Adrián Quitiaquez. 2024. Revisión Teórica de la Imputación de datos con una aplicación en R. Revista Sigma, 20 (2). Páginas 35–42.

**REVISTA SIGMA**

Departamento de Matemáticas y Estadística

*Universidad de Nariño*

*Volumen XX Nº 2 (2024), páginas 35–42*

# Revisión Teórica de la Imputación de datos con una aplicación en R

Adrián Arturo Quitiaquez [1]

Abstract: This article makes a brief bibliographic review of the evolution and importance of data imputation methods and their applications in recent years, since the appearance of missing data is a common problem in surveys and research carried out in different areas of study, as most studies relate to methods of replacing missing values with estimated values to complete the data set and allow for a more robust statistical analysis on the structure of the data. In order that the imputation methods described in this work can be implemented and used more easily, codes are provided in the R programming languages.

*Keywords.* Data imputation, variables, bias, random data, methods.

Resumen: Este articulo hace una breve revisión bibliográfica de la evolución y la importancia de los métodos de imputación de datos y sus aplicaciones en los últimos años, ya que la aparición de datos faltantes es un problema común en las encuestas e investigaciones realizadas en los diferentes ámbitos de estudio, ya que la mayoría de los estudios se relacionan con los métodos de reemplazo de valores faltantes por valores estimados para completar el conjunto de datos y permitir un análisis estadístico más robusto sobre la estructura de los datos. Con el fin de que los métodos de imputación descritos en este trabajo se puedan implementar y usar con mayor facilidad se proporcionan códigos en los lenguajes de programación R.

*Palabras Clave.* Imputación de datos, variables, sesgo, datos aleatorios, métodos.

## 1. Introducción

La utilización de encuestas ha servido como un medio de recopilación de información para el estudio de diferentes temas lo que ha derivado problemas importantes en la aparición de valores perdidos o la falta de respuesta, es por ello que se puede utilizar una técnica para el

[1] Maestría en Estadística Aplicada, Facultad Ciencias Exactas y Naturales, Universidad de Nariño. Correo: aaquitiaquez@udenar.edu.co

tratamiento de estos datos perdidos que es denominada imputación de datos, esta técnica ha tenido un papel muy importante para el desarrollo en el ámbito estadístico y la investigación científica en los últimos años debido a que se han desarrollado procedimientos que tienen mejores propiedades estadísticas que las opciones tradicionales que fueron: la imputación media imputación de valores vecinos más cercanos con modelos predictivos más avanzados (Medina & Galvan, 2007), destacando que los datos perdidos son un problema significativo en las investigaciones, ya que implica una pérdida importante de información por lo que las estimaciones de parámetros pueden ser ineficientes, según (Horton & Lipsitz, 2001) se puede generar la pérdida de eficiencia, complicaciones en el análisis de datos faltantes, y además estimadores sesgados que ponen en riesgo la validez del proceso.

Las metodologías usadas hoy en día para sustituir los datos faltantes pueden ser procedimientos sin considerar sus fundamentos teóricos y limitaciones prácticas que generan estimadores sesgados que distorsionan las relaciones de causalidad entre variables, subestiman la varianza y alteran el valor de los coeficientes de correlación, según Rubín (1987) menciona que los procedimientos de imputación múltiple (IM) deben aplicarse de forma intensiva, pero no aclara que en la práctica no se satisfagan los supuestos por ende se debe tener en claro el tipo de variable a imputar si es continua, considerando el intervalo para el que se define, y si es cualitativa, tanto nominal como ordinal, las categorías de variables para la eficiencia del uso de una técnica de imputación.

Este articulo presenta una metodología de revisión teórica y la recolección de información de las técnicas más utilizadas de imputación de datos, además se proporcionan códigos de algunos paquetes de datos de R para la imputación de datos.

**¿Que son datos faltantes?**

Los datos faltantes, también conocidos como valores faltantes o valores perdidos, son aquellos registros dentro de un conjunto de datos que no tienen un valor o una observación asociada en una variable particular (Molenberghs, 2015) estos valores pueden ocurrir por errores humanos al ingresar la información recolectada en los sistemas informáticos, al trascribir datos de un formato a otro, como también fallas de equipos o sensores donde pueden presentar fallas de graduación o fallas mecánicas.

Estos datos faltantes se pueden observar en la base de datos como (NA), además existen varios tipos de datos faltantes, MCAR valores faltantes completamente aleatorio, donde un valor falte no está relacionada con ninguna variable del conjunto de datos ni con el valor real que se espera observar, MAR valores faltantes aleatorio, este valor puede estar relacionado con las variables observadas y NMAR valores no aleatorios los cuales estas relacionados directamente con la variable faltante. (Araneda, 2021)

## 2. Antecedentes históricos de métodos de imputación

En la búsqueda de los primeros aportes a la imputación de datos se encontró que Wilks en 1932 realizo aportes esenciales para el reemplazo de los valores faltantes por la media de los datos presentes de la variable, con el objetivo principal de reducir el sesgo y mejorar la precisión de los análisis estadísticos que se realizaban con un conjunto de datos completo (Castro et al., 2006), este método era utilizado cuando se presentaban pocos datos faltantes, ya que tienden a alterar la distribución de las variables.

De acuerdo a los avances tecnológicos y los estudios realizados para los años setenta se dice que la imputación de datos solo significaba identificar y sustituir los registros sin información

es decir la utilización de dos métodos hot-deck, y cold-deck (Medina & Galvan, 2007), donde el primero hace referencia a el remplazo o la duplicación de los valores faltantes con valores similares con base a fuentes de información que se hayan observado, este era utilizado para sustituir información en censos y encuestas lo que significa que no se necesitaba datos relevantes para estimar los valores individuales de las respuestas faltantes, mientras que el segundo método estaba enfocado en seleccionar valores externos relevantes (datos históricos) sin tener en cuenta las características del conjunto o población. (ESARA, 2016), sin embargo existieron autores importantes que aportaron grandes investigaciones a la imputación de datos como Rubín, Kalton y Kasprzyk, Little y Helmel.

Por otra parte el investigador Rubín en 1976 realizo una diferenciación de aspectos relevantes MAR (valores perdidos o faltantes de manera aleatoria) y MCAR (valores perdidos o faltantes de manera completamente aleatoria), para MAR significa que la pérdida de información de una variable no depende de ella sino de los valores observados en un conjunto de datos mientras que para MCAR los datos faltantes son completamente aleatorios y no están relacionados con ninguna variable observada o no observada en el conjunto de datos (Costas et al., 2012), mas adelante tomo el concepto de IM (imputacion multiple) una tecnica que pudo facilitar el remplazo de los datos faltantes por m¿1 con valores simulados lo que permitió hacer un uso eficiente de los datos, obtener estimadores no sesgados y reflejar adecuadamente la incertidumbre que la no-respuesta parcial introduce en la estimación de parámetros (Goicoechea, 2002) también hizo aportaciones con un enfoque llamado Bayesiano al utilizarlo se obtiene más precisión del intervalo de probabilidad para la variable a imputar (Medina & Galvan, 2007).

Debido a la presencia de un gran conjunto de datos a analizar Helmen en 1987 fue una de las personas quien se permitió manejar la falta de respuesta mediante un método llamado Listwise. Este método se usa cuando su conjunto de datos es grande y puede obtener un pequeño pero básico eliminando filas o columnas con datos faltantes, aunque este método no se recomienda para datos pequeños por la pérdida de información, hay otras investigaciones que buscaron cómo crear nuevos métodos o mejorarlos.

Desde el año 2000 se han implementado árboles de clasificación para mejorar los procedimientos de imputación de datos, pero la literatura considera que el desarrollo del método IM (imputación múltiple) ha solucionado el debate sobre la mejor forma de imputar datos omitidos debe considerar la distinción en los enfoques que existir y dependiendo del conjunto de datos a analizar. (Castro et al., 2006) Pero se acepta que, con o sin datos, el objetivo del análisis estadístico es generar una inferencia válida. No se trata solo de obtener estimadores insesgados de mínima varianza, ni de ajustar modelos para sustituir la información faltante. (Medina & Galvan, 2007).

# 3. Métodos de imputación

Varios estudios (Goicoechea, 2002; Platek, 1986; Estadística gubernamental;, 1996), indican que las técnicas de imputación se pueden clasificar de la siguiente manera:

1. **Técnicas fundamentadas en información externa**: Cuando son basadas en variables relacionadas con una encuesta perteneciente a otras bases de datos o reglas previas.

    - Métodos deductivos: Cuando los datos faltantes se deducen con cierto grado de certidumbre de otros registros completos del mismo caso.

2. **Técnicas determinísticas**: Cuando al repetir la imputación en varias unidades bajo las mismas condiciones, producirá las mismas respuestas.

   - **Imputación de la media o modo**: El dato faltante de cada variable se cambia con la media de los registros no faltantes en caso de variables cuantitativas, o con la moda en caso de variables cualitativas.
   - **Imputación de media de clases**: Las respuestas de cada variable son agrupadas en clases disjuntas con diferentes medias, y a cada registro faltante se le imputará con la media respectiva de su grupo.
   - **Imputación por regresión**: Se ajusta un modelo lineal que describa a y, variable a imputar, para un conjunto X de variables auxiliares que se deben disponer.
   - **Imputación por el vecino más cercano**: Se identifica la distancia entre la variable a imputar y, y cada una de las unidades restantes (x o variables auxiliares) mediante alguna medida de distancia, entonces se determina la unidad más cercana a y, usando el valor de esta unidad cercana para imputar el faltante.
   - **Algoritmo EM (Expectation Maximization)**: Basada en la función de máxima verosimilitud, permite obtener estimaciones máximo-verosímiles (MV) de los parámetros cuando hay datos incompletos con unas estructuras determinadas.

3. **Técnicas aleatorias o estocásticas**: Son aquellas que cuando se repite el método de imputación bajo las mismas condiciones para una unidad, producen resultados diferentes.

   - **Imputación aleatoria de un caso seleccionado**: Para cada caso con una celda faltante, se selecciona un donante aleatoriamente para ser asignado al dato faltante.
   - **Imputación secuencial Hot-Deck**: Cada caso es procesado secuencialmente. Si el primer registro tiene un dato faltante, este es reemplazado por un valor inicial para imputar, pudiendo ser obtenido de información externa, si el valor no está perdido, éste será al valor inicial y es usado para imputar el subsiguiente dato faltante.
   - **Imputación jerárquica Hot-Deck**: Similar al método secuencial anterior. En esta se organizan dentro de clases haciendo uso de variables auxiliares en forma de una estructura jerárquica.
   - **Imputación por regresión aleatoria**: Se hace primero un procedimiento de regresión, luego se adiciona un término residual para imputar los valores de y. Este término de error se obtiene de diferentes maneras, una de ellas es mediante los residuos del modelo de regresión, generado con registros completos, eligiendo uno aleatoriamente.
   - **Imputación por regresión logística**: similar a la técnica anterior, pero para imputar variables binarias.
   - **Imputación simple**: Es un método para completar los valores faltantes con valores estimados, se puede cambiar el valor faltante por la media, moda, mediana o al azar. El objetivo es emplear relaciones conocidas que puedan identificarse en los valores válidos del conjunto de datos para ayudar a estimar los valores faltantes.
   - **Imputación múltiple**: En lugar de imputar un único valor para cada valor faltante, la imputación múltiple genera múltiples conjuntos de datos completos, cada uno con valores faltantes imputados de manera diferente.

### 3.1. Como seleccionar un método de imputación de datos

La selección de un método adecuado para la imputación de datos puede depender de varios factores debido a que algunos métodos en comparación con otros pueden generar mayor precisión y valores verdaderos en los resultados, es decir que la elección de la técnica correcta puede ser adecuada para algunas variables, pero para otras no y será decisión del investigador seleccionar la técnica que menos afecte las estimaciones de las variables lo cual dependerá del tipo del conjunto de datos, tamaños del archivo, tipo de no respuesta, patrón de pérdida de respuesta, de los objetivos de la investigación, características específicas de la población, características generales de la organización del estudio y software disponible (Eltinge, 1996).

Antes de seleccionar un buen método de imputación de datos se debe tener en cuenta:

- La naturaleza de los datos: Comprender el contexto del problema o investigación, la distribución, la estructura o el tipo de variable a imputar, si son datos faltantes aleatoriamente al azar (MCAR), datos faltantes al azar (MAR) o no al azar (MNAR).
- Conocer los diferentes métodos de imputación: Algunas técnicas incluyen la imputación por media, mediana, moda, regresión, vecinos más cercanos, múltiples y otras técnicas más avanzadas como MICE (Ecuaciones de Estimación Múltiple) y algoritmos de aprendizaje automático que se pueden tener en cuenta dentro de la identificación de los parámetros que se desean estimar (Goicoechea, 2002).
- Determinar las tasas de no respuesta y exactitud necesaria: cuando el porcentaje de no respuesta es alto en una base de datos, que se considera que no hay confiabilidad en los resultados obtenidos, según Goicochea (2002) menciona que es bueno hacer uso de la información auxiliar disponible, ya que con ella se puede deducir información de los valores ausentes de una variable o permite hallar grupos homogéneos respecto a una variable auxiliar que se encuentre altamente correlacionada con la variable a imputar, y de esta manera encontrar un donante adecuado que sea similar al registro receptor.

## 4. Paquetes más útiles en de R

El conjunto de datos base propia de R contiene 150 observaciones de tres especies (Iris setosa, Iris virginica e Iris versicolor), con cinco variables, cuatro son cuantitativas y una variable tipo categórica.

**Tabla 1.***Resumen estadístico de la base de datos original.*

| | Sepal.Length | Sepal.Width | Petal.Length | Petal.Widt | Especie |
|---|---|---|---|---|---|
| Min | 4.2 | 2.000 | 1.000 | 0.100 | Setosa: 50 |
| 1st Qu | 5.1 | 2.800 | 1.600 | 0.300 | Versicolor: 50 |
| Median | 5.8 | 3.000 | 4.350 | 1.300 | Virginica: 50 |
| Mean | 5.843 | 3.057 | 3.758 | 1.199 | |
| 3rd Qu | 6.400 | 3.300 | 5.100 | 1.800 | |
| Max | 7.900 | 4.400 | 6.900 | 2.500 | |

Elaboración propia en programa R.

Resumen base de datos iris con datos faltantes con la función prodNA de R introducimos aleatoriamente un 10 % (0.1) aproximado de datos faltantes a la base original de iris

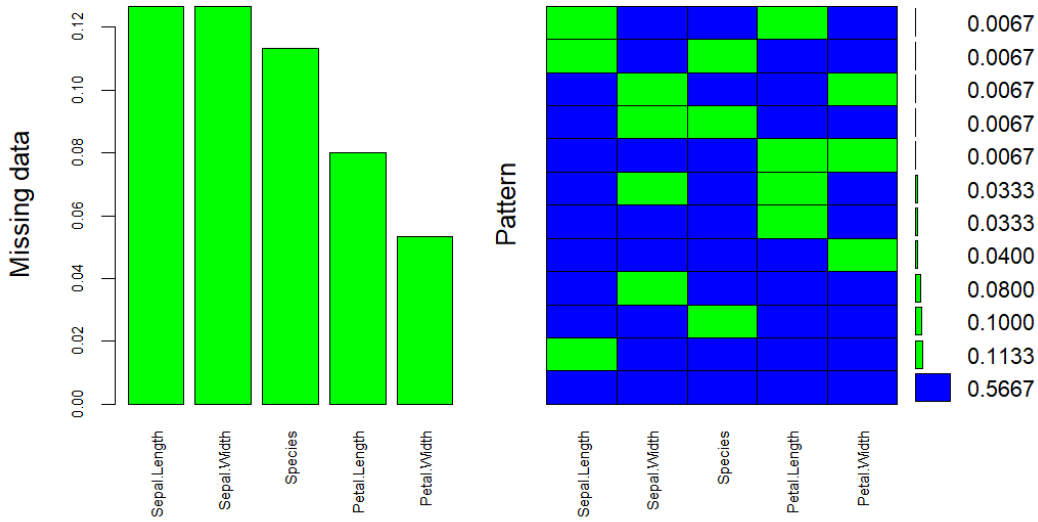
iris.mis<-prodNA(iris, noNA = 0.1)

**Tabla 2.***Resumen estadístico de la base con datos faltantes.*

| | SEPAL LENGTH | SEPAL WIDTH | PETAL LENGTH | PETAL WIDTH | SPECIES |
|---|---|---|---|---|---|
| MIN | 4.2 | 2.000 | 1.000 | 0.100 | Setosa: 50 |
| 1ST QU | 5.1 | 2.800 | 1.600 | 0.300 | Versicolor: 50 |
| MEDIAN | 5.8 | 3.000 | 4.350 | 1.300 | Virginica: 50 |
| MEAN | 5.843 | 3.057 | 3.758 | 1.199 | |
| 3RD QU | 6.400 | 3.300 | 5.100 | 1.800 | |
| MAX | 7.900 | 4.400 | 6.900 | 2.500 | |
| VALORES FALTANTES (NA'S) | 13 | 22 | 16 | 15 | 9 |

Elaboración propia en programa R.

**Gráfica 1.***Resumen da valores faltantes.*




Elaboración propia en programa R.

Los valores perdidos se identifican con el color verde donde aproximadamente el 56 % no tiene ningún valor faltante en las variables, la variable Petal Length tiene aproximadamente el 8 % de valores faltantes, la variable que presenta menos valores perdidos es Petal Width con un 5 %.

Imputación con el paquete mice implementa un método para tratar los datos que faltan el paquete crea imputaciones múltiples. Donde cada variable incompleta se imputa mediante un modelo separado. El algoritmo puede imputar mezclas de datos continuos, binarios, categóricos desordenados y ordenados.

**Gráfica 2.** *Códigos en R.*

```
install.packages("VIM")
library(VIM)

iris.mis <- subset(iris.mis,select = -c(Species))
summary(iris.mis)

md.pattern(iris.mis)

mice_plot <- aggr(iris.mis, col=c('blue','green'),
                  numbers=TRUE, sortVars=TRUE,
                  labels=names(iris.mis), cex.axis=.7,
                  gap=3,ylab=c("Missing data","Pattern"))


imputed_Data <- mice(iris.mis, m=5, maxit=50, method='mean', seed = 500)
summary(imputed_Data)
```

Elaboración propia en programa R.

Parámetros:

- m: Se refiere a 5 conjuntos de datos imputados
- matix: Numero de iteraciones tomadas para imputar los valores perdidos
- method: Método utilizado en la imputación (media, moda, emparejamiento, etc.)

## 5. Conclusiones

- Los avances de la imputación de datos a través de la historia han contribuido al desarrollo de nuevas maneras y métodos de afrontar los problemas de los valores faltantes, es por ello por lo que todos los métodos estudiados tienen sus limitaciones y su aplicación depende de cómo se comporten los datos faltantes. El número de valores faltantes o la magnitud de observaciones y patrones de las variables determinan el método de imputación, la eficacia de todas las metodologías se debilita aún en procedimientos estadísticamente robustos como imputación múltiple y de máxima verosimilitud.
- La imputación de datos es un proceso esencial en el análisis de datos que nos permite manejar valores faltantes de manera efectiva, preservar la integridad de nuestros datos y mejorar la precisión de nuestros análisis. Un manejo adecuado de los valores faltantes es fundamental para obtener conclusiones confiables y fundamentadas en nuestros datos.
- La técnica de imputación simple puede generar sesgos en los datos, especialmente si la imputación se realiza utilizando medidas de tendencia central como la media o la mediana, subestimando la variabilidad de los datos, ya que no se tiene en cuenta la incertidumbre asociada con la imputación.

## Referencias

[1] Araneda, P. (2021). *Missing data*. Obtenido de Pubs Rstudio: https://rpubs.com/paraneda/missingdata

[2] Castro, U., Lelly , M., Mesa, Á., & Dulce, M. (2006). Una introducción a la Imputación de Valores Perdidos. *Terra Nueva Etapa*, 12.

[3] Costas, C., Hernandez , J., & Ramirez, G. (2012). *Estimación de datos perdidos por máxima verosimilitud en patrones MAR y MCAR*. Oviedo-España: Psicothema. Obtenido de https://www.redalyc.org/pdf/727/72790117.pdf

[4] Eltinge, J. (1996). *Discusión de documentos de imputación*. Asocion americana de estadistica.

[5] ESARA. (2016). Informe de Imputacion de datos de la operacion Estadística. *Instituto Nacional de Estadistica y Sensos*, 4.

[6] Estadística gubernamental;. (1996). *Informe del Grupo de Trabajo sobre Imputacion*. Reino Unido.

[7] Goicoechea, A. (2002). Imputaciòn basada en arboles de clasificaciòn. *Eustat*, 8. Obtenido de https://www.eustat.eus/document/datos/ct_04_c.pdf

[8] Horton, N., & Lipsitz, S. (2001). *Imputación múltiple en la práctica: comparación de paquetes de software para modelos de regresiòn con variables faltante*. Asociación Americana de Estadistica.

[9] Medina, F., & Galvan, M. (2007). Imputacion de datos: Teoria y practica. *Naciones Unidas CEPAL: Division de Estadistica y Proyecciones economicas.*, 7.

[10] Molenberghs, G. (2015). Manual de metodología de datos faltantes. 5.

[11] Platek, k. (1986). Metodologia y tratamiento de la no respuesta. *Seminario Internacional de Estadística.*, 44-50.