Tecnura
ISSN: 0123-921X
ISSN: 2248-7638
Universidad Distrital Francisco José de Caldas

Romero Duque, Gustavo Andrés; González Prieto, Cristian Andrés;
Díaz Barriosnuevos, María Angélica; Rueda Menjura, Nataly Alejandra

Revisión y perspectivas para la construcción de bases de datos robustas con datos faltantes: caso aplicado a información financiera

Tecnura, vol. 27, núm. 75, 2023, Enero-Marzo, pp. 14-37
Universidad Distrital Francisco José de Caldas

DOI: https://doi.org/10.14483/22487638.18268

Disponible en: https://www.redalyc.org/articulo.oa?id=257074909002

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Tecnura

https://revistas.udistrital.edu.co/ojs/index.php/Tecnura/issue/view/1136

DOI: https://doi.org/10.14483/22487638.18268

INVESTIGACIÓN

Revisión y perspectivas para la construcción de bases de datos robustas con datos faltantes: caso aplicado a información financiera

Review and perspectives for the construction of robust databases with missing data: case applied to financial information

Gustavo Andrés Romero Duque [1], Cristian Andrés González Prieto [2], María Angélica Díaz Barriosnuevos [3], Nataly Alejandra Rueda Menjura [4]

Fecha de Recepción: 14 de julio de 2022 **Fecha de Aceptación**: 26 de septiembre de 2022

Cómo citar: Romero-Duque., G.A. González-Prieto., C.A. Día-Barriosnuevos., M.A y Rueda-Menjura., N.A. (2023). Revisión y perspectivas para la construcción de bases de datos robustas con datos faltantes: caso aplicado a información financiera. *Tecnura*, **27**(75), 14-37. https://doi.org/10.14483/22487638.18628

Resumen

Contexto: Se propone un conjunto de opciones que ayudan a determinar el método más adecuado para subsanar en bases de datos de tamaño apreciable, condiciones iniciales de datos faltantes y que serán utilizadas en procesos de investigación.
Metodología: El presente artículo aborda una propuesta para el desarrollo y manejo de bases de datos robustas como el caso de registros financieros, enfocándose desde el proceso *knowledge discovery in databases* (KDD).
Resultados: Se desarrolla y prueba una metodología utilizando tres técnicas de imputación en una base de datos construida a partir de 1 253 280 registros financieros de 2238 empresas y que representan siete años de su actividad económica en la localidad de Chapinero, en la ciudad de Bogotá D. C.
Conclusiones: Se realiza un comparativo de los métodos de imputación como factor determinante para la elección del método de imputación y consolidación de la base para su posterior uso.
Financiamiento: Fundación Universitaria Los Libertadores.

Palabras clave: base de datos, métodos de imputación, KDD, valores faltantes.

[1]Magíster en Ingeniería Industrial, ingeniero de Producción. Docente Fundación Universitaria Los Libertadores. Bogotá, Colombia. Email: garomerod@libertadores.edu.co
[2]Estadístico, magíster en Ciencias: Estadística. Docente Fundación Universitaria Los Libertadores. Bogotá, Colombia. Email: cagonzalezp01@libertadores.edu.co
[3]Ingeniera industrial. Fundación Universitaria Los Libertadores. Bogotá, Colombia. Email: madiazb02@libertadores.edu.co
[4]Ingeniera industrial. Fundación Universitaria Los Libertadores. Bogotá, Colombia. Email: naruedam@libertadores.edu.co

Abstract

Context: A set of options is proposed to help determine the most appropriate method to correct in databases of appreciable size, initial conditions of missing data and that will be used in research processes.
Methodology: This article addresses a proposal for the development and management of robust databases such as financial records, focusing from the Knowledge Discovery in Data bases (KDD) process.
Results: A methodology is developed and tested using three imputation techniques in a database built from 1,253,280 financial records of 2,238 companies that represent seven years of their economic activity in the town of Chapinero in the city of Bogotá D.C.
Conclusions: A comparison of the imputation methods is carried out as a determining factor for the choice of the imputation method and consolidation of the base for later use.
Financing: Fundación universitaria Los Libertadores

Keywords: database, imputation methods, KDD, missing values.

Tabla de contenidos

INTRODUCCIÓN

En una gran mayoría de estudios y en todos los campos del conocimiento se está constantemente recolectado datos que, con el tiempo, se convierten en grandes volúmenes de información. Para el manejo de estos, es esencial crear una base en la cual se pueda almacenar y, además, operar la información de forma adecuada para los fines pertinentes (Giraldo *et al.*, 2013).

Este artículo tiene como propósito proporcionar una metodología para quienes se encuentren construyendo bases de datos, principalmente aquellos que se enfoquen en aspectos financieros. En este sentido, se exponen algunas técnicas para hallar y suplementar los datos faltantes, y se suministra una secuencia para su desarrollo, a partir de la búsqueda de información y la recolección de los datos hasta la consolidación final de la base.

Al mismo tiempo, en esta propuesta se tiene como objetivo determinar, de forma empírica, la confiabilidad y precisión de tres técnicas de imputación para datos faltantes con comportamiento longitudinal para el estudio de un fenómeno en casos particulares. Considerando los diferentes caminos que el investigador pueda tomar, se busca responder a las siguientes preguntas: ¿Cómo definir si la base está incompleta? ¿Qué pasa cuando la base está incompleta? ¿Cuándo se deben eliminar o imputar datos? ¿Cómo elegir el método más adecuado de imputación para la base de datos?

Implicaciones teóricas y empíricas

El *knowledge discovery in databases* (KDD) es, según Timaran y Yépez (2012), un proceso automático, el cual combina el descubrimiento y el análisis dentro de una base de datos, y se centra en la extracción de patrones inferidos a partir de los datos para ser analizados por el interesado. La mayoría de los autores tienden a resaltar los siguientes pasos en la metodología KDD, aplicables a la creación de bases de datos:

1. *Recopilar e integrar datos.* Como lo menciona Detours *et al.* (2003), esta fase ayuda a los investigadores a formar una visión integral de los datos existentes necesarios y priorizar mejor los esfuerzos experimentales.

2. *Limpieza de datos.* Según Kim *et al.* (2003), los datos deben ser limpiados para reparar datos sucios, es decir, todos aquellos incorrectos o que no sean acordes al comportamiento estándar. Esto garantizará un análisis más preciso.

3. *Transformación de datos.* Para Lin (2002), la transformación de los datos o atributos es necesaria para el descubrimiento del conocimiento aplicando métodos matemáticos diferentes dependiendo el tipo de datos que de manejen.

4. *Reducción de datos.* Se suprime la información irrelevante en el estudio; según Booth *et al.* (2019), se descarta información de cualquier tipo antes de que esta sea evaluada o tomada en cuenta dentro de un estudio.

Por otro lado, en la creación de bases de datos, Brintha Rajakumari y Nalini (2014) mencionan *la agregación* como concepto valioso y de gran importancia en el diseño. En esta, los datos son conocidos como objetos y pueden ser modelados mediante el diseño de aplicaciones de bases de datos.

Cuando se están ejecutando análisis en bases de datos, su resultado dependerá en gran medida de la integridad y precisión con que estos cuenten (Witten *et al.*, 2016). Sin embargo, el análisis se puede topar datos faltantes y valores atípicos. Los datos de proceso con entradas faltantes, generalmente denominados *incorrectos* o *contaminados*, representan un gran desafío para la minería de datos y el monitoreo de ejercicios estadísticos (Imtiaz y Shah, 2008, Kadlec *et al.*, 2009), debido a su complejidad de manejo (Ge, 2018).

Los datos faltantes se precisan como valores no disponibles que serían útiles o significativos para el análisis de los resultados (Dagnino, 2014), lo cual podría afectar directamente los resultados en un ejercicio de análisis e investigación. Los datos faltantes se refieren a casos en los que hay una o más entradas de datos incompletas para las variables observadas en una base, lo que reduce la representatividad de las muestras de datos y puede dar lugar a una inferencia estadística inadecuada; este aspecto se abordará más adelante.

Es importante resaltar que, en la literatura, no se encuentra un criterio que muestre cuál es el método más adecuado y eficaz para generar datos faltantes sin que el resultado final de la investigación se vea gravemente afectado. Algunos autores proponen diferentes metodologías para generar lo descrito.

Medina y Galván (2007) comentan que, cuando esto sucede, existen procedimientos para sustituir la información, pero nunca una cifra imputada será mejor que una observada. También explican la diferencia entre la falta de respuesta total y la no respuesta parcial, donde no se obtiene respuesta en algunos ítems.

Con respecto del tratamiento para datos faltantes, Cañizares *et al.* (2003) dan una idea de cómo se ha intentado solucionar esta problemática a lo largo del tiempo:

> En los años setenta, la regla general era olvidarlos, por lo que su tratamiento consistía en la eliminación de la información incompleta. En los años ochenta se generalizó el tratamiento de los datos incompletos a través de la búsqueda de un valor que posteriormente sería asignado al dato faltante. En la década de los noventa se produjo un cambio en la filosofía del tratamiento de los datos incompletos: ya no importa buscar un valor, sino modelar la incertidumbre alrededor de él, y se comienzan a realizar las primeras imputaciones múltiples. (p. 59)

Existen trabajos que analizan el proceso tanto de la etapa de alistamiento como de modelado (Allison, 2002, Carpenter y Kenward, 2013, Enders, 2010, Graham, 2012, Kodamana *et al.*, 2018, Van Buuren *et al.*, 2006, Xu *et al.*, 2015) y en la literatura se han definido metodologías que permiten estos análisis. Las principales metodologías sobre minería de datos se pueden dividir en dos: a) minería robusta

para preprocesamiento de datos, y b) minería robusta para modelado estadístico. La primera se preocupa por tratar y limpiar los datos atípicos y faltantes, lo que se puede ejecutar con algunas técnicas tradicionales de minería de datos (Ge y Song, 2013) que también permiten apoyar el problema de normalización de datos, acción a considerar dentro de esta etapa. En la segunda, el análisis de datos comprende utilizar, entre otras, técnicas como análisis de componentes principales (ACP), modelos bayesianos, no lineales o dinámicos.

Algo a tener en consideración es que si bien tener una base de datos completa es ideal, se debe ser muy cuidadoso con el método de imputación a utilizar, pues, como mencionan Medina y Galván (2007), este es parte de la investigación que busca llegar a conclusiones sustentadas en evidencia empírica sólida; aplicar los métodos inapropiados traería más inconvenientes que soluciones.

Consideraciones antes de los métodos de imputación

Para romper las limitaciones de modelos a utilizar en completar los datos faltantes, primero deben tenerse en cuenta tres cuestiones específicas: la proporción de datos faltantes, sus patrones y sus mecanismos.

- *Proporción de datos faltantes.* Da un primer vistazo a los datos empíricos antes de tomar las mediciones válidas. Aunque no hay un criterio estricto, se sugiere que las tasas de faltante sean extremadamente bajas (menor al 5 %), ya que así no harán una interferencia significativa por inferencia. Sin embargo, si se cuenta con faltantes entre el 5 % y 10 %, se podrá trabajar teniendo presente que dará como resultado inferencias sesgadas significativas (Dong y Peng, 2013). Con más de 10 % de datos faltantes dentro de una data, es mejor eliminar algunas variables (Wood *et al.*, 2004, Peugh y Enders, 2004, Jelicic *et al.*, 2009), para así llegar al máximo faltante de 10 %.

- *Patrones de datos faltantes.* Hay dos patrones comunes de datos faltantes, a saber, el de tasa múltiple y el general. El primero se define cuando en la base faltan datos en diferentes niveles o variables, y el segundo se define cuando los valores faltantes son de un mismo nivel o variable.

- *Mecanismos de datos perdidos.* Proporcionan un marco probabilístico sobre las relaciones de datos perdidos. El saber por qué faltan datos es necesario para el diseño y la aplicación adecuada de los métodos de análisis estadístico (Graham, 2012). En la literatura, se pueden encontrar tres tipos comunes de faltas, asumiendo diferentes relaciones probabilísticas entre la parte faltante y la parte observada: a) fallar completamente al azar (MCAR); b) faltar al azar (MAR), y c) no faltar al azar (NMAR) (Schafer y Graham, 2002). El mecanismo MCAR supone que los datos faltantes deben ser independientes de la parte observada y la parte no observada. El MAR relaja el MCAR asumiendo que la parte faltante solo está relacionada con la parte observada y es ampliamente aceptada. El NMAR supone que los datos faltantes están relacionados tanto con la parte observada como con la parte faltante; debido a esto, apenas puede manejarse para

inferencia estadística. En consecuencia, con MCAR y MAR, los datos faltantes se pueden inferir de la parte observada.

La pregunta que surge entonces es ¿qué método es el más adecuado a usar? Y la respuesta dependerá del tipo de dato con que se cuente, ya que cada base tiene su propia estructura de variación que se podría ver afectada por la imputación utilizada. Los siguientes problemas deben resolverse (Sande (1982)): a) el de la edición y la imputación: búsqueda de la consistencia entre la información y las repuestas a imputar o editar; b) las distribuciones marginales y conjuntas de las respuestas son ciertamente diferentes para cada tipo de población, por lo que asumir normalidad no es una buena práctica, ¿qué hacer entonces?; c) identificación de los patrones de los campos faltantes; d) tiempo del que se dispone para la imputación; e) la estimación de muchos más parámetros (los datos faltantes) hace que los métodos se *esfuercen* más.

Algunas técnicas de imputación

Los acercamientos de Sande (1982) y Olinsky *et al.* (2003) definen un primer criterio sobre cómo completar bases dependiendo de la naturaleza de sus datos: a) aquellos que provienen de información correlacionada en el tiempo y en el espacio o b) los que provienen de información transversal como encuestas de satisfacción, de evaluación de productos, entre otras. Por su parte, García Reinoso (2015) expone que se pueden clasificar los métodos de imputación en tres categorías: a) determinísticos, referentes a un modelo matemático que produce una respuesta única (Useche y Mesa, 2016, Herrera *et al.*, 2017); b) estocásticos, que ofrecen una estimación probabilística para el dato imputado (Benítez y Álvarez, 2008, Ingsrisawang y Potawee, 2012), y c) los de inteligencia artificial, basados en modelos matemáticos complejos.

Complementando, a través del avance científico se han desarrollado varios métodos que se podrían clasificar en dos grupos (Liu *et al.*, 2020): la imputación simple, que tiene que ver con métodos que proporcionan un número para que se reemplace el espacio del dato faltante, y la imputación múltiple, que se basa en la incertidumbre de los datos y proporciona varios posibles valores simulados para el dato a imputar, los cuales pueden ser generados, como lo comentó Jarrett (1978), con un método estándar de mínimos cuadrados.

Son muchas las técnicas de imputación que han sido desarrolladas hasta la fecha, entre ellas, una de las primeras, es la propuesta por Wilks (1932), que busca reemplazar (pocos) datos faltantes con datos existentes en la data. En décadas posteriores, los adelantos computacionales permitieron la propuesta de técnicas de imputación más perfeccionadas.

Dentro de las propuestas para imputación de datos, se tienen las consideradas en diferentes momentos por: a) Rubin 1976, que distingue cuando los valores faltantes tienen o no relación con los existentes [MAR, MCAR]; b) en 1983, clasificados como enfoque basado en la aleatorización, y el enfoque bayesiano; c) Little y Rubin, en 1987 desarrollan la técnica de imputación múltiple, en la que, mediante valores simulados, se sustituyen los datos faltantes (Puerta Goicoechea, 2002).

Otras propuestas son las de Kalton y Kasprzyk (1982), quienes establecen las diferencias entre las técnicas de ajuste ponderado y las de imputación para los casos de (pocos) valores faltantes. Helmel *et al.* (1987) aportó el método *listwise*, que es usado con bases de datos de gran tamaño y que busca eliminar un bloque completo disminuyendo la data, pero teniendo una información completa. Todeschini (1990) propuso un k-vecino más cercano como método de estimación de valores perdidos, y Mesa *et al.* (2000) realizaron un estudio de imputación mediante el uso de árboles de clasificación, aunque se ha mostrado que sus resultados son muy pobres.

Otras investigaciones han buscado mejorar técnicas existentes de imputación, como las basadas en ACP (Gleason y Staelin, 1975), descomposición GH-Biplot (Vásquez, 1995), redes neuronales (Koikkalainen, 2002), análisis factorial (Geng y Li, 2013), entre otras (Useche y Mesa, 2016). A continuación, se desglosan otros métodos de imputación.

- *Sustitución media.* Considera la sustitución de los valores faltantes por el promedio de la variable. Para el caso de la imputación de procesos multimodo, la sustitución se toma del valor medio de la distribución dentro del modo. La sustitución media proporcionará estimaciones eficientes e imparciales para ubicaciones en aquellas situaciones cercanas a MCAR. Sin embargo, la sustitución media tiene efectos secundarios como las distorsiones de las variaciones y correlaciones. Por tanto, la sustitución media no es recomendada en la mayoría de los casos.

- *Sustitución en caliente.* Para preservar la distribución durante la imputación, la sustitución en caliente (*hot-deck*) reemplaza una entrada faltante a la vez con el valor disponible de un ítem similar en el mismo estudio. Al hacerlo, obtiene la mejor estimación de varianza en comparación con la contraparte de imputación media. De hecho, la sustitución en caliente es uno de los métodos más utilizado. Sin embargo, el problema surgirá para este enfoque cuando ocurran varios registros faltantes juntos en el archivo. La sustitución en caliente está diseñada para trabajar en sustituciones MAR.

- *Sustitución de regresión.* También conocida como *imputación media condicional*, intenta sumergir las entradas faltantes con una estimación de regresión de otras variables auxiliares correlacionadas. Este método, al igual que el de sustitución en caliente, está diseñado para trabajar en sustituciones MAR. A través de la sustitución, los valores imputados son tan buenos como el modelo de regresión utilizado para predecirlos. Por tanto, este método puede distorsionar los análisis de varianzas y correlaciones, ya que la regresión exagerará la fuerza de la relación de riesgo. Otro inconveniente es que a veces puede producir resultados improbables que pueden ser inválidos o del dominio razonable.

- *Sustitución basada en la distribución condicional.* Se imputa mediante el sorteo aleatorio de entradas faltantes de la distribución condicional de incertidumbres (Schafer y Graham, 2002). Para este tipo de imputación, se tiene que definir la distribución condicional explícita de la variable faltante dadas esas variables observadas para hacer una mejor sustitución. Posteriormente, esta

sustitución aliviará el problema de la distorsión de las distribuciones. Sin embargo, el principal problema es cómo inferir las distribuciones adecuadas con parámetros desconocidos. En algunos casos, la distribución puede ser bastante complicada, lo que hace que el método sea, engorroso.

- *Imputación con variables ficticias.* En esta metodología se crea una variable ficticia Z para estimar los datos faltantes, que puede asumir 0 o 1. Medina y Galván (2007) sustentan que al usar este método se generarían inconsistencias en la capacidad explicativa de los estimadores. Por ello, es pertinente evitar su ejecución, ya que pareciesen resolver la situación, pero generan sesgos al momento de ser interpretada.

- *Estimación por máxima verosímil.* Se asume que los datos faltantes siguen un esquema MAR y los valores son imputados mediante iteraciones. Medina y Galván (2007) explican que este algoritmo se aplica hasta lograr la convergencia, es decir, en cada iteración se anexará más información y el procedimiento terminará cuando los valores de la matriz de covarianza sean similares a los obtenidos en la iteración anterior.

- *Imputación múltiple.* Es un método de imputación relativamente moderno (Rubin, 2004), manejando los datos faltantes en tres pasos: a) imputa esos datos faltantes varias veces para generar varios conjuntos de datos completos; b) analiza cada conjunto de datos utilizando un procedimiento estadístico estándar; c) los resultados se combinan usando reglas simples para generar estimaciones, errores estándar y valores p que incorporan formalmente la incertidumbre de los datos faltantes.

  La imputación múltiple puede lograr una mejor imputación que otras técnicas. Sin embargo, el problema también es obvio, ya que se tiene que imputar varias veces para lograr buena inferencia estadística, que es computacionalmente intensiva. También debe tenerse en cuenta que el objetivo de la imputación es aliviar el deterioro de los valores faltantes a la inferencia estadística en lugar de recuperar los datos verdaderos.

METODOLOGÍA

El manejo y estudio de bases de datos se considera un insumo básico para la generación de conocimiento y solución de problemas. Por esto, es necesaria una metodología que ilustre una manera efectiva de construcción de aquellas, en especial, cuando estas cuentan con un gran volumen de datos, como los relacionados con registros financieros que usualmente suelen presentar altos índices de valores faltantes. Se proponen los siguientes pasos para su creación y manejo teniendo como punto de referencia la metodología KDD, como se puede apreciar en la figura **1**.

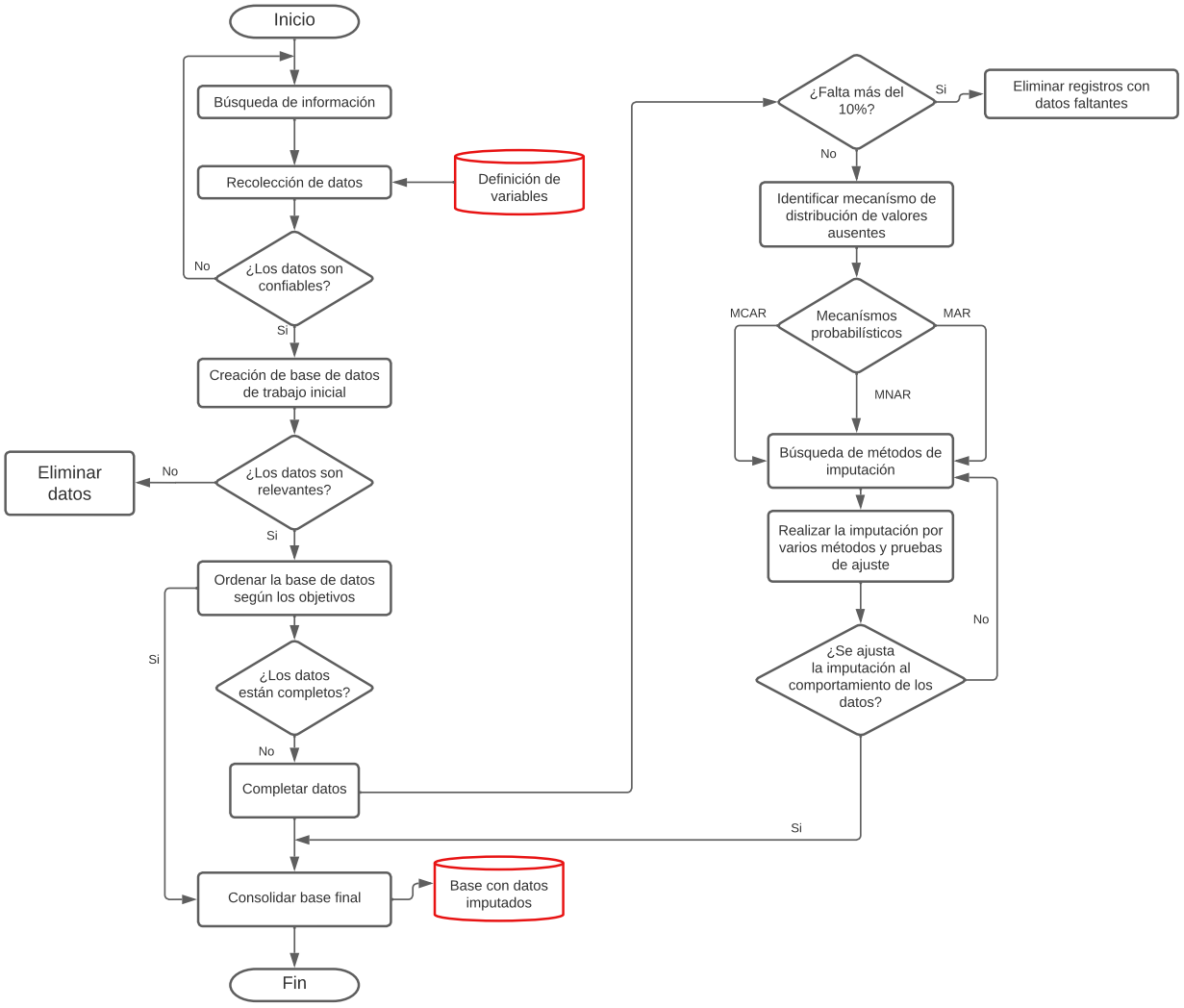

Figura 1. Metodología para la construcción de bases de datos

Búsqueda de información

En el caso de las bases de datos financieras, la búsqueda de información debe centrarse en su confiabilidad y veracidad: "Identificar las fuentes especializadas es el paso inicial para elegir la de mayor pertinencia al tema" (Moncada-Hernández, 2014, p. 111).

Además, se debe tener en cuenta los objetivos de la investigación, ya que así se sabrá dónde centrar la búsqueda de información; porque es en ellos donde se describe qué se quiere hacer con la información contenida en la base y cómo lograrlo (Manterola y Otzen, 2013).

Particularmente, en este trabajo se busca proporcionar una metodología para la creación y manejo de una base de datos, tomando como guía la construcción de una con información financiera centrada en empresas ubicadas en la localidad de Chapinero en Bogotá, que servirá para posteriores

investigaciones en esta zona. La fuente de la información fue la Superintendencia de Industria y Comercio, que en su plataforma SIREM (Superintendencia de Sociedades, 2020) registra la información financiera de las empresas que están sometidas a su inspección y vigilancia.

Recolección de datos

Es indispensable realizar la recolección de datos de una forma organizada y programada, teniendo en cuenta el objetivo de la investigación (Torres *et al.*, 2014), el cual será científicamente válido al estar soportado por información comprobable. En este paso se debe tener en cuenta el tipo de fuentes a manejar, ya sean estas primarias o secundarias: en las primarias, los datos proceden directamente de la población y para su recolección existen diferentes técnicas, algunas de ellas son: entrevistas, encuestas, experimentos y observación directa. En el caso de las fuentes secundarias, la recolección se realiza a partir de datos existentes recopilados por terceros, como es el caso del ejercicio práctico de este documento: desde entidades oficiales. Para ello, al utilizar estas fuentes, (Torres *et al.*, 2014, p. 3) recomiendan analizar cuatro preguntas básicas:

- ¿Es pertinente? Se adapta a los objetivos.
- ¿Es obsoleta? No es actual.
- ¿Es fidedigna? No es cuestionada.
- ¿Es digna de confianza? Ha sido obtenida con la metodología adecuada, con objetividad, naturaleza continuada y exactitud.

En este paso, en el caso dado, se recopilan todos los datos relevantes y confiables de la fuente externa SIREM (Superintendencia de Sociedades, 2020). Tales datos contienen información referente a las finanzas de aproximadamente 30 375 empresas en Colombia.

Adicionalmente, se utilizó otra fuente externa para la recolección de datos necesarios para el estudio, la plataforma Mapas Bogotá (Alcaldía de Bogotá, 2021), desde la cual se obtienen las coordenadas de cada una de las empresas.

Creación de la base de datos inicial

Se recopila toda la información proveniente de las fuentes externas en un solo formato, en el cual se tenga toda la información disponible. Para la recopilación de información proveniente de diversas fuentes, se debe primero definir y construir la estructura de la base de datos, definir los parámetros que permitan consignar la información relevante de acuerdo con el objetivo del estudio. Para la base de datos en cuestión, se recopila la información de los tres estados financieros más importantes para el análisis: estado de la situación financiera (Balance General), estado de resultados y flujo de efectivo;

se separaron los estados en tres archivos diferentes con la información financiera del periodo 2012-2018 de todas las empresas contenidas en la base SIREM.

Con la información disponible, se debe revisar su relevancia para el estudio, si no es así, se debe pasar a la etapa de limpieza de datos. El análisis se centraliza en la ciudad de Bogotá; se eliminan las empresas que (a) no se encuentran registradas en la ciudad capital (15 375); (b) aquellas que no se ubican en la localidad de Chapinero (10 700), para un total de 4300 empresas. Con los datos restantes, se debe realizar una limpieza de estos: "El objetivo es tener datos limpios, sin valores nulos o anómalos que permitan obtener patrones de calidad" (Timaran y Yépez (2012), p. 121).

Seguidamente, se procede a eliminar las empresas con datos atípicos o con un porcentaje de información faltante elevado (>10 %), que no tenían información ni registros en más de tres años, entre el 2012 y el 2018, debido la imposibilidad de adquirir esta información de fuentes externas, para un total de 2400 empresas, de las cuales, se determina eliminar de la data de aquellas empresas que contienen valores nulos, lo que arroja como resultado final total 2238 empresas.

Ordenar la base de datos según los objetivos

En este paso se establece el orden que se le quiere dar a la información, para que al momento de ser consultada se facilite su tratamiento, según los objetivos planteados por los investigadores. Para el caso en cuestión, se consideraron 80 atributos de las cuentas más relevantes de los tres estados financieros más importantes: balance general, estado de resultados y flujo de efectivo, para los años comprendidos entre 2012 y 2018.

Se procede a unificar las cuentas en un solo documento Excel, anexando la información de cada empresa de 2012 a 2018, lo que arroja como resultado total 2238 empresas con registros financieros bajo 80 atributos, para un total de 1 253 280 datos.

Completar datos

Al organizar la base de datos, puede surgir el inconveniente de *missing values*, o datos faltantes. Estos pueden tener varios orígenes, ya sea por error humano o problemas del programa que se utilice para manejar la base. Estos *missing values* pueden afectar los resultados del estudio de investigación y su posterior análisis. En un proceso investigativo, lo ideal es tener datos completos, pero si se encuentra con este inconveniente se debe tener las siguientes consideraciones:

Según Dagnino, 2014, se tienen tres alternativas cuando se cuenta con datos faltantes: a) omitir algunas variables. Como ya se mencionó, en el caso particular, se eliminaron algunas cuentas de los estados financieros que tenían poca relevancia para el estudio. b) Omitir los individuos, volviendo al caso, fueron eliminadas empresas con información faltante en más de tres de los siete años. c) Imputar los datos faltantes, los cuales se obtienen por diferentes metodologías, utilizando datos existentes.

Cabe mencionar que, para la imputación de valores faltantes, el porcentaje máximo de estos debe ser del 10 %; si estos valores exceden el porcentaje, se puede optar por cualquiera de las otras dos opciones mencionadas. Con énfasis en las alternativas, a partir del caso expuesto y la experiencia adquirida en su desarrollo, a continuación, se expone la metodología para la imputación de datos.

Búsqueda de métodos

Puerta Goicoechea, 2002 nombra cinco criterios para la elección del método: a) importancia de la variable a imputar, b) tipo de variable, c) estadísticos que se desean estimar, d) tasa de no respuesta y exactitud necesaria, e) información auxiliar disponible. En el caso expuesto se tienen los siguientes supuestos:

- Los datos tienen un comportamiento longitudinal y una correlación temporal: datos financieros de empresas a través del tiempo.
- La información es creciente, ya que esta al ser contable se ajusta por el valor del dinero en el tiempo.

Se opta por diferentes métodos de imputación, para un análisis comparativo y objetivo del método que proporcione mayor confiabilidad.

Imputación

Se procede a aplicar los métodos definidos en el paso anterior. Para el ejercicio y el caso dado se realizó la imputación de datos como se detalla a continuación:

- *Imputación simple.* Actualmente es aplicado por algunas entidades del Estado colombiano, como el Departamento Administrativo Nacional de Estadística, "entidad responsable de la planeación, levantamiento, procesamiento, análisis y difusión de las estadísticas oficiales de Colombia" (DANE, 2020). Consiste en completar la información faltante, haciendo uso de la existente. Se recomienda su aplicación en información del tipo de serie temporal para cada empresa de la base. Como paso inicial se debe calcular la variación entre los periodos. Dado que estos son anuales, la variación de la empresa i en el año t para la variable x se calcula utilizando la fórmula (1):

$$VA_i = \left(\frac{X_{it}}{x_{i(t-1)} - 1}\right) \cdot 100 \tag{1}$$

  Esta indicará la variación porcentual en los valores de las variables para las cuales se realice el cálculo. Se procede a realizar la imputación de la siguiente manera: si la información que se desea imputar corresponde al tiempo $t + 1$ se utiliza la fórmula (2):

$$X_{i(t+1)} = X_{it} \cdot \left(1 + \frac{VA_{it}}{100}\right) \tag{2}$$

Si la información a imputar corresponde al tiempo $t-1$ se utiliza la fórmula (3)

$$X_{i(t-1)} = X_{it} \cdot \left(1 + \frac{-VA_{it}}{100}\right) \tag{3}$$

- *Imputación suavizamiento.* Su intención es usar toda la información para ir corrigiendo el pronóstico del periodo siguiente, por lo que se trabajan dos tipos de información: a) un pronóstico realizado y b) la definición de demanda (anterior).

$$F_{t+1} = \alpha D_t + (1-\alpha)F_t \tag{4}$$

Donde, α corresponde a un *suavizador* que define el peso que se desea suministrar a la corrección del pronóstico y la técnica aplica un conjunto de ponderaciones decrecientes a todos los datos pasados (despliegue de F_{t+1}):

$$\begin{aligned} F_{t+1} &= \alpha D_t + (1-\alpha)F_t \\ F_{t+1} &= F_t - \alpha(F_t - D_t) \\ F_{t+1} &= F_t - \alpha(e_t) \\ F_t < D_t &\rightarrow e_t < 0 \end{aligned} \tag{5}$$

- *Imputación múltiple.* Es un método estadístico propuesto por Little y Rubin (2002) que permite completar datos faltantes a partir de la distribución de los valores conocidos de la variable. Tal generación se realiza de manera bayesiana donde los nuevos valores se estimarán de la distribución posterior de los datos, utilizando alguna distribución *a priori* no informativa. Ahora bien, los procesos computacionales podrían complicarse por la dificultad en las operaciones de integración que se deben ejecutar, por lo que (Little y Rubin, 2019, p. 214) han propuesto algunas alternativas que podrían facilitar ese proceso: *imputación múltiple impropia*, uso de una distribución posterior de una subbase, uso de la distribución asintótica del estimador vía máxima verosimilitud de la distribución, entre otros. La imputación mantiene la incertidumbre de los datos haciendo que, en cada iteración, los datos que se generan difieran, pero los originales permanecen intactos, lo que implica que existirán M versiones de las bases de datos completas.

Selección del método

Como lo mencionan Cañizares *et al.* (2003), la elección del método es una tarea dispendiosa, ya que un mismo método, dependiendo la situación, puede generar estimaciones precisas o no. Por ello, se aconseja considerar más de una opción para tratarlos y realizar un análisis de sensibilidad que facilite la elección del método a implementar.

Para la selección del método de imputación se tomó una muestra de la base principal, de empresas cuya información estaba completa (*sin missing values*) y se construyó una base de prueba. De esta

última fueron eliminados el 10 % de los datos al azar. Con este porcentaje se estaría dentro del criterio de decisión nombrado en el quinto paso de la construcción de bases de datos con valores faltantes. Adicionalmente, se tomaron solo las cuentas principales de cada estado financiero. Para el estado de la situación financiera en la base de prueba se tuvieron en cuenta el *total activo, total pasivo* y *total patrimonio*. En el caso del estado de resultados, se eligió la cuenta de *ingresos operaciones* y, por último, del flujo de efectivo se consideró la *utilidad del periodo*, como se detalla en la tabla 1.

Tabla 1. Información cuentas estados financieros

Año	Empresa	Total activos	Total pasivo	Patrimonio total	Ingresos operacionales	Utilidad del periodo
2012	Empresa 1	4.814.762	2.406.238	2.408.524		275.573
2013	Empresa 1	6.014.391	3.373.579	2.640.812	5.068.855	248.046
2014	Empresa 1	4.866.573	1.833.748		4.868.595	413.093
2015	Empresa 1	4.286.232	1.104.455	3.181.777	4.212.500	197.968
2016	Empresa 1		1.189.886	2.148.906	4.948.623	
2017	Empresa 1	5.687.179	3.098.735	2.588.444	4.948.623	138.952
2018	Empresa 1	7.032.359		2.806.865	6.639.704	592.421
2012	Empresa 2	6.009.183	3.073.443	2.935.740	472.097	21.305
2013	Empresa 2	8.634.809	5.035.621	3.599.188	7.519.957	203.448
2014	Empresa 2		2.440.624		777.469	344.047
2015	Empresa 2	11.331.690	875.164	10.456.526		332.578
2016	Empresa 2	17.277.080	5.840.722	11.436.358	1.379.318	825.591
2017	Empresa 2	17.884.263	6.583.533	11.300.730	1.379.318	
2018	Empresa 2	16.868.004	5.622.842	11.245.162	1.336.496	75.009

Luego, se realizó la imputación por los tres métodos elegidos previamente, se obtuvieron los nuevos registros para los valores *faltantes* y se procedió con la evaluación de la efectividad de cada método. Se compararon valores reales versus valores estimados en cada uno de los métodos. Posteriormente, se propuso la suma de diferencias de cuadrados y la suma de diferencias de desvíos entre la base de datos completa y la imputada, con el fin de determinar qué método se acercaba más a la realidad, teniendo en cuenta que este contenía las diferencias de menor magnitud (tablas 2, 3, 4, 5, 6 y 7).

Tabla 2. Sumas de diferencias de cuadrados imputación enfoque simple

Años	Total activos	Total pasivo	Patrimonio total	Ingresos operacionales	Utilidad del periodo
2012	2,31896E+14	77942030761	5,92378E+11	2,10935E+11	1,08945E+13
2013	1,19675E+12	7,28231E+12	1,44973E+12	3,55757E+12	3,02106E+12
2014	9,17863E+12	1,0821E+13	1,1202E+14	2,09258E+13	7,7902E+11
2015	5,34942E+13	577877647,3	1,3199E+14	2,10564E+12	5,99707E+12
2016	1,90362E+11	0	4,32023E+11	20085995920	2,86173E+12
2017	2,22592E+14	481848461,4	9,32223E+11	2,79348E+12	1,85265E+13
2018	0	0	2,30779E+13	1,6656E+13	0
Total general	5,18548E+14	1,81823E+13	2,70494E+14	4,62695E+13	4,20799E+13

Fuente: Elaboración de los autores

Tabla 3. Sumas de diferencias de cuadrados imputación múltiple

Años	Total activos	Total pasivo	Patrimonio total	Ingresos operacionales	Utilidad del periodo
2012	3,28566E+14	1,29913E+11	1,06622E+14	1,836E+13	1,84578E+11
2013	9,19493E+13	1,21351E+14	4,0429E+13	1,74135E+14	1,35908E+12
2014	1,29722E+14	4,41893E+14	1,34428E+15	9,35793E+14	3,22467E+12
2015	5,27001E+15	8,12311E+11	3,61766E+15	7,22822E+12	7,52124E+12
2016	1,30198E+11	0	1,65615E+12	1,02623E+14	3,00798E+12
2017	2,70198E+13	39400662016	8,25397E+12	4,65825E+14	9,11697E+12
2018	0	1,30757E+13	1,15287E+14	1,32484E+14	0
Total general	5,8474E+15	5,77302E+14	5,23419E+15	1,83645E+15	2,44145E+13

Fuente: Elaboración de los autores

Consolidación de la base de datos

La consolidación corresponde a la fase final de la construcción de bases de datos. En este punto, lo siguiente es aplicar el método de imputación elegido a la base real. Siempre y cuando sea pertinente

Tabla 4. Sumas de diferencias de cuadrados imputación por suavizamiento exponencial simple

Años	Total activos	Total pasivo	Patrimonio total	Ingresos operacionales	Utilidad del periodo
2012	1,06375E+14	1,07956E+12	3,84438E+11	1,50349E+12	2,57196E+12
2013	1,01541E+14	1,55424E+13	1,01306E+14	1,12726E+14	1,76992E+12
2014	3,45799E+12	2,18763E+14	7,21516E+13	1,71086E+13	4,42495E+11
2015	2,12529E+14	6,88528E+11	2,67109E+14	1,0934E+12	4,114E+12
2016	1,18656E+12	0	7,29183E+11	2,17893E+11	6,15716E+11
2017	1,55322E+13	693900964	3,2655E+12	36932516477	4,94295E+12
2018	0	1,72384E+12	2,28306E+13	1,20592E+13	0
Total general	4,40622E+14	2,37798E+14	4,67776E+14	1,44746E+14	1,4457E+13

Fuente: Elaboración de los autores

Tabla 5. Sumas desvíos absolutos imputación enfoque simple

Años	Total activos	Total pasivo	Patrimonio total	Ingresos operacionales	Utilidad del periodo
2012	16.098.720	279.181	769.661	568.384	3.300.683
2013	1.974.250	4.830.572	1.410.439	3.455.681	3.289.651
2014	5.380.898,9	4.353.543,9	18.637.634,14	6.287.465,03	1.358.644,29
2015	1 5211 399,3	24.039,086	13.021.164,2	1.532.140,94	3.408.136,48
2016	436.304,9	0	657.284,9	176.310,95	2.962.059,31
2017	23.929.089,14	21.951,047	965.516,75	1.675.484,49	6.196.932,32
2018	0	3.844.319,92	4.960.025,31	4.369.856,95	0
Total general	$ 63 030 662	$ 13 353 607	$ 40 421 725	$ 18 065 323	$ 20 516 106

Fuente: Elaboración de los autores

y se hayan considerado los criterios nombrados en este documento, con esto se obtendrán los nuevos registros para la base y así se complementará para los fines previstos de las partes interesadas.

Tabla 6. Sumas desvíos absolutos imputación múltiple

Años	Total activos	Total pasivo	Patrimonio total	Ingresos operacionales	Utilidad del periodo
2012	25.323.307	360.435	10.325.795	5.759.849	429.626
2013	15.769.304	18.039.272	7.200.356	18.820.005	2.048.473
2014	22.896.467	33.952.088	57.211.134	34.884.383	3.375.229
2015	106.535.097	901.283	96.590.384	2.927.137	4.609.698
2016	360.830	0	12 86 914	10.362.171	3.353.312
2017	6.801.983	198.496	2.872.973	30.511.369	3.560.348
2018	0	3.616.035	11.192.092	11.881.453	0
Total general	$ 177 686 988	$ 57 067 609	$ 186 679 648	$ 115 146 367	$ 17 376 686

Fuente: Elaboración de los autores.

Tabla 7. Sumas desvíos absolutos imputación suavizamiento exponencial simple

Años	Total activos	Total pasivo	Patrimonio total	Ingresos operacionales	Utilidad del periodo
2012	12.162.674	1.039.017	620.031	1.548.068	1.603.734
2013	14.788.592	7.765.903	12.124.084,0	12.956.121	2.855.564
2014	3.969.371	20.525.147	16.305.912	5.892.704	1.198.022
2015	18.402.427	829.776	18.963.373	1.123.166	2.821.226
2016	1.089.292	0	853.922	653.346	1.305.584
2017	6.734.167	26.342	1.807.070	271.777	2.343.499
2018	0	1.312.951	4.850.672	3.593.630	0
Total general	$ 57 146 523	$ 31 499 136	$ 55 525 064	$ 26 038 812	$ 12 127 629

Fuente: Elaboración de los autores.

Conclusiones

En la literatura correspondiente a bases de datos con faltantes, existe poca información que exprese el proceso de su reconstrucción con métodos de imputación para su consolidación final. La existente propone trabajar con la metodología *knowledge discovery in data bases* (KDD), pero la ma-

yoría de los textos destinados a la sustitución de datos faltantes no dan claridad de que método de imputación utilizar en casos particulares.

La metodología propuesta se determina para ser utilizada en campos variados de la investigación; se enmarcan las implicaciones teóricas necesarias y relevantes para orientar al investigador en el proceso; se presentan pautas para llevar a cabo cada uno de los pasos en adquirir conocimiento por medio de bases de datos. Por último, se proporcionan herramientas para determinar la factibilidad de los métodos de imputación a utilizar, en caso de ser necesarios para lograr una base de datos consolidada.

Adicionalmente, es necesario tener en cuenta que cuando la base está incompleta se deben considerar uno de tres caminos, dependiendo del estado de la base (Altman y Bland, 2007): a) omitir individuos: a realizarse cuando la información faltante de dicho individuo sea elevada en relación con la existente de los demás individuos; b) omitir variables: a realizarse cuando la mayor cantidad de registros faltantes se presenten en la misma variable; c) imputar: recomendable cuando se tenga hasta 10 % de datos faltantes. Los dos primeros caminos deben ser la última opción, ya que se puede sufrir una pérdida significativa de información. En contraste, las imputaciones de datos tienen el mérito comparativamente deseable sin sacrificio de datos.

Para el caso aquí desarrollado, se aplicaron tres métodos de imputación de datos (imputación simple, imputación por suavizamiento exponencial e imputación múltiple), a dichos métodos se les realizó pruebas de ajuste, suma de diferencias de cuadrados y suma de desvíos absolutos. Por medio de estas se concluyó que el método de imputación simple es el más adecuado para este caso, puesto que, al revisar los promedios arrojados en cada una de las cuentas, es el que presenta el valor más cercano a cero en comparación a los otros dos métodos sometidos.

Lo anterior puede deberse al comportamiento longitudinal de los datos y, por ende, a una correlación temporal, lo cual lleva a pensar que la imputación múltiple rompió la secuencia en el tiempo y por ello es más efectiva cuando los datos se comportan de manera transversal. En cambio, el método de imputación simple manejó la información de años anteriores de la misma empresa para predecir el nuevo registro, lo que permitió mantener la independencia entre cada una de las empresas.

Esta propuesta es una invitación a que, en entornos y procesos de investigación que manejen bases de datos robustas, se genere un mayor rigor en el tratamiento de aquellos, e igualmente que, como parte de dichos procesos, los investigadores participen en el desarrollo o aplicación nuevas metodologías o técnicas de tratamientos de bases de datos con faltantes; es un tema que todavía tiene mucho por ser desarrollado.

Financiamiento

Artículo de investigación científica derivado del proyecto de investigación “Economías de aglomeración y modelos de asociación en la localidad de Chapinero. Estudio mediante técnicas PCA y

AEDE", financiado por Fundación Universitaria Los Libertadores, año de inicio: 2021, año de finalización: 2021.

REFERENCIAS

[Alcaldía de Bogotá, 2021] Alcaldía de Bogotá. (7 de 10 de 2021). *Infraestructura de datos espaciales para el distrito capital*. https://www.ideca.gov.co/sobre-ideca/la-ide-de-bogota. ↑Ver página 23

[Allison, 2002] Allison, P. (2002). *Missing data*. Sage. https://doi.org/10.4135/9781412985079 ↑Ver página 17

[Altman y Bland, 2007] Altman, D. G. y Bland, J. M. (2007). Missing data. *British Medical Journal*, *334*(7590), 424. https://doi.org/10.1136/bmj.38977.682025.2C. ↑Ver página 31

[Benítez y Álvarez, 2008] Benítez, M. y Álvarez, M. (2008). Reconstrucción de series temporales en ciencias ambientales. *Revista Latinoamericana de Recursos Naturales*, *4*(3), 326-335. ↑Ver página 19

[Booth *et al.* (2019)] Booth, B. G., Keijsers, N. L. W., Sijbers, J. y Huysmans, T. (2019). An assessment of the information lost when applying data reduction techniques to dynamic plantar pressure measurements. *Journal of Biomechanics*, *87*, 161-166. https://doi.org/10.1016/j.jbiomech.2019.02.008. ↑Ver página 16

[Brintha Rajakumari y Nalini (2014)] Brintha Rajakumari, S. y Nalini, C. (2014). An efficient data mining dataset preparation using aggregation in relational database. *Indian Journal of Science and Technology*, *7*, 44-46. https://doi.org/10.17485/ijst/2014/v7iS5/50381. ↑Ver página 17

[Cañizares *et al.* (2003)] Cañizares, M., Barroso, I. y Alfonso, K. (2003). Datos incompletos: una mirada crítica para su manejo en estudios sanitarios. *Gaceta Sanitaria*, *18*(1), 58-63. https://doi.org/10.1016/s0213-9111(04)72000-2. ↑Ver página 17, 26

[Carpenter y Kenward, 2013] Carpenter, J. y Kenward, M. (2013). *Multiple imputation and its application*. Wiley. https://doi.org/10.1002/9781119942283 ↑Ver página 17

[Dagnino, 2014] Dagnino, J. (2014). Bioestadística y epidemiología. Datos faltantes (missing values). *Revista Chilena de Anestesia*, *43*(4), 332-334. https://doi.org/10.25237/revchilanestv43n02.03 ↑Ver página 17, 24

[DANE, 2020] Departamento Nacional de Estadística (DANE). (22 de 08 de 2020). *Estadísticas por tema*. https://www.dane.gov.co/index.php/estadisticas-por-tema. ↑Ver página 25

[Detours *et al.* (2003)] Detours, V., Dumont, J. E., Bersini, H. y Maenhaut, C. (2003). Integration and cross-validation of high-throughput gene expression data: Comparing heterogeneous data sets. *FEBS Letters, 546*(1), 98-102. https://doi.org/10.1016/S0014-5793(03)00522-2. ↑Ver página 16

[Dong y Peng, 2013] Dong, Y. y Peng, C. Y. J. (2013). Principled missing data methods for researchers. *SpringerPlus, 2*(1), 1-17. https://doi.org/10.1186/2193-1801-2-222. ↑Ver página 18

[Enders, 2010] Enders, C. (2010). *Applied missing data analysis*. Guilford Press. ↑Ver página 17

[García Reinoso (2015)] García Reinoso, P. L. (2015). Imputación de datos en series de precipitación diaria caso de estudio cuenca del río Quindío. *Ingeniare, 5*, 73-86. https://doi.org/10.18041/1909-2458/ingeniare.18.539. ↑Ver página 19

[Ge, 2018] Ge, Z. (2018). Process data analytics via probabilistic latent variable models: A tutorial review. *Industrial and Engineering Chemistry Research, 57*(38), 12646-12661. https://doi.org/10.1021/acs.iecr.8b02913. ↑Ver página 17

[Ge y Song, 2013] Ge, Z. y Song, J. (2013). Non-gaussian process monitoring. En *Multivariate statistical process control process monitoring methods and applications* (pp. 13-27). Springer. https://doi.org/10.1007/978-1-4471-4513-4. ↑Ver página 18

[Geng y Li, 2013] Geng, Z. y Li, K. (2003). Factorization of posteriors and partial imputation algorithm for graphical models with missing data. *Statistics and Probability Letters, 64*, 369-379. https://doi.org/10.1016/S0167-7152(03)00181-0 ↑Ver página 20

[Giraldo *et al.*, 2013] Giraldo, F., León, E. y Gómez, J. (2013). Caracterización de flujos de datos usando algoritmos de agrupamiento. *Tecnura, 17*(37), 153-166. https://doi.org/10.14483/udistrital.jour.tecnura.2013.3.a13 ↑Ver página 16

[Gleason y Staelin, 1975] Gleason, T. y Staelin, R. (1975). A proposal for handling missing data. *Psychometrika, 40*(2), 229-252. https://doi.org/10.1007/BF02291569 ↑Ver página 20

[Graham, 2012] Graham, J. (2012). *Missing data: Analysis and design*. Springer. https://doi.org/10.1007/978-1-4614-4018-5 ↑Ver página 17, 18

[Hemel *et al.*, 1987] Hemel, J., Van der Voet, H., Hindriks, F. R. y Van der Slik, W. (1987). Stepwise deletion: A technique for missing data handling in multivariate analysis. *Analytical Chemical Acta, 193*, 255-268. https://doi.org/10.1016/S0003-2670(00)86157-7 ↑Ver página

[Herrera *et al.*, 2017] Herrera, C., Campos, J. y Carrillo, F. (2017). Estimación de datos faltantes de precipitación por el método de regresión lineal: caso de estudio Cuenca Guadalupe, Baja California, México. *Redalyc, 25*(71), 34-44. https://doi.org/10.33064/iycuaa201771598 ↑Ver página 19

[Imtiaz y Shah, 2008] Imtiaz, S. A. y Shah, S. L. (2008). Treatment of missing values in process data analysis. *Canadian Journal of Chemical Engineering*, *86*(5), 838-858. https://doi.org/10.1002/cjce.20099. ↑Ver página 17

[Ingsrisawang y Potawee, 2012] Ingsrisawang, L. y Potawee, D. (2012). Multiple imputation for missing data in repeated measurements using MCMC and Copulas. *Proceedings of the Internacional Multiconference of Engineers and Computer Scientists*, *II*, 1-5. ↑Ver página 19

[Jarrett (1978)] Jarrett, R. G. (1978). The analysis of designed experiments with missing observations. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *27*(1), 38-46. https://www.jstor.org/stable/2346224. ↑Ver página 19

[Jelicic *et al.*, 2009] Jelicic, H., Phelps, E. y Lerner, R. (2009). Use of missing data methods in longitudinal studies: The persistence of bad practices in developmental psychology. *Developmental Psychology*, *45*(4), 1195-1199. 10.1037/a0015665. PMID: 19586189. https://doi.org/10.1037/a0015665 ↑Ver página 18

[Kadlec *et al.*, 2009] Kadlec, P., Gabrys, B. y Strandt, S. (2009). Data-driven soft sensors in the process industry. *Computers and Chemical Engineering*, *33*(4), 795-814. https://doi.org/10.1016/j.compchemeng.2008.12.012. ↑Ver página 17

[Kalton y Kasprzyk (1982)] Kalton, G. y Kasprzyk, D. (1982). *Imputing for Missing Survey Responses*. American Statistical Association. Proceeding of the Section on Survey Research Methods. ↑Ver página 20

[Kim *et al.* (2003)] Kim, W., Choi, B. J., Hong, E. K., Kim, S. K. y Lee, D. (2003). A taxonomy of dirty data. *Data Mining and Knowledge Discovery*, *7*(1), 81-99. https://doi.org/10.1023/A:1021564703268. ↑Ver página 16

[Kodamana *et al.*, 2018] Kodamana, H., Huang, B., Ranjan, R., Zhao, Y., Tan, R. y Sammaknejad, N. (2018). Approaches to robust process identification: A review and tutorial of probabilistic methods. *Journal of Process Control*, *66*, 68-83. https://doi.org/10.1016/j.jprocont.2018.02.011. ↑Ver página 17

[Koikkalainen, 2002] Koikkalainen, P. (2002). *Neural network for editing and imputation*. University of Jyvâskylâ. ↑Ver página 20

[Lin, 2002] Lin, T. Y. (2002). Attribute transformations for data mining I: Theoretical explorations. *International Journal of Intelligent Systems*, *17*(2), 213-222. https://doi.org/10.1002/int.10017. ↑Ver página

[Little y Rubin, 1987] Little, R. y Rubin, D. (1987). *Statistical analysis with missing data. Series in Probability and Mathematical Statistics*. John Wiley & Sons. ↑Ver página

[Little y Rubin (2002)] Little, R. J. A. y Rubin, D. B. (2002). *Statistical analysis with missing data*. Wiley & Sons. https://doi.org/10.1002/9781119013563 ↑Ver página 26

[Little y Rubin, 2019] Little, R. J. y Rubin, D. (2019). *Statistical analysis with missing data*. John Wiley & Sons. https://doi.org/10.1002/9781119482260 ↑Ver página 26

[Liu *et al.*, 2020] Liu, X., Wang, X., Zou, L., Xia, J. y Pang, W. (2020). Spatial imputation for air pollutants data sets via low rank matrix completion algorithm. *Environment International, 139*, 105713. https://doi.org/10.1016/j.envint.2020.105713. ↑Ver página 19

[Manterola y Otzen, 2013] Manterola, C. y Otzen, T. (2013). Por qué investigar y cómo conducir una investigación. *International Journal of Morphology, 31*(4), 1498-1504. https://doi.org/10.4067/S0717-95022013000400056. ↑Ver página 22

[Medina y Galván (2007)] Medina, F. y Galván, M. (2007). *Imputación de datos: teoría y práctica*. Serie "Estudios estadísticos y prospectivos". Comisión Económica para América Latina y el Caribe (Cepal). https://doi.org/978-92-1-323101-2. ↑Ver página 17, 18, 21

[Mesa *et al.* (2000)] Mesa, D., Tsai, P. y Chambers, R. (2000). *Using tree-based models for missing data imputation: An evaluation using Uk Census Data. Reporte técnico*. Universidad de Southampton. ↑Ver página 20

[Moncada-Hernández, 2014] Moncada-Hernández, S. (2014). Cómo realizar una búsqueda de información eficiente. Foco en estudiantes, profesores e investigadores en el área educativa. *Investigación en Educación Médica, 3*(10), 106-115. http://www.riem.facmed.unam.mx/index.php/riem/article/view/362. ↑Ver página 22

[Olinsky *et al.* (2003)] Olinsky, A., Chen, S. y Harlow, L. (2003). The comparative efficacy of imputation methods for missing data in structural equation modeling. *European Journal of Operational Research, 151*(1), 53-79. https://doi.org/10.1016/S0377-2217(02)00578-7. ↑Ver página 19

[Peugh y Enders, 2004] Peugh, J. y Enders, C. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research, 74*, 525e556. https://doi.org/10.3102/00346543074004525 ↑Ver página 18

[Puerta Goicoechea, 2002] Puerta Goicoechea, A. (2002). *Imputación basada en árboles de clasificación*. Eustat. ↑Ver página 19, 25

[Timaran y Yépez (2012)] Timaran, R. y Yépez, M. C. (2012). La minería de datos aplicada al descubrimiento de patrones de supervivencia en mujeres con cáncer invasivo de cuello uterino. *Universidad y Salud, 14*(2), 117-129. ↑Ver página 16, 24

[Rubin, 1976] Rubin D.B., (1976). Inference and missing data. *Biometrika, 63*, 581-592. https://doi.org/10.1093/biomet/63.3.581 ↑Ver página

[Rubin, 2004] Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys*. John Wiley & Sons. ↑Ver página 21

[Sande (1982)] Sande, I. G. (1982). Imputation in Surveys: Coping with reality. *The American Statistician*, *36*(3a), 145-152. https://doi.org/10.1080/00031305.1982.10482816. ↑Ver página 19

[Schafer y Graham, 2002] Schafer, J. L. y Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147-177. https://doi.org/10.1037/1082-989X.7.2.147. ↑Ver página 18, 20

[Superintendencia de Sociedades, 2020] Superintendencia de Sociedades. (08 de abril de 2020). *Asuntos económicos y societarios*. https://www.supersociedades.gov.co/delegatura_aec/Paginas/Base-completa-EF-2019.aspx. ↑Ver página 23

[Timarán-Pereira *et al.*, 2016] Timarán-Pereira, S. R., Hernández-Arteaga, I., Caicedo-Zambrano, S. J., Hidalgo-Troya, A. y Alvarado-Pérez, J. C. (2016). El proceso de descubrimiento de conocimiento en bases de datos. *Ingenierías*, *8*(26), 63-86. ↑Ver página

[Todeschini (1990)] Todeschini, R. (1990). Weighted k-nearest neighbour method for the calculation of missing values. *Chenometrics and Intelligent Laboratory Systems*, *9*, 201-205. https://doi.org/10.1016/0169-7439(90)80098-Q ↑Ver página 20

[Torres *et al.*, 2014] Torres, M., Paz, K. y Salazar, F. G. (2014). Métodos de recolección de datos para una investigación. *Boletín electrónico*, *3*, 1-21. http://bit.ly/2uhM4ot. ↑Ver página 23

[Useche y Mesa, 2016] Useche, L. y Mesa, D. (2006). Una introducción a la imputación de valores perdidos. *Terra Nueva Etapa*, *12*(31), 127-151. ↑Ver página 19, 20

[Van Buuren *et al.*, 2006] Van Buuren, S., Brand, J., Groothuis-Oudshoorn, C. y Rubin, D. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, *76*, 1049e1064. https://doi.org/10.1080/10629360600810434 ↑Ver página 17

[Vásquez, 1995] Vásquez, M. (1995). *Aportación al análisis biplot: un enfoque algebraico* [Tesis doctoral]. Universidad de Salamanca. ↑Ver página 20

[Wilks (1932)] Wilks, S. (1932): Moments and distributions of estimates of population parameters from fragmentary simple. *Annals of Mathematical Statistics*, *B*, 163-195. https://doi.org/10.1214/aoms/1177732885 ↑Ver página 19

[Witten *et al.*, 2016] Witten, I. H., Frank, E., Hall, M. A. y Pal, C. J. (2016). *Data mining: Practical machine learning tools and techniques*. 4.ª ed. Morgan Kaufmann. ↑Ver página 17

[Wood *et al.*, 2004] Wood, A., White, I. y Thompson, S. (2004). Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. *Clinical Trials*, *1*, 368e376. https://doi.org/10.1191/1740774504cn032oa ↑Ver página 18

[Xu *et al.*, 2015] Xu, S., Lu, B., Baldea, M., Edgar, T. F., Wojsznis, W., Blevins, T. y Nixon, M. (2015). Data cleaning in the process industries. *Reviews in Chemical Engineering*, *31*(5), 453-490. https://doi.org/10.1515/revce-2015-0022. ↑Ver página 17