victor Márquez Pérez, Ph.D.
Docente, Universidad Técnica de Manabí, UTM, Instituto de Ciencias Básicas, Departamento de Matemática y Estadística, Portoviejo, Ecuador, e-mail: victor.marquez@utm.edu.ec,
https://orcid.org/0000-0003-2458-2415

vestigación presenta una propuesta de metodología para la imputación de las pérdidas o falta de los datos, mediante la combinación de un Análisis de segmentación, específicamente un Árbol de Regresión y la técnica clásica de imputación Hot Deck secuencial. Se calcula los estimadores para la media, totales y varianza, así como la validación empírica de la propuesta, en la que se obtuvieron estimadores insesgados. Se considera que la técnica mejora la robustez entre pérdidas del 5 y 30 % de los datos, se mantiene la variabilidad de los datos y las relaciones entre las variables, mejora las estimaciones con respecto a la técnica Hot Deck sin el uso de la segmentación.
Palabras Claves Imputación, datos ausentes, árboles de Regresión, Hot Deck.

1 Introducción

Juster y Smith (1998) afirman que uno de los grandes males en el análisis de la información de los datos económicos y en las investigaciones de las ciencias sociales, es justamente la falta de información y la presencia de datos atípicos.

Sin embargo, hay muchas formas de minimizar la cantidad de datos faltantes, por ejemplo, mediante reglas estandarizadas para optimizar la recopilación de datos, capacitar al personal para coordinar la recopilación de datos, también, los estudios piloto pueden ayudar a identificar variables particularmente susceptibles a valores perdidos, y tomar medidas para mejorar la integridad (Pedersen et al., 2017). Por lo general, hay dos formas de manejar los valores perdidos del conjunto de datos. Algunas investigaciones intentan modelar directamente el conjunto de datos con valores perdidos. La segunda forma es imputar los valores faltantes para obtener el conjunto de datos completo y luego utilice métodos convencionales para analizar el conjunto de datos (Luo, Cai, Zhang, Xu, y xiaojie, 2018). Es frecuente que en la recolección y procesamiento de datos los usuarios conozcan de estos datos faltantes, pero no manejen correctamente las técnicas de imputación y no sean conscientes de las consecuencias como resultados deficientes o incorrectos aunque las técnicas estadísticas sean las correctas. La imputación de valor perdido (MVI) para conjuntos de datos incompletos es un problema muy importante en minería de datos y análisis de big data. Si los conjuntos de datos incompletos no están bien imputados, el resultado final del análisis podría verse afectado (Lin y Tsai, 2020).

Adicionalmente, Medina y Galván (2007) afirman que cuando se imputa información de manera errónea, conlleva además de introducir sesgos, reducción del poder explicativo, disminución de la eficiencia llevando a conclusiones incorrectas.

Los métodos de Hot Deck con sus distintos tipos surgen su uso para sustituir información faltante desde los años setenta principalmente en censos y encuestas. La imputación hoy en día forma parte del análisis exploratorio de datos o de la preparación de los datos, es decir, se debe tomar en cuenta como parte del proceso de investigación que permita resultados empíricos confiables, ya que, es bien conocido que las técnicas y métodos estadísticos en su mayoría funcionan con bases de datos completas. En base a ello, surge la necesidad de mejorar las técnicas clási-

cas de imputación con metodologías de clasificación, que bien, pueden haber sido aplicadas empíricamente pero no han sido estudiadas teóricamente. El objetivo de la imputación es obtener un archivo de datos completos y consistentes para que puedan ser analizados mediante técnicas estadísticas tradicionales. Es usado como tratamiento de la falta de respuesta parcial. (Useche y Mesa, 2006). Cuando tenemos pérdida parcial de datos, en la imputación de datos el objetivo es obtener una base de datos completa de manera tal de generar estimadores insesgados en el cual la base de datos final sea lo más parecida posible a la base de datos original.

1.1 Ausencia de datos

La ausencia de datos o falta de respuesta puede presentarse de dos manera. la ausencia de todo una fila o individuo conocida como ausencia o falta de respuesta total, por ejemplo; en la aplicación de una encuesta la persona no se encontró o no quiso responder todo el cuestionario, no se pudo tener acceso a la unidad de medición o deficiencias en el marco de muestreo. Por otra parte, cuando la ausencia no es de toda la fila sino de algunos datos de la misma, se considera ausencia o falta de respuesta parcial, por ejemplo, puede ser porque la persona no quiso responder algunas preguntas, no se pudo obtener una medición de varias unidades experimentales. La no respuesta total se corrige mediante reponderación y en el caso de la no respuesta parcial se utilizan los procedimientos de imputación, siendo muy común las técnicas Hot Deck.

1.2 Descripción del modelo de imputación (Little y Rubin, 2019).

Sean x una variable aleatoria observada y y una variable aleatoria parcialmente observada. Sea ϕ un vector y θ un vector de parámetros. Se define A como la respuesta definida como una variable aleatoria con distribución $f(A/x, y, \theta, \phi)$.
Se define $y = (y_o, y_f)$, donde y_o es un vector de tamaño $n \times 1$ que representa los valores observados de y y; y_f es un vector de tamaño $(N - n) \times 1$ que representa los valores perdidos de y.
La distribución de $f(x, y, A/\theta, \phi)$ puede ser escrita como: $f(x, y_o, y_f, A/\theta, \phi)$, donde:

$$f(x, y_o, y_f, A/\theta, \phi) = f(x, y_o, y_f, /\theta) \times f(A/x, y_o, y_f, /\phi).$$

Se puede integrar en función de y_f para obtener la distribución de los datos observados:

$$f(x, y_o, A/\theta, \phi) = \int f(x, y_o, y_f, \theta) \times f(A/x, y_o, y_f, \phi)\, dy_f.$$

Asumiendo que A es independiente de y dado X, la data es llamada pérdida aleatoria, MAR, en el cual: $f(A/x, y_o, y_f, \phi) = f(A/x, y_o, \phi)$.

Los datos observados son, por lo tanto:

$$f(x, y_o, A/\theta, \phi) = f(x, y_o, /\theta) \times f(A/x, y_o, /\phi).$$

Entonces, se usa el procedimiento de máxima verosimilitud para completar los datos observados, método que puede ser usado para estimar el parámetro θ cuando los datos están incompletos.

1.3 Método Hot Deck

La metodología Hot Deck, es una metodología clásica de imputación, ampliamente usada en los censos y encuestas Andridge, R. and Little, R. (2010). Esta metodología consiste en reemplazar el primer dato perdido con un valor inicial, bien sea aleatorio o inclusive con información no perteneciente a los datos, si éste está completo, será el donante o el valor utilizado para imputar el próximo que sí esté faltando. No es recomendable cuando la ausencia de los datos es muy alta ya que puede tender a usarse el mismo donante para varias imputaciones, perdiendo precisión las estimaciones.
(García y Sancho, 2005) indican que el método de imputación Hot Deck puede ser más parsimonioso que otros métodos de imputación univariante y que la regresión paramétrica, preservando la distribución de los datos siendo una buena selección cuando se requiere imputar en diferentes bases de datos.
Como lo manifiesta Montaquila y Ponikowski (1993), cuando hay poca información auxiliar disponible o la información auxiliar no está altamente correlacionada con las características de interés o con la variable propensa a pérdida, el procedimiento Hot Deck secuencial puede ser usado. Es allí, la importancia de clasificar los datos en grupos homogéneos, en el cual las correlaciones entre las variables dejan de ser el desafío de la técnica, más bien ayudaría a la clasificación de los grupos.
En Little (2019) se explica la técnica de la siguiente manera.

Suponga que una muestra de n fuera de las N unidades es seleccionada, y m fuera de los n valores muestrales de una variable Y son recolectados, donde n, N y m son tratados como datos ajustados. Para simplicidad, sean las primeras n unidades, $i = 1, \ldots, n$ como muestra, y la primera $m < n$ unidades como respondientes. Dado un esquema de probabilidades iguales en la muestra, la media Y podría ser estimada como la media de los respondientes y de las unidades imputadas. Esto podría estar escrito de la forma:

$$\overline{y}_{HD} = \{m\overline{y}_A + (n-m)\overline{y}^*_{NA}\}/n,$$

donde $\overline{y}_A$ es la media de las unidades respondientes, y

$$\overline{y}^*_{NA} = \sum_{i=1}^{m} \frac{H_i y_i}{n-m},$$

donde H_i es el número de veces en el que y_i es utilizado como un sustituto para un valor perdido de Y. Note que $\sum_{i=1}^{m} H_i = n - m$, es el número de unidades perdidas.

1.4 Árboles de Regresión

Para definir los árboles de regresión, se parte de los análisis de segmentación el cual son, técnicas recursivas que han venido desarrollándose desde hace ya varios años, con mayor crecimiento con el surgimiento de la tecnología Escobar, M. (2007). El objetivo principal de los análisis de segmentación son divisiones jerárquicas de un grupo de datos con la ayuda de variables auxiliares con la finalidad de obtener grupos más homogéneos dentro de sí y más heterogéneos entre ellos. Su uso actual es bien reconocido en la minería de datos o Big data, por las ventajas ante el manejo de grades volúmenes de datos, adicionalmente, han sido de gran provecho en procesos de muestreo, para la identificación de estratos, o para estudiar inclusive grupos atípicos, para clasificar individuos, caracterizarlos y recientemente el uso de esta metodología para la estimación de la data faltante.
Pero existen diferentes técnicas de segmentación basada en árboles, entre ellas se encuentran divisiones binarias o no binarias, para datos cualitativos como los árboles de clasificación, y para los datos cuantitativos como los árboles de regresión.
El algoritmo del árbol de clasificación y regresión contiene tres tareas importantes. La primera tarea es ¿cómo segmentar los datos en cada paso?, la segunda tarea es ¿cuándo detener la segmentación?, y la última es ¿cómo predecir el valor Y para cada X en un segmento? (Olfaz, Tirink, y Önder, 2019). Por otra parte, un modelo de árbol tiene el mismo objetivo que un modelo lineal, que es estimar una función de regresión (Loh, Zhang, Zhang, y Zhou, 2020). Sin embargo, la regresión lineal falla si hay multicolinealidad y la regresión logística falla si hay una separación cuasi completa (Loh, Eltinge, Cho, y Li, 2018). Un modelo de árbol es una aproximación lineal por partes a la función verdadera, no menos que un modelo lineal es una aproximación lineal. No hay ninguna razón por la que no se pueda utilizar un modelo lineal si el modelo real no es lineal (Little, 2019).
Este árbol de decisiones es normalmente aplicable en datos minería para producir un marco que prediga el valor de objeto o de su variable dependiente, establecida sobre las diversas entrada o variable independiente (Bhargava, Dayma, Kumar, y Singh, 2017).

2 Desarrollo metodológico

La metodología de la investigación se basa en proponer un método de imputación de datos mediante la combinación de una técnica clásica como lo es el procedimien-

to Hot Deck Secuencial y los árboles de Regresión. Se inicia con la obtención de los estimadores para la media y totales de sus varianzas y sesgos, con base a propiedades de insesgadez, invarianza y consistencia. En caso de que alguna de estas propiedades no se cumpla, se propone una corrección. Luego se evalúa de manera empírica con una base de datos cuantitativa en la que se evalúa los patrones de pérdida de la misma y se elimina los perdidos por filas, obteniendo una base de datos más pequeña pero completa. Luego, se realiza una pérdida artificial de datos en la base completa, de manera que, al imputarse, se conoce el dato real para poder hacer comparaciones. Las pérdidas se hacen con un 5 %, 10 %, 15 %, 20 % y 30 %, con la finalidad de determinar la robustez del método.
Se lleva acabo un análisis de varianza (ANOVA) de comparación entre los diferentes porcentajes de pérdida, entre las bases de datos con y sin el uso de árboles de clasificación y regresión (CART) para determinar la efectividad de la propuesta y comparaciones de matrices de varianzas y covarianzas con el objeto de observar los cambios en la variabilidad de los datos si hay subestimación o sobreestimación de la varianza.

3 Resultados y discusión

Para esta técnica de imputación Hot Deck Secuencial asumimos que el primer registro de la variable está presente, así como ya tiene un orden dentro del nodo y se sustituye con el valor del registro anterior y_{i-1}, por tanto;

Estimador de la media:

$$\overline{Y} = \frac{1}{Q}\sum_{q=1}^{Q}\left[\frac{1}{N_q}\left(\sum_{i=1}^{N_q} y_i \cdot a_i + \sum_{i=2}^{N_q} y_{(i-1)} \cdot (1 - a_i)\right)\right], \tag{1}$$

siendo Q el número total de nodos y a_i se define como un vector indicador de las variables para Y, es decir,

$$a_i = \begin{cases} 1, & \text{si } y_i \text{ esta observada} \\ 0, & \text{si no es observada.} \end{cases}$$

Hallando un estimador insesgado para la Media dentro de cada nodo. Se calcula la esperanza del estimador de (1).

Esperanza del estimador de la media:

$$E(\overline{Y}) = \frac{1}{N_q}\left(\sum_{i=1}^{N_q} y_i \cdot a_i + \sum_{i=2}^{N_q} y_{(i-1)} \cdot (1 - a_i)\right)$$
$$= \frac{1}{N_q} E\left[\sum_{i=1}^{N_q} y_i \cdot a_i\right] + \frac{1}{N_q} E\left[\sum_{i=2}^{N_q} y_{(i-1)} \cdot (1 - a_i)\right].$$

Luego, como $E(y_i) = \mu$, $E(a_i) = \varphi$, siendo μ el parámetro de la media, φ el parámetro de la media de los a_i y asumiendo independencia entre y_i y a_i, se tiene;

$$E(\overline{Y}) = \frac{1}{N_q}\left[\sum_{i=1}^{N_q} E(y_i) \cdot E(a_i)\right] + \frac{1}{N_q}\left[\sum_{i=2}^{N_q} E(y_{(i-1)}) \cdot E(1 - a_i)\right]$$
$$= \frac{N_q}{N_q}\mu\varphi + \frac{N_q - 1}{N_q}(1 - \varphi)\mu$$
$$= \mu\varphi + \frac{N_q - 1}{N_q}(\mu - \varphi\mu)$$
$$= \mu\varphi + \frac{N_q\mu - N_q\varphi\mu - \mu + \varphi\mu}{N_q}$$
$$= \mu\varphi + \mu - \mu\varphi - \frac{\mu}{N_q} + \frac{\mu\varphi}{N_q}$$
$$= \mu - \frac{\mu}{N_q} + \frac{\mu\varphi}{N_q}$$
$$= \mu\left(\frac{N_q - (1 - \varphi)}{N_q}\right). \tag{2}$$

Ahora bien, obteniendo un estimador insesgado de (2) se tiene:

$$\overline{Y'} = \frac{N_q}{N_q - (1 - \varphi)} \cdot E(\overline{Y})$$
$$= \frac{N_q}{N_q - (1 - \varphi)} \cdot \frac{1}{N_q}\left(\sum_{i=1}^{N_q} y_i \cdot a_i + \sum_{i=2}^{N_q} y_{(i-1)} \cdot (1 - a_i)\right)$$
$$= \frac{1}{N_q - (1 - \varphi)}\left(\sum_{i=1}^{N_q} y_i \cdot a_i + \sum_{i=2}^{N_q} y_{(i-1)} \cdot (1 - a_i)\right).$$

Finalmente, dentro del árbol queda;

$$\overline{Y'} = \frac{1}{Q}\sum_{q=1}^{Q}\left[\frac{1}{N_q - 1 + \varphi}\left(\sum_{i=1}^{N_q} y_i \cdot a_i + \sum_{i=2}^{N_q} y_{(i-1)} \cdot (1 - a_i)\right)\right]. \tag{3}$$

Obteniendo el estimador insesgado de la media para la imputación por Hot Deck. Al obtener un estimador insesgado, su sesgo es igual a cero.

Varianza del estimador de la media:

$$
\begin{aligned}
V(\overline{Y'}) &= V\left[\frac{1}{(N_q-1+\varphi)}\left(\sum_{i=1}^{N_q} y_i\cdot a_i+\sum_{i=2}^{N_q} y_{(i-1)}\cdot(1-a_i)\right)\right]\\
&= \frac{1}{(N_q-1+\varphi)^2}V\left[\left(\sum_{i=1}^{N_q} y_i\cdot a_i+\sum_{i=2}^{N_q} y_{(i-1)}\cdot(1-a_i)\right)\right]\\
&= \frac{1}{(N_q-1+\varphi)^2}\cdot E\left(\left(\sum_{i=1}^{N_q} y_i\cdot a_i+\sum_{i=2}^{N_q} y_{(i-1)}\cdot(1-a_i)\right)\right)^2\\
&- \frac{1}{(N_q-1+\varphi)^2}\cdot E^2\left(\left(\sum_{i=1}^{N_q} y_i\cdot a_i+\sum_{i=2}^{N_q} y_{(i-1)}\cdot(1-a_i)\right)\right). \qquad (4)
\end{aligned}
$$

Desarrollando la esperanza en el primer término de (4) obtenemos:

$$
\begin{aligned}
&E\left(\left(\sum_{i=1}^{N_q} y_i\cdot a_i+\sum_{i=2}^{N_q} y_{(i-1)}\cdot(1-a_i)\right)\right)^2 = E\left[\left(\sum_{i=1}^{N_q} y_i\cdot a_i\right)^2+2\left(\sum_{i=1}^{N_q} y_i\cdot a_i\right)\cdot\right.\\
&\left.\cdot\ \sum_{i=2}^{N_q} y_{(i-1)}\cdot(1-a_i)+\left(\sum_{i=2}^{N_q} y_{(i-1)}\cdot(1-a_i)\right)^2\right]. \qquad (5)
\end{aligned}
$$

Del primer término de (5) obtenemos:

$$
[N_q(\mu^2\sigma_r^2+\varphi\sigma^2+\sigma^2\sigma_r^2)+N_q^2(\mu^2\varphi^2)]. \qquad (6)
$$

Resolviendo el segundo término de (5) obtenemos:

$$
\begin{aligned}
&E\left[2\left(\sum_{i=1}^{N_q} y_i\cdot a_i\right)\cdot\sum_{i=2}^{N_q} y_{(i-1)}\cdot(1-a_i)\right] = 2E\left(\sum_{i=1}^{N_q} y_i\cdot a_i\right)\cdot E\left(\sum_{i=2}^{N_q} y_{(i-1)}\cdot(1-a_i)\right)\\
&= 2\left(\sum_{i=1}^{N_q} E(y_i)\cdot E(a_i)\right)\cdot\left(\sum_{i=2}^{N_q} E(y_{(i-1)})\cdot E(1-a_i)\right)\\
&= 2N_q\mu\varphi(N_q-1)\mu(1-\varphi)\\
&= 2N_q(N_q-1)\mu^2\varphi(1-\varphi). \qquad (7)
\end{aligned}
$$

La ecuación (7) se deduce tomando en cuenta la independencia entre y_i y a_i y entre los y_{i-1} y $(1-a_i)$.
Resolviendo el tercer término de (5) obtenemos:

$$E\left(\sum_{i=2}^{N_q} y_{(i-1)} \cdot (1-a_i)\right)^2 = E[(y_1(1-a_2)+y_2(1-a_3)+\cdots+y_{N_q}(1-a_{N_q}))\cdot$$
$$\cdot\ (y_1(1-a_2)+y_2(1-a_3)+\cdots+y_{N_q}(1-a_{N_q}))]$$
$$= E\left(\sum_{i=2}^{N_q} y_{i-1}^2(1-a_i)^2 + 2\sum_{i=2}^{N_q}\sum_{i=2}^{N_q} y_{i-1}y_{j-1}(1-a_i)(1-a_j)\right)$$
$$= E\left[\sum_{i=2}^{N_q} y_{i-1}^2(1-2a_i+a_i^2)\right] + E\left[2\sum_{i=2}^{N_q}\sum_{i=2}^{N_q} y_{i-1}y_{j-1} - y_{i-1}y_{j-1}a_i - y_{i-1}y_{j-1}a_j +\right.$$
$$\left.+\ y_{i-1}y_{j-1}a_ia_j\right]$$
$$= \left(\sum_{i=2}^{N_q} E[y_{i-1}^2 - 2y_{i-1}^2a_i + y_{i-1}^2a_i^2] + 2\sum_{i=2}^{N_q}\sum_{i=2}^{N_q} E[y_{i-1}y_{j-1} - y_{i-1}y_{j-1}a_i - y_{i-1}y_{j-1}a_j +\right.$$
$$\left.+\ y_{i-1}y_{j-1}a_ia_j]\right)$$
$$= \left(\sum_{i=2}^{N_q} E(y_{i-1}^2) - E(y_{i-1}^2a_i) + 2\sum_{i=2}^{N_q}\sum_{i=2}^{N_q} E(y_{i-1})E(y_{j-1}) - E(y_{i-1})E(y_{j-1})E(a_i) -\right.$$
$$\left.-\ E(y_{i-1})E(y_{j-1})E(a_j) +\ E(y_{i-1})E(y_{j-1})E(a_i)E(a_j)\right)$$
$$= (N_q-1)(\mu-\mu\varphi) + 2\binom{N_q-1}{2}(\mu^2-\mu^2\varphi-\mu^2\varphi+\mu^2\varphi^2)$$
$$= (N_q-1)(\mu-\mu\varphi) + 2\binom{N_q-1}{2}(\mu^2-2\mu^2\varphi+\mu^2\varphi^2)$$
$$= (N_q-1)(\mu-\mu\varphi) + 2\binom{N_q-1}{2}(\mu-\mu\varphi)^2$$
$$= (N_q-1)(\sigma^2+\mu^2) - (\sigma^2+\mu^2)\varphi + (N_q-1)\cdot(N_q-2)(\mu-\mu\varphi)^2. \tag{8}$$

Sustituyendo (6), (7) y (8) en (5) obtenemos:

$$E\left(\sum_{i=1}^{N_q} y_i\cdot r_i + \sum_{i=2}^{N_q} y_{(i-1)}\cdot(1-a_i)\right)^2 = N_q(\mu^2\sigma_r^2+\varphi\sigma^2+\sigma^2\sigma_r^2) + N_q^2(\mu^2\varphi^2) +$$
$$+\ 2N_q(N_q-1)\mu^2\varphi(1-\varphi) + (N_q-1)\cdot(\sigma^2+\mu^2)(\sigma^2+\mu^2)\varphi + (N_q-1)(N_q-2)\cdot$$
$$\cdot\ (\mu-\mu\varphi)^2$$
$$= N_q\mu^2\sigma_r^2 + N_q\varphi\sigma^2 + N_q\sigma^2\sigma_r^2 + N_q^2\mu^2\varphi^2 + 2N_q^2\mu^2\varphi - 2N_q^2\mu^2\varphi^2 + 2N_q\mu^2\varphi^2 -$$
$$-\ 2N_q\mu^2\varphi + N_q\sigma^2 + N_q^2\mu^2 - \sigma^2 - \mu^2 - \sigma^2\varphi - \mu^2\varphi + N_q^2\mu^2 - 2N_q^2\mu^2\varphi + N_q^2\mu^2\varphi^2 -$$
$$-\ 3N_q\mu^2 + 6N_q\mu^2\varphi - 3N_q\mu^2\varphi^2 + 2\mu^2 - 4\mu^2\varphi + 2\mu^2\varphi^2$$
$$= N_q\mu^2\sigma_r^2 + (N_q\varphi + N_q\sigma_r^2 + N_q - 1 - \varphi)\sigma^2 + \left(4N_q\varphi - 2N_q + 1 - 5\varphi + N_q^2\right)\mu^2 +$$
$$+\ \left(2\mu^2 - 3N_q\mu^2\right)\varphi^2. \tag{9}$$

Desarrollando la esperanza en el segundo término de (4) obtenemos,

$$
\begin{aligned}
E^2\left(\left(\sum_{i=1}^{N_q} y_i \cdot a_i + \sum_{i=2}^{N_q} y_{(i-1)} \cdot (1-a_i)\right)\right) &= \left(\sum_{i=1}^{N_q} E(y_i)\cdot E(a_i) + \sum_{i=2}^{N_q} E(y_{(i-1)})\cdot E(1-a_i)\right)^2 \\
&= [N_q\mu\varphi + (N_q-1)\mu(1-\varphi)]^2 \\
&= N_q^2\mu^2\varphi^2 + 2N_q^2\mu^2\varphi - 2N_q\mu^2\varphi - 2N_q^2\mu^2\varphi^2 + 2N_q\mu^2\varphi^2 + (N_q^2 - 2N_q + 1)(\mu^2 - \\
&- 2\mu^2\varphi + \mu\varphi^2) \\
&= N_q^2\mu^2\varphi^2 + 2N_q^2\mu^2\varphi - 2N_q\mu^2\varphi - 2N_q^2\mu^2\varphi^2 + 2N_q\mu^2\varphi^2 + N_q^2\mu^2 - 2N_q\mu^2 + \mu^2 - \\
&- 2N_q^2\mu^2\varphi + 4N_q\mu^2\varphi - 2\mu^2\varphi + N_q^2\mu^2\varphi^2 - 2N_q\mu^2\varphi^2 + \mu^2\varphi^2 \\
&= \mu^2\left(N_q^2\varphi^2 + 2N_q^2\varphi - 2N_q\varphi - 2N_q^2\varphi^2 + 2N_q\varphi^2 + N_q^2 - 2N_q 2N_q^2\varphi + 4N_q\varphi - 2\varphi\right) + \\
&+ \varphi^2\left(N_q^2\mu^2 - N_q\mu^2 + \mu^2\right). \qquad (10)
\end{aligned}
$$

Reemplazando (9) y (10) en la ecuación (4) obtenemos:

$$
\begin{aligned}
V(\overline{Y'}) &= \frac{1}{(N_q - 1 + \varphi)^2}\left[N_q\mu^2\sigma_r^2 + (N_q\varphi + N_q\sigma_r^2 + N_q - 1 - \varphi)\sigma^2 + \left(4N_q\varphi - 2N_q + \right.\right. \\
&+ 1 - 5\varphi + N_q^2\Big)\mu^2 + \left(2\mu^2 - 3N_q\mu^2\right)\varphi^2 - \mu^2\left(N_q^2\varphi^2 + 2N_q^2\varphi - 2N_q\varphi - 2N_q^2\varphi^2 + \right. \\
&+ \left.\left. 2N_q\varphi^2 + N_q^2 - 2N_q 1 - 2N_q^2\varphi + 4N_q\varphi - 2\varphi\right) - \varphi^2\left(N_q^2\mu^2 - N_q\mu^2 + \ \mu^2\right)\right] \\
&= \frac{1}{(N_q - 1 + \varphi)^2}\left[N_q\mu^2\sigma_r^2 + \sigma^2 N_q\varphi + N_q\sigma_r\sigma^2 + \sigma^2 N_q - \sigma^2 - \varphi\sigma^2 + \mu^2 N_q^2\varphi^2 + \right. \\
&+ \left. 2\mu^2 N_q\varphi + 2\mu^2 N_q\varphi^2 - 3\mu^2\varphi - \varphi^2 N_q^2\mu^2 + 4\varphi^2 N_q\mu^2 - 3\varphi^2\mu^2\right] \\
&= \frac{1}{(N_q - 1 + \varphi)^2}\left[N_q\mu^2\sigma_r^2 + \sigma^2 N_q\varphi + N_q\sigma_r\sigma^2 + \sigma^2 N_q - \sigma^2 - \varphi\sigma^2 + 2\mu^2 N_q\varphi + \right. \\
&+ \left. 6\varphi^2 N_q\mu^2 - 3\mu^2 - 3\varphi^2\mu^2\right]. \qquad (11)
\end{aligned}
$$

Y así se obtiene la varianza del estimador de la media para la técnica de imputación por Hot Deck.

Consistencia:

A partir de la ecuación 11 se obtiene que $V(\overline{Y'}) \to 0$ cuando $N_q \to \infty$. El cálculo de este límite se realizó en Matlab R2017b. En vista de que el resultado es cero, se concluye que el estimador es consistente.

Estimador del total:

Como se ha desarrollado en la técnica dada por Useche, Parra, Mendoza y Chacón (2021) se tiene que,

$$\hat{T} = \sum_{q=1}^{Q}\left(\sum_{i=1}^{N_q} y_i \cdot a_i + \sum_{i=2}^{N_q} y_{(i-1)} \cdot (1 - a_i)\right). \tag{12}$$

Hallando un estimador insesgado para (12) se calcula lo siguiente:

Esperanza del estimador del total:

$$\begin{aligned}
E(\hat{T}) &= \sum_{q=1}^{Q}\left(E\sum_{i=1}^{N_q} y_i \cdot a_i + E\sum_{i=2}^{N_q} y_{(i-1)} \cdot (1 - a_i)\right) \\
E(\hat{T}) &= N_q\mu\varphi + (N_q - 1)\mu(1 - \varphi) \\
&= N_q\mu\varphi + (N_q\mu - \mu)(1 - \varphi) \\
&= N_q\mu\varphi + N_q\mu - N_q\mu\varphi - \mu + \mu\varphi \\
&= N_q\mu - \mu + \mu\varphi \\
&= \mu(N_q - 1 + \varphi). \qquad (13)
\end{aligned}$$

Ahora bien, obteniendo un estimador insesgado de (13):

$$\begin{aligned}
\hat{T}' &= \frac{1}{N_q - 1 + \varphi}\hat{T} \\
\hat{T}' &= \frac{1}{N_q - 1 + \varphi}\left(\sum_{i=1}^{N_q} y_i \cdot a_i + \sum_{i=2}^{N_q} y_{(i-1)} \cdot (1 - a_i)\right). \qquad (14)
\end{aligned}$$

Varianza del estimador del total:

$$V(\hat{T}') = \frac{1}{(N_q - 1 + \varphi)^2} V\left(\sum_{i=1}^{N_q} y_i \cdot a_i + \sum_{i=2}^{N_q} y_{(i-1)} \cdot (1 - a_i)\right).$$

Basándose en la expresión:

$$V(\hat{T}') = E(x^2) - [E(x)]^2.$$

Sustituyendo:

$$V(\hat{T}') = \frac{1}{(N_q - 1 + \varphi)^2}\left[E\left[\left(\sum_{i=1}^{N_q} y_i \cdot a_i + \sum_{i=2}^{N_q} y_{(i-1)} \cdot (1 - a_i)\right)^2\right] -\right.$$

$$- \left[E\left(\sum_{i=1}^{N_q} y_i \cdot a_i + \sum_{i=2}^{N_q} y_{(i-1)} \cdot (1 - a_i) \right) \right]^2 \Bigg]. \tag{15}$$

En base al desarrollo obtenido en el estimador de la media en (11) se tiene finalmente, la varianza del estimador del total:

$$\begin{aligned} V(\hat{T}') = \frac{1}{(N_q - 1 + \varphi)^2} [N_q \mu \sigma_r^2 + (N_q \varphi + N_q \sigma_r^2 + N_q - 1 - \varphi)\sigma^2 + \\ + (2N_q \varphi + 5\varphi - 4)\mu^2 + (N_q - 3)\varphi^2]. \end{aligned} \tag{16}$$

Los sesgos para las medias y los totales son iguales a cero por ser insesgados.

Consistencia:

A partir de la ecuación (16) se obtiene que $V(\hat{T}') \to 0$ cuando $N_q \to \infty$. El cálculo de este límite se realizó en Matlab R2017b. En vista de que el resultado es cero, se concluye que el estimador es consistente.

Estimador de la varianza:

Al igual que con las dos técnicas anteriores, en base a la fórmula general para la estimación de la varianza,

$$s^2 = \frac{\sum_{i=1}^{N} y_i^2 - N\bar{y}^2}{N - 1}. \tag{17}$$

Para el método de imputación por Hot Deck se tiene

$$\sum_{i=1}^{N_q} y_i^2 = \left(\sum_{i=1}^{N_q} y_i \cdot a_i + \sum_{i=2}^{N_q} y_{(i-1)} \cdot (1 - a_i) \right)^2 = \left(\sum_{i=1}^{N_q} y_i^2 \cdot a_i + \sum_{i=2}^{N_q} y_{(i-1)}^2 \cdot (1 - a_i) \right)$$

$$\begin{aligned} N_q \bar{y}^2 &= N_q \left[\frac{1}{N_q} \left(\sum_{i=1}^{N_q} y_i \cdot a_i + \sum_{i=2}^{N_q} y_{(i-1)} \cdot (1 - a_i) \right) \right]^2 \\ &= \frac{1}{N_q} \left[\left(\sum_{i=1}^{N_q} y_i \cdot a_i \right)^2 + 2 \left(\sum_{i=1}^{N_q} y_i \cdot a_i \right) \left(\sum_{i=2}^{N_q} y_{(i-1)} \cdot (1 - a_i) \right) + \right. \\ &\quad \left. + \left(\sum_{i=2}^{N_q} y_{(i-1)} \cdot (1 - a_i) \right)^2 \right] \\ &= \frac{1}{N_q} \left[\sum_{i=1}^{N_q} y_i^2 \cdot a_i + 2 \left(\sum_{i=1}^{N_q} y_i \cdot a_i \right) \left(\sum_{i=2}^{N_q} y_{(i-1)} \cdot (1 - a_i) \right) + \sum_{i=2}^{N_q} y_{(i-1)}^2 \cdot (1 - a_i) \right]. \end{aligned}$$

Sustituyendo en (17);

$$
\begin{aligned}
s^2 = & \frac{1}{(N_q-1)}\left\{\sum_{i=1}^{N_q} y_i^2\cdot a_i + 2 + \left(\sum_{i=2}^{N_q} y_{(i-1)}^2\cdot(1-a_i)\right) - \frac{1}{N_q}\left[\sum_{i=1}^{N_q} y_i^2\cdot a_i\ -\right.\right. \\
& \left.\left. -\ 2\left(\sum_{i=1}^{N_q} y_i\cdot a_i\right)\cdot\left(\sum_{i=2}^{N_q} y_{(i-1)}\cdot(1-a_i)\right) - \sum_{i=2}^{N_q} y_{(i-1)}^2\cdot(1-a_i)\right]\right\} \\
= & \frac{1}{N_q(N_q-1)}\left[N_q\sum_{i=1}^{N_q} y_i^2\cdot a_i + N_q\sum_{i=2}^{N_q} y_{(i-1)}^2\cdot(1-a_i) - \sum_{i=1}^{N_q} y_i^2\cdot a_i - 2\left(\sum_{i=1}^{N_q} y_i\cdot a_i\right)\cdot\right. \\
& \left.\cdot\left(\sum_{i=2}^{N_q} y_{(i-1)}\cdot(1-a_i) - \sum_{i=2}^{N_q} y_{(i-1)}^2\cdot(1-a_i)\right)\right] \\
= & \frac{1}{N_q(N_q-1)}\left[(N_q-1)\sum_{i=1}^{N_q} y_i^2\cdot a_i + (N_q-1)\sum_{i=2}^{N_q} y_{(i-1)}^2\cdot(1-a_i) - 2\left(\sum_{i=1}^{N_q} y_i\cdot a_i\right)\cdot\right. \\
& \left.\cdot\left(\sum_{i=1}^{N_q} y_{(i-1)}\cdot(1-a_i)\right)\right] \\
= & \frac{1}{N_q(N_q-1)}\left[(N_q-1)\left(\sum_{i=1}^{N_q} y_i^2\cdot a_i + \sum_{i=2}^{N_q} y_{(i-1)}^2\cdot(1-a_i)\right) - 2\left(\sum_{i=1}^{N_q} y_i\cdot a_i\right)\cdot\right. \\
& \left.\cdot\left(\sum_{i=1}^{N_q} y_{(i-1)}\cdot(1-a_i)\right)\right]. \qquad (18)
\end{aligned}
$$

Así se obtiene el estimador de la varianza de la variable para la técnica de imputación por Hot Deck secuencial dentro de cada nodo.

Esperanza del estimador de la varianza:

$$
\begin{aligned}
E(s^2) = E\left\{\frac{1}{N_q(N_q-1)}\left[(N_q-1)\left(\sum_{i=1}^{N_q} y_i^2\cdot a_i + \sum_{i=2}^{N_q} y_{(i-1)}^2\cdot(1-a_i)\right) - 2\left(\sum_{i=1}^{N_q} y_i\cdot a_i\right)\cdot\right.\right. \\
\left.\left.\cdot\left(\sum_{i=1}^{N_q} y_{(i-1)}\cdot(1-a_i)\right)\right]\right\}
\end{aligned}
$$

$$= \left\{ \frac{1}{N_q(N_q-1)} \left[(N_q-1)\left(E\left(\sum_{i=1}^{N_q} y_i^2 \cdot a_i \right) + E\left(\sum_{i=2}^{N_q} y_{(i-1)}^2 \cdot (1-a_i) \right) \right) - \right.$$

$$\left. - \; 2E\left(\sum_{i=1}^{N_q} y_i \cdot a_i \right) \cdot E\left(\sum_{i=1}^{N_q} y_{(i-1)} \cdot (1-a_i) \right) \right] \right\}. \tag{19}$$

Como:

$$E\left(\sum_{i=1}^{N_q} y_i^2 \cdot a_i \right) = N_q(\sigma^2+\mu^2)\varphi$$

$$= E\left(\sum_{i=2}^{N_q} y_{(i-1)}^2 \cdot (1-a_i) \right)$$

$$= (N_q-1)(\sigma^2+\mu^2)(1-\varphi),$$

y,

$$E\left(\sum_{i=1}^{N_q} y_i \cdot a_i \right) = N_q\mu\varphi E\left(\sum_{i=2}^{N_q} y_{(i-1)} \cdot (1-a_i) \right) = (N_q-1)\mu(1-\varphi),$$

sustituyendo en (19) se tiene:

$$E(s^2) = \frac{1}{N_q(N_q-1)} \left[N_q(N_q-1)(\sigma^2+\mu^2)\varphi + (N_q-1)^2(\sigma^2+\mu^2)(1-\varphi) - \right.$$

$$\left. - \; 2N_q\mu^2\varphi \cdot (N_q-1)(1-\varphi) \right]$$

$$E(s^2) = (\sigma^2+\mu^2)\varphi + \frac{(N_q-1)(1-\varphi)}{N_q}(\sigma^2+\mu^2) - 2\mu^2\varphi(1-\varphi). \tag{20}$$

Varianza del estimador de la varianza:

$$V(s^2) = V\left\{ \frac{1}{N_q(N_q-1)} \left[(N_q-1)\left(\sum_{i=1}^{N_q} y_i^2 \cdot a_i + \sum_{i=2}^{N_q} y_{(i-1)}^2 \cdot (1-a_i) \right) - \right.\right.$$

$$\left.\left. - \; 2\left(\sum_{i=1}^{N_q} y_i \cdot a_i \right)\left(\sum_{i=1}^{N_q} y_{(i-1)} \cdot (1-a_i) \right) \right] \right\}$$

$$= \frac{1}{N_q^2(N_q-1)^2} \left\{ E\left[\left((N_q-1)\left(\sum_{i=1}^{N_q} y_i^2 \cdot a_i + \sum_{i=2}^{N_q} y_{(i-1)}^2 \cdot (1-a_i) \right) - \right.\right.\right.$$

$$- \ 2\left(\sum_{i=1}^{N_q} y_i \cdot a_i\right) \cdot \left(\sum_{i=1}^{N_q} y_{(i-1)} \cdot (1-a_i)\right)\Bigg)^2\Bigg] - E^2\Bigg[(N_q-1)\Bigg(\sum_{i=1}^{N_q} y_i^2 \cdot a_i +$$

$$+ \ \sum_{i=2}^{N_q} y_{(i-1)}^2 \cdot (1-a_i)\Bigg) - 2\left(\sum_{i=1}^{N_q} y_i \cdot a_i\right)\left(\sum_{i=1}^{N_q} y_{(i-1)} \cdot (1-a_i)\right)\Bigg]\Bigg\}. \tag{21}$$

Se resuelve por separado cada uno de los factores de (21), es decir, el primer factor viene dado por la expresión:

$$E\Bigg\{(N_q-1)^2\left[\sum_{i=1}^{N_q} y_i^2 \cdot a_i + \sum_{i=2}^{N_q} y_{(i-1)}^2 \cdot (1-a_i)\right]^2 - 4\Bigg[(N_q-1)\Bigg[\sum_{i=1}^{N_q} y_i^2 \cdot a_i +$$

$$+ \ \sum_{i=2}^{N_q} y_{(i-1)}^2 \cdot (1-a_i)\Bigg]\left[\sum_{i=1}^{N_q} y_i a_i \sum_{i=2}^{N_q} y_{(i-1)} \cdot (1-a_i)\right] + 4\left(\sum_{i=1}^{N_q} y_i a_i\right)^2 \cdot$$

$$\cdot \ \left(\sum_{i=2}^{N_q} y_{(i-1)} \cdot (1-a_i)\right)^2\Bigg]\Bigg\} \tag{22}$$

De manera análoga se tiene que el segundo factor de (21) viene dado por la expresión:

$$E\left[(N_q-1)[N_q(\sigma^2+\mu^2)\varphi + (1-\varphi)(\sigma^2+\mu^2)] - 2N_q\mu^2\varphi(1-\varphi)\right]^2. \tag{23}$$

Sustituyendo (22) y (23) en (21) se obtiene:

$$V(s^2) = E\Bigg\{(N_q-1)^2\left[\sum_{i=1}^{N_q} y_i^2 \cdot a_i + \sum_{i=2}^{N_q} y_{(i-1)}^2 \cdot (1-a_i)\right]^2 - 4\Bigg[(N_q-1)\Bigg[\sum_{i=1}^{N_q} y_i^2 \cdot a_i +$$

$$+ \ \sum_{i=2}^{N_q} y_{(i-1)}^2 \cdot (1-a_i)\Bigg]\left[\sum_{i=1}^{N_q} y_i a_i \sum_{i=2}^{N_q} y_{(i-1)} \cdot (1-a_i)\right] + 4\left(\sum_{i=1}^{N_q} y_i a_i\right)^2 \cdot$$

$$\cdot \ \left(\sum_{i=2}^{N_q} y_{(i-1)} \cdot (1-a_i)\right)^2\Bigg]\Bigg\} - E\Big[\Big[(N_q-1)[N_q(\sigma^2+\mu^2)\varphi + (1-\varphi)(\sigma^2+\mu^2)] -$$

$$- \ 2N_q\mu^2\varphi(1-\varphi)\Big]^2\Big] \tag{24}$$

Al seguir desarrollando se obtendrá una expresión de cuarto orden, por tanto se limita a expresarse aqui.

Desarrollo empírico.

Para comprobar los resultados algebraicos obtenidos de los estimadores se hace uso de una base de datos del VI Censo Agrícola de Venezuela, en la cual en la investigación de Márquez et al. (2017), se realizó pérdidas artificiales de los datos, con pérdidas entre uno y cinco variables simultáneamente, así mismo, se construyeron árboles CART, específicamente árboles de regresión ya que las variables son cuantitativas.
De la base de datos, se tomó un 60 % de muestra de entrenamiento y 40 % de validación de los árboles. Dentro de la muestra de entrenamiento, el 20 % se selecciona para la poda del árbol. La validación se tomó como índices AIC (Akaike Information Criterion) y BIC (Bayesian Information Criterion).
Los criterios utilizados en la investigación son los mismos de Márquez et al. (2017), se toma como criterio 30 elementos mínimos a segmentar, máximo 10 niveles y 33 % como muestra de aprendizaje y 33 % de validación.

Comprobación de los supuestos fundamentales de Análisis de la Varianza.

Para poder establecer interpretaciones más efectivas, lo más recomendable es aplicar un ANOVA. Se probaron los supuestos de independencia de los errores, normalmente distribuidos con media cero y varianza constante, se requirió de transformación logarítmica en los datos.

ANOVA de comparaciones de data transformada (LOG) por variable según porcentaje de pérdida.

Se contrasta las medias entre los porcentajes de pérdida para comprobar si son significativamente iguales, o la metodología propuesta podría ser vulnerable ante cambios, principalmente aumentos en los porcentajes de pérdidas, todos los porcentajes de pérdida se comparan junto con la base de datos original.
$Ho: \ \mu_{BDO} = \mu_{5\,\%} = \mu_{10\,\%} = \mu_{20\,\%} = \mu_{30\,\%}$
H_1 : al menos uno de los porcentajes de pérdida difiere del resto.

Tabla 1: Significancia de los ANOVAS según el procentaje de pérdida. imputación Hot Deck secuencial

Variable	**Significancia**
Vacas paridas	.999
Vacas en ordeño	.998
Novillo	.983
Gallinas reproductoras	.948

Fuente: Elaboración Propia.

Tal como se aprecia en el cuadro 1 con un error del 5 % no hay diferencias significativas entre las medias de la base de datos original y la base de datos según cada porcentaje de pérdida para las 4 variables de estudio, ya que ninguna significancia es menor a 0.05. Lo cual indica que la metodología propuesta se mantiene robusta ante los porcentajes de pérdida de los datos, hasta un 30 %.

Análisis de la varianza entre los sesgos según pérdida.

Se contrasta la siguiente hipótesis.
$Ho: \ \mu_{s5\%} = \mu_{s10\%} = \mu_{s20\%} = \mu_{s30\%}$
H_1 : Hay diferencia en al menos uno de los porcentajes.

Tabla 2: Significancia de los ANOVA entre los sesgos según porcentajes de pérdida. Imputación Hot Deck Secuencial.

Variables	**Significancia**
Vacas paridas	.382
Vacas en ordeño	.497
Novillo	.658

Fuente: Elaboración Propia.

Matrices de varianza y covarianza para la comparación de las relaciones de las variables.

Se comparan las matrices de varianzas y covarianzas de las variables sometidas a imputación junto con la base de datos completa y de esta manera se estudia si

mantienen o no la variabilidad de los datos o si estamos ante la presencia de sobreestimación o subestimación de la varianza.

Prueba M de Box para la igualdad de matrices de covarianza.

Sean $n_1, n_2, \ldots, n_k$ muestras independientes de distribuciones normal multivariante. La hipótesis de igualdad de matrices de covarianza para $k = 2$ poblaciones multivariantes es;

$$H_0 : \Sigma_1 = \Sigma_2.$$

Comparando con un valor teórico para $k = 5$ poblaciones, $p = 5$ variables y $v = 268$ g.l., obtenemos los siguientes resultados:

Tabla 3: Valores T obtenidos y valores teóricos.

Valor Obtenido		**Valor Teórico**	
T2	1.7507	MT2	1.7420
T3	5.9572	MT3	5.9274
T4	7.1460	MT4	7.1103
T5	11.5811	MT5	11.5232

Buscando una aproximación a F, se rechaza si $MT > F_\alpha$.
Fuente: Elaboración Propia.

Como el valor calculado es mayor al valor teórico, se rechaza la hipótesis de igualdad de matrices de covarianzas, es decir, se rechaza $H_0 : \Sigma_0 = \Sigma_{tx}$, donde Σ_0 es la matriz de varianzas y covarianzas de la base de datos completa y Σ_{tx} es la matriz de varianzas y covarianzas de la base de datos con la técnica de imputación selecionada. Es decir, la técnica no preserva la relación entre las variables de manera simultánea.

Uso de los CART. Comparaciones.

Con el objeto de poder determinar la influencia del uso de los CART sobre la estimación de los datos faltantes, se evalúa comparando la base de datos completa con la base de datos imputada, usando la técnica de imputación Hot Deck Secuencial, en presencia y ausencia de los CART. Se usó un porcentaje de pérdida de 30 % que sería el peor de los caso, es decir, cuando hay mayor pérdida de información. Los resultados se aprecian en el cuadro 4.

Tabla 4: Medias y desviaciones estándar de las bases de datos imputadas con y sin CART

Variable	Vacas Paridas		Vacas en Ordeño		Novillos		Gallinas reproductoras	
Parámetro	Media	D.E	Media	D.E	Media	D.E	Media	D.E
BDC	1152.87	3239.31	1686.31	4207.22	926.91	2186.29	4136.02	22977.68
Sin CART	1069	2798	1742	4018	955.49	1996.69	4261	25824
CART	1134.69	3181.34	1635.47	4027.45	881.19	2039.19	4228.89	23195.59

BDC: Base de datos completa.
SIN CART: Base de datos imputada sin el uso de árboles de regresión.
CART: Base de datos imputada con el uso de árboles de regresión.
Fuente: Elaboración Propia.

Se lleva a cabo un ANOVA con la finalidad de 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.

$$H_0 : \mu_{BDC} = \mu_{sinCART} = \mu_{conCART}$$

$$H_1 : \text{Hay diferencias en al menos uno.}$$

Tabla 5: Significancia de ANOVAS entre bases de datos imputadas con y sin el uso de CART y la base de datos completa.

Variable	**Significancia**	**Resultado**
Vacas Paridas	.00	Difiere BD Hot Deck sin CART
Vacas en Ordeño	.001	Difiere BD Hot Deck sin CART
Novillos	.001	Difiere BD Hot Deck sin CART

Fuente: Elaboración Propia.

Por obtenerse valores de significancia inferiores a 0.05 conlleva a rechazar la hipótesis de igualdad de medias, es decir, si hay diferencias significativas entre las medias obtenidas en la base de datos completa, sin el CART (método clásico) y con el uso de CART (metodología propuesta). Dicho resultado conlleva a realizar un análisis posteriori lo cual arrojó diferencias significativas entre la base de datos completa y la metodología sin CART, y igualdad de medias entre la base de datos completa y la metodología propuesta, lo cual es un resultado satisfactorio que permite concluir que la metodología propuesta mejora la estimación de la media ante la metodología clásica.

4 Conclusiones

El diseño de la metodología propuesta, es decir, la segmentación a priori de los datos mediante al árbol CART y la metodología clásica de imputación Hot Deck Secuencial, permite obtener estimadores para la Media y Totales insesgados, obtenidos mediante la esperanza del estimador, observándose insesgamiento al evaluar las propiedades de los estimadores. La metodología propuesta se conserva robusta ante pérdida de datos entre el 5 y el 30 %, tanto en sus medias como en sus sesgos. Adicionalmente, la técnica mantiene la variabilidad de los datos y las relaciones entre las variables, al no encontrarse diferencias significativas en las comparaciones de las matrices de varianzas y covarianzas. Los ANOVA para las diferentes variables establecen que por medio de la segmentación las medias se mantienen iguales con respecto a la base de datos originales, mientras que sin el uso de la segmentación se presenta alteraciones significativas, indicado una ventaja de la metodología propuesta ante la metodología clásica de imputación Hot Deck.

5 Bibliografía

References

Andridge, R., y Little, R. (2010). A review of hot deck imputation for survey non-response. *International Statistical Review*, *1*(78), 40-64. doi: 10.1111/j.1751-5823.2010.00103.x

Bhargava, N., Dayma, S., Kumar, A., y Singh, P. (2017). An approach for classification using simple cart algorithm in weka. En *Proceedings of 2017 11th International Conference on Intelligent Systems and Control (ISCO)* (pp. 212–216). Coimbatore, India.

Escobar, M. (2007). *El análisis de segmentación: técnicas y aplicaciones de los árboles de clasificación*. Madrid: CIS.

Juster, F., y Smith, J. (1998). Enhancing the quality of data on income and wealth: recent developments in survey methodology. En *Proceedings of 25th General Conference of the International Association for Research in Income and Wealth.* Cambridge, England.

Lin, W. C., y Tsai, C. F. (2020). Missing value imputation: a review and analysis of the literature (2006–2017). *Artificial Intelligence Review*, *53*(2), 1487–1509. doi: 10.1007/s10462-019-09709-4

Little, R. J. A. (2019). On algorithmic and modeling approaches to imputation in

large data sets. *Statistica Sinica*.

Little, R. J. A., y Rubin, D. B. (2019). *Statistical analysis with missing data*. John Wiley & Sons.

Loh, W.-Y., Eltinge, J. L., Cho, M. J., y Li, Y. (2018). Classification and regression trees and forests for incomplete data from sample surveys. *Statistica Sinica*, *29*(2019), 431-453. doi: 10.5705/SS.202017.0225

Loh, W.-Y., Zhang, Q., Zhang, W., y Zhou, P. (2020). Missing data, imputation and regression trees. *Statistica Sinica*, *30*(2020), 1697-1722. doi: 10.5705/ss.202019.0122

Luo, Y., Cai, X., Zhang, Y., Xu, J., y xiaojie, Y. (2018). Multivariate time series imputation with generative adversarial networks. En S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, y R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 31). Curran Associates, Inc. Descargado de https://proceedings.neurips.cc/paper/2018/file/96b9bff013acedfb1d140579e2fbeb63-Paper.pdf

Medina, F., y Galván, M. (2007). *Imputación de datos: teoría y práctica*. Cepal. Descargado de http://hdl.handle.net/11362/4755

Montaquila, J. M., y Ponikowski, C. H. (1993). Comparison of methods to impute missing answers in a survey of establishments. En *Proceedings of the Survey Research Methods Section* (p. 446-451).

Márquez, V., Useche, L., Chacón, A. I., y Mesa, D. (2017). Estrategia de imputación con la media bajo el uso de árboles de regresión. *Comunicaciones en Estadística*, *10*(1), 9-40.

Olfaz, M., Tirink, C., y Önder, H. (2019). Use of cart and chaid algorithms in karayaka sheep breeding. *Kafkas Univ Vet Fak Derg*, *25*(1), 105-110. doi: 10.9775/kvfd.2018.20388

Pedersen, A. B., Mikkelsen, E. M., Cronin-Fenton, D., Kristensen, N. R., Pham, T. M., Pedersen, L., y Petersen, I. (2017). Missing data and multiple imputation in clinical epidemiological research. *Clinical Epidemiology*, *2017*(9), 157-166. doi: 10.2147/CLEP.S129785

Useche, L., y Mesa, D. (2006). Una introducción a la imputación de valores perdidos. *Terra. Nueva Etapa*, *XXII*(31), 127-151. Descargado de https://www.redalyc.org/articulo.oa?id=72103106

Useche, L., Parra, J. P., Mendoza, C. G., y Chacón, A. I. (2021). Design of an imputation methodology by random selection usign regression trees. *Bull. Comput. Appl. Math*, *9*(2), 97-121.