UNIVERSIDAD DE EL SALVADOR

FACULTAD DE CIENCIAS NATURALES Y MATEMÁTICA

ESCUELA DE MATEMÁTICA

TESIS:

ANÁLISIS Y ALGORITMO DE SELECCIÓN DE TÉCNICAS DETERMINÍSTICAS Y ESTOCÁSTICAS DE IMPUTACIÓN DE DATOS

PRESENTADO POR:
BR. REINA EMPERATRIZ GARCÍA FLORES
BR. DANIEL DE JESÚS PALACIOS HERNÁNDEZ

PARA OPTAR AL GRADO DE:
LICENCIADO(A) EN ESTADÍSTICA

CIUDAD UNIVERSITARIA, Mayo de 2013

UNIVERSIDAD DE EL SALVADOR

FACULTAD DE CIENCIAS NATURALES Y MATEMÁTICA

ESCUELA DE MATEMÁTICA

TESIS:

ANÁLISIS Y ALGORITMO DE SELECCIÓN DE TÉCNICAS DETERMINÍSTICAS Y ESTOCÁSTICAS DE IMPUTACIÓN DE DATOS

PRESENTADO POR:
BR. REINA EMPERATRIZ GARCÍA FLORES
BR. DANIEL DE JESÚS PALACIOS HERNÁNDEZ

ASESOR:
DR. JOSÉ NERYS FUNES TORRES

CIUDAD UNIVERSITARIA, Mayo de 2013

AUTORIDADES

RECTOR UNIVERSITARIO:
ING.MARIO ROBERTO NIETO LOVO

SECRETARIA GENERAL:
DRA. ANA LETICIA ZAVALETA DE AMAYA

FISCAL GENERAL:
LIC. FRANCISCO CRUZ LETONA

FACULTAD DE CIENCIAS NATURALES Y MATEMÁTICA

DECANO:
MSC. MARTÍN ENRIQUE GUERRA CÁCERES

SECRETARIO:
LIC. CARLOS QUINTANILLA

ESCUELA DE MATEMÁTICA

DIRECTOR:
DR. JOSÉ NERYS FUNES TORRES

SECRETARIA:
MSC. ALBA IDALIA CÓRDOVA CUÉLLAR

CIUDAD UNIVERSITARIA, Mayo de 2013

UNIVERSIDAD DE EL SALVADOR
FACULTAD DE CIENCIAS NATURALES Y MATEMÁTICA
ESCUELA DE MATEMÁTICA

ASESOR:
DR. JOSÉ NERYS FUNES TORRES

CIUDAD UNIVERSITARIA, Mayo de 2013

Agradecimientos

A Dios y María Auxiliadora, por estar conmigo a cada instante, por fortalecer mi corazón e iluminar mi mente y por haber puesto en mi camino a aquellas personas que han sido mi soporte y compañía durante estos años de estudio.

A mi esposo Gilberto, por confiar siempre en mí, por animarme a seguir adelante, por todo tu apoyo y cariño de siempre, te amo, fuiste mi motor todos estos años.

A mi hija Alejandra, gracias por ser mi inspiración, mi razón de ser y mis deseos de seguir.

A mis padres Sonia y Alfredo, por su cariño y sus oraciones, que me acompañaron siempre especialmente en los momentos más difíciles, gracias porque siempre quisieron lo mejor para mí.

A mis segundos padres Yolanda y Gilberto, por el enorme esfuerzo de su trabajo en mi educación, por ser incondicionales siempre y porque soñaron con que fuera una profesional y dieron lo mejor de sí para que este sueño se volviera realidad.

A todos mis educadores y en especial a mi asesor Dr. Nerys Funes, gracias por iluminar mi camino con su enorme sabiduría.

A mis amigos y compañeros de toda la carrera en especial a Dany, gracias por todo lo que compartimos, siempre riéndonos en las buenas y en las malas.

A Sammy, mi amigo y compañero fiel en esas largas noches de desvelo.

En fin sería difícil expresar en estas líneas todo mi agradecimiento, pero dicen que cuando quieres algo, todo el universo conspira para que realices tu deseo, así que gracias a todos los que en estos años conspiraron a mi favor.

Con todo mi cariño, Reina.

Dedicatoria

A mi esposo y mi hija, que son la luz de mis ojos y mi razón de ser, porque gracias a ustedes nunca me di por vencida. Especialmente a mi hija, nunca olvides que... detrás de cada línea de llegada, hay una de partida. Detrás de cada logro, hay otro desafío. Sigue aunque todos esperen que abandones. Haz que en vez de lástima, te tengan respeto. Cuando por los años no puedas correr, trota. Cuando no puedas trotar, camina. Cuando no puedas caminar, usa el bastón. ¡Pero nunca te detengas!

Reina

A mi madre y mis hermanos incluyendo a Claudia, que han estado apoyándome a pesar de todo.

Daniel

Índice general

Introducción

En la mayoría de los estudios muestrales o censales, encontramos múltiples obstáculos y entre los más comunes se encuentra perder una medición, lo que genera espacios vacíos, en la estructura de los datos. De hecho, los datos completos constituyen más una excepción que la regla. Esta situación es una severa limitante, puesto que los métodos estadísticos tradicionales están diseñados para ser aplicados sobre conjuntos de datos completos y además las rutinas de los paquetes estadísticos también asumen que se trabaja con datos completos e incorporan opciones que no siempre son las más adecuadas para imputar observaciones sin que el usuario se dé cuenta de ello. Está ampliamente documentado que la aplicación de procedimientos inapropiados de sustitución de información introduce sesgos y reduce el poder explicativo de los métodos estadísticos, le resta eficiencia a la fase de inferencia y puede incluso invalidar las conclusiones del estudio.

Desde hace ya varias décadas, se ha venido estudiando la forma de "llenar" estos espacios vacíos, con el fin de obtener un conjunto de datos completos para analizarse por la vía de los métodos estadísticos tradicionales, en los últimos años gracias a los continuos avances de la informática se ha hecho posible el surgimiento y puesta en práctica de nuevas metodologías para el tratamiento de información con datos faltantes, los cuales, en su mayoría, producen resultados aceptables cuando hay pocos valores perdidos. Aún así, todavía son muchas las deficiencias que enfrentan las técnicas actuales, como los sesgos en las estimaciones, alteración de la relación entre las variables, cambios en las varianzas, entre otros y a pesar de la variedad de métodos existentes, el problema permanece abierto, sin que hasta ahora parezca haberse hallado una solución definitiva; además esta situación se complica cuando los datos se presentan en una matriz formada por diversas variables sobre la cual se realizan estudios multivariantes, haciéndose necesario la aplicación de métodos que convenientemente imputen conjuntamente los datos.

Los procedimientos de imputación que se utilizan con mayor frecuencia limitan o sobredimensionan el poder explicativo de los modelos (Acock 2005), y generan estima-

dores sesgados que distorsionan las relaciones de causalidad entre las variables, generan subestimación en la varianza y alteran el valor de los coeficientes de correlación. Por ello para lograr buenos resultados en procesos de limpieza de datos, la elección de la técnica es fundamental, pero hasta ahora no se conoce de alguna metodología que detalle la forma de realizar dicha selección de técnicas.

Entre los estadísticos de encuestas existe consenso de que la mejor forma de enfrentar la falta de respuesta es evitándola. No obstante, se reconoce que eliminarla es imposible, por lo que toda vez que esta se presenta existen procedimientos para sustituir información, pero bajo ninguna circunstancia es adecuado afirmar que una cifra imputada es mejor que el dato observado. Los métodos de imputación consisten en utilizar la información correlacionada con la variable que se desea imputar para generar la información faltante, el objetivo de la imputación no es generar información artificial, sino explotar exhaustivamente la información existente para obtener un conjunto de datos completos.

En este trabajo, se analizan los fundamentos teóricos de un conjunto amplio de métodos de imputación desde los métodos clásicos hasta los métodos robustos de imputación múltiple. En la primera parte se describe la teoría en la que se sustentan y la forma en que se aplican, haciendo énfasis en sus bondades y limitaciones, así como en los sesgos que se generan cuando se utilizan de manera acrítica, además se construye una guía metodológica que orienta al analista de los datos hacia una selección, con mayor rigor científico, de las técnicas adecuadas para aplicar a un conjunto de datos particular.

Luego se realiza una implementación práctica en la que se aplican algoritmos de imputación con el objetivo de sustituir valores omitidos en los datos de variables hidrogeoquímicas recolectadas por el Sistema Nacional de Estudios Territoriales (SNET), en una de las principales estaciones de monitoreo del volcán de San Salvador. Las muestras se tomaron de lagunas, manantiales o pozos cercanos al volcán, en este caso se hará uso de muestras tomadas en el manantial El Jabalí, situado en las faldas del volcán.

En dicha base de datos se observan registros faltantes, esto debido a que el monitoreo se ve suspendido cuando se presenta un fenómeno que requiere mayor atención por parte de los expertos, por lo que la falta de recursos humanos y materiales imposibilita la toma de las muestras de manera periódica, es por eso que se analizan y aplican las técnicas más adecuadas de imputación y de esta manera lograr completar la base de datos para luego poder aplicarles una técnica clásica de análisis multivariante que permita dar una explicación del comportamiento de dichas variables y por consiguiente conocer el comportamiento del fenómeno en estudio.

Justificación

Los datos faltantes en la recolección de información es un problema con el que se debe lidiar casi siempre en las investigaciones estadísticas, sin importar el área en la que se realice la recolección de información, siempre hay factores que dificultan y en ocasiones imposibilitan la toma de datos, sin mencionar que en la toma de los mismos es posible cometer errores que al momento de realizar un análisis exploratorio de los éstos obligue a su eliminación, lo que además constituye un punto en común es la búsqueda de los investigadores de mecanismos que les posibilite obtener datos completos; por lo que el contar con una técnica que permita completar estos espacios vacíos y obtener así un conjunto de datos al que se le pueda aplicar una técnica estadística clásica de análisis de datos es muy importante.

Pero además de conocer las técnicas es importante conocer las virtudes y debilidades de cada una y tener criterios estadísticos que permitan seleccionar dicha técnica para aplicarla de manera adecuada a un conjunto de datos en particular, de tal forma que los resultados sean estadísticamente válidos y permitan realizar inferencias certeras sobre los datos en estudio.

El presente trabajo es de vital importancia ya que constituye una guía metodológica que permite conocer con base en criterios estadísticos válidos las técnicas clásicas y robustas de imputación de datos y además cómo debe seleccionarse la más adecuada de acuerdo a sus ventajas y desventajas para un estudio en particular, permitiendo de esta manera resolver el problema de la falta de datos y constituye una solución práctica al crear un instrumento de selección de técnicas que orienta a investigadores de distintas áreas ante las diversas condiciones de patrones de pérdida de datos que presente su fenómeno en estudio.

Además contiene una implementación práctica que desarrolla la metodología de selección y la ejecución paso a paso de una técnica de imputación aplicada sobre un conjunto de datos de emanaciones Hidrogeoquímicas recolectadas en el volcán de San Salvador.

Planteamiento del Problema

En los estudios de diferente índole que se trabaja con grandes cantidades de información a menudo se encuentra el fenómeno de datos faltantes. En muchas ocasiones el recolector de datos con el analista no son la misma persona y esto imposibilita que en la recolección se evite datos faltantes, además en diferentes situaciones no es error sino imposibilidad por causas de naturaleza que no se puede efectuar esta recolección de datos, en casos como mal tiempo, en entrevistas que la pregunta incomode al entrevistado, etc.

En general las técnicas estadísticas para el análisis de datos requieren la información completa de una variable y muchas veces no se puede trabajar con datos faltantes ya que existe el riesgo de llegar a conclusiones equivocadas; debido a esta situación se vuelve relevante el conocer una metodología que permita ajustar los datos faltantes para generar un conjunto de datos completos que permita realizar análisis estadísticos posteriores de cualquier índole, pudiendo de esta manera dar respuesta a los objetivos que se plantean en cualquier investigación teniendo así la posibilidad de obtener el máximo aprovechamiento de los datos y una buena calidad en los resultados.

Se pretende en el trabajo de investigación:

- Estudiar y sistematizar las técnicas estocásticas de imputación de datos.
- Desarrollar una metodología para la selección de la mejor técnica estocástica de imputación de datos.
- Aplicar las técnicas estocásticas de imputación de datos para depurar y validar la base de datos del volcán de San Salvador, la cual contiene registros desde Julio del año 2000 hasta Febrero de 2010, para un conjunto de veintiún variables, de las cuales se analizarán únicamente aquellas que presentan una tasa de menor del 10 % de datos faltantes, éstas son: Temperatura (T), pH in situ, Dureza y Alcalinidad.

En dicha base de datos se observan registros faltantes, debido a que el monitoreo se

ve suspendido cuando se presenta un fenómeno que requiere mayor atención por parte de los expertos, por lo que la falta de recursos humanos y materiales imposibilita la toma de las muestras de manera periódica, además de la falta de equipo por desperfectos en los mismos al momento de tomar las muestras.

Por lo descrito anteriormente, la base de datos en estudio es idónea para realizar la aplicación del algoritmo para seleccionar la técnica de imputación adecuada siendo posible obtener una base de datos consistente y al abordar mediante una técnica estadística clásica que permita dar una explicación del comportamiento de las variables en estudio. Todo esto se realizará haciendo uso del software estadístico R versión 32/64 bit 2.15.2.

Objetivos

Objetivo general

Analizar y sistematizar las técnicas de imputación de datos, para luego elaborar un algoritmo de selección de la mejor técnica estocástica de imputación aplicada a un conjunto de datos en particular.

Objetivos específicos

- Definir y caracterizar las técnicas de imputación de datos, con el fin de conocer sus propiedades, funcionalidad, fortalezas y debilidades.
- Diseñar la guía metodológica para la selección de la técnica más adecuada.
- Realizar una aplicación práctica que sirva como ejemplo de la selección y aplicación de técnicas de imputación de datos.
- Mostrar que mediante la aplicación de la técnica adecuada de imputación de datos es posible obtener un conjunto de datos completo que puede ser analizado mediante una técnica estadística clásica

Capítulo 1

Análisis de Datos Faltantes y Técnicas de Imputación

1.1. Introducción

En los estudios de diferente índole que se trabaja con grandes cantidades de información a menudo se encuentra el fenómeno de datos faltantes. En muchas ocasiones el recolector de datos con el analista no son la misma persona y esto imposibilita que en la recolección se eviten los datos faltantes, además en diferentes situaciones no es error sino imposibilidad por causas de naturaleza incontrolable que no se puede efectuar esta recolección de datos, en casos como mal tiempo, en entrevistas que la pregunta incomode al entrevistado, etc.

En general las técnicas estadísticas para el análisis de datos requieren la información completa de una variable y muchas veces no se puede trabajar con datos faltantes ya que existe el riesgo de llegar a conclusiones equivocadas;*debido a esta situación se vuelve relevante el conocer una metodología que permita ajustar los datos faltantes para generar un conjunto de datos completos que permita realizar análisis estadísticos posteriores de cualquier índole*, pudiendo de esta manera dar respuesta a los objetivos que se plantean en cualquier investigación teniendo así la posibilidad de obtener el máximo aprovechamiento de los datos y una buena calidad en los resultados.

1.2. Tipos de Errores en las Investigaciones Estadísticas

Las investigaciones estadísticas pueden ser de diversos tipos; investigaciones por muestreo, investigaciones exhaustivas, experimentos comparativos y estudios observacionales. Todas estas investigaciones son estructuradas por diversas fases en las cuales pueden estar presentes fuentes de errores; exactitud es un término general que denota la ausencia de error de todo tipo, aunque ésta condición raramente la encontramos en los diferentes estudios. Las fuentes de errores presentes en las diferentes fases de una investigación estadística, pudieran producir dos tipos de errores; errores muestrales y errores no muestrales, o también llamados errores ajenos al muestreo.

Los tipos de errores muestrales son los producidos al observar una muestra de la población y no la totalidad de ella. Este tipo de error está compuesto por la variabilidad del estimador ante muestras repetidas y su sesgo, llamado sesgo técnico. (Mesa, 2004) y se pueden clasificar de la siguiente manera:

- **Error aleatorio:** es el producido por el sistema de realización de la medición. Desde el punto de vista estadístico, el error aleatorio también puede ser considerado como la variabilidad del muestreo. Aún cuando no esté involucrado un procedimiento de muestreo formal, como por ejemplo, una única medición de presión sanguínea en un sólo individuo. Este es un error constante que está presente en todas y cada una de las repeticiones que se efectúen. Su valor no afecta al valor real ni al promedio. Lo inverso del error aleatorio es la precisión, que es por lo tanto un atributo deseable de la medición y de la estimación.
- **Error sistemático:** aparece repetidamente debido al error del aparato o la poca destreza del experimentador. Por ejemplo, en una investigación donde se mide el peso de un objeto, el error sistemático es el producido por la medición de cada uno de dichos pesos. No es un error constante, es el error de redondeo que se lleva a cabo en cada una de las mediciones que se efectúan. Es lo que se conoce en estadística como sesgo.
- **Error accidental:** error por azar, cuando el experimentador comete puntualmente un fallo; con muchas medidas éstas se elimina.

Los errores no muestrales son aquellos errores presentes en una investigación, no atribuibles al observar una muestra. Pueden ser aleatorios o sistemáticos. Otra forma de clasificar los errores no muestrales es según su fuente; los cuales pueden ser:

- **Errores de cobertura:** son producidos por problemas en el marco muestral; por

la no inclusión de algunas unidades de observación, quedando excluidas del proceso de selección, y por ende tienen probabilidad nula de ser seleccionadas. Otro problema puede ser por exceso o duplicación, el cual ocurre cuando algunas unidades de observación aparecen más de una vez en el marco de muestreo.

- **Errores de respuesta:** ocurren cuando la información que se obtiene de la unidad de observación es incorrecta, estos errores se producen en la fase de recolección de datos.
- **Errores de falta de respuesta:** surgen cuando las unidades de observación seleccionadas para la encuesta no proporcionan todos los datos que deberían recogerse.

1.2.1. Tipos de No Respuesta o Falta de Datos

Se reconoce en general que la no respuesta es una importante medida de la calidad de los datos, cuando en una muestra aparecen valores perdidos por razones fuera del control del investigador, es necesario establecer unos supuestos sobre el proceso que los ha generado. Estos supuestos serán en general no verificables, y por ello deberán hacerse explícitos y deberá analizarse la sensibilidad del procedimiento general de estimación frente a desviaciones de los mismos, es por esta razón que se hace necesario distinguir la ausencia de datos.

Como punto de partida, es útil distinguir entre los patrones de datos perdidos y los mecanismos de datos faltantes. Estos términos en realidad tienen significados muy diferentes, pero los investigadores suelen utilizarlos indistintamente. Un patrón de datos faltantes se refiere a la configuración observada de los valores perdidos en un conjunto de datos, mientras que los mecanismos de falta de datos describen las posibles relaciones entre las variables medidas y la probabilidad de los datos que faltan.

Se debe tener en cuenta que un patrón de falta de datos se limita a describir la ubicación de los 'agujeros' en los datos y no explica por qué los datos faltan. En cambio los mecanismos de datos faltantes representan relaciones matemáticas genéricas entre los datos disponibles y los datos ausentes. Resulta útil distinguir el patrón de ausencia de datos y el mecanismo de ausencia de datos, debido a que algunos métodos de análisis dependen de estos patrones (Little R, Rubin D.B. 2000).

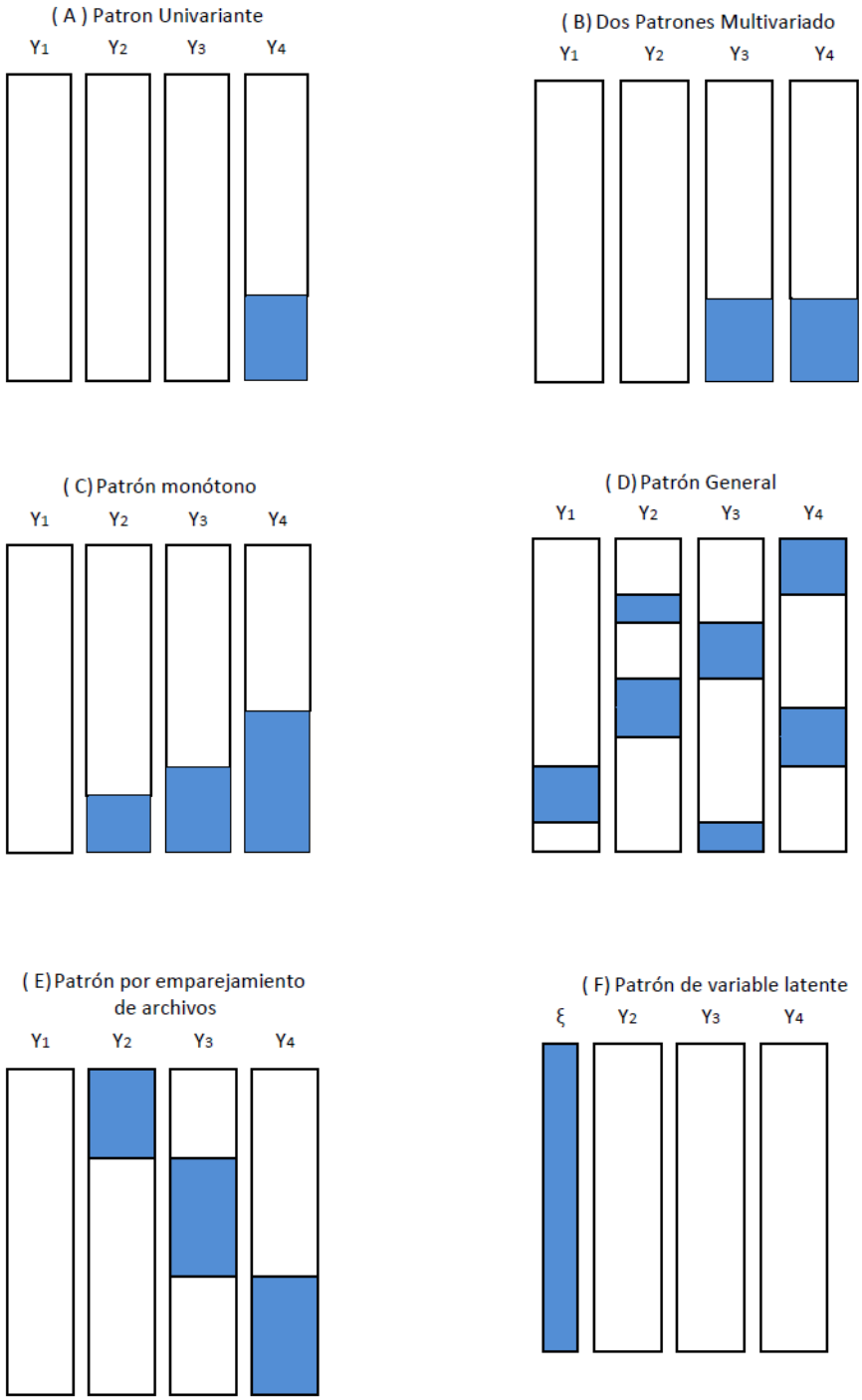

Figura 1.1: Prototipos de patrones de datos perdidos que puede encontrarse en una base de datos, las áreas sombreadas representan la ubicación de los valores que faltan en un conjunto de datos con cuatro variables.

En la **Figura 1.1** se observa el patrón univariante en el panel A que tiene valores perdidos aislados para una sola variable, los patrones univariantes es relativamente raros en algunas disciplinas, pero puede surgir en estudios experimentales.

El panel B muestra una configuración de valores perdidos conocido como un patrón de no respuesta único, este patrón ocurre a menudo en las investigaciones tipo encuesta, donde Y_1 e Y_2 son características que están disponibles para todos los miembro del marco de muestreo y Y_3 e Y_4 son las encuestas que algunos encuestados se nieguen a responder. En el panel C se observa un patrón de datos monótono faltantes esto se asocia típicamente con un estudio longitudinal donde los participantes abandonan y no vuelven nunca, por ejemplo en los estudios de supervivencia o ensayos clínicos. Un patrón general de datos perdidos es quizás la configuración más común, esto se observa en el panel D, un patrón general es el que posee valores perdidos dispersos a través de la matriz datos de una manera casual.

La falta de datos planeada en el patrón de datos faltantes en el panel E corresponde al cuestionario de tres formas de diseño descrito por Graham, Hofer, y MacKinnon (1996). Por ejemplo, el diseño en el panel E distribuye los cuatro cuestionarios a través de tres formas, de manera que cada formulario incluye Y_1 pero falta Y_2, Y_3, o Y_4. Finalmente, el patrón de una variable latente en el panel F es único para análisis de variables latentes tales como los modelos de ecuaciones estructurales. Este patrón es interesante porque los valores de la variable latente faltan para toda la muestra Históricamente, los investigadores han desarrollado técnicas analíticas que se ocupan de un determinado modelo de falta de datos. Por ejemplo, Little y Rubin (2002) dedican un capítulo entero a mayores métodos que fueron desarrollados específicamente para los estudios experimentales con un patrón de falta de datos univariado.

1.3. Mecanismos de Pérdida de datos

En el libro de Rubin(1987) se establece que la falta de datos consiste en dos conjuntos de parámetros: los parámetros que se consideran al inicio asumiendo que se obtendrán datos completos y los parámetros que describen la probabilidad de perder datos. Los investigadores rara vez saben por qué los datos faltan, por lo que es imposible de describir con certeza los mecanismos de pérdida.

El punto importante es que en general no hay manera de determinar o estimar los parámetros que describen la propensión a la falta de datos. Los parámetros que describen

la probabilidad de que los datos que faltan son un problema y no tienen valor sustantivo; sin embargo, en algunas situaciones, estos parámetros pueden influir en la estimación de los parámetros reales de los datos.

Para el análisis de los Mecanismos de pérdida de datos se considera, con carácter general, un vector aleatorio Y k-dimensional y sea M la matriz k-dimensional indicadora de ausencia de datos con $m_{ij} = 1$ si el valor es ausente y 0 si está presente, y define el patrón de ausencia de datos. Si denotamos mediante Y a una muestra multidimensional de Y podemos hacer una partición de forma que $Y = (Y_{obs}, Y_{aus})$, donde Y_{obs} y Y_{aus} denotan la parte observada y la no observada. Además ϕ representa el vector de parámetros desconocidos del mecanismo de no respuesta.

- Si: $f(M|Y, \phi) = f(M|\phi)$ para todo Y, ϕ MCAR (missing completely at random) Si la probabilidad de que el valor de una variable Y_j sea observado para un individuo i no depende ni del valor de esa variable, y_{ij}, ni del valor de las demás variables consideradas, $y_{ik}, k \neq j$.
- Si: $f(M|Y, \phi) = f(M|Y_{obs}, \phi)$ para todo Y_{aus}, ϕ MAR (missing at random) Si la probabilidad de que el valor de una variable Y_j sea observado para un individuo i no depende del valor de esa variable, y_{ij}, pero tal vez del que toma alguna otra variable observada y_{ik}, $k \neq j$.
- Si: el mecanismo de ausencia depende de y_i, NMAR (not missing at random) Si la probabilidad de que un valor y_{ij} sea observado depende del propio valor y_{ij}.

Uno de los puntos a considerar en el tratamiento de la no respuesta, es justamente el mecanismo de pérdida de los datos faltantes, ya que estos pueden influir en la selección del método de imputación. Los mecanismos de pérdida pueden ser ignorables o no ignorables, desde un punto de vista práctico.

Se presenta el caso de no respuesta ignorable cuando la probabilidad de que un hogar entrevistado no responda no depende de las características del hogar, en tanto que en la falta de respuesta no ignorable existe correlación entre la omisión de datos y las características de las unidades que no quisieron colaborar en la investigación.

Los mecanismos de pérdida pueden ser ignorables si ocurren de manera completamente aleatoria (MCAR, Missing Completely At Random) o de manera aleatoria (MAR, Missing At Random). El primer caso (MCAR), ocurre cuando la ausencia de información depende de alguna variable presente en la matriz de datos ya sea X o Y. Para el segundo caso (MAR), ocurre cuando la ausencia de los datos depende de variables presentes en la

matriz de datos, excluyendo la variable perdida.

Los patrones de pérdida no ignorables (NMAR) son los que ocurren cuando la ausencia de los datos depende de la variable perdida, esto traería como consecuencia estudiar el patrón de pérdida de los datos ausentes para luego imputar tomando en cuenta dicho patrón.

1.3.1. Cómo probar la existencia de un mecanismo de pérdida de datos en una matriz de datos

Se define como mecanismo de pérdida (proceso de no respuesta) al origen, causas, momento,relaciones, características, que producen la falta de información. Es importante tratar de establecer si las observaciones han sido perdidas al azar o su falta se asocia a causas definibles. Algunas veces el mecanismo está bajo el control del analista, otras no puede controlarlo pero sí comprenderlo y en muchos casos al no considerarlo explícitamente, se está suponiendo que el mecanismo es ignorable.

La idea de descubrir el mecanismo de pérdida de datos en una matriz de datos es sumamente compleja, por una parte están las matrices que poseen patrones de datos no ignorables, de los cuales el investigador posee información a priori para identificarlos pero en el caso de los mecanismos de pérdida de datos ignorables la identificación de este mecanismo no es tan evidente, solamente existe forma de identificar aquellos que poseen patrones de tipo MCAR pues este es el único mecanismo que produce proposiciones contrastables. Es claro que habiendo descartado al menos dos de los mecanismos de ausencia de datos el tercero es evidentemente la única opción verificable.

La prueba para definir el mecanismo de pérdida de información se realiza dependiendo si el conjunto de datos es un problema de tipo univariante o multivariante; en el primer caso la prueba es más sencilla pues se emplea una prueba t de Student, pero en el segundo los cálculos son más complejos y debe emplearse la prueba multivariante propuesta por Little(1988) que no es sino una extensión de la prueba t en la que se coparan simultáneamente las diferencias de medias en todas las variables en el conjunto de datos, a continuación se define cada una de las pruebas mencionadas anteriormente iniciando con el estudio del caso univariante.

1.3.1.1. Prueba t de Student para contrastar el mecanismo de pérdida de información (MCAR)

El método más simple para evaluar la existencia de un mecanismo del tipo MCAR es utilizar una prueba de tipo **t de student** para comparar los datos que faltan en subgrupos de datos (Dixon, 1988), esta prueba se emplea con una variante en la cual se hace uso de la prueba t de Welch que no asumen varianzas iguales entre los grupos en comparación, esta anotación será importante pues el estadístico de contraste con el que se trabajará, es una variación del estadístico usual de la t de Student empleado usualmente.

Este enfoque separa los datos faltantes y los datos completos para una variable en particular y utiliza una prueba t para examinar las diferencias de medias de grupo con otras variables en el conjunto de datos. El mecanismo MCAR implica que en promedio los casos con datos observados deben comportarse de la misma forma que los datos no observados; por consiguiente la no significancia en el resultado de la prueba t provee evidencia de que los datos son MCAR, mientras que una prueba t estadísticamente significativa sugiere que los datos son MAR o MNAR.

La hipótesis a probar es:

H_0: La media de los datos observados es igual a la media de los datos no observados en la variable de interés. *si esta hipótesis se rechaza entonces el mecanismo de pérdida de los datos es de tipo MCAR*

H_1: La media de los datos observados no es igual a la media de los datos no observados en la variable de interés. El estadístico de contraste empleado en la prueba es:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \tag{1.1}$$

Donde:
t = Estadístico equivalente a t de Student.
$\bar{X}_1$ = Media aritmética del grupo 1.
$\bar{X}_2$ =Media aritmética del grupo 2.
S_1^2 = Varianza del grupo 1.
S_2^2 = Varianza del grupo 2.
n_1 = Tamaño de la muestra del grupo 1.
n_2 = Tamaño de la muestra del grupo 2.

Ejemplo 1: Para mostrar el procedimiento de aplicación de la prueba t de Student, se analiza la pérdida de información en un estudio realizado para la selección de empleados donde contienen las variables de el coeficiente intelectual (IQ) y el rendimiento en el trabajo de un grupo de individuos, los datos se muestran en la tabla siguiente:

Tabla 1. Conjunto de datos de selección de Empleados

IQ	Rendimiento en el Trabajo	Prueba Psicológica
78	NA	13
84	NA	9
84	NA	10
85	NA	10
87	NA	NA
91	NA	3
92	NA	12
94	NA	3
94	NA	13
96	NA	NA
99	7	6
105	10	12
105	11	14
106	15	10
108	10	NA
112	10	10
113	12	14
115	14	14
118	16	12
134	12	11

El ejemplo se hace con la ayuda del software R versión 32/64 bit 2.15.2, para facilitar procesos, se utilizan para este fin dos librerías que son: *BaylorEdPsych* y *mvnmle* que se encuentran en http://cran.r-project.org/web/packages/available_packages_by_name.html, y las cargamos con:

```
> library(BaylorEdPsych)
> library(mvnmle)
```

Se cargan los datos de la tabla 1, que están incluidos en la librería *BaylorEdPsych*, y se utilizarán únicamente en el ejercicio la variable IQ y el rendimiento de trabajo:

```
> data(EndersTable1_1)
> ejemplo<-data.frame(EndersTable1_1$IQ,EndersTable1_1$JP)
```

El primer paso para la realización de la prueba consiste en separar los valores perdidos de los observados y calcular la media para cada uno de estos subgrupos en la variable IQ. Lo cual se logra con la siguiente sintaxis en R:

```
> prueba<-LittleMCAR(ejemplo)
> prueba$data$DataSet1
> prueba$data$DataSet2
```

y se visualiza la siguiente información:

```
> prueba$data$DataSet1
   EndersTable1_1.IQ EndersTable1_1.JP
11                99                 7
12               105                10
13               105                11
14               106                15
15               108                10
```

```
16                112                 10
17                113                 12
18                115                 14
19                118                 16
20                134                 12

> prueba$data$DataSet2
   EndersTable1_1.IQ EndersTable1_1.JP
1                 78                NA
2                 84                NA
3                 84                NA
4                 85                NA
5                 87                NA
6                 91                NA
7                 92                NA
8                 94                NA
9                 94                NA
10                96                NA
```

Se recuerda la hipótesis a contrastar la cual es:
$H_0 : \bar{X}_1 - \bar{X}_2 = 0$ y en este caso los datos son MCAR
$H_1 : \bar{X}_1 - \bar{X}_2 \neq 0$
En este caso es MAR o MNAR
Ahora se puede aplicar la prueba t de la siguiente forma:

```
> t.test(prueba$data$DataSet1[1],prueba$data$DataSet2[1])
```

y se obtiene la siguiente salida:

```
                Welch Two Sample t-test
data:  prueba$data$DataSet1[1] and prueba$data$DataSet2[1]
```

```
t = 6.4427,    df = 14.675,   p-value = 1.231e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  15.37614     30.62386
sample estimates:
mean of x     mean of y
    111.5       88.5
```

El test indica que la medias tienen diferencia estadísticamente significativa, t(14.68) = 6.44, con un p-valor< .001, por lo tanto no existe suficiente información en la muestra para sostener que los datos son MCAR, se rechaza H_0 y por lo tanto los datos se pueden suponer que son MCAR.

Un problema con el enfoque de prueba t es la posibilidad de tamaños de grupo muy pequeño, esto puede disminuir la potencia de la prueba y hacer que sea imposible realizar determinadas comparaciones.

Finalmente, es importante señalar que las comparaciones de medias no proporcionan una prueba concluyente de MCAR porque MAR y MNAR también pueden producir subgrupos de datos que faltan con medias iguales.

1.3.1.2. Prueba de Little MCAR

Little (1988) propuso una extensión multivariante del enfoque t-test que simultáneamente evalúa las diferencias de medias en cada variable del conjunto de datos. A diferencia de las pruebas t univariado, el procedimiento de Little es una prueba global de MCAR que se aplica al conjunto de datos. Al igual que el enfoque t-test, prueba de Little evalúa las diferencias de medias entre los subgrupos de casos que comparten el mismo patrón de datos perdidos. La estadística de prueba es una suma ponderada de las diferencias estandarizadas entre las medias de subgrupo y de una gran media global, el estadístico de contraste es el siguiente:

$$d^2 = \sum_{j=1}^{J} n_j \left(\hat{\mu}_j - \hat{\mu}_j^{(MI)}\right)^T \hat{\Sigma}_j^{-1} \left(\hat{\mu}_j - \hat{\mu}_j^{(MI)}\right) \tag{1.2}$$

Donde:
n_j : El número de casos en los datos perdidos en el patron j.

$\hat{\mu}_j$: Contiene la media de las variable para los casos de datos perdidos en el patrón j.
$\hat{\mu}_j^{(MI)}$: Contiene la estimación por máximo verosimilitud de la gran media calculada para el conjunto de datos con valores completos.
$\hat{\Sigma}_j$: Contiene las estimaciones máximo verosímiles de la matriz de varianzas y covarianzas.

Las Hipótesis a probar son: H_0 : Los datos son MCAR y H_1 : Los datos son MAR, d^2 se distribuye aproximadamente como el estadístico chi-cuadrado con $\sum k_j - k$ grados de libertad, donde k_j es el número de variables completas del patrón j, y k es el número total de variables.

Ejemplo: Para ilustrar la prueba de Little MCAR, se reconsidera la Tabla 1, los datos contienen 4 patrones de datos perdidos a continuación se detallan, se obtienen con el siguiente código en R:

```
> ejemplo1<-LittleMCAR(EndersTable1_1)
 #Patrones de datos
> ejemplo1$data
```

y se obtiene:

```
> ejemplo1$data
$DataSet1
    IQ JP WB
11  99  7  6
12 105 10 12
13 105 11 14
14 106 15 10
16 112 10 10
17 113 12 14
18 115 14 14
19 118 16 12
20 134 12 11
```

```
$DataSet2
  IQ JP WB
1 78 NA 13
2 84 NA  9
3 84 NA 10
4 85 NA 10
6 91 NA  3
7 92 NA 12
8 94 NA  3
9 94 NA 13

$DataSet3
    IQ JP WB
15 108 10 NA

$DataSet4
   IQ JP WB
5  87 NA NA
10 96 NA NA
```

El caso son solo IQ(DataSet4) se tiene un $n_j = 2$, con IQ y prueba psicológica(WB) (DataSet2)se tiene un $n_j = 8$, el caso de IQ y Rendimiento de Trabajo(JP, DataSet3) un $n_j = 1$ y caso del patrón con los datos completos(DataSet1) de las 3 variables $n_j = 9$. Ahora bien este paso se entenderá mejor hasta que se tenga una idea de un algoritmo de Máxima Verosimilitud que se verá en las técnicas de imputación en el apartado 1.3.2.8 Algoritmo EM(Expectation Maximization). Se necesita estimar los parámetros de la gran media $\hat{\mu}$ y la matriz de covarianzas $\hat{\Sigma}$ obtenidas con máxima verosimilitudes, en R se obtienen con los siguientes comandos:

```
>gmean <- mlest(EndersTable1_1)$muhat
>gmean
>gcov <- mlest(EndersTable1_1)$sigmahat
>gcov
```

y se obtiene la gran media:

$$\hat{\mu} = \begin{bmatrix} \hat{\mu}_{IQ} \\ \hat{\mu}_{JP} \\ \hat{\mu}_{WB} \end{bmatrix} = \begin{bmatrix} 100 \\ 10.23 \\ 10.27 \end{bmatrix}$$

y la $\hat{\Sigma}$:

$$\hat{\Sigma} = \begin{bmatrix} \hat{\sigma}^2_{IQ} & \hat{\sigma}^2_{IQ,JP} & \hat{\sigma}^2_{IQ,WB} \\ \hat{\sigma}^2_{JP,IQ} & \hat{\sigma}^2_{JP} & \hat{\sigma}^2_{JP,WB} \\ \hat{\sigma}^2_{WB,IQ} & \hat{\sigma}^2_{WB,JP} & \hat{\sigma}^2_{WB} \end{bmatrix} = \begin{bmatrix} 189.60 & 22.30 & 12.21 \\ 22.30 & 8.68 & 5.61 \\ 12.21 & 5.61 & 11.04 \end{bmatrix}$$

Para empezar, se considera el grupo que contienen los casos con los datos solo con IQ $n_j = 2$. Este patrón(DataSet4) tiene una media de IQ de 91.50 y la contribución al estadístico d^2 es el siguiente:

```
> mean(ejemplo1$data$DataSet4)
> d2_1=2*t(91.5-100.00)%*%solve(189.60)%*%(91.50-100)
> d2_1
```

que se obtiene: $d_1^2 = 2(91.50 - 100)(189.60^{-1})(91.50 - 100) = 0.76$

Ahora se considera el subgrupo de casos con los datos de IQ y prueba psicológica(WB) (DataSet2)se tiene un $n_j = 8$, las medias para este grupo son 87.75 y 9.13 respectivamente, y la contribución al estadístico d^2 es:

```
> mean(ejemplo1$data$DataSet2)
> d2_2=8*t(matrix(c(87.75,9.13),nrow=2)-matrix(c(100.00,10.27),nrow=2))%*%
solve(matrix(c(189.60,12.21,12.21,11.04),nrow=2))%*%(matrix(c(87.75,9.13),
nrow=2)-matrix(c(100.00,10.27),nrow=2))
> d2_2
```

que se obtiene:

$$d_2^2 = 8\left(\begin{bmatrix} 87.75 \\ 9.13 \end{bmatrix} - \begin{bmatrix} 100.00 \\ 10.27 \end{bmatrix}\right)^T \begin{bmatrix} 189.60 & 12.21 \\ 12.21 & 11.04 \end{bmatrix}^{-1} \left(\begin{bmatrix} 87.75 \\ 9.13 \end{bmatrix} - \begin{bmatrix} 100.00 \\ 10.27 \end{bmatrix}\right) = 6.43$$

Para el caso del subgrupo de casos con los datos del caso de IQ y Rendimiento de Trabajo(JP)(DataSet3) un $n_j = 1$, y la contribución al estadístico d^2 es:

```
> mean(ejemplo1$data$DataSet3)
> d2_3=1*t(matrix(c(108,10),nrow=2)-matrix(c(100.00,10.23),nrow=2))%*%
solve(matrix(c(189.60,22.31,22.31,8.68),nrow=2))%*%(matrix(c(108,10),
nrow=2)-matrix(c(100.00,10.23),nrow=2))
> d2_3
```

que se obtiene:

$$d_3^2 = 1\left(\begin{bmatrix} 108 \\ 10 \end{bmatrix} - \begin{bmatrix} 100.00 \\ 10.23 \end{bmatrix}\right)^T \begin{bmatrix} 189.60 & 22.31 \\ 22.31 & 8.68 \end{bmatrix}^{-1} \left(\begin{bmatrix} 108 \\ 10 \end{bmatrix} - \begin{bmatrix} 100.00 \\ 10.23 \end{bmatrix}\right)$$

$= 0.56$

Y en el último caso cuando se toman todas las variables(DataSet1) un $n_j = 9$, y la contribución al estadístico d^2 es:

```
>  mean(ejemplo1$data$DataSet1)
>  d2_4=9*t(matrix(c(111.89,11.89,11.44),nrow=3)-matrix(c(100.00,10.23,
10.27),nrow=3))%*%solve(matrix(c(189.60,22.31,12.21,22.31,8.68,6.50,12.21,
5.61,11.04),nrow=3))%*%(matrix(c(111.89,11.89,11.44),nrow=3)matrix(c(100.00,
10.23,10.27),nrow=3))
>  d2_4
```

que se obtiene:

$$d_4^2 = 9\left(\begin{bmatrix} 111.89 \\ 11.89 \\ 11.44 \end{bmatrix} - \begin{bmatrix} 100.00 \\ 10.23 \\ 10.27 \end{bmatrix}\right)^T \begin{bmatrix} 189.60 & 22.31 & 12.21 \\ 22.31 & 8.68 & 5.61 \\ 12.21 & 5.61 & 11.04 \end{bmatrix}^{-1} \left(\begin{bmatrix} 111.89 \\ 11.89 \\ 11.44 \end{bmatrix} - \begin{bmatrix} 100.00 \\ 10.23 \\ 10.27 \end{bmatrix}\right)$$

$= 6.87$

Entones el valor de d^2 es: $d^2 = 0.76 + 6.43 + 0.56 + 6.87 = 14.63$ simplificando los procesos se encuentra todo estos cálculos con:

```
> LittleMCAR(EndersTable1_1)
```

donde se obtiene parte de la salida:

```
> LittleMCAR(EndersTable1_1)
this could take a while$chi.square
[1] 14.63166

$df
[1] 5

$p.value
[1] 0.01205778

$missing.patterns
[1] 4
```

Este es similar al procedimiento realizado anteriormente paso a paso, ahora con el estadístico chi-cuadrado que se encuentra los grados de libertad sumando las variables en todos los patrones(DataSet1,DataSet2,DataSet3 y DataSet4) donde estaban completas $K_j = 8$ menos el número total de variables principales $k = 3$, por lo tanto $X^2_{5,0.95}$ nos proporciona un p-valor de menos de 0.01 con esto no existe evidencia para no poder rechazar la hipótesis H_0 y no lo consideraríamos MCAR.

Al igual que el enfoque de la prueba t, la prueba de Little tiene una serie de problemas a considerar. En primer lugar, la prueba no identifica las variables específicas que violan MCAR, por lo que sólo es útil para probar una hipótesis general de que es poco

probable que mantenga en el primer lugar. En segundo lugar, la versión de la prueba esbozada anteriormente supone que los patrones de datos faltantes comparten una matriz de covarianza común.

Mecanismos MAR y MNAR puede producir patrones de datos perdidos con diferentes variaciones y covarianzas, y el estadístico de prueba en la ecuación 1.4 no necesariamente detecta covarianza basada en desviaciones de patrón MCAR. En tercer lugar, los estudios de simulación sugieren que la prueba de Little sufre de baja potencia, en particular cuando el número de variables que violan MCAR es pequeña y la relación entre los datos observados y los datos perdidos es débil, o los datos son MNAR(Thoemmes & Enders, 2007). En consecuencia, la prueba tiene una tendencia a producir errores de Tipo II y puede llevar a una falsa sensación de seguridad sobre el mecanismo de datos faltantes. Finalmente, la comparación de medias no provee un test definitivo y concluyente de patrón MCAR, ya que el patrón MAR y MNAR pueden producir subgrupos de datos perdidos con medias iguales.

1.4. Técnicas de Imputación

Todos los investigadores necesitan depurar los datos que reciben a través de recolecciones de datos antes de proceder a extraer conclusiones, este proceso de depuración consiste en verificar si los valores de cada encuesta satisfacen un conjunto de reglas de consistencias, típicamente conocidas; en el caso que este supuesto no se cumpla el investigador esta ente un problema que se conoce como Edición e Imputación: "edición" es localizar los campos a modificar e "imputación" es determinar los nuevos valores para tales campos.

Hay una diferencia fundamental entre depuración previa e imputación. Consideremos el conjunto de todas las combinaciones posibles de códigos en un cuestionario, la depuración previa se puede definir como la división del conjunto en dos subconjuntos disjuntos. Las combinaciones que se consideran aceptables y las que se consideran inaceptables, las últimas contienen espacios en blanco no validos y entradas inconsistentes. Así, la depuración previa es básicamente un diagnóstico y operativamente se puede definir mediante un conjunto de reglas. Por otro lado, la imputación pertenece por naturaleza al tratamiento de datos y es el proceso de asignar valores a datos que falten produciendo así un conjunto de datos completo. No hay un método insesgado conocido de imputación pero algunos métodos son más adecuados que otros.

Es posible, en lugar de imputar la no-respuesta en el momento en que se preparan

las tabulaciones de la encuesta, presentar estas informaciones sobre el tamaño de la no-respuesta. En este caso los usuarios podrían elegir entre diversos métodos de imputación a partir de los datos tabulados.

La situación más sencilla se da cuando hay solo un valor que se puede imputar, en un campo de forma que después de la imputación el valor sea consistente. A este caso se le denomina imputación determinista. Por ejemplo, si la esposa aparece codificada como masculino solo hay un valor posible a imputar al sexo que sea consistente con el resto de la información. A veces hay más de un valor que lo hace consistente. Si es este el caso, se elegirá aquel valor particular que es más predominante en relación a la frecuencia total o más recomendable. Un ejemplo de este tipo se encuentra en la encuesta sobre mano de obra. Así, si una persona entre 15 y 16 años no ha rellenado la característica sobre su actividad laboral en los meses que van de otoño a primavera se le asigna como asistiendo a la escuela, aunque es posible que no asista a la escuela.

Mientras la proporción de tales casos sea pequeña, el efecto de esta imputación será un incremento pequeño en el sesgo pero habrá reducción en la varianza. En otras situaciones cuando se puede razonablemente imputar un intervalo de valores, necesitamos otros criterios. Uno sería el minimizar el error medio cuadrático de las estimaciones resultantes. La cuestión, es que no se sabe que error cuadrático medio hay que minimizar. Además no se conocen los diferentes agregados a los que unos datos pueden contribuir y sus diferentes formas de tabulación.

En otras palabras, ¿Cómo se puede predecir el mejor valor de un campo sobre la base de conocer los otros campos del conjunto? Un buen ejemplo de este tipo de imputación es el uso de los datos del mes previo en la encuesta sobre mano de obra: para una determinada persona, difícilmente se encontrará un valor imputado mejor, particularmente en aquellos casos en los que las características demográficas cambien lentamente. Si no disponemos de información pasada se tiene que recurrir a otros métodos de imputación.

1.4.1. Clasificación de las Técnicas de Imputación

Existen diversas clasificaciones para los distintos métodos de imputación pues cada uno de ellos se aplica para diferentes patrones de pérdida de respuesta, atendiendo a la clasificación de Goicoechea, 2002; Platek 1986; y Government Statistical Service 1996 y complementando con algunas otras técnicas estudiadas se presenta la siguiente clasificación:

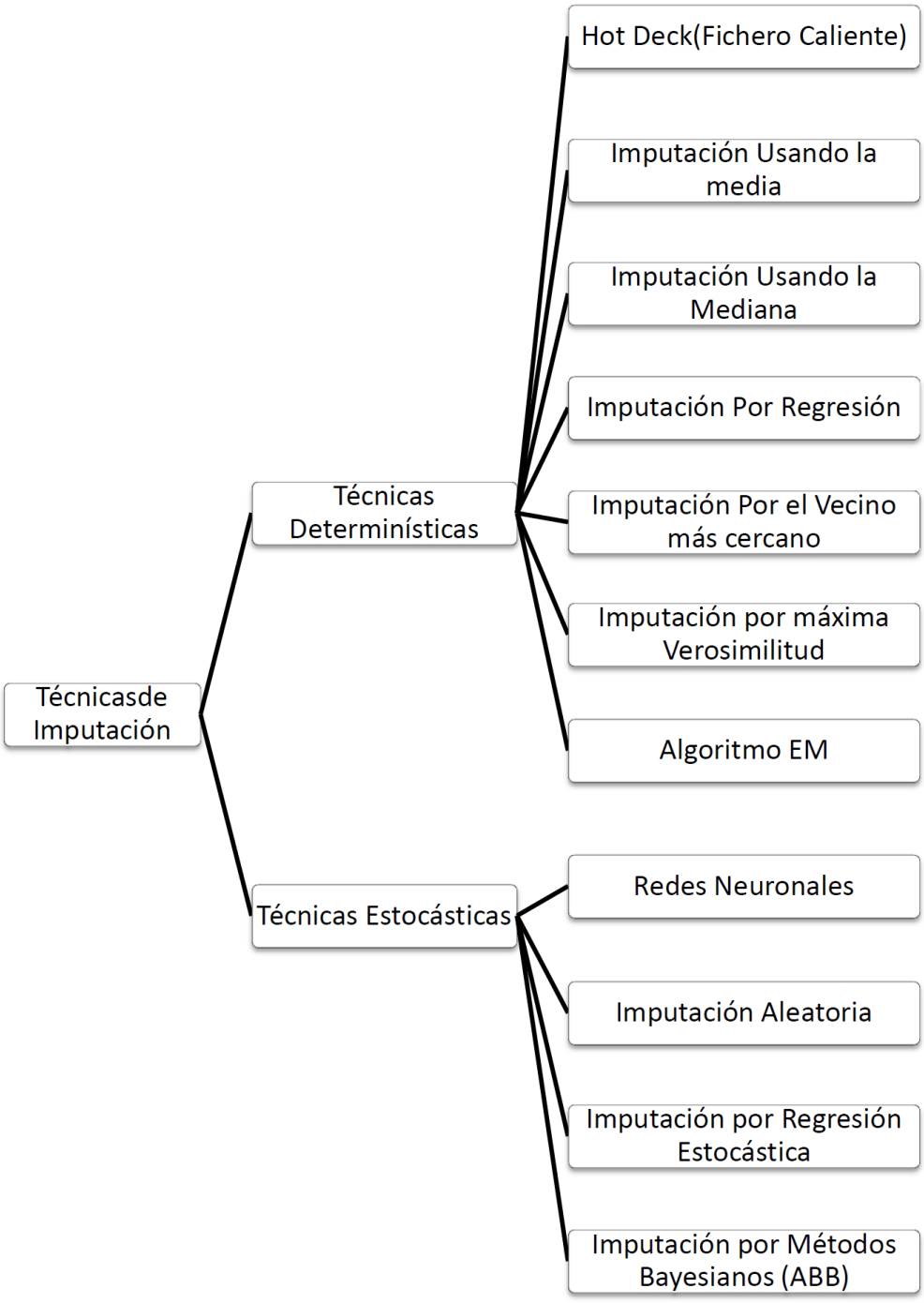

Figura 1.2: Clasificación de Técnicas de Imputación.

1.4.2. Técnicas Determinísticas

Este tipo de técnicas se emplea cuando al repetir la imputación en varias unidades bajo las mismas condiciones, producirán las mismas respuestas. Algunas de las técnicas que conducen a estos resultados son:

1.4.2.1. Fichero Caliente (Hot Deck)

Es un método usual de ajustar conjuntos de datos para valores no observados y admite diversas variantes. Generalmente el fichero caliente es un procedimiento de duplicación. Cuando falta un valor, se duplica un valor ya existente en la muestra para reemplazarlo. El principal propósito del fichero caliente es reducir el sesgo debido a la no respuesta. Para reducir este sesgo, el procedimiento de fichero caliente incorpora un método de clasificación. Todas las unidades muestrales se clasifican en grupos disjuntos de forma que sean lo más homogéneas posible dentro de los grupos. A cada valor que falte, se le asigna un valor del mismo grupo. De modo que el supuesto implícito que se está utilizando es que dentro de cada clasificación la no respuesta sigue la misma distribución que los que responden. Tal supuesto impone una fuerte restricción para las variables de clasificación. Estas variables han de estar correladas con los valores que falten y con los valores de los que contestan. Si esto no se mantiene el fichero-caliente reduce solo en parte el sesgo debido a la no-respuesta i) produce un conjunto de datos limpios, esto es, un conjunto de datos completo y claro; ii) reduce el sesgo mientras preservemos las distribuciones conjuntas y marginales. Por ejemplo si sustituyéramos un valor que falte por la media, la distribución de los valores muestrales resultaría afectada. Y si escogiéramos aleatoriamente un valor entre los datos se reduciría la distorsión de la distribución pero no el sesgo.

Como método de imputación los procedimientos de fichero caliente tienen ciertos rasgos atractivos entre los que se encuentran los siguientes:

1. Los procedimientos conducen a una post-estratificación sencilla;
2. No se presentan problemas especiales de encajar conjuntos de datos;
3. No se necesitan supuestos fuertes para estimar los valores individuales de las respuestas que falten.

Ejemplo: En la tabla siguiente se presenta un conjunto de datos con valores faltantes. Nótese que el caso tres tiene un dato faltante en el atributo Item 4. Usando técnicas

de tipo Hot Deck, cada uno de los otros casos con los datos completos es examinado y el valor del caso más similar es sustituido por el valor faltante. En este ejemplo, el caso uno, dos y cuatro son examinados. El caso cuatro es fácilmente eliminado, ya que no tiene nada en común con el caso 3. Los casos uno y dos tienen similitudes con el caso 3. El caso uno tiene un ítem en común mientras que el caso dos tiene dos ítems en común. Por tanto, el caso dos es más similar al caso 3.

Tabla 2. Conjunto de datos incompletos

Caso	Item 1	Item 2	Item 3	Item 4
1	10	2	3	5
2	13	10	3	13
3	5	10	3	**?**
4	2	5	10	2

Fuente: [Wang, 2003]

Una vez el caso más similar es identificado, la imputación Hot Deck sustituye el valor del caso más completo por el valor faltante. Ya que el segundo caso contiene el valor 13 en el ítem cuatro, el valor de 13 reemplaza el valor faltante en el caso tres (Tabla 3).

Tabla 3. Conjunto de datos completos

Caso	Item 1	Item 2	Item 3	Item 4
1	10	2	3	5
2	13	10	3	13
3	5	10	3	**13**
4	2	5	10	2

Fuente: [Wang, 2003]

Sin embargo, los procedimientos de fichero caliente tienen ciertas desventajas. Carecen de un mecanismo de probabilidad, lo que imposibilita calcular su confianza sin algún procedimiento de modelización depende demasiado de la experiencia del investigador. Existe la posibilidad de usar varias veces a una misma unidad que ya ha respondido.

Al evaluar los métodos de fichero caliente sería conveniente saber come afectan al sesgo y a la confianza de las estimaciones principales: el tamaño de los grupos de clasificación (clases de ponderación), la frecuencia de los datos que faltan y la elección de los ítems de encaje.

1.4.2.2. Imputación Haciendo uso de la Media

Este método, propuesto por primera vez por Wilks (1932), es posiblemente uno de los procedimientos de imputación más antiguo y más sencillo. Existen dos variantes las medias incondicionadas y las medias condicionadas.

Imputación haciendo uso de Medias incondicionadas

La forma más simple de imputación no aleatoria de un valor desconocido consiste en asignar el valor promedio de la variable que lo contiene, calculado en los casos que tienen valor. Si se trata de una variable categórica se imputa la moda de la distribución. Consiste en estimar los valores perdidos de la j-ésima variable mediante la media de sus valores observados, la cual ha sido llamada por Little y Rubin (1987), la expresión clásica para el cálculo es la siguiente:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{1.3}$$

Donde:
x_i : valores observados de la variable x
n: cantidad de individuos

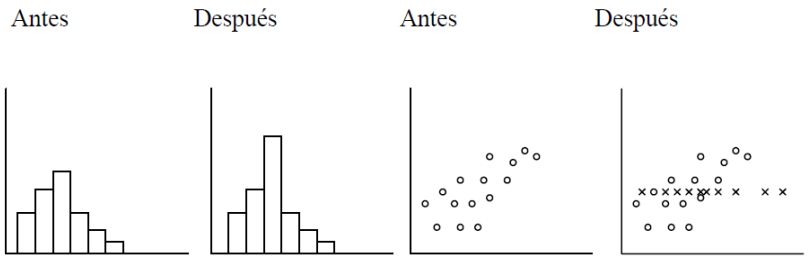

Figura 1.3: Características de la imputación por medias.

Por ejemplo en la **Figura 1.3**, en el histograma se observa que el uso de la imputación haciendo uso de la media incondicional condujo a la modificación de la distribución de los

datos, así como también en el diagrama de dispersión es posible observar que el uso de imputación por la media no es una técnica recomendable cuando posteriormente se desea realizar un análisis estadístico mediante técnicas de regresión.

Bajo este procedimiento de imputación, el valor medio de la variable se preserva, pero otros estadísticos que definen la forma de la distribución varianza, covarianza, quartiles, sesgo, curtosis, entre otros, pueden ser afectados [Acock, 2005].

Ejemplo En la siguiente tabla se presenta una serie de remesas de El Salvador desde enero de 1991 hasta diciembre de 2010.

Tabla 4. Conjunto de datos incompletos

	Ene	Feb	Mar	Abr	May	Jun	Jul	Ago	Sep	Oct	Nov	Dic
1991	63.1	58.4	67.6	77.8	77.4	67.8	70.0	53.5	53.1	64.0	64.3	73.1
1992	65.0	66.0	75.6	74.5	76.3	75.6	77.7	62.0	60.3	65.5	70.5	89.3
1993	57.7	65.3	81.0	76.4	75.6	71.6	76.0	68.7	62.9	66.6	74.1	88.2
1994	69.4	72.9	81.1	79.0	88.2	77.1	75.1	86.4	80.4	73.5	80.1	99.3
1995	82.1	74.4	86.2	76.2	98.1	91.6	90.6	93.1	85.0	89.1	89.1	105.9
1996	90.6	74.0	89.6	84.6	100.9	86.1	105.3	96.3	88.2	94.0	80.0	96.9
1997	89.2	77.8	84.2	103.3	100.9	106.8	117.1	98.2	105.8	106.9	88.7	120.6
1998	98.9	86.6	110.2	113.1	112.5	111.3	116.2	114.7	114.2	114.6	115.1	130.9
1999	106.7	97.2	115.5	117.9	119.4	108.6	119.1	106.5	106.4	113.9	121.5	141.1
2000	132.1	125.9	140.7	121.7	153.4	143.6	152.0	156.2	142.7	159.8	155.6	167.0
2001	147.6	147.2	149.6	139.7	179.1	157.8	162.9	166.8	146.7	169.6	158.7	184.8
2002	143.4	146.2	157.8	174.2	180.3	167.9	162.2	160.1	150.5	156.5	160.9	175.2
2003	146.0	149.1	170.1	177.4	186.1	178.1	175.8	172.8	180.4	181.1	174.8	213.6
2004	171.3	170.3	218.4	213.8	220.5	212.6	210.1	224.4	213.5	215.9	230.6	246.2
2005	189.7	199.3	250.4	245.5	272.2	250.8	240.9	272.2	245.2	261.5	266.1	323.3
2006	237.7	249.8	309.4	274.0	330.8	289.8	284.9	293.7	271.4	301.0	279.5	348.9
2007	270.9	269.0	320.2	310.3	338.0	**?**	324.6	**?**	281.6	323.8	283.5	351.1
2008	270.5	295.9	338.4	334.4	343.6	332.9	328.8	299.5	303.1	303.0	262.7	329.3
2009	248.6	270.9	309.3	281.8	296.3	286.9	275.4	285.2	269.4	278.1	259.0	326.2
2010	228.1	263.2	337.0	296.3	319.2	294.2	286.8	287.0	260.5	269.8	262.7	326.1

Se utilizará en estos datos la técnica de imputación por medias incondicionadas, por medio de la sintaxis en R, se insertan los datos y se encuentra la media omitiendo los datos faltantes:

```
> remesas <-c(63.1,58.4, 67.6,.....,262.7,326.1)
> n=length(remesas[[1]])
> media<-mean(na.omit(remesas[[1]]))
```

se obtiene $\mu = 167.6676$ ahora solo para automatizar el reemplazo de los valores faltantes con la media encontrada se tiene en R:

```
> b<-is.na(remesas[[1]])
> for(i in 1:n) {if(b[i]==TRUE) remesas[[1]][i]<-media}
> remesas
```

Tiene como desventaja la modificación de la distribución de la variable modificándose creándola más estrecha ya que reduce su varianza, además, no conserva la relación entre variables y se debe asumir una variable MCAR. Si bien uno puede imputar todos los valores ausentes X_i, por la manera en que se realiza la sustitución de los datos omitidos, la suma de cuadrados de las desviaciones de las observaciones respecto de la media permanece inalterada, es decir la varianza de X se contraerá debido a que todos los valores X_i adicionados no contribuirán en nada a la misma pero se incrementa el tamaño de muestra, lo cual origina que la varianza de la variable disminuya y se generen, en forma artificial, intervalos de confianza más estrechos.
El uso de este método, afectará la correlación entre la variable imputada y cualquiera otra, reduciendo su variabilidad. Esto es, la sustitución de la media en una variable, puede llevar a perjudicar estimaciones de los efectos de otra o todas las variables en un análisis de regresión, porque el perjuicio en una correlación puede afectar los pesos de todas las variables. Adicionalmente, si se imputa un gran número de valores usando la media, la distribución de frecuencias de la variable imputada puede ser engañosa debido a demasiados valores localizados centralmente creando una distribución más alargada o leptocúrtica [Rovine y Delaney, 1990].

Imputación por medias condicionadas

Imputa medias condicionadas a valores observados. Un método común consiste en agrupar los valores observados y no observados en clases ajustadas e imputar los valores faltantes de los valores observados en la misma clase.

Una variante del procedimiento anterior se presenta cuando 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. La sustitución de los datos faltantes por la media reduce la amplitud del intervalo de confianza debido a la disminución de la varianza del estimador. Al igual que el procedimiento de medias, en este caso se asume que los datos faltantes siguen un patrón MCAR y existirán tantos promedios como categorías se formen, lo cual contribuye a atenuar los sesgos en cada celda pero de ninguna manera los elimina. Este procedimiento tiene las mismas desventajas que el caso anterior, pero en menor proporción por estar agrupadas. Igualmente es de fácil aplicación.

En la medida que la falta de información por categoría sea baja, los sesgos disminuyen pero no desaparecen. No obstante, no se sugiere utilizar este procedimiento en la medida de que se disponga de una mejor alternativa para sustituir la información omitida.

Ejemplo: Utilizando de nuevo la tabla 4, se hará una imputación por medias condicionadas, en este punto se observa que los datos están agrupados en año y en meses resulta más lógico crear la media en función del año, entonces los datos faltantes pertenecen al año 2007, se encuentra la media con los datos completos de este años de la siguiente manera:

```
> a2007<-c(269.0,320.2,310.3,338.0,NA,324.6,NA,281.6,323.8,283.5,351.1,270.5)
> media_2007<-mean(na.omit(a2007))
> media_2007
```

se obtiene $\mu = 307.26$ y comparando ambos métodos(con la media incondicional) claramente con la media condicionada se obtiene mejor resultado.

La sustitución de datos utilizando promedios es una vieja práctica entre investigadores de diversas disciplinas, a pesar de que por sus limitaciones teóricas no se considera un procedimiento apropiado. Su ventaja es la facilidad de la aplicación del método.

1.4.2.3. Imputación usando la mediana

Según Acuña y Rodríguez, dado que la media es afectada por la presencia de valores extremos, parece natural usar la mediana en vez de la media con el fin de asegurar robustez. En este caso el valor faltante de una característica dada es reemplazado por la mediana de todos los valores conocidos de ese atributo.

Este método es también una opción recomendada cuando la distribución de los valores de una característica es sesgada [Acuña y Rodríguez, 2009]. Obviamente técnicas como la imputación de la media y la mediana, sólo son aplicables a variables cuantitativas y no pueden usarse con valores faltantes en una característica categórica, en cuyo caso puede usarse la imputación de la moda. Estos métodos de imputación son aplicados separadamente en cada característica que contiene valores faltantes. Nótese que la estructura de correlación de los datos no está siendo considerada en los métodos anteriores.

Ejemplo: Usando los datos de la tabla 4 y el año 2007, se procede a la imputación por la mediana, con el siguiente código en R se encuentra:

```
#insertamos como vector el grupo con datos faltante
a2007<-c(269.0,320.2,310.3,338.0,NA,324.6,NA,281.6,323.8,283.5,351.1,270.5)
median(na.omit(a2007))
```

por lo que se encuentra que el valor de la mediana es: 315.25 y este valor se sustituye en los datos faltantes por lo que se obtiene:

	Ene	Feb	Mar	Abr	May	Jun	Jul	Ago	Sep	Oct	Nov	Dic
2007	270.9	269.0	320.2	310.3	338.0	**315.25**	324.6	**315.25**	281.6	323.8	283.5	351.1

Por lo tanto, una medida alternativa de tendencia central representa mejor la distribución subyacente y por tanto una mejor estimación para los valores faltantes. La mediana, en particular, frecuentemente funciona bien como una medida de tendencia central cuando las distribuciones se desvían considerablemente de la distribución normal estándar. El procedimiento para sustituir la mediana para los valores faltantes para una variable particular sigue la misma lógica y protocolo que la sustitución de la media [Mcknight et. al., 2007].

1.4.2.4. Imputación por Regresión

Este método, propuesto por Buck (1960), supone que las filas de la matriz de datos constituyen una muestra aleatoria de una población normal multivariante. El vector de medias y la matriz de varianzas y covarianzas de los datos completos son utilizados como estimaciones de los parámetros poblacionales, con los cuales se ajustan ecuaciones en regresión para cada una de las variables con datos perdidos, en término de las restantes. Ante la presencia de un patrón de datos faltantes MAR es posible utilizar modelos de regresión para imputar información en la variable Y, a partir de un grupo de covariables $(X_1, X_2, ..., X_p)$ correlacionadas.

Se considera una variable Y_i que presenta n_{per} valores perdidos y $n_i = n - n_{per}$ valores observados. Se supone que las $k-1$ restantes variables X_j, con $i \neq j$, no presentan valores perdidos. Con este método se estima la regresión de la variable Y_i sobre las variables X_j, $\forall j \neq i$, a partir de los n_i casos completos y se imputa cada valor perdido con la predicción dada por la ecuación de regresión estimada. Esto es, si para el caso I el valor y_{li} no se observa, entonces se imputa mediante:

$$\hat{y}_{li} = \hat{\beta}_{0.obs} + \sum_{j \neq i} \hat{\beta}_{j.obs} x_{lj} \tag{1.4}$$

Donde $\hat{\beta}_{0.obs}$ y $\hat{\beta}_{j.obs}$, $j \neq i$ representan los coeficientes de la regresión de X_i, $\forall j \neq i$, basada en las n_i observaciones completas

Ejemplo En la Tabla 5 se presentan dos variables: estatura y peso, con estos datos se obtendrá un modelo de regresión para encontrar los valores faltantes.

El ejemplo se realizará con la ayuda el software R, para facilitar procesos que están fuera de los objetivos de esta investigación como es el explicar la realización de un modelo de regresión, lo cual se automatizará mediante un código que lo calcula eficientemente, para el desarrollo de los ejemplos programables, para esta tesis se asume que el lector tiene las mínimas bases de programación en el paquete R, ya que no ser profundizará en cuestiones básicas del programa. Además que este código se incluirá en el Anexo una función simplificada que hace las mismas rutinas del que se explicará a continuación.

Tabla 5. Conjunto de datos Incompletos de Estatura(cm)-Peso(Kg)

Estatura(cm)	Peso (Kg)	Estatura(cm)	Peso (Kg)
185	85	170	?
185	75	176	68
180	70	174	75
178	68	177	70
159	44	170	68
172	?	161	57
176	72	170	63
183	85	190	80
185	95	185	?
179	70	162	54
186	75	165	54
169	59		

Para empezar se tiene que introducir los datos al entorno R como a continuación se detalla en el siguiente código:

```
>peso<-c(82,75,70,68,44,NA,72,85,95,70,75,59,69,68,75,70,NA,57,63,80,NA,54,54)
>estatura<-c(185,185,180,178,159,172,176,183,185,179,186,169,176,176,174,177,
 170,161,170,190,185,162,165)
>datos<-data.frame(estatura,peso)
```

Con el código anterior se obtiene en "datos" la variable peso y estatura en un solo conjunto de datos, ahora se debe contabilizar el número de individuos y luego el conjunto de individuos completos sin datos faltantes para encontrar el modelo de regresión lineal; se consigue con el siguiente código:

```
>n=length(estatura)
>nx<- na.omit(datos)
```

En “n” se guarda el número de individuos contabilizados incluyendo los que están incompletos y en “nx” los individuos completos, que servirán para el modelo de regresión, que a continuación se calcula con la instrucción siguiente:

```
>reg<-lm(nx$peso~nx$estatura)
>cat("Modelo de Regresión: y=",reg$coefficients[[1]],"+",reg$coefficients[[2]],
 "* X \n")
```

La cual genera la siguiente salida:

```
> Modelo de Regresión: y= -137.9772 + 1.178767 *x
```

Ahora se procede a la imputación de los datos en base al modelo de regresión encontrado, lo primero es identificar donde están los datos perdidos y se consigue con un vector llamado “b” el cual contiene en “Verdadero” los datos faltantes, se crea un bucle que sustituirá los valores faltantes con los valores del modelo que se encontró todo este procedimiento, esta simplificado en el siguiente código:

```
>b<-is.na(peso)
>b
>for (i in 1:n) {if(b[i]==TRUE)
 datos[i,2]<-reg$coefficients[[1]]+datos[i,1]*reg$coefficients[[2]]}
```

Con esto se tiene el conjunto de datos completos, obteniéndose la siguiente tabla:

Tabla 6. Conjunto de datos Completos Estatura(cm)-Peso(Kg)

Estatura(cm)	Peso (Kg)	Estatura(cm)	Peso (Kg)
185	85	170	**62.4132**
185	75	176	68
180	70	174	75
178	68	177	70
159	44	170	68
172	**64.7707**	161	57
176	72	170	63
183	85	190	80
185	95	185	**80.0947**
179	70	162	54
186	75	165	54
169	59		

Frente a la imputación mediante la media, este método incorpora la información que sobre Y_i contienen el resto de variables.

No se sugiere aplicar este método cuando el análisis secundario de datos involucra técnicas de análisis de covarianza o de correlación, ya que sobreestima la asociación entre variables, y en modelos de regresión múltiple puede sobredimensionar el valor del coeficiente de determinación R^2.

1.4.2.5. Imputación por series de tiempo

Se asume que los datos perdidos ocurren de tal forma, y en tal sistema, que el problema se reduce a una situación, en la cual, hay una serie de tiempo, donde una(s) serie(s) de observaciones están perdidas, haciendo óptimo el uso de interrelaciones entre sucesivas observaciones en cada serie de tiempo, mediante el uso de un modelo adecuado para estas series. Se utilizan comúnmente las metodologías Holt-Winters y Box-Jenkins, se pueden obtener como pronósticos hacia delante o hacia atrás con los valores completos para sustituir los valores que faltan. Para una mejor comprensión de estos temas se recomienda leer Análisis de Series Temporales de Daniel Peña(2005), Análisis de series temporales de Antonio Aznar Grasa(1993) y http://www.itl.nist.gov/div898/handbook/pmc/section4/

pmc4.htm(Engineering Statistics Handbook).

La metodología de suavizado exponencial ha comprobado a través de los años ser muy útil en muchas situaciones de pronósticos. Fue sugerido primero por C.C. Holt en 1957 y fue significativo para ser usado en series temporales no estacionales que no muestran tendencia. El ofreció un procedimiento más tarde (1958) que maneja tendencias. Winters (1965) generalizó el método para incluir estacionalidad, de ahí el nombre "Método Holt-Winters".

El Método Holt-Winters tiene tres ecuaciones actualizadas, cada una con una constante que va de 0 a 1. Las ecuaciones están pensadas para dar más ponderación a las observaciones recientes y menos ponderación a las observaciones más allá en el pasado. Dichas ecuaciones son:
Suavizado General

$$S_t = \alpha \frac{y_t}{I_{t-L}} + (1-\alpha)(S_{t-1} + b_{t-1}) \tag{1.5}$$

Suavizado de Tendencia

$$b_t = \gamma(S_t - St - 1) + (1-\gamma)b_{t-1} \tag{1.6}$$

Suavizado Estacional

$$I_t = \beta \frac{y_t}{S_t} + (1-\beta)I_{t-L} \tag{1.7}$$

Predicción

$$F_{t+m} = (S_t + mb_t)I_{t-L+m} \tag{1.8}$$

Donde:

- y es la observación
- S es la observación suavizada
- b es el factor de tendencia
- I es el índice de estacionalidad
- F es el pronóstico en m periodos hacia adelante
- t es un índice denotando un período de tiempo

α, β y γ son las constantes que deben ser estimadas de tal manera que el MSE*(Medias al cuadrado de los errores)* sea minimizado. Esto es mejor dejarlo a un buen software estadístico.

Para inicializar el método Holt-Winters se necesita al menos unos datos de estación completa para determinar estimaciones iniciales de los índices estacionales I_{t-L}.

Usando la tabla 4 de datos del ejemplo de imputación por la media, se introduce los datos en R:

```
> remesas <- read.csv(file="remesas1.csv",head=TRUE,sep=",")
> remesas
> remesas[[1]][0:197]
```

Se acorta la serie hasta 197 donde se tienen los datos completos, ahora se crea la serie de tiempo iniciando en enero de 1991:

```
> remesa_st=ts(remesas[[1]][0:197],start=1991,frequency=12)
> remesa_st
```

Ahora se aplica el método Holt-Winters con estimación automática de las constantes α, β y γ, luego se observan los valores:

```
> pred.hw2=HoltWinters(remesa_st)

#Observamos el mejor valor del alpha
> pred.hw2$alpha

#Observamos el mejor valor del beta
> pred.hw2$beta

#Observamos el mejor valor del gamma
> pred.hw2$gamma
```

y se obtienen los valores de las constantes $\alpha = 0.2721$, $\beta = 0.0402$ y $\gamma = 0.7908$ se puede observar asimismo los valores predichos con los valores reales en la siguiente gráfica:

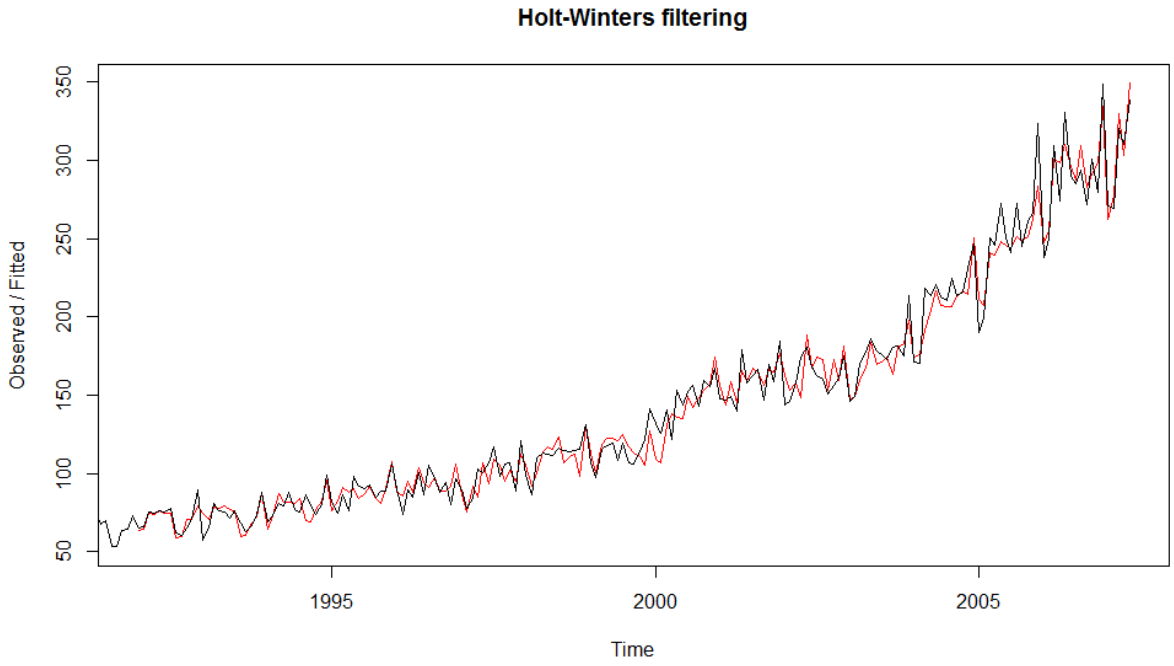

Figura 1.4: Valores Predichos-Línea roja y valores reales-Línea negra

Ahora se procede a encontrar los datos que serán usados en la imputación que se obtiene por medio de una predicción:

```
#Predicción hasta agosto de 2007
> predict(pred.hw2,n.ahead=3)
```

Obteniéndose :

```
> predict(pred.hw2,n.ahead=3)
          Jun         Jul         Aug
```

```
2007  310.5363     305.3315   319.9310
```

El procedimiento Holt-Winters puede hacerse totalmente automático por software de usuario amigable. Una Desventaja a este método es que deben existir suficientes datos antes, entre o después de los valores que faltan.

1.4.2.6. Imputación usando el vecino más cercano

En este método 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, los datos deben de presentar un mecanismo de pérdida no ignorable.

Este es un método de imputación que sustituye a cada valor perdido por el de un donante elegido a partir de una determinada distancia calculada a través de una variable con información completa.

La búsqueda del vecino más cercano que se define de la siguiente manera: Dado un conjunto de puntos $P = \{p_1, ..., p_n\}$ en un espacio métrico X con función de distancia d, permitiendo algún preprocesamiento en P de manera eficiente, se desea responder a dos tipos de solicitudes:

- Vecino más cercano: localizar el punto en P más cercano a $q \in X$.
- Rango: Dado un punto $q \in X$, y $r > 0$, regresar todos los puntos $p \in P$ que satisfagan $d(p, q) \leq r$

Regla de Decisión y Selección de la Distancia

El método de decisión está relacionado con la noción de proximidad o similitud entre los individuos. El índice de similitud entre individuos más utilizado es la métrica de Minkowski:

$$d(i, j) = [|x_{i1} - x_{j1}|^q + |x_{i2} - x_{j2}|^q + ... + |x_{ip} - x_{jp}|^q]^{1/p} \tag{1.9}$$

Con $p = 2$, se tiene la distancia euclídea clásica:

$$d(i,j) = \sqrt{|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + ... + |x_{ip} - x_{jp}|^2} \tag{1.10}$$

Otra métrica conocida es la distancia Manhattan ($p = 1$):

$$d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + ... + |x_{ip} - x_{jp}| \tag{1.11}$$

Estas métricas satisfacen los requisitos matemáticos de una función de distancia:

- $d(i,j) \geq 0$
- $d(i,i) = 0$
- $d(i,j) = d(j,i)$
- $d(i,j) \geq d(i,h) + d(h,j)$

Ejemplo: Se tiene la siguiente tabla de Datos artificiales y se realizará el método de imputación del vecino más cercano usando la distancia euclídea clásica en R:

Tabla 7. Conjunto de datos artificiales Incompletos Var 1-Var 2

Posiciones	Variable 1	Variable 2
1	3	1
2	4	4
3	5	1
4	2	3
5	3	5
6	1	4
7	4	2
8	6	6
9	2	3
10	6	4
11	3	1
12	4	**?**
13	5	3
14	**?**	6
15	3	4

Se introducen los datos en R como una matriz de dimensión 15x2 con el siguiente código:

```
>dato<-matrix(c(3,4,5,2,3,1,4,6,2,6,3,4,5,NA,3,
1,4,1,3,5,4,2,6,3,4,1,NA,3,6,4),nrow=15,ncol=2)
```

Para iniciar la técnica de imputación se divide en 2 partes los datos completos y los incompletos de la siguiente forma:

```
>x <- as.matrix(dato)
>N <- dim(x)
>p <- N[2]
>N <- N[1]
>nas <- is.na(drop(x %*% rep(1, p)))
>xcomplete <- x[!nas, ]
>xbad <- x[nas, , drop = FALSE]
```

Donde **xbad** nos guarda los datos incompletos y **xcomplete** los datos completos y se obtienen las tablas siguientes:

Tabla 8. Conjunto de datos artificiales columnas incompletas Var 1-Var 2

Posiciones	Variable 1	Variable 2
12	4	**?**
14	**?**	6

La tabla con los datos completos:

Tabla 9. Conjunto de datos artificiales sin columnas de datos faltantes Var 1-Var 2

Posiciones	Variable 1	Variable 2
1	3	1
2	4	4
3	5	1
4	2	3
5	3	5
6	1	4
7	4	2
8	6	6
9	2	3
10	6	4
11	3	1
13	5	3
15	3	4

Entonces el primer paso de la técnica es asignar la distancia a cada una de las observaciones, en base al valor "observado" valga la redundancia, para el caso I se tiene que utilizar el valor **4** que sería el punto **q** es decir que es el punto de referencia y la variable 1 será el conjunto de puntos **p**, y se utiliza la distancia euclidiana con el siguiente código:

```
>missing <- c()
>for (i in seq(nrow(xbad))) {
        missing[i] <- sum(is.na(xbad[i, ]))
    }
>missingorder <- order(missing)
>xnas <- is.na(xbad)
>xbadhat <- xbad
>cat(nrow(xbad), fill = TRUE)
>j <- order(missingorder[1])
```

```
>xinas <- xnas[missingorder[1], ]
>xd <- as.matrix(scale(xcomplete, xbad[missingorder[1],], FALSE)[, !xinas])
>dd <- drop(xd^2 %*% rep(1, ncol(xd)))
```

Se obtiene la tabla siguiente:

Tabla 10. Conjunto de datos artificiales distancia Euclideana - Variable 1

Posiciones	$(p-q)$	$(p-q)^2$
1	-1	1
2	0	0
3	1	1
4	-2	4
5	-1	1
6	-3	9
7	0	0
8	2	4
9	-2	4
10	2	4
11	-1	1
13	1	1
15	-1	1

Ahora para empezar a encontrar el vecino más cercano se recomienda un promedio de k-vecinos que se tomarán en cuenta para estar en los candidatos, Hastie, et al. (1999) han demostrado que k de 5 a 10 es adecuado. Tomando $k = 10$ se ordenan por la distancia con el siguiente código:

```
>K=10
>od <- order(dd)[seq(K)]
```

```
>od <- od[!is.na(od)]
>K <- length(od)
>distance <- dd[od]
```

y se obtiene las siguientes posiciones en la tabla 11:

Tabla 11. Conjunto de datos artificiales distancia Euclideana ordenados - Variable 1

Posiciones	$(p-q)$	$(p-q)^2$
2	0	0
7	0	0
1	-1	1
3	1	1
5	-1	1
11	-1	1
13	1	1
15	-1	1
4	-2	4
8	2	4

Entonces la tabla anterior ubicará los vecinos que son más cercanos según la distancia euclidiana y se puede observar que los dos primeros tienen la misma distancia(distancia cero) se hará un promedio de estos, que son en la variable 2, las observaciones en las posiciones 2 y 7, los valores de ellas son 4 y 2 respectivamente, de donde se obtienen que el valor imputado es: $\frac{4+2}{2} = 3$. En este caso se hizo el promedio ya que hubo dos valores con la misma distancia menor. Para facilitar el cálculo este proceso se obtiene bajo la siguiente sintaxis:

```
>s <- sum(1/(distance + 1e-15))
>weight <- (1/(distance + 1e-15))/s
> xbadhat[missingorder[1], ] <- drop(weight %*% xcomplete[od,
```

```
 xinas, drop = FALSE])
> xbadhat[missingorder[1], ]
```

Para el siguiente caso se omitirán los códigos de R ya que tienen el mismo proceso anterior, y que además en Anexos está incluida una función que facilita el proceso. Para el caso 2 de la imputación del vecino más cercano se tiene que el valor $q = 6$, es el punto de referencia, y se obtiene con la variable 2 la siguiente tabla:

Tabla 12. Conjunto de datos artificiales distancia Euclideana - Variable 2

Posiciones	$(p-q)$	$(p-q)^2$
1	-5	25
2	-2	4
3	-5	25
4	-3	9
5	-1	1
6	-2	4
7	4	16
8	0	0
9	-3	9
10	-2	4
11	-5	25
13	-3	9
15	-2	4

Como se tomó el k=10 se obtienen las posiciones de las distancias en la Tabla 13.

La Tabla 13 ubicará los vecinos que son más cercanos según la distancia euclidiana y se puede observar que en la posición 8 está el vecino con menos distancia, por lo tanto la observación de la variable 1 en la octava posición es **6** y este sería el valor imputado.

Tabla 13. Conjunto de datos artificiales distancia Euclideana ordenados - Variable 2

Posiciones	$(p-q)$	$(p-q)^2$
8	0	0
5	-1	1
2	-2	4
6	-2	4
10	-2	4
15	-2	4
4	-3	9
9	-3	9
13	-3	9
7	4	16

Y la tabla con los datos ya imputados por el método del vecino más cercano es la siguiente:

Tabla 14. Conjunto de datos artificiales Completos Var 1-Var 2

Posiciones	Variable 1	Variable 2
1	3	1
2	4	4
3	5	1
4	2	3
5	3	5
6	1	4
7	4	2
8	6	6
9	2	3
10	6	4
11	3	1
12	4	**3**
13	5	3
14	**6**	6
15	3	4

En la práctica este método resulta poco aconsejable. Se han ido desarrollando algoritmos eficientes que evitan recorrer exhaustivamente todo el conjunto de entrenamiento. En comparación a los métodos paramétricos los métodos del vecino más cercano no hacen suposiciones de la distribución del modelo y pueden retener la estructura varianza/covarianza en la salida.

1.4.2.7. Imputación por Máxima Verosimilitud

Los métodos de imputación por máxima verosimilitud tienen como objetivo realizar estimaciones máximo verosímiles de los parámetros de una distribución cuando existen datos faltantes. Se asume que los datos faltantes siguen un patrón MAR y que la distribución marginal de los registros observados está asociada a una función de verosimilitud para un parámetro θ desconocido, siempre que el modelo sea adecuado para el conjunto de datos completos. Se resume el procedimiento para estimar los parámetros de un modelo utilizando una muestra de datos faltantes, de la siguiente manera:

1. Estimar los parámetros del modelo con los datos completos con la función de máxima verosimilitud.
2. Utilizar los parámetros estimados para predecir los valores omitidos.
3. Sustituir los datos por las predicciones y obtener nuevos valores de los parámetros maximizando la verosimilitud de la muestra completa.
4. Aplicar el algoritmo hasta lograr la convergencia, la que se obtiene cuando el valor de los parámetros no cambia entre dos iteraciones sucesivas.

Un procedimiento eficiente para maximizar la verosimilitud cuando existen datos faltantes es el algoritmo EM, que fue proporcionado por Dempster, Laird y Rubin(1977)

1.4.2.8. Algoritmo EM(Expectation Maximization)

El algoritmo EM es un algoritmo iterativo muy general para la estimación de parámetros por máxima verosimilitud cuando algunas de las variables aleatorias que intervienen no se observan es decir, considerada falta o está incompleto. El algoritmo EM formaliza la idea intuitiva de obtener estimaciones de parámetros cuando algunos de los datos que faltan, de la siguiente forma:

i. Reemplazar valores perdidos por los valores estimados.

ii. Estimar parámetros.

iii. Repetir de la siguiente forma:

- Paso (i) usando los valores estimados de los parámetros como valores verdaderos, y
- Paso (ii) usando los valores estimados como valores "observados", iterando hasta la convergencia.

Esta idea ha estado en uso durante muchos años antes de Orchard y Woodbury (1972) en su falta información proporcionada principio el fundamento teórico de la idea subyacente.

El término fue introducido en EM Dempster, Laird y Rubin (1977) donde la prueba de generar resultados sobre el comportamiento del algoritmo se le dio RST, así como un gran número de aplicaciones.

Para esta discusión, se supone que se tiene un vector aleatorio cuya densidad conjunta $f(y;\theta)$ se indexan con un parámetro p-dimensional $\theta \in \Theta \subseteq R^p$. Si los datos completos del vector y son observados es de interés calcular la estimación máxima verosimilitud de θ basado en la distribución de y. La función de log-verosimilitud de y es:

$$logL(\theta;y) = l(\theta;y) = logf(y;\theta) \tag{1.12}$$

que requiere entonces ser maximizada.

En presencia de los datos faltantes, sin embargo, sólo una función de datos completos del vector y, es observada. Se denota esta expresión y como (y_{obs}, y_{mis}), donde y_{obs} y y_{mis} contiene los no observados o "falta de datos". Por simplicidad de descripción, se supone que los datos que faltan se faltan al azar (MAR) (Rubin, 1976), de modo que

$$f(y;\theta) = f(y_{obs}, y_{mis};\theta) = f_1(y_{obs};\theta) \cdot f_2(y_{mis}|y_{obs};\theta) \tag{1.13}$$

donde f_1 es la densidad conjunta de y_{obs} y f_2 es la densidad conjunta de y_{mis} dada la observada y_{obs} de datos, respectivamente. Por lo tanto se deduce que

$$l_{obs}(\theta;y_{obs}) = l(\theta;y) - logf_2(y_{mis}|y_{obs};\theta) \tag{1.14}$$

Donde $l_{obs}(\theta; y_{obs})$ son los datos completos log-verosimilitud.

Algoritmo EM es útil cuando se dificulta maximizar l_{obs}, pero maximizar los datos completos log-verosimilitud es simple. Sin embargo, puesto que y no se observa, l no pueden ser evaluados y por lo tanto maximizado. El algoritmo EM intenta maximizar $l(\theta; y)$ iterativamente, mediante la sustitución que por su expectativa condicional dados los y_{obs} de datos observados. Esta expectativa se calcula con respecto a la distribución de los datos completos a evaluarse en la estimación actual de θ.

Más específicamente, si $\theta^{(0)}$ es un valor inicial para θ, a continuación, en la primera iteración se requiere calcular

$$Q(\theta; \theta^{(0)}) = E_{\theta^{(0)}}[l(\theta; y)|y_{obs}] \tag{1.15}$$

$Q(\theta; \theta^{(0)})$ es ahora maximizado con respecto a θ, esto es, $\theta^{(1)}$ es encontrado talque

$$Q(\theta^{(1)}; \theta^{(0)}) \geq Q(\theta; \theta^{(0)}) \tag{1.16}$$

para todo $\theta \in \Theta$. Así, el algoritmo EM se compone de un E-paso (paso Estimación) seguido un M-step (paso de maximización) definido como:

Paso-E : Calcular $Q(\theta; \theta^{(t)})$, donde

$$Q(\theta; \theta^{(t)}) = E_{\theta^{(t)}}[l(\theta; y)|y_{obs}] \tag{1.17}$$

Paso-M : Encontrar $\theta^{(t+1)}$ en Θ de tal manera que

$$Q(\theta^{(t+1)}); \theta^{(t)}) \geq Q(\theta; \theta^{(t)}) \tag{1.18}$$

para todo $\theta \in \Theta$

El Paso-E y el Paso-M se repiten alternativamente hasta que la diferencia $L(\theta^{(t+1)}) - L(\theta^{(t)})$ es menor que ξ, donde ξ es una pequeña cantidad prescrita.

Una desventaja de EM es que su tasa de convergencia puede ser muy lento si una gran cantidad de datos que faltan. Dempster, Laird y Rubin (1977) muestran que la convergencia es lineal con velocidad proporcional a la fracción de la información acerca de θ en $l(\theta; y)$ que se observa.

Ejemplo: Ejemplo para el ajuste de datos provenientes de una normal univariante, la aplicación práctica en este ejercicio será con los siguientes datos:

Tabla 15. Conjunto de datos Incompletos Obtenidos aleatoriamente de una distribución Normal con $\mu = 5$

5.226416	?	?	3.852736
4.246948	4.720129	4.401418	4.353384
5.719032	?	5.273086	?
5.062461	4.199522	4.567361	6.328042
6.621925	5.607660	4.205563	?

Para la realización de este ejercicio se basó en un código fuente del Profesor Alan L. Yuille (http://www.stat.ucla.edu/~yuille/), y para insertar los datos en R se hace de la siguiente forma:

```
>x<-c(5.226416,4.246948,5.719032,5.062461,6.621925,NA,4.720129,NA,4.199522,
5.607660,NA,4.401418,5.273086,4.567361,4.205563,4.353384,3.852736,NA,6.328042,NA)
```

Sea $Y_1, ..., Y_n$ una muestra aleatoria de una distribución normal. La función de verosimilitud es:

$$f(\mathbf{y}; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_1-\mu)^2}{2\sigma^2}} ... \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_n-\mu)^2}{2\sigma^2}} \tag{1.19}$$

$$= \left(\tfrac{1}{2\pi\sigma^2}\right)^{n/2} e^{\frac{-\sum_{i=1}^{n}(y_i-\mu)^2}{2\sigma^2}}$$

$$= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} e^{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n} y_i^2 - 2\mu\sum_{i=1}^{n} y_i + n\mu^2\right)} \tag{1.20}$$

lo cual implica que $(\sum_{i=1}^{n} y_i, \sum_{i=1}^{n} y_i^2)$ son suficientes para calcular el estadístico $\theta = (\mu, \sigma^2)$ la función completa de log-verosimilitud es:

$$l(\mu, \theta^2; \mathbf{y}) = -\frac{n}{2} log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\left(\sum_{i=1}^{n} y_i^2 - 2\mu\sum_{i=1}^{n} y_i + n\mu^2\right) \tag{1.21}$$

Entonces con los datos que se tienen es posible definir la función de verosimilitud con el siguiente código:

```
>ll <- function(y, mu, sigma2, n){
      -.5*n*log(2*pi*sigma2)-(1/(sigma2))*(sum(y^2)-2*mu*sum(y)+n*mu^2)
         }
```

Se supone que $y_i, i = 1, ..., m$ son datos observados y y_i, $i = m+1, ..., n$ son datos faltantes al azar (MAR) donde y_i se asume como i.i.d. $N(\mu, \sigma^2)$. Se denota el vector de datos por $y_{obs} = [y_1, ..., y_m]^T$. Ya que los datos completos y vienen de una familia exponencial, el paso de estimación (E-Step) requiere el cálculo de:

$$E_\theta[\textstyle\sum_{i=1}^n y_i | y_{obs}] \text{ y } E_\theta[\textstyle\sum_{i=1}^n y_i^2 | y_{obs}]$$

En lugar de calcular la esperanza de los datos completos de la función de verosimilitud, mostrada arriba, se calcula las n iteraciones del Paso de estimación, se calcula:

$$s_1^{(t)} = E_{\mu^{(t)}, \sigma^{2(t)}} \left(\sum_{i=1}^{n} y_i | y_{obs} \right) \tag{1.22}$$

$$= \sum_{i=1}^{m} y_i + (n-m)\mu^{(t)} \tag{1.23}$$

Se ejecuta con el código siguiente:

```
> Yobs <- x[!is.na(x)]
> Ymis <- x[is.na(x)]
> n <- length(c(Yobs, Ymis))
> r <- length(Yobs)

# Valores iniciales
> mut <- mean(Yobs)
> sit <- var(Yobs)*(r-1)/r
```

```
#Se calcula la función de verosimilitud con los valores iniciales
>  lltm1 <- ll(Yobs, mut, sit, n)
>  lltm1

# Paso-E (Estimación)
> EY <- sum(Yobs) + (n-r)*mut
```

Pues $E_{\mu^{(t)},\sigma^{2(t)}}(y_i) = \mu^{(t)}$ donde $\mu^{(t)}$ y $\sigma^{2^{(t)}}$ son las estimaciones de μ y σ^2, y

$$s_2^{(t)} = E_{\mu^{(t)},\sigma^{2(t)}}\left(\sum_{i=1}^{n} y_i^2 | y_{obs}\right) \tag{1.24}$$

$$= \sum_{i=1}^{m} y_i^2 + (n-m)\left[\sigma^{2^{(t)}} + \mu^{(t)^2}\right] \tag{1.25}$$

Pues $E_{\mu^{(t)},\sigma^{2(t)}}(y_i^2) = \sigma^{2^{(t)}} + \mu^{(t)^2}$.

y para esto se utiliza en la parte de estimación el complemento del código:

```
> EY2 <- sum(Yobs^2) + (n-r)*(mut^2 + sit)
```

Para el paso de Maximización (M-Step), hay que tener en cuenta que la estimación de máxima verosimilitud de los datos completos de μ y σ^2 son:

$$\hat{\mu} = \frac{\sum_{i=1}^{n} y_i}{n} \text{ y } \hat{\sigma}^2 = \frac{\sum_{i=1}^{n} y_i^2}{n} - \left(\frac{\sum_{i=1}^{n} y_i}{n}\right)^2$$

El paso de maximización se define mediante la sustitución de las estimaciones calculadas en el paso de estimación de los datos completos con suficientes estadísticos, que se mostraron arriba, para nuevas iteraciones para μ y σ^2. Tenga en cuenta que los estadísticos suficientes de todos los datos en sí no se puede calcular directamente desde $y_{(m+1)}, ..., y_n$ ya que no están presentes. Se obtiene las expresiones:

$$\mu^{(t+1)} = \frac{s_1^{(t)}}{n} \tag{1.26}$$

y

$$\sigma^{2^{(t+1)}} = \frac{s_2^{(t)}}{n} - \mu^{(t+1)^2} \tag{1.27}$$

Para el ejemplo con los datos considerados, el paso de Maximización se realiza con el siguiente código:

```
# Paso-M (Maximización)
>  mut1 <- EY / n
>  sit1 <- EY2 / n - mut1^2
```

Por lo tanto, el paso de estimación implica la evaluación del cálculo de (1.23) y (1.25) que comienza con valores iniciales $\mu^{(0)}$ y $\sigma^{2^{(0)}}$. El paso de Maximización implica la sustitución de éstos en (1.26) y (1.27) para calcular nuevos valores $\mu^{(1)}$ y $\sigma^{2^{(1)}}$, etc. Por lo tanto, el algoritmo EM se repite sucesivamente entre (1.23) y (1.25), (1.26) y (1.27). El algoritmo se detiene hasta que $\mu^{(t+1)} - \mu^{(t)}$ y $\sigma^{2^{(t+1)}} - \sigma^{2^{(t)}}$) es menor que $\xi = 1x10^{-3}$, donde ξ es una pequeña cantidad prescrita para este ejemplo. Entonces para completar el ejemplo con la primera iteración se obtiene:

```
# Se actualiza los  valores de los parámetros
> mut <- mut1
> sit <- sit1

#Se calcula la función de Máxima Verosimilitud con los valores iniciales
>llt <- ll(Yobs, mut, sit, n)

#Se detiene la convergencia según el error prescrito
>abs(lltm1 - llt)
```

Donde se obtuvo que el error fue de $6.82121e-13$ es decir que con una iteración convergió el Algoritmo EM, obteniéndose $\hat{\mu} = 4.959046$ y $\hat{\sigma}^2 = 0.6388194$ para automatizar

la imputación se usa la siguiente rutina:

```
>b<-is.na(x)
>for (i in 1:n){if(b[i]==TRUE) x[i]<-mut}
>x
```

Y obtenemos la tabla 16:

Tabla 16. Conjunto de datos Completados por medio de Imputación por Máxima Verosimilitud y Algoritmo EM

5.226416	**4.959046**	**4.959046**	3.852736
4.246948	4.720129	4.401418	4.353384
5.719032	**4.959046**	5.273086	**4.959046**
5.062461	4.199522	4.567361	6.328042
6.621925	5.607660	4.205563	**4.959046**

En Anexos se encuentra la función con la rutina completa en R para este ejemplo.

1.4.3. Técnicas Aleatorias o Estocásticas

Son aquellas que cuando se repite el método de imputación bajo las mismas condiciones para una unidad, producen resultados diferentes.

1.4.3.1. Imputación por redes Neuronales

Son sistemas de información de procesamiento, que reconocen patrones de los datos sin algún valor perdido para aplicarlo a bases de datos incompletas. Estas redes son más usadas para variables cualitativas que cuantitativas, siendo más adecuadas cuando la distribución es no lineal. No es aconsejable cuando hay registros atípicos que distorsionan la red, además son costosos y requieren de capacitación del analista así como de "software" adecuado.

A la hora de implementar una red neuronal como parte de un programa o sistema informático, se pueden distinguir 3 fases básicas:

- **Diseño**: en esta fase se elige el tipo de red neuronal a usar (la arquitectura o topología), el número de neuronas que la compondrán, etc.
- **Entrenamiento**: en esta fase se le presentan a la red neuronal una serie de datos de entrada y datos de salida (resultados), para que a partir de ellos pueda aprender.
- **Uso**: se le suministran las entradas pertinentes a la red, y esta genera las salidas en función de lo que ha aprendido en la fase de entrenamiento.

Funcionamiento Básico

Las redes neuronales están formadas por un conjunto de neuronas artificiales interconectadas. Las neuronas de la red se encuentran distribuidas en diferentes capas de neuronas, de manera que las neuronas de una capa están conectadas con las neuronas de la capa siguiente, a las que pueden enviar información. La arquitectura más usada en la actualidad de una red neuronal (como la presentada en la figura 1.5) consistiría en:

- Una primera capa de entradas, que recibe información del exterior.
- Una serie de capas ocultas (intermedias), encargadas de realizar el trabajo de la red.
- Una capa de salidas, que proporciona el resultado del trabajo de la red al exterior.

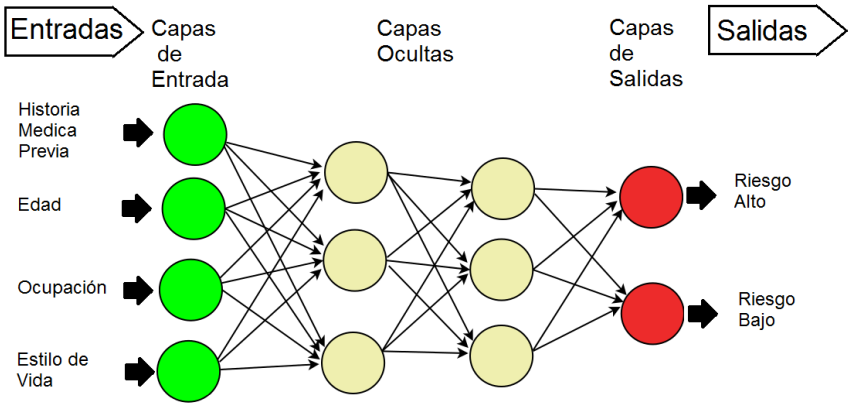

Figura 1.5: Esquema de una red neuronal antes del entrenamiento. Los círculos representan neuronas, mientras las flechas representan conexiones entre neuronas

El número de capas intermedias y el número de neuronas de cada capa dependerá del tipo de aplicación al que se vaya a destinar la red neuronal, comúnmente se aplica el criterio de parsimonia al de las capas ocultas.

Neuronas y conexiones

Cada neurona de la red es una unidad de procesamiento de información; es decir, recibe información a través de las conexiones con las neuronas de la capa anterior, procesa la información, y emite el resultado a través de sus conexiones con las neuronas de la capa siguiente, siempre y cuando dicho resultado supere un valor umbral.

En una red neuronal ya entrenada, las conexiones entre neuronas tienen un determinado peso ("peso sináptico"). Un ejemplo de una neurona sobre la que convergen conexiones de diferente peso sináptico (W_i) sería el de la figura 1.6:

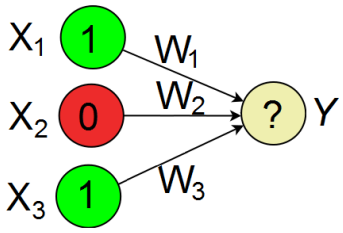

Figura 1.6: Conexiones de diferentes peso sináptico($W_1 > W_2 > W_3$)convergen sobre la misma neurona Y

El procesamiento de la información llevado a cabo por cada neurona Y, consiste en una función (F) que opera con los valores recibidos desde las neuronas de la capa anterior (X_i, generalmente 0 o 1), y que tiene en cuenta el peso sináptico de la conexión por la que se recibieron dichos valores (W_i). Así, una neurona dará más importancia a la información que le llegue por una conexión de peso mayor que a aquella que le llegue por una conexión de menor peso sináptico.

Un modelo simple de la función F seria:

$$F = X_1W_1 + X_2W_2 + ... + X_iW_i \tag{1.28}$$

Si el resultado de la función F es mayor que el valor umbral (U), la neurona se activa y emite una señal (1) hacia las neuronas de la capa siguiente. Pero, si por el contrario, el resultado es menor que el valor umbral, la neurona permanece inactiva (0) y no envía

ninguna señal:

$$X_1W_1 + X_2W_2 + ... + X_iW_i \leq U \leftrightarrow \text{Inactivación} \leftrightarrow Y = 0 \tag{1.29}$$

$$X_1W_1 + X_2W_2 + ... + X_iW_i > U \leftrightarrow \text{Activación} \leftrightarrow Y = 1 \tag{1.30}$$

De esta forma, definido un conjunto inicial de pesos en las conexiones, al presentar un estímulo (conjunto de ceros y unos que representa un dato, perfil u objeto) a la capa de entradas, cada neurona en cada capa realiza la operación descrita anteriormente, activándose o no, de manera que al final del proceso las neuronas de la capa de salidas generan un resultado (otro conjunto de ceros y unos), que puede coincidir o no con el que se desea asociar el estímulo.

En el entrenamiento de una red neuronal tanto el peso sináptico de las conexiones como el valor umbral para cada neurona se modifican (según un algoritmo de aprendizaje), con el fin de que los resultados generados por la red coincidan con (o se aproximen a) los resultados esperados.

Y para simplificar el sistema de entrenamiento, el valor umbral (U) pasa a expresarse como un peso sináptico más ($-W_0$), pero asociado a una neurona siempre activa (X_0). Esta neurona siempre activa, se denomina "*bias*", y se sitúa en la capa anterior a la neurona Y, tal como se muestra en la figura 1.7.

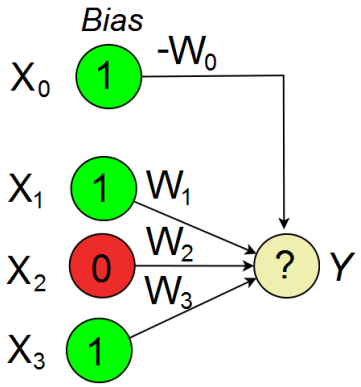

Figura 1.7: Neurona bias y su peso sináptico asociado ($-W_0$, en sustitución del valor umbral)

Así, la condición de activación puede reescribirse como:

$$X_0W_0 + X_1W_1 + X_2W_2 + ... + X_iW_i > 0 \leftrightarrow \text{Activación} \leftrightarrow Y = 1 \tag{1.31}$$

Tipos de Aprendizajes básicos

Para poder aprender, las redes neuronales se sirven de un algoritmo de aprendizaje. Estos algoritmos están formados por un conjunto de reglas que permiten a la red neuronal aprender (a partir de los datos que se le suministran), mediante la modificación de los pesos sinápticos de las conexiones entre las neuronas (recordar que el umbral de cada neurona se modificará como si fuera un peso sináptico más).

Generalmente los datos que se usan para entrenar la red se le suministran de manera aleatoria y secuencial. Los tipos de aprendizaje pueden dividirse básicamente en tres, atendiendo a como esta guiado este aprendizaje:

- **Aprendizaje supervisado**: se introducen unos valores de entrada a la red, y los valores de salida generados por esta se comparan con los valores de salida correctos. Si hay diferencias, se ajusta la red en consecuencia.
- **Aprendizaje de refuerzo**: se introducen valores de entrada, y lo único que se le indica a la red si las salidas que ha generado son correctas o incorrectas.
- **Aprendizaje no supervisado**: no existe ningún tipo de guía. De esta manera lo único que puede hacer la red es reconocer patrones en los datos de entrada y crear categorías a partir de estos patrones. Así cuando se le entre algún dato, después del entrenamiento, la red será capaz de clasificarlo e indicará en qué categoría lo ha catalogado.

Un modelo simple de red neuronal

Se considera una red neuronal formada por 2 capas:

- Una capa de entradas formada por 2 neuronas: 1 y 2.
- Una capa de salidas formada por una sola neurona: 3.

Las conexiones entre las dos neuronas de entrada y la neurona de salida presentan pesos sinápticos ajustables mediante el entrenamiento. Y a su vez, el valor umbral (U) de la neurona 3 puede ser ajustado como un peso sináptico más, al considerar $U = -W_0$ asociado a una neurona bias (siempre activa: $X_0 = 1$).

De esta manera el algoritmo de aprendizaje puede ajustar el umbral como si ajustara un peso sináptico más.

El esquema de este modelo sería el presentado en la figura 1.8:

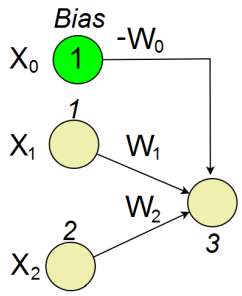

Figura 1.8: Modelo Simple, formado por tres neuronas(1,2 y 3), más una neurona bias

Y la condición de activación de la neurona 3 sería:

$$X_0W_0 + X_1W_1 + X_2W_2 > 0 \leftrightarrow \text{Activación} \tag{1.32}$$

Además, en este modelo, cada neurona de una capa "transmite" su estado de activación (0 o 1) a la siguiente capa de neuronas, y después deja de estar activa.

Ejemplo: Se tiene un conjunto de datos con datos faltantes, y se supone que los datos tienen un patrón MAR, la información se muestra en la Tabla 17:

Tabla 17. Conjunto de datos Incompletos de clasificación de puntos en el plano cartesiano

X_1	X_2	X_3	X_1	X_2	X_3
0	1	0	2	0	**?**
0	0	**?**	5	1	0
1	1	0	4	4	1
0	2	1	0	4	1
2	3	1	2	1	0
4	2	0	1	5	**?**
2	5	1	0	3	**?**
1	3	1	4	5	**?**
2	4	1			

Donde las variables X_1 y X_2 son puntos en el espacio y se clasifican en "0" ó "1", es decir que se están clasificando en 2 grupos, gráficamente se podrían representar así:

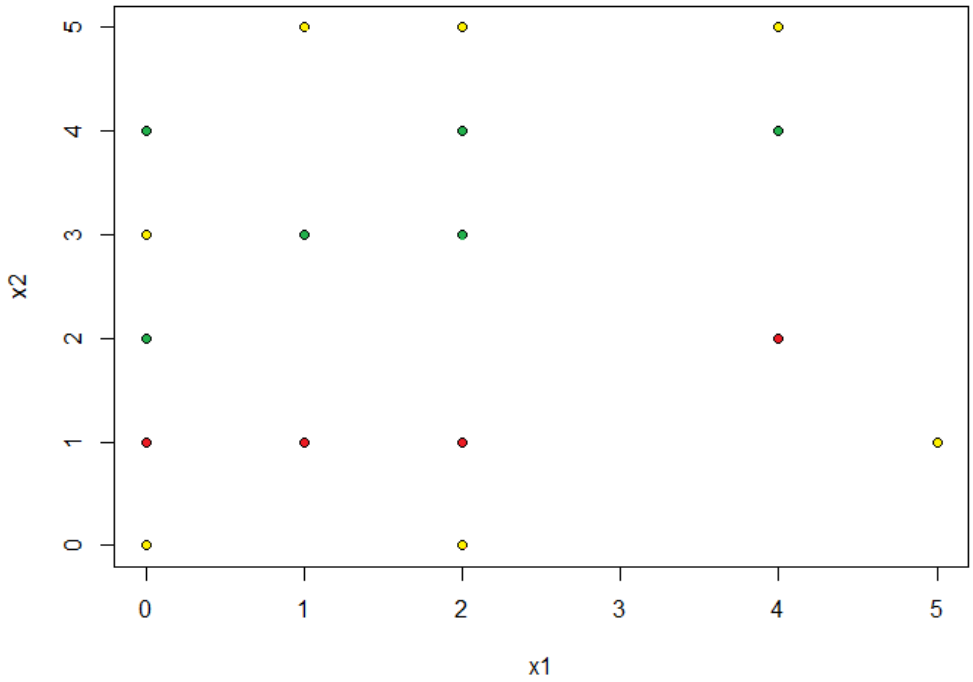

Figura 1.9: Donde los puntos color verde representan los clasificados como "1", los rojos como "0" y los puntos amarillos son los no clasificados, y la red neuronal será quien los clasifique.

Ahora bien para este ejemplo se utiliza la libraría nnet(que se encuentra en http://cran.r-project.org/src/contrib/nnet_7.3-5.tar.gz) del paquete estadístico R, para dar inicio al ejercicio práctico se debe introducir los datos en R de la siguiente forma:

```
>x1<-c(0,1,0,2,4,2,1,2,5,4,4)
>x2<-c(1,1,2,3,2,5,3,4,1,4,4)
>x3<-c(0,0,1,1,0,1,1,1,0,1,0)
>datos<-data.frame(x1,x2,x3)
```

Se debe tener en cuenta que los datos que se introducen en primera instancia son los completos para entrenar la red, luego se carga la librería y se inicia la creación de la red neuronal con aprendizaje no supervisado:

```
>library(nnet)
>and.nn<-nnet(x3~x1+x2, datos, size=0, decay=1e-4, linout=T,skip=T, maxit=1000,
 Hess=T)
>summary(and.nn)
```

En la instrucción anterior se puede ver que se tiene un parecido con la creación de un modelo de regresión, ya que la variable X_3, en este caso, depende de X_1 y X_2 se ha buscado ser lo más simple posible por lo que en $size = 0$ se está especificando que no se busca ninguna neurona capa oculta, esto se quiere decir que se esta creando una red neuronal 2-0-1. Se obtienen los pesos siguientes:

```
a 2-0-1 network with 3 weights
options were - skip-layer connections  linear output units  decay=1e-04
b->o i1->o i2->o
0.13 -0.13  0.26
```

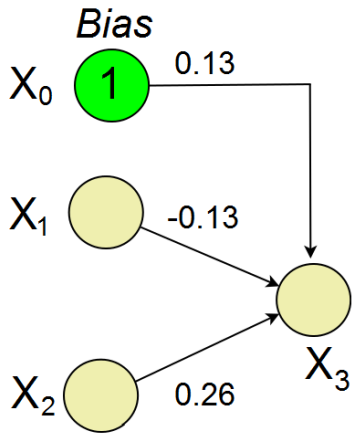

Figura 1.10: Modelo obtenido con los datos de entrenamiento, de las variables X_1, X_2 y X_3

Ahora ya con los pesos de las neuronas, se procede a poner en práctica la red con los datos incompletos, en primer lugar se deben introducir mediante el código siguiente:

```
>x1<-c(0,2,1,0,4)
>x2<-c(0,0,5,3,5)
>x3<-c(NA,NA,NA,NA,NA)
>pruebadatos<-data.frame(x1,x2,x3)
```

Con los datos de prueba se crean predicciones y se observa la efectividad de la red:

```
>predicciones=predict(and.nn,newdata=pruebadatos)
>round(predicciones, 0)
```

En la última parte del código se redondearon los datos para obtener datos enteros ya que la clasificación en este caso es binaria se obtuvieron los siguientes resultados:

```
> round(predicciones, 0)
  [,1]
1    0
2    0
3    1
4    1
5    1
```

Con lo que los datos quedan bien predichos con una efectividad del 100 %, se debe recordar que ya que este es un procedimiento estocástico si se repite n-veces hay una probabilidad que los pesos de las neuronas varíen y sucede más a menudo cuando hay neuronas en las capas ocultas.

1.4.3.2. Imputación por Regresión Aleatoria o Estocástica

En este método se hace primero un procedimiento de regresión, luego un término residual es adicionado para imputar los valores de y. Este término de error puede ser obtenido de diferentes maneras, una de ellas es a través de los residuos del modelo de regresión, generado con registros completos, eligiendo uno de éstos residuos aleatoriamente, con lo que la ecuación para el cálculo de un y_i viene dada por:

$$\widehat{y_i} = \widetilde{\beta_0} + \sum_{j=1}^{r} \widetilde{\beta_j} x_j + Z \tag{1.33}$$

Donde: $\widetilde{\beta} = (X'X)^{-1}(X'Y)$;

Z=Representa el residual distribuido normalmente $N(0, S^2)$.

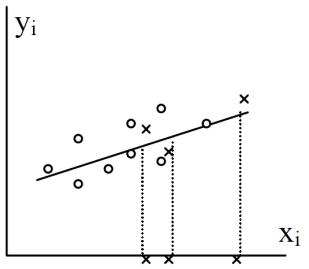

Figura 1.11: Imputación por Regresión Aleatoria

Retomando los datos de la tabla 5, y al hacer el cambio en la función de creación de los datos, con base en el modelo de regresión, agregando el elemento estocástico que viene a ser un residuo distribuido normalmente, entonces la secuencia del código en R quedará de la siguiente manera:

```
>library("MASS")
>ImputacionRegA<-function(x) {
  n=length(x[[2]])
  nx<- na.omit(x)
  reg<-lm(nx[[2]]~nx[[1]])
  residuos.est<-stdres(reg)
  b<-is.na(x[[2]])
  for (i in 1:n) {if(b[i]==TRUE){
        residuo<-sample(residuos.est,1)
```

```
      x[i,2]<-reg$coefficients[[1]]+x[i,1]*reg$coefficients[[2]]+residuo[[1]]}}
  cat("Modelo de Regresión: y=",reg$coefficients[[1]],"+",reg$coefficients[[2]],
  "*x \n")
   x }}
```

y para insertar y utilizar la función anterior se ejecuta el siguiente código:

```
> peso<-c(82,75,70,68,44,NA,72,85,95,70,75,59,69,68,75,70,NA,57,63,
  80,NA,54,54)
> estatura<-c(185,185,180,178,159,172,176,183,185,179,186,169,176,
  176,174,177,170,161,170,190,185,162,165)
>datos<-data.frame(estatura,peso)
>ImputacionRegA(datos)
```

Se sabe que al ejecutar el código una y otra vez siempre el resultado será distinto por el componente aleatorio, en una iteración se obtuvieron mostrados en la Tabla 18.

Tabla 18. Conjunto de datos Completos Estatura(cm)-Peso(Kg)

Estatura(cm)	Peso (Kg)	Estatura(cm)	Peso (Kg)
185	85	170	**62.2911**
185	75	176	68
180	70	174	75
178	68	177	70
159	44	170	68
172	**65.2324**	161	57
176	72	170	63
183	85	190	80
185	95	185	**79.3877**
179	70	162	54
186	75	165	54
169	59		

y si se compara la tabla 18 con la tabla 6 de la imputación por regresión que no es estocástica se puede ver la diferencia claramente en la variación, como ilustración se observa la figura 1.12 donde se comparan ambos métodos gráficamente, los puntos rojos representan los datos imputados.

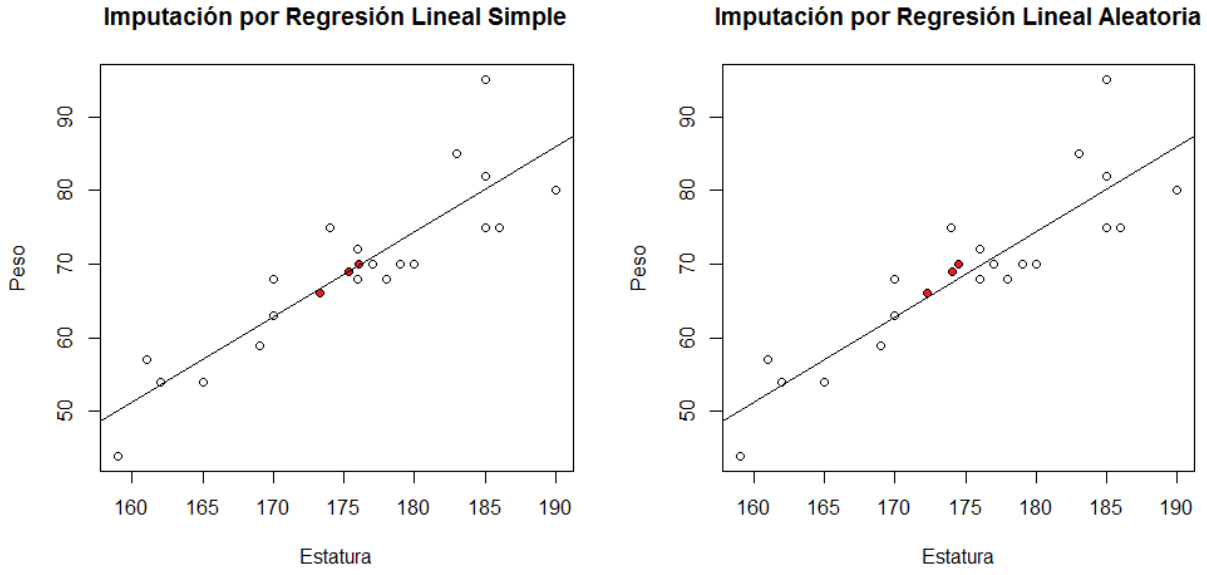

Figura 1.12: Comparación entre imputación por medio de regresión lineal simple y regresión aleatoria

1.4.3.3. Imputación Aleatoria de un Caso Seleccionado

La imputación de un valor aleatorio a los datos consiste en reemplazar cada valor ausente por un valor aleatorio. Este procedimiento puede distinguir dos formas de generar el valor aleatorio:

1. Generación de un valor aleatorio dentro del rango de valores de la variable, asignando la misma probabilidad a todos los valores (VADU, Valor Aleatorio de una Distribución de probabilidad Uniforme)

2. Generación de un valor aleatorio a partir de la función de probabilidad que caracteriza la variable (binomial, multinomial, normal, etc.) (VADE, valor aleatorio de una distribución de probabilidad estimada para la variable).

Para tener clara esta técnica se usarán nuevamente los datos de la Tabla 15, que corresponden a un conjunto de datos incompletos obtenidos aleatoriamente de una distribución normal con $\mu = 5$, se parte del supuesto que no se conoce la distribución de estos datos, entonces para encontrar los datos perdidos dentro del rango de la variable se hace bajo la siguiente sintaxis:

```
> x<-c(5.226416,4.246948,5.719032,5.062461,6.621925,NA,4.720129,NA,4.199522,
 5.607660,NA,4.401418,5.273086,4.567361,4.205563,4.353384,3.852736,NA,
 6.328042,NA)

#VADU
> N<-length(x)
> Xobs<-x[!is.na(x)]
> Xmis<-x[is.na(x)]
> n<-length(Xmis)
> i=min(Xobs)
> f=max(Xobs)
> muestra<- seq(from = i, to = f, by = 0.0001)
> vade<-sample(muestra,n)
> sust<-vade
#Sustituir valores
> b<-is.na(x)
> j=1
> for (i in 1:N){
     if(b[i]==TRUE){
     x[i]<-sust[j]
     j=j+1}
     }
> x  # vector completo
```

En este caso se obtuvo la siguiente salida:

```
> x  # vector completo
 [1] 5.226416 4.246948 5.719032 5.062461 6.621925 4.954736 4.720129 6.584136
```

```
 [9] 4.199522 5.607660 3.893636 4.401418 5.273086 4.567361 4.205563 4.353384
[17] 3.852736 6.104536 6.328042 5.882236
```

Ahora no se conoce la distribución de la variable se empezará tratando de caracterizar la variable, iniciando por un histograma de los datos:

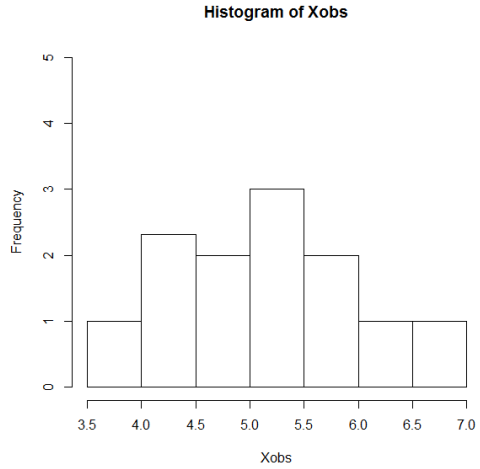

Figura 1.13: Histograma de los datos completos

Con ese histograma se intuye que se trata de una distribución normal pero para reforzar este resultado también se puede ver otro tipo de gráfico como el de normalidad:

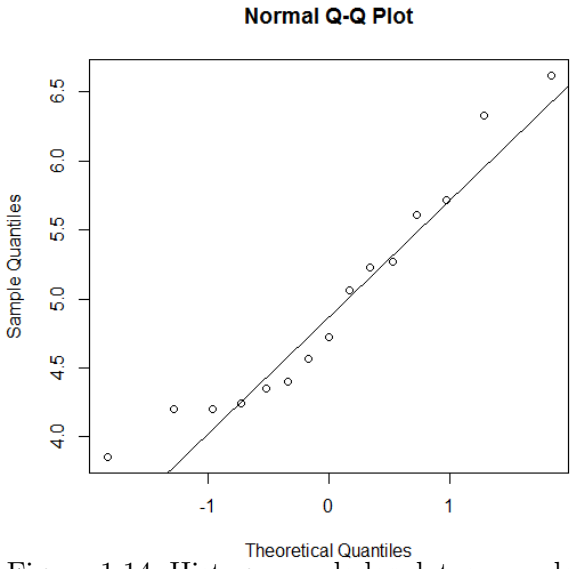

Figura 1.14: Histograma de los datos completos

Ya con la Figura 1.13 y Figura 1.14 se intuye que los datos siguen una distribución normal pero no está de más una prueba de hipótesis que respalde lo observado, para este fin se aplica la prueba de normalidad Shapiro-Wilk, donde la hipótesis a contrastar son:

H_0:Los datos provienen de una distribución Normal.

H_1:Los datos no provienen de una distribución Normal.

y se aplica el test con el código siguiente:

```
> shapiro.test(Xobs)
```

se obtiene la siguiente salida:

```
> shapiro.test(Xobs)

        Shapiro-Wilk normality test

data:  Xobs
W = 0.9278, p-value = 0.2531
```

Tomando el p-valor de referencia para contrastar la hipótesis con un $\alpha = 0.05$ no existe suficiente evidencia en los datos para rechazar la hipótesis nula, entonces los datos se supone que proviene de una distribución normal, ahora se procede a utilizar la imputación estocástica VADE para la variable. Para este ejemplo se creó la siguiente sintaxis:

```
#VADE
> x<-c(5.226416,4.246948,5.719032,5.062461,6.621925,NA,4.720129,NA,4.199522,
 5.607660,NA,4.401418,5.273086,4.567361,4.205563,4.353384,3.852736,NA,
 6.328042,NA)
> Xobs<-x[!is.na(x)]
```

```
> Xmis<-x[is.na(x)]
> sust<-rnorm(length(Xmis),mean=mean(Xobs))
> sust
#Sustituir valores
> b<-is.na(x)
> j=1
> for (i in 1:N){
     if(b[i]==TRUE){
     x[i]<-sust[j]
     j=j+1}
     }
> x  # vector completo
```

y se obtuvo la siguiente salida con una ejecución:

```
> x
 [1] 5.226416 4.246948 5.719032 5.062461 6.621925 3.935090 4.720129 4.604740 4.199522
[10] 5.607660 4.730841 4.401418 5.273086 4.567361 4.205563 4.353384 3.852736 5.140833
[19] 6.328042 5.511266
```

Tanto la imputación VADU como la imputación VADE son opciones poco utilizadas en la práctica, ya que acostumbran a provocar un importante sesgo, especialmente en el caso de variables con una elevada dispersión.

1.4.3.4. Bootstrap, Imputación por Métodos Bayesianos (ABB)

El bootstrap es un método de remuestreo propuesto por Bradley Efron en 1979. Se utiliza para aproximar la distribución en el muestreo de un estadístico. Se usa frecuentemente para aproximar el sesgo o la varianza de un análisis estadístico, así como para construir intervalos de confianza o realizar contrastes de hipótesis sobre parámetros de interés. Se emplea para imputar variables binarias por lo que sigue, el mecanismo perdida se supone que es ignorable en el sentido de que $(y_1, ..., y_r)$ es una muestra aleatoria de $(y_1, ..., y_n)$ y cada y_i, $i = 1, 2, ..., n$ es independiente e idénticamente distribuido con μ y

σ^2. Para inferencias con respecto a μ, la ABB se lleva a cabo a través de IM(Imputación Múltiple) con los siguientes pasos:

1. Se extraen r observaciones independientes con sustitución de los datos observados, $(y_1, ..., y_r)$.
2. Se extraen $n-r$ valores perdidos con reemplazo de la y_i extraídos en (1). Los valores imputados son denotado como $(\bar{y}_{r+1}, .., \bar{y}_n)$
3. A partir de los pasos 1 y 2, la media estimada es:

$$\widehat{\mu} = n^{-1}\left(\sum_{i=1}^{r} y_i + \sum_{r+1=1}^{n} \bar{y}_i\right) \tag{1.34}$$

   Y la varianza estimada es:

$$U = \widehat{V}ar(\widehat{\mu}) = [n(n-1)]^{-1}\left[\sum_{i=1}^{r}(y_i - \widehat{\mu})^2 + \sum_{r+1=1}^{n}(\bar{y}_i - \widehat{\mu})^2\right] \tag{1.35}$$

4. Repita los pasos 1 y 2 de forma independiente m veces. Analice el rendimiento de los m conjuntos de datos imputados $\widehat{\mu}$ y $U^k = \widehat{V}ar(\widehat{\mu}^k)$, para $k = 1, ..., m$.

   El estimador de IM de μ es

$$\widehat{\mu}_{IM} = m^{-1}\sum_{k=1}^{m}\widehat{\mu}^k \tag{1.36}$$

y la varianza estimada es:

$$\widehat{V} = \widehat{W} + \left(\frac{m+1}{m}\right)\widehat{B} \tag{1.37}$$

Donde: $\widehat{W} = m^{-1}\sum_{k=1}^{m} u^m$ es el promedio dentro de la imputación de la varianza y $\widehat{B} = (m-1)^{-1}\sum_{k=1}^{m}(\widehat{\mu}^m - \widehat{\mu}_{IM})^2$ es la varianza entre la imputación.

Para ilustrar esta técnica de imputación se creó el siguiente código en R, como en los ejemplos de las técnicas anteriores se encontrará en anexos una función automatizada que realiza esta mismo proceso, pero aquí se explicará paso a paso cada parte de la sintaxis.

Tabla 19. Conjunto de datos Incompletos Obtenidos al azar

4	5	3	7	4	1	8	4
1	**?**	4	6	1	**?**	6	**?**
7	2	7	2	1			

Como primer paso se introducen los datos artificiales con mecanismo de pérdida ignorable de la tabla 19 en R.

```
> Y<-c(4,5,3,7,4,1,8,4,1,NA,4,6,1,NA,6,NA,7,2,7,2,1)
```

En primera instancia se debe separar los valores completos y los faltantes en el siguiente código Y_{obs} se almacenan los datos completos para extraer r observaciones que se especifican en el paso 1, de la imputación múltiple explicada anteriormente.

```
>   library(LaplacesDemon)
>   m=7
>   Yobs<-Y[!is.na(Y)]
>   Ymis<-Y[is.na(Y)]
>   k <- length(Yobs)
>   r <- length(Ymis)
>   n <- length(c(Yobs, Ymis))
>   bootstrap<-ABB(Y,m)
```

Se observa una constante m que en este apartado se fija por orden lógico en la sintaxis, pero en el procedimiento de la imputación múltiple es el paso 4 que es el número de veces que se repite el paso 1 y 2, se fijo con $m = 7$ pero independientemente se puede incrementar para hacer unos parámetros más robustos, asimismo se utilizó una librería de R, la cual nos facilita la extracción de los elementos, es decir el remuestreo, dicha librería es *LaplacesDemon* (http://cran.r-project.org/web/packages/LaplacesDemon/index.html) ahora para encontrar la media estimada $\hat{\mu}$ y la varianza estimada $U = \hat{V}ar(\hat{\mu})$ del paso 3, se creo el siguiente ciclo:

```
> for(i in 1:m){
     sumobs<-0
     sumboot<-0
```

```
    mu_estimada[i]<-(1/n)*(sum(Yobs)+sum(bootstrap[[i]]))
    cat("Mu",i," ",mu_estimada[i], "\n")
    boot<-bootstrap[[i]]
    for (j in 1:k){sumobs<-sumobs+(Yobs[j]-mu_estimada[i])^2}
    for (l in 1:r){sumboot<-sumboot+(boot[l]-mu_estimada[i])^2}
    U[i]<-(1/(n*(n-1)))*(sumobs+sumboot)
    cat("U",i," ",U[i], "\n")
  }
```

Y que con el código anterior se obtuvo en este caso particular, para $m = 7$ el total de m medias estimadas $\hat{\mu}$ y las varianzas estimadas $U = \hat{V}ar(\hat{\mu})$ ahora para completar la técnica se necesita el estimador de IM de μ es $\widehat{\mu}_{IM}$ y la varianza estimada es: $\widehat{V}$ que lo se obtiene así:

```
> mu_IM=mean(mu_estimada)
> cat("mu_IM",mu_IM, "\n")
>  v_estimada<-mean(U)+((m+1)/m)*var(mu_estimada)
>  cat("Varianza(mu_IM)",v_estimada, "\n")
```

y se obtuvo la siguiente salida:

```
Mu 1   4.095238
U 1   0.2614512
Mu 2   4.190476
U 2   0.293424
Mu 3   3.809524
U 3   0.2600907
Mu 4   4.047619
U 4   0.2498866
Mu 5   3.857143
U 5   0.2537415
```

```
Mu 6   3.952381
U 6   0.2403628
Mu 7   4.190476
U 7   0.293424
>mu_IM=mean(mu_estimada)
>cat("mu_IM",mu_IM, "\n")
mu_IM 4.020408
>v_estimada<-mean(U)+((m+1)/m)*var(mu_estimada)
>cat("Varianza(mu_IM)",v_estimada, "\n")
Varianza(mu_IM) 0.2912814
```

Cabe recordar que si se vuelve a ejecutar no darán estos resultados por ser una técnica estocástica.

Resumen de Técnicas

Tabla 20. Resumen de Técnicas

Método	**Supuestos y patrón de datos faltantes**	**Aplicación**	**Ventajas**	**Desventajas**
Hot-Deck	Patrón de datos faltantes condicionado (MNAR)	En las encuestas de hogares continuas se utiliza información pasada de la misma unidad de observación.	Se hace uso de información de las mismas unidades de observación.	Carecen de un mecanismo de probabilidad, Existe la posibilidad de usar varias veces a una misma unidad que ya ha respondido.
Imputación haciendo uso de Medias incondicionadas	Patrón de datos faltantes MCAR.	Las observaciones faltantes se reemplazan por el valor medio de la variable de análisis.	Fácil de entender y aplicar	El uso de este método, afectará la correlación entre la variable imputada y cualquiera otra, reduciendo su variabilidad.
Imputación por medias condicionadas	Patrón de datos faltantes MCAR.	Se divide la base de datos en subgrupos Utilizando variables correlacionadas. Las observaciones faltantes en el subgrupo de interés se reemplazan por el valor medio de la variable	Fácil de entender y aplicar.	El uso de este método, afectará la correlación entre la variable imputada y cualquiera otra, reduciendo su variabilidad.
Imputación usando la mediana	Patrón de datos faltantes MCAR.	En este caso el valor faltante de una característica dada es reemplazado por la mediana de todos los valores conocidos de ese atributo.	Una medida alternativa de tendencia central representa mejor la distribución subyacente y por tanto una mejor estimación para los valores faltantes.	
Imputación por Regresión	Patrón de datos faltantes MAR. Se requiere especificar un modelo en donde las covariables estén altamente correlacionadas con la variable a imputar.	Se divide la base de datos en subgrupos utilizando variables correlacionadas. Los valores faltantes se sustituyen con el valor medio estimado por la regresión .	Frente a la imputación mediante la media, este método incorpora la información que sobre Y_i contienen el resto de variables.	Sobreestima la asociación entre variables, y en modelos de regresión múltiple puede sobredimensionar el valor del coeficiente de determinación R^2.

Series Temporales	Patrón de datos faltantes MCAR.	Se utilizan comúnmente las metodologías Holt-Winters y Box-Jenkins, se pueden obtener como pronósticos hacia delante o hacia atrás.		Una Desventaja a este método es que deben existir suficientes datos antes, entre o después de los valores que faltan.
Imputación usando el vecino más cercano	Patrón de datos faltantes MAR.	Se sustituye a cada valor perdido por el de un donante elegido a partir de una determinada distancia calculada a través de una variable con información completa.	En comparación a los métodos paramétricos los métodos del vecino más cercano no hacen suposiciones de la distribución del modelo y pueden retener la estructura varianza/covarianza en la salida.	
Imputación por Máxima Verosimilitud/ Algoritmo EM	Patrón de datos faltantes MAR.	Se resume el procedimiento para estimar los parámetros de un modelo utilizando una muestra de datos	Genera estimaciones robustas basadas en la muestra observada. No efectúa simulaciones.	No siempre están disponibles. Hay que programar el algoritmo que se desea aplicar.
Imputación por redes Neuronales	Patrón de datos faltantes MAR.	Reconocen patrones de los datos sin algún valor perdido para aplicarlo a bases de datos incompletas	No hacen suposiciones de la distribución del modelo y pueden retener la estructura varianza/covarianza en la salida.	Carece de un mecanismo de probabilidad.
Imputación por Regresión Aleatoria	Patrón de datos faltantes MAR. Se requiere especificar un modelo en donde las covariables estén altamente correlacionadas con la variable a imputar	Se divide la base de datos en subgrupos utilizando variables correlacionadas. Los valores faltantes se sustituyen con el valor medio estimado por la regresión más un valor aleatorio .	Frente a la imputación mediante la media, este método incorpora la información que sobre Y_i contienen el resto de variables.	La asociación entre variables, y en modelos de regresión múltiple puede sobredimensionar el valor del coeficiente de determinación R^2.
Imputación Aleatoria de un Caso Seleccionado	Patrón de datos faltantes ignorable	Generación de un valor aleatorio dentro del rango de valores de la variable o de una distribución según sea el caso.	Fácil implementación.	Acostumbran a provocar un importante sesgo
Bootstrap, Imputación por Métodos Bayesianos (ABB)	Patrón de datos faltantes ignorable	Método de remuestreo.	Se hace uso de información de las mismas unidades de observación se obtienen estimadores robustos.	

Capítulo 2

Guía Metodólogica

2.1. Como seleccionar el método adecuado de imputación

Seleccionar un método de imputación es una decisión de gran importancia, ya que para un conjunto de datos determinado, algunas técnicas de imputación podrán dar mejores aproximaciones a los valores verdaderos que otras. La selección del método de imputación adecuado dependerá del tipo de datos, tamaño del archivo, mecanismo de datos faltantes, características específicas de la población, software disponible, distribuciones de frecuencias de cada variable marginal o conjunta, entre otros factores.

Puede suceder que la técnica de imputación seleccionada sea adecuada para algunas variables pero para otras no y será decisión del investigador seleccionar el método que menos afecte a las estimaciones de las variables más importantes. Fellegi y Holt (1971), plantean que: "La técnica de imputación seleccionada debe superar las reglas de validación, cambiando lo menos posible los registros, manteniendo la frecuencia de la estructura de los datos."

Goicoechea (2002), resume los criterios a tomar en consideración al momento de seleccionar el modelo de imputación adecuado:

1. La importancia de la variable a imputar. Si la variable es de elevada importancia, es natural que se elija más cuidadosamente la técnica de imputación a aplicar.
2. Tipo de variable a imputar. Si es continua o categórica, tanto nominal como ordinal. Teniendo en cuenta para el primer grupo el intervalo para el cual está definido y para los segundos las distintas categorías de la variable.

3. Parámetros que se desean estimar. En el caso que solamente interese conocer el valor medio y el total, se pueden aplicar los métodos deterministicos más sencillos. En el caso en el que se requiera la distribución de frecuencias de la variable, la varianza y asociaciones entre las distintas variables, se deben emplear métodos mas elaborados y analizar el fichero de datos a profundidad; el problema en este caso se incrementa cuando hay una elevada tasa de no respuesta.
4. Tasas de no respuesta. No se debe abusar de los métodos de imputación y menos cuando se tiene una elevada tasa de no respuesta de la cual no se conoce el mecanismo.
5. Información auxiliar disponible. La imputación puede mejorar al emplear información auxiliar disponible. En el caso de no disponer de información auxiliar una técnica recomendada aplicar es la imputación aleatoria Hot Deck.

Con base en los criterios mencionados, el proceso a seguir en la toma de decisión y aplicación de una técnica de imputación adecuada es mostrada en la Figura 2.1.

Para la ejecución de la guía metodológica se definen a continuación cada uno de los pasos:

Paso 1. Base de datos con valores faltantes. Es importante tener en cuenta la estructura de la base de datos, teniendo en consideración la cantidad de datos faltantes para la variable que se desea imputar, pues la eficiencia de las técnicas de imputación depende en gran medida de este factor.

Paso 2. Identificación del patrón de pérdida de datos. En este paso se ubican los puntos o fragmentos de la variable en los cuales se focaliza la perdida de información, para tener una idea general de las particularidades que presenta la serie de datos faltantes como repetitividad, periodo, estacionalidad, entre otras, que podrían ayudar a establecer posteriormente el mecanismo de datos faltantes.

Paso 3. Cálculo de la tasa de no respuesta presente en los datos. En este paso el investigador deberá estimar el valor de la tasa de no respuesta como el porcentaje de espacios vacios en la base de datos para cada variable de acuerdo a la cantidad de registros totales de la base de datos para esa variable, así:

$$\text{Tasa de no respuesta } X_i = \frac{\text{cantidad de espacios vacios para la variable } x_i}{\text{total de registros completos y vacios de la variable } x_i} * 100\,\%$$

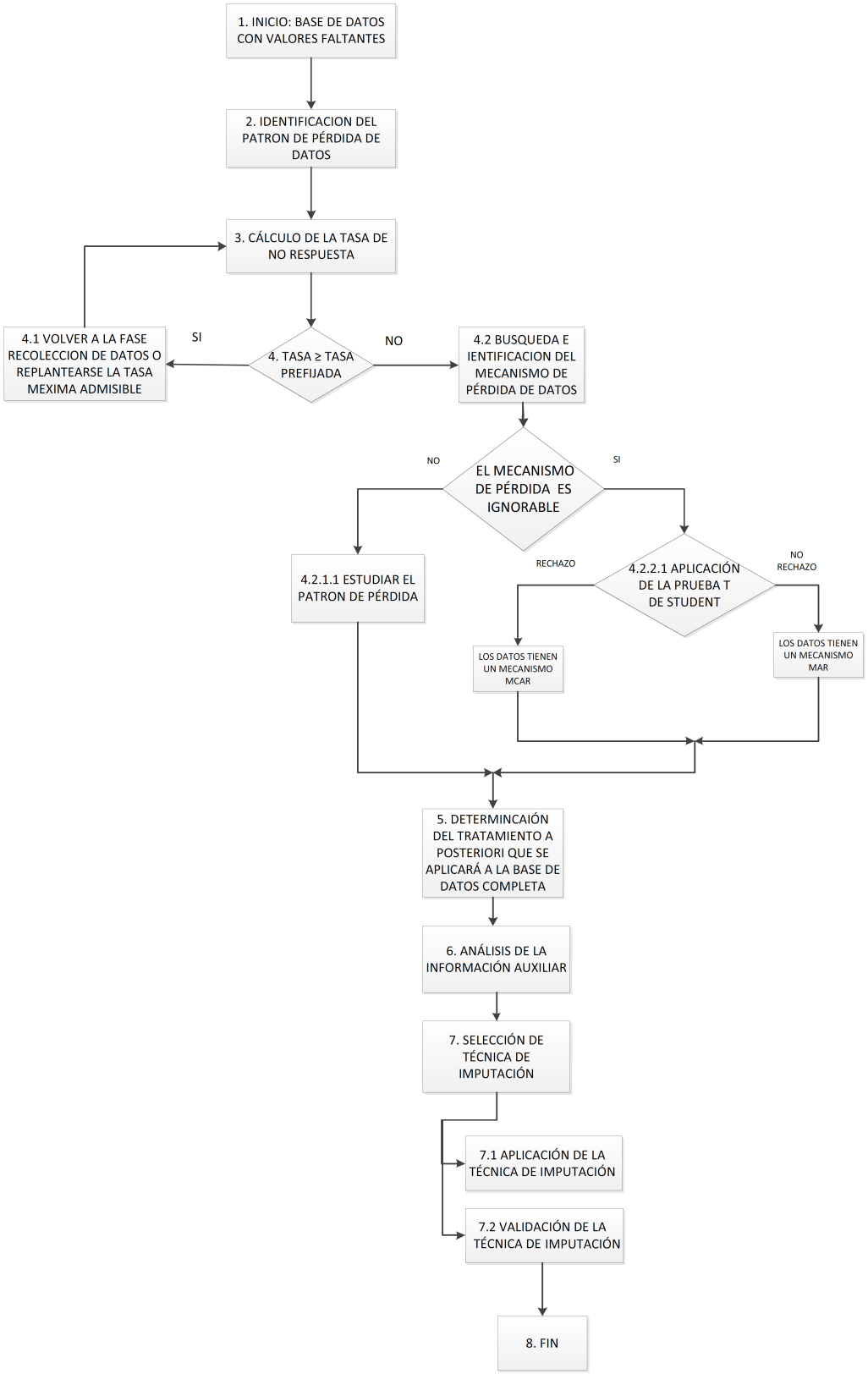

Figura 2.1: Guía para el proceso de Imputación

Paso 4. Comparación de la tasa de no respuesta con el valor prefijado. Es claro que no existe un criterio uniforme entre la comunidad académica que determine la tasa máxima de no respuesta admisible en un trabajo de investigación y que este valor en muchos casos se determina siguiendo el criterio del investigador con base en su experiencia en el tema o al nivel máximo admisible que este desea considerar y siendo que este valor debe ser fijado al inicio de la investigación en este punto de la misma en el que se conoce la cantidad de espacios vacios es posible determinar si esta tasa es acorde al máximo admisible planteado al inicio por el investigador, el resultado de este análisis conducirá a dos caminos diferentes los cuales se definen a continuación:

4.1 Caso 1: La tasa de no respuesta es mayor al nivel prefijado por el investigador En este caso el investigador deberá volver a la fase de campo a recolectar mas información que le permita reducir el tamaño de la tasa de no respuesta y en caso de no ser posible deberá replantear una tasa de pérdida admisible acorde a la información disponible para el fenómeno en estudio.

4.2 Caso 2: La tasa de no respuesta es menor o igual al nivel prefijado por el investigador En este caso el investigador puede proceder al siguiente paso que es la búsqueda y determinación del mecanismo de pérdida de datos presente en su base de datos. **Búsqueda del mecanismo de pérdida de datos.** En este paso se debe realizar una caracterización de la variable en estudio, para conocer su distribución y las diferentes relaciones que pueda tener con el resto de las variables, para poder establecer de manera clara el mecanismo que genera la perdida de información en la matriz de datos, ya que dependiendo de si el mecanismo identificado es ignorable o no ignorable así será la gama de técnicas de imputación que se podrán aplicar posteriormente para obtener un conjunto de datos completos. En este caso el investigador nuevamente tendrá dos caminos a tomar, dependiendo del tipo de mecanismo que identifique, así:

Paso 4.2.1: El mecanismo de pérdida de datos es no ignorable (NMAR) En este caso la ausencia de los datos depende de la variable perdida, esto traerá como consecuencia estudiar el patrón de pérdida de los datos ausentes para luego imputar tomando en cuenta dicho patrón.

Paso 4.2.2: El mecanismo de pérdida de datos es ignorable (MAR o MCAR) En este caso el investigador puede encontrarse con dos mecanismos de pérdida diferentes MAR ó MCAR y para determinar cuál de ellos se apega a su investigación de verá aplicar una prueba T de Student univariante o multivariante según sea el fenómeno en estudio para la identificación del mismo de acuerdo a lo que se planteo en los apartados 1.3.1.1 ó 1.3.1.2.

Paso 5. Definir el procedimiento a posteriori que se aplicara al conjunto de datos. En este punto es de vital importancia conocer que se desea hacer con la base de datos completa, es decir que parámetros se desean estimar o qué tipo de análisis se desea aplicar, pues cada técnica de imputación es compatible con diferentes procesos de análisis que se aplican sobre los conjuntos de datos, por lo que al encontrar esta compatibilidad se maximizara la eficiencia de la técnica de imputación y se obtendrán resultados que sean lo más cercanos posibles al valor real de la variable imputada.

Paso 6. Análisis de información auxiliar. En algunos casos el investigador tiene conocimiento a priori sobre factores externos que pueden afectar los valores que toman las variables, por lo que es importante tener en cuenta esta información que en algunos casos contribuye a acercar los datos estimados a los verdaderos valores de la variable imputada.

Paso 7. Selección de la técnica de imputación. Una vez se ha determinado el mecanismo de pérdida de datos ,el tratamiento posterior que se desea aplicar y la información auxiliar que pueda influir en los datos, se debe elegir una técnica de imputación que se acople a estos resultados, para lograr estimaciones plausibles de los parámetros de interés. En caso que dos o más técnicas sean adecuadas, deberá evaluarse la posibilidad de aplicarlas para luego validar los resultados obtenidos mediante la comparación de los respectivos vectores de parámetros estimados para cada una de ellas.

Paso 7.1. Aplicación de la técnica de imputación. Una vez se ha seleccionado la técnica de imputación que ayuda a resolver el problema de falta de información en la variable de interés, se puede hacer uso de las rutinas para el Software R definidas a lo largo de este trabajo de investigación o emplear algún otro software que permita la aplicación de la técnica seleccionada a la base de datos para obtener un conjunto de datos completo. En caso de haber seleccionado más de una técnica de imputación deben desarrollarse todas en este paso para posteriormente validar la más idónea y emplear los valores obtenidos con esta técnica para completar la variable en estudio.

Paso 7.2. Validación de la técnica de imputación. En este paso se debe comparar el vector de parámetros obtenido de los valores observados originalmente en la variable, con el vector de parámetros de la variable con valores imputados, para observar que tan similares son, es deseable que estos no se alejen significativamente. En caso de haber aplicado más de una técnica de imputación se elegirá aquella cuyos parámetros se acerquen más a los valores de los parámetros calculados para los valores observados.

Paso 8. FIN. El proceso finaliza al sustituir los valores encontrados en los espacios vacíos de la base de datos y completar la variable en estudio.

Todo el proceso anterior se realiza para elegir un método de imputación que sea capaz de reproducir eficientemente un fichero de datos completos al cual se le pueda aplicar un análisis estadístico para datos completos.

A continuación se proponen una serie de medidas para obtener una buena imputación, el proceso de imputación debe:

- Resultar un valor imputado que sea lo más cercano posible al valor real.
- Para variables numéricas o categóricas ordinales, debe resultar una ordenación que relacione el valor imputado con el valor real o sea muy similar.
- Preservar la distribución de los valores reales.
- Producir parámetros insesgados e inferencias eficientes de la distribución de los valores reales.
- Y finalmente conducir a valores imputados que sean plausibles. Estas medidas dependen del tipo de variable que se estén considerando.

Capítulo 3

Aplicación Práctica de la Guía Metodológica

3.1. Introducción

El propósito del presente capítulo es ilustrar las técnicas de imputación para el manejo de valores incompletos en una matriz de datos, para lo cual, en los capítulos anteriores se definió que es "Imputación de datos", y una gran variedad de técnicas determinísticas y estocásticas para la implementación de la imputación atendiendo cada una de ellas a diferentes características que puedan presentar las variables de interés; además para ilustrar la implementación paso a paso de algunas de las técnicas, se hará uso de la base de datos de emanaciones Hidrogeoquímicas de la estación El Jabalí provenientes del Volcán de San Salvador para depurarla y validarla; ésta contiene registros desde Julio del año 2000 hasta Febrero de 2010, para un conjunto de veintiún variables de las que la mayoría no poseen más de 25 % de datos por lo que solo usaremos 12 de ellas y se imputarán aquellas que presenten una tasa de no respuesta menor o igual al 10 %, éstas variables son: Temperatura (T), pH in situ (tomado en el lugar de recolección de la muestra), Dureza y Alcalinidad, haciendo uso además de aquellas variables dentro de la base de datos que puedan aportan información por tener relación con alguna de las variables en estudio.

Es importante recordar que uno de los factores que contribuye en la elección de una técnica adecuada de imputación es la información a priori que posea el investigador de la base de datos, por lo que antes de implementar los pasos de la guía metodológica es recomendable explorar las causas de pérdida de información.
Por ejemplo, de la base de datos sujeto de estudio en esta investigación, se sabe que en algunos casos debido a la presencia de actividad volcánica en otros puntos del país era

necesario movilizar los esfuerzos de monitoreo hacia estas zonas que más lo demandaban y se dejaba de recolectar información en la estación El Jabalí.

En otros casos el monitoreo se veía interrumpido por la falta de equipo o recurso humano disponible, por lo anteriormente expuesto es válido aclarar que la imputación se realizará sobre aquellos valores perdidos presentes en los días de toma de muestra registrados, teniendo en cuenta que los muestreos no se realizaban de manera diaria y aunque en el planteamiento metodológico de los investigadores se proponía la realización de un muestreo mensual, este no se cumplió al pie de la letra a lo largo del período en estudio, pues por diversas causas los muestreos no se realizaron de manera sistemática con la misma amplitud de tiempo, sino más bien se realizaban de acuerdo a la disponibilidad y a las necesidades de monitoreo de todos los volcanes en observación, es así que por ejemplo para algunos años como el año 2000 solamente se reportan dos observaciones registradas en igual número de monitoreos, y el año que más observaciones reporta únicamente tiene diez observaciones siendo el año 2009, lo que pone de manifiesto las irregularidades en todo el período. Es de recordar que las técnicas de imputación buscan dar tratamiento a los errores de respuesta y de no respuesta, más no a los errores de cobertura de la muestra.

Una vez aclarado el corte temporal del fenómeno se establece que la base de datos a imputar está formada por 60 observaciones para cada una de las variables, recolectadas en igual número de muestras tomadas entre el año 2000 y el año 2010. Además dada la irregularidad en el espaciado de la toma de muestra es preciso aclarar que sobre esta base de datos no es posible aplicar técnicas que involucren series de tiempo, ya que se incumplen los supuestos de dicho método y su implementación sería errónea y traería consigo resultados inválidos.

3.2. Implementación de la guía metodológica

3.2.1. Base de datos con valores faltantes

En este apartado se presenta la base de datos que se tratará en este capítulo la cual se observa en la Tabla 21.

Los componentes Hidrogeoquímicos son:

- **pH in situ:** Corresponde a la medida del potencial de hidrógeno del agua que se toma en el lugar al momento de recoger la muestra haciendo uso de una sonda

multiparámetros que registra esta medición, esta magnitud no tiene unidades de medida, por lo cual se considera adimensional. Se define como el logaritmo negativo en base 10 de la actividad de los iones hidrógeno:
$pH = -log_{10}[a_{H+}]$

- **Ca:** Corresponde a la medida del calcio del agua para obtener esta medición se hace uso del laboratorio de aguas del SNET, las unidades de medición de esta variable son los miligramos por litro (mg/l).
- **Na:** Es un metal alcalino blando, untuoso, de color plateado, muy abundante en la naturaleza, encontrándose en la sal marina y el mineral halita. Es muy reactivo, arde con llama amarilla, se oxida en presencia de oxígeno y reacciona violentamente con el agua.
- **Dureza** Es la capacidad del agua para producir espuma, es decir es la medida del consumo de jabón (detergente) del agua. Los minerales removidos por el jabón se vuelven espuma (Romero, 1996). Esta propiedad es causada por los iones metálicos divalentes, es decir los cationes de Calcio (Ca^{+}) y Magnesio (Mg^{++}). Esta propiedad es muy mportante en acuíferos cársticos que tienen la característica de formarse en rocas carbónicas como la dolomita y la caliza, lo cual, hace que se diluyan rocas formando carbonatos y bicarbonatos. La dureza se puede clasificar en dos grandes grupos: la carbonatada y la no carbonatada. La dureza carbonatada es también llamada temporal, porque se remueve con evaporación y precipitado del calcio y del magnesio. Esta se mide en términos del carbonato de calcio (mg/L). La dureza no carbonatada es igual a la diferencia entre la dureza total y la carbonatada. Esta indica la cantidad de calcio y magnesio combinados con sulfatos, cloruros, nitratos y algunas veces hierro; esta clase de dureza no se puede remover por evaporación. El agua se puede clasificar como dura o blanda según la normativa. Para efectos prácticos aguas con dureza menor de 50 mg/L son consideradas blanda, con dureza entre los 50 y los 150 mg/L, son de uso no objetable, y para durezas mayores de 150 mg/L se consideran duras. La concentración de carbonato de calcio es cinco veces mayor que la de carbonato de Magnesio (Driscoll, 1986).
- **Alcalinidad** La alcalinidad del agua puede definirse como la capacidad del agua para neutralizar ácidos, para reaccionar con iones hidrógeno, para aceptar protones, o como la medida del contenido total de sustancias alcalinas (OH^{-}). La determinación de la alcalinidad total y de las distintas formas de alcalinidad es importante en los procesos de coagulación química, ablandamiento, control de corrosión y evaluación de la capacidad tamponamiento del agua. En el ablandamiento del agua por métodos de precipitación, la alcalinidad es un dato necesario para el cálculo de la cantidad de cal y carbonato de sodio necesaria para el proceso. En aguas naturales

la alcalinidad es debida generalmente a la presencia de tres clases de compuestos: bicarbonatos, carbonatos e hidróxidos. En algunas aguas es posible encontrar otras clases de compuestos (boratos, silicatos y fosfatos) que contribuyen a su alcalinidad; sin embargo en la práctica la contribución de estos es insignificante y puede ignorarse. La alcalinidad del agua puede determinarse por titulación con ácido sulfúrico 0.02 N y se expresa como mg/L de carbonato de calcio equivalente a la alcalinidad determinada. Los iones H^+ procedentes de la solución 0.02 N de H_2SO_4 neutralizan los iones de OH^- libres y los disociados por concepto de la hidrólisis de carbonatos y bicarbonatos (Romero, 1996).

- **Hierro y Manganeso**: Tanto el hierro como el manganeso crean problemas en suministros de agua (Romero, 1996). En general estos problemas son más comunes en aguas subterráneas y en aguas de hipolimnio anaeróbico de lagos estratificados. El hierro existe en suelos y minerales principalmente como óxido férrico insoluble y sulfuro dehierro (FeS_2, pirita). En algunas áreas se presenta también como carbonato ferroso (siderita), la cual es muy poco soluble. Como las aguas subterráneas contienen cantidades apreciables de CO_2, producidas por la oxidación bacterial de la materia orgánica con la cual el agua entra en contacto, se puede disolver cantidades apreciables de carbonato ferroso. Sin embargo, los problemas con el hierro predominan cuando éste está presente en el suelo como compuestos férricos insolubles. Si existe oxígeno disuelto en el agua, la solución del hierro en tales suelos con el agua no ocurre, aún en presencia de suficiente CO_2, pero en condiciones anaeróbicas, el hierro férrico es reducido a hierro ferroso y la solución ocurre sin dificultad.

  Las aguas subterráneas que contienen cantidades apreciables de hierro o, manganeso carecen siempre de Oxígeno disuelto y poseen un alto contenido de CO_2. El hierro y el manganeso están presentes como Fe^{++} y Mn^{++}. El alto contenido de CO_2 indica que ha existido oxidación bacterial de la materia orgánica; la ausencia de oxígeno disuelto indica que se han desarrollado condiciones anaeróbicas. A los pozos que producen agua de buena calidad, con bajo contenido de hierro y manganeso, se les deteriora la calidad del agua cuando se han descargado residuos orgánicos sobre el suelo alrededor del pozo y se generan condiciones anaeróbicas.

- **Cloruros**: El ion cloruro es una de las especies de cloro de importancia en aguas. Las aguas subterráneas en áreasadyacentes al océano están en equilibrio hidrostático con el agua de mar. Un sobrebombeo de las aguas subterráneas produce una diferencia de cabeza hidrostática a favor del agua de mar haciendo que ésta se introduzca en el área de agua dulce.

- **Fluoruros**: El ingeniero tiene un doble interés en la determinación de fluoruros; por una parte es responsable del diseño y operación de unidades de tratamiento para

remoción de fluoruros, en aguas que contienen cantidades excesivas y por otra parte, es responsable supervisar y fomentar la adición de fluoruros en dosis óptimas a los suministros de agua para la salud dental de la población. La mayoría de los fluoruros son de baja solubilidad; por ello la concentración de fluoruros en aguas naturales es normalmente baja, generalmente menor de 1 mg/L en aguas superficiales, raras veces mayor de 10 mg/L y excepcionalmente más de 50 mg/L.

- **Grupo del Azufre (SO_4)**: Tanto en la purificación de aguas como en el tratamiento de aguas residuales se presentan diferentes formas químicas del azufre que son de interés. Estas son: sulfatos, sulfuros y sulfitos. Según Doménico (1990), la química del agua subterránea varía con la profundidad de las cuencas sedimentarias. En la parte alta, existen grandes movimientos de aguas que remueven sales minerales por la disolución de las rocas, provocando que existan pocos sólidos disueltos y sus contenidos sean altos en bicarbonatos (HCO_3). En la parte media, el movimiento es más lento, y el agua gana mayor cantidad de sólidos disueltos, y el ion sulfato **($SO4^-$)** es el dominante. En la parte baja, el movimiento es casi nulo, lo que hace que haya una remoción de sales por disolución, además se incrementan los sólidos diluidos. El ion predominante es el ion cloruro (Cl^-).
- **T:** La temperatura es una magnitud referida a las nociones comunes de caliente, tibio o frío que puede ser medida con un termómetro. En física, se define como una magnitud escalar relacionada con la energía interna de un sistema termodinámico, definida por el principio cero de la termodinámica.
- **pH:** Igual que pH in situ solo cambia que es medido en laboratorio.
- **K:** Es un metal alcalino de color blanco-plateado, que abunda en la naturaleza en los elementos relacionados con el agua salada y otros minerales. Se oxida rápidamente en el aire, es muy reactivo, especialmente en agua, y se parece químicamente al sodio.

Tabla 21. Emanaciones Hidrogeoquímicas de la estación El Jabalí provenientes del Volcán de San Salvador

pH	Ca	Na	Mg	Cl	SO_4	T	Dureza	pH in situ	Alcalinidad	Fluor	K
6.3	217.86	0.0	43.71	7.78	487.5	25.0	724.00	6.50	411.70	0.00	0.00
6.3	121.44	0.0	121.00	14.18	503.0	29.5	751.00	6.30	505.00	0.00	0.00
6.3	123.00	NA	123.00	15.24	445.0	25.8	639.00	6.50	517.00	NA	NA
6.4	233.00	NA	27.00	9.22	445.0	25.8	692.00	6.30	504.00	NA	NA
6.7	171.00	NA	64.00	9.22	530.0	25.8	712.00	6.10	491.00	NA	NA
6.6	217.00	NA	40.00	8.51	538.0	25.4	702.00	6.30	549.00	NA	NA
6.3	224.85	NA	35.23	10.63	538.0	25.3	706.00	NA	555.00	NA	NA
6.3	137.29	NA	92.43	67.99	463.0	24.5	722.94	6.40	566.00	NA	NA
6.5	144.47	NA	86.99	20.39	538.0	21.3	716.97	6.40	553.00	NA	NA
6.4	139.68	NA	90.01	28.16	450.0	21.4	718.96	6.60	562.00	NA	NA
6.4	131.71	NA	96.79	23.31	513.0	29.4	726.93	6.50	224.00	NA	NA
6.7	141.21	NA	97.77	17.00	450.0	21.4	754.67	5.80	426.00	NA	NA
6.6	149.91	NA	86.53	7.29	450.0	25.0	730.66	6.60	NA	NA	NA
6.1	90.50	NA	124.83	35.00	513.0	26.0	740.00	5.60	409.74	NA	NA
7.0	65.70	NA	133.10	17.50	500.0	26.0	712.00	5.60	403.70	NA	NA
6.7	116.94	NA	101.03	17.50	475.0	25.8	708.00	5.60	383.75	NA	NA
6.2	123.35	NA	113.66	19.45	375.0	25.7	776.00	6.40	416.73	NA	NA
6.4	48.06	NA	24.77	62.23	415.0	26.1	222.00	6.40	195.87	1.00	NA
6.3	217.86	NA	43.71	7.78	487.5	25.0	724.00	6.50	411.74	NA	NA
6.6	134.56	NA	101.03	11.66	487.5	25.0	752.00	7.10	509.32	NA	NA
6.7	126.55	NA	101.99	17.48	525.0	25.5	736.00	6.60	507.03	NA	NA
6.2	136.02	NA	125.96	13.92	385.0	NA	858.31	NA	503.76	NA	NA
6.0	102.33	NA	149.24	16.90	462.5	25.0	870.04	7.00	589.62	NA	NA
6.2	149.28	NA	112.75	16.90	425.0	23.0	837.05	6.68	493.30	NA	NA
6.2	148.33	NA	124.24	15.91	375.0	25.5	881.94	6.00	545.81	0.60	NA
6.2	139.32	NA	113.56	15.88	375.0	24.5	814.93	6.10	527.50	0.91	NA
6.6	100.92	NA	116.57	15.61	475.0	25.4	732.00	6.00	552.00	0.95	NA
6.1	121.75	NA	107.83	15.91	437.5	25.4	NA	6.78	281.00	1.07	NA
5.9	131.36	NA	100.06	13.97	400.0	25.4	740.00	6.65	505.80	0.93	NA
6.2	126.55	NA	160.77	23.95	425.0	26.9	978.00	6.60	600.68	0.94	NA
6.1	124.95	NA	105.89	17.92	310.0	27.2	748.00	6.90	526.85	0.91	NA
6.0	126.55	94.8	100.54	15.93	387.5	24.0	730.00	6.20	522.01	0.75	15.20
6.3	184.22	69.9	68.97	17.96	385.0	25.5	744.00	6.54	615.44	0.80	14.60
6.3	155.39	73.3	84.51	17.96	400.0	24.5	736.00	6.44	611.82	0.80	13.80
6.0	259.51	94.8	24.29	9.98	365.0	24.4	748.00	5.98	514.44	0.87	14.50
6.0	137.76	158.0	97.14	18.96	435.0	25.4	744.00	5.69	NA	0.84	12.80
6.0	232.28	22.5	30.11	21.03	425.0	25.6	704.00	6.25	556.79	0.93	16.20
6.1	72.09	94.8	133.08	19.87	550.0	26.0	728.00	6.01	514.64	0.81	16.50
6.2	221.06	59.7	72.86	13.91	490.0	25.0	852.00	6.06	520.12	0.78	4.80
5.9	120.14	59.9	105.89	15.91	460.0	25.7	736.00	6.00	515.64	0.79	18.20
5.9	126.55	58.8	94.23	19.86	440.0	25.7	704.00	5.95	530.95	0.71	12.60
5.9	120.14	71.2	101.51	15.89	400.0	28.0	718.00	5.90	525.83	0.45	15.80
5.8	128.15	63.6	101.50	17.87	400.0	25.4	738.00	5.70	515.56	0.98	14.00
5.6	123.35	68.6	97.14	13.90	445.0	25.9	NA	6.00	519.41	1.04	13.60
5.8	267.52	63.3	1.94	17.87	430.0	25.5	676.00	6.40	273.17	0.86	15.10
6.1	103.96	73.1	108.99	20.26	460.0	25.6	708.40	6.03	532.87	1.00	18.30
5.9	118.10	68.4	99.96	19.06	415.0	25.5	705.20	6.20	513.21	0.89	14.90
5.8	165.16	73.4	61.78	17.87	435.0	25.9	666.80	6.10	523.37	0.73	7.37
5.8	119.42	67.3	98.45	8.94	445.0	25.8	703.60	5.93	511.38	0.90	38.50
5.8	125.11	71.4	98.40	8.94	480.0	25.5	717.60	6.00	415.17	0.72	41.10
5.9	115.34	66.2	110.35	9.53	435.0	25.5	742.40	6.20	420.39	1.00	13.30
5.8	127.99	32.3	96.95	10.33	390.0	25.8	718.80	6.30	472.63	0.48	19.30
6.3	126.07	65.5	95.01	9.93	485.0	25.6	706.00	6.10	449.49	0.87	9.32
6.1	120.06	68.5	103.46	9.83	425.0	25.7	725.80	6.10	473.05	0.84	16.80
6.0	125.43	65.3	99.38	8.34	425.0	26.0	722.40	6.20	448.40	0.88	41.50
6.0	119.34	69.6	99.47	8.54	395.0	25.6	707.60	6.10	469.68	0.90	26.70
6.0	120.46	73.0	98.45	20.28	420.0	25.6	706.20	6.20	407.88	0.83	51.40
6.0	118.38	69.0	105.69	20.64	405.0	25.5	730.80	6.30	368.26	1.14	38.50
5.7	117.90	69.4	98.41	18.20	445.0	25.7	699.60	6.06	377.91	0.98	15.50
5.8	70.00	68.2	126.87	19.11	415.0	25.7	697.20	6.20	377.14	0.77	14.20

Fuente: Servicio Nacional de Estudios Territoriales (SNET) de El Salvador.

3.2.2. Identificación del patrón de pérdida de datos

El análisis de la base de datos se inicia con la identificación del patrón de pérdida de datos para la matriz en estudio, se obtiene de la siguiente manera:

```
#Se introducen los Datos
> base_jab<-read.csv(file="jabali.csv",head=TRUE,sep=";")
> attach(base_jab)
> base_jab0<-data.frame(pH,Ca,Na,Mg,Cl,S04,T,Dureza,ph.in.situ,Alcalinidad,Fluor,K)
#Patrón de Datos
> library("colorspace")
> library(VIM)
> incvars<-c("pH","Ca","Na","Mg","Cl","S04","T","Dureza","ph.in.situ","Alcalinidad","Fluor","K")
> aggr(base_jab0[, incvars], numbers=TRUE, prop = c(TRUE, FALSE))
```

Se obtiene la Figura 3.1:

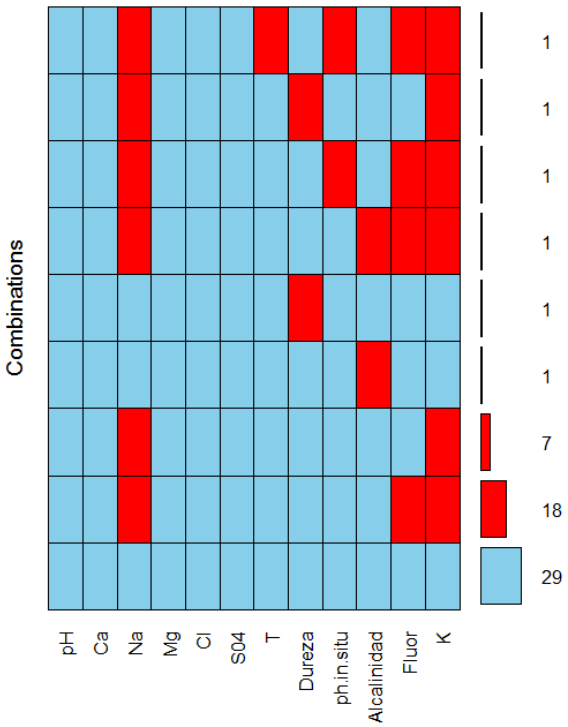

Figura 3.1: Patrón de datos perdidos

Como se observa, si se contabilizan los valores de los bloques a la derecha, se obtiene

un total de 60 datos los cuales son los casos totales y además resumen los valores perdidos, por ejemplo que las variables completas son pH, Ca, Mg, Ml y SO_4, la variable que tiene 1 valor faltante es la Temperatura(T), los que tienen 2 valores faltantes son Dureza, pH in situ y Alcalinidad, Fluor tiene 21 valores faltantes y finalmente K y Na tienen 29 valores faltantes. Con el patrón de datos perdidos se observa que es un "Patrón General", como el mostrado en la Figura 1.1 del Capítulo I.

3.2.3. Cálculo de la tasa de no respuesta presente en los datos

Se procede al cálculo de la cantidad de registros faltantes y el porcentaje de pérdida para cada una de las variables en el estudio lo cual se muestra en la Tabla 22.

Tabla 22. Tasa de no respuesta para cada una de las variables

Variable	Cantidad de registros Faltantes	Tasa de datos faltantes
pH	0	0.00 %
Ca	0	0.00 %
Na	29	48.33 %
Mg	0	0.00 %
Cl	0	0.00 %
SO4	0	0.00 %
T	1	1.66 %
Dureza	2	3.33 %
pH in situ	2	3.33 %
Alcalinidad	2	3.33 %
Fluor	21	35.00 %
K	29	48.33 %

3.2.4. Comparación de la tasa de no respuesta con el valor prefijado

En la tabla 22 es posible observar que ninguna de las variables de las que se plantearon para imputar (pH in situ, Dureza, Temperatura y Alcalinidad) presenta una tasa de no respuesta superior a la tasa prefijada al inicio que es del 10 %, por lo que a las que presentan valores faltantes es posible aplicarles una técnica de imputación para generar registros completos. Con respecto a la variable Na (Sodio), flúor y K (potasio) aventurarse a una imputación con casi el 50 % de los datos ausentes en dos variables y el 35 % para

el flúor, representaría un importante sesgo, se incluyó en la base ya que tiene relaciones con otras y servirán sus partes completas como covariables. Una vez se conoce que las variables cumplen con la tasa de pérdida admisible se procede a la implementación de los pasos sucesivos para determinar la técnica de imputación más adecuada para cada caso.

3.2.4.1. Mecanismo de pérdida de datos

En este caso es posible afirmar que el mecanismo de pérdida es ignorable ya que al observar el patrón de datos faltantes es posible notar aleatoriedad en las omisiones. Siendo que el mecanismo es ignorable es preciso aplicar una prueba que revele si este es de tipo MAR ó MCAR.

Ya que se analizará toda la base de datos se precisa una prueba a todas las variables, la prueba que ayudará a reconocer el mecanismo de pérdida, como se estudio en el capítulo I es la prueba de Little's para contrastar la hipótesis nula de existencia de un mecanismo MCAR, la cual se aplica con el siguiente código:

```
> library(BaylorEdPsych)
> library(mvnmle)
> LittleMCAR(base_jab0)
```

de donde se obtiene parte de la salida:

```
this could take a while$chi.square
[1] 106.2021

$df
[1] 73

$p.value
[1] 0.006773522
```

```
$missing.patterns
[1] 9

$amount.missing
                pH Ca         Na Mg Cl S04          T     Dureza ph.in.situ
Number Missing   0  0 29.0000000  0  0   0 1.00000000 2.00000000 2.00000000
Percent Missing  0  0  0.4833333  0  0   0 0.01666667 0.03333333 0.03333333
                Alcalinidad Fluor          K
Number Missing   2.00000000 21.00 29.0000000
Percent Missing  0.03333333  0.35  0.4833333
```

Proporciona un p-valor de menos de 0.007 con una significancia al 5 % no existe evidencia para sostener la hipótesis H_0 y se supondrá un mecanismo MAR. Se hace un recordatorio que muchas ocasiones la prueba tiene resultados erróneos ya que puede llegarse a un mecanismo como el obtenido (MAR), pero la debilidad de la prueba se basa en decir cuales variables están relacionadas.

3.2.5. Definir el procedimiento a posteriori que se aplicara al conjunto de datos

El objetivo de la imputación de esta base de datos es utilizar las mejores técnicas de imputación que no generen sesgos significativos y que cualquier tipo de técnica estadística arroje valores consistentes, por lo que el dato imputado se acercará al dato no observado. La validación de las técnicas se realizará en base a los parámetros más comunes tales como la media y varianza, así como un gráfico de tipo histograma para comprobar la no modificación de su distribución de los datos.

3.2.6. Análisis de información auxiliar

Para el análisis de esta base de datos se ha consultado con varios expertos en química y se ha buscado de igual forma relaciones teóricas entre las variables que nos ayudará a la imputación correcta de los datos ya que es lógico pensar que existen relaciones entre ellas debido a el mecanismo de pérdida MAR y la prueba de little no especifica que variables

están relacionadas.

Se encontraron diversos modelos que siguen algunas variables y podrán ser de mucha ayuda a la hora de definir el mejor valor estimado en la imputación. Como inicio se consulto a un vulcanólogo para poder llegar a establecer relaciones entre las variables y se determino que el **pH in situ** es inversamente proporcional al aumento de Na (Sodio), Cl (Cloro), SO_4 y el Flúor, así su medida de acidez aumenta.

La **Dureza** del agua es la concentración de compuestos minerales de cationes polivalentes (principalmente divalentes y específicamente los alcalinotérreos) que hay en una determinada cantidad de agua, en particular sales de magnesio y calcio, entonces se sabe que:
Dureza (mg/l de $CaCO_3$)= $2.50[Ca^{++}] + 4.16[Mg^{++}]$

Donde:

$[Ca^{++}]$: Concentración de ion Ca^{++} expresado en mg/l.

$[Mg^{++}]$: Concentración de ion Mg^{++} expresado en mg/l.

La alcalinidad del agua se puede definir como una medida de su capacidad para neutralizar ácidos. Se determinó que la **Alcalinidad** es proporcional al aumento de Na (Sodio), Mg (Magnesio), K (Potasio) y Ca (Calcio).

3.2.7. Selección de la técnica de imputación

Se estableció que se imputarán las variables con menos del 10 % de valores perdidos, se iniciará con la variable "**pH in situ**" se analizó anteriormente que tiene relación con otra variable lo cual se deberá comprobar si en los datos de la estación El Jabalí cumplen. En algunas pláticas los expertos en ese campo aseguraban que algunas no se cumplían por la poca actividad del volcán de San Salvador, se partirá del supuesto de un modelo de regresión lineal múltiple y se comparará los resultados, sino se buscará otra alternativa.

Se parte del modelo de regresión siguiente:

```
> ph_insitu<-lm(ph.in.situ~Na+Cl+S04+Fluor)
```

y se obtienen los parámetros siguientes:

```
Call:
lm(formula = ph.in.situ ~ Na + Cl + S04 + Fluor)

Coefficients:
(Intercept)           Na           Cl          S04        Fluor
   6.889846    -0.003998     0.004526    -0.001174    -0.070916
```

de la salida de R anterior, se tienen el siguiente modelo de regresión:

$$ph.in.situ = 6.8898 - 0.0039 * Na + 0.0045 * Cl - 0.0011 * SO4 - 0.0709 * Fluor \quad (3.1)$$

Al poner en práctica el modelo se obtienen los valores estimados, en la tercera columna se tiene la diferencia del valor real y del valor estimado por el modelo, así:

```
> comp<-data.frame(estim,ph.in.situ,ph.in.situ-estim)
> na.omit(comp)
      estim    ph.in.situ   ph.in.situ-estim
1  6.352502       6.50      0.147498369
2  6.363263       6.30     -0.063262668
32 6.074657       6.20      0.125343218
33 6.182780       6.54      0.357220445
34 6.151570       6.44      0.288430020
35 6.065644       5.98     -0.085643719
36 5.773540       5.69     -0.083539853
37 6.329971       6.25     -0.079971065
38 5.897382       6.01      0.112618211
39 6.083326       6.06     -0.023326057
40 6.126103       6.00     -0.126103192
41 6.177541       5.95     -0.227540583
42 6.175418       5.90     -0.275417574
```

```
43 6.177176       5.70     -0.477176271
44 6.082113       6.00     -0.082113494
45 6.151651       6.40      0.248348612
46 6.078127       6.03     -0.048127335
47 6.152138       6.20      0.047861991
48 6.114620       6.10     -0.014620476
49 6.074791       5.93     -0.144790752
50 6.030058       6.00     -0.030058212
51 6.086512       6.20      0.113488343
52 6.315385       6.30     -0.015385161
53 6.041616       6.10      0.058384065
54 6.101766       6.10     -0.001765835
55 6.104979       6.20      0.095021391
56 6.122509       6.10     -0.022509136
57 6.137652       6.20      0.062347683
58 6.150906       6.30      0.149094200
59 6.102631       6.06     -0.042631347
60 6.161674       6.20      0.038326180

> mean(abs(na.omit(ph.in.situ-estim)))
[1] 0.1189666
```

La media de la diferencia en valor absoluto entre los valores estimados y los valores reales es de 0.1189, que es relativamente bajo, pero existe un pequeño inconveniente para la imputación con esta técnica ya que donde se necesita el valor imputado están ausentes el valor del Flúor y del Sodio por lo que por la localización de los valores ausentes esta técnica se convierte en inviable, pero es relativamente confiable para estimar valores futuros.

Dado que esta base de datos viene de un proceso en la naturaleza muchas de estas variables tienen distribuciones normales, tal es el caso de la variable **pH in situ** observamos su histograma en la Figura 3.2, asimismo el Gráfico Q-Q Plot de Normalidad en la Figura 3.3.

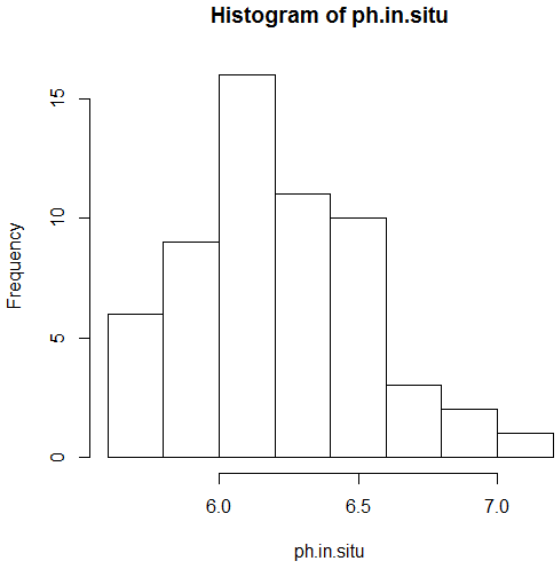

Figura 3.2: Histograma de la variable pH in situ

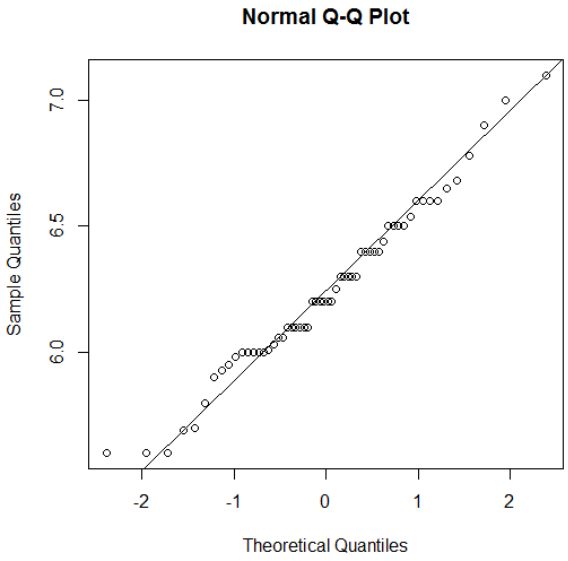

Figura 3.3: Q-Q Plot de la variable pH in situ

Para respaldar el análisis gráfico de normalidad se aplica la prueba Shapiro de normalidad teniendo la hipótesis H_0: Que los datos provienen de una distribución normal.

```
> shapiro.test(ph.in.situ) #test de normalidad
```

```
        Shapiro-Wilk normality test

data:  ph.in.situ
W = 0.9793, p-value = 0.4221
```

Por lo que con la salida anterior con un p-valor de 0.4221 no hay suficiente evidencia en la muestra de datos para rechazar la hipóteiss H_0 y se supone que los datos provienen de una distribución normal. Con estás afirmaciones se puede hacer uso de varias técnicas de imputación para la variable **pH in situ** tales como: Imputación por Máxima Verosimilitud - Algoritmo E-M e Imputación por un valor aleatorio (VADE).

Para la variable **Dureza** con la información auxiliar se determinó que existe un modelo que sigue esta variable y es de tener en cuenta si se tienen los valores de las covariables disponibles, para el valor ausente, con lo que en este caso si están disponibles. Se aplicará el siguiente modelo para este caso:

Dureza (mg/l de $CaCO_3$)= $2.50[Ca^{++}] + 4.16[Mg^{++}]$

Para la variable **Alcalinidad** que tiene 2 valores ausentes se tienen completas las covariables para un caso, pero no para otro dato ausente, por lo que se aplicará el siguiente modelo de regresión múltiple:

```
> Alcalinidad_imp

Call:
lm(formula = Alcalinidad ~ Na + Ca + K + Mg)

Coefficients:
(Intercept)           Na           Ca            K           Mg
    73.1255       0.6785       1.3033      -1.7555       2.3623
```

Respecto a la variable **Alcalinidad** en cuanto al valor donde no existen valores para las covariables del resultado de los análisis omitiendo los valores extremos se obtienen las Figuras 3.4 y 3.5.

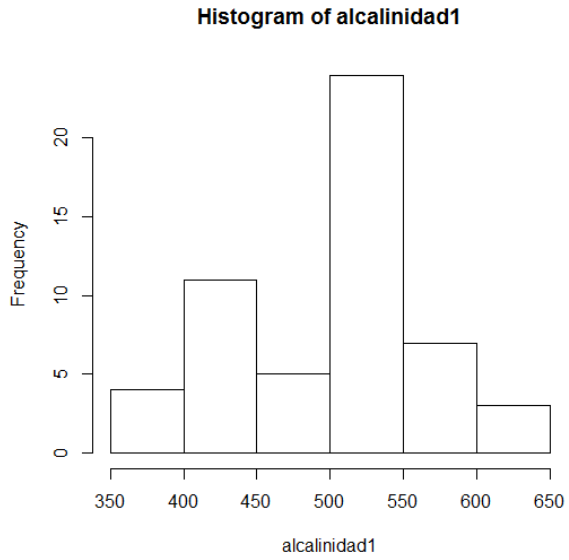

Figura 3.4: Histograma de la variable Alcalinidad

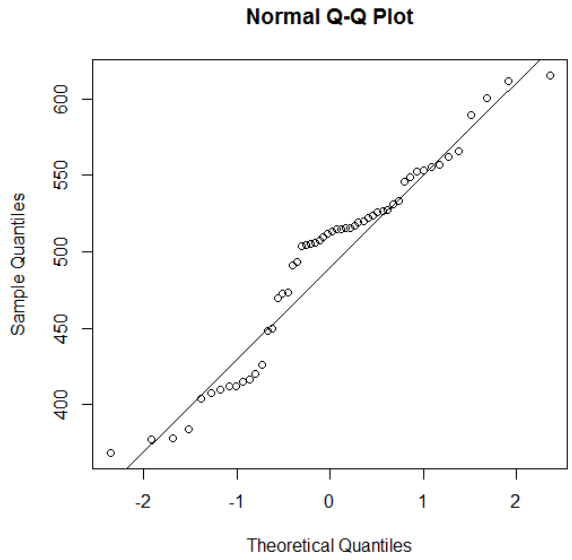

Figura 3.5: Q-Q Plot de la variable Alcalinidad

Al analizar el histograma y el Q-Q Plot de la variable **Alcalinidad** se puede suponer con cierta reserva que dicha variable proviene de una distribución normal. En base al análisis anterior se podrá utilizar los siguientes métodos Imputación por Máxima Verosimilitud - Algoritmo E-M y Imputación por un valor aleatorio (VADE).

Para la variable **Temperatura** que solo presenta un valor ausente usaremos técnicas

deterministas básicas tales como de vecino más cercano y la mediana y una estocástica como imputación por Métodos Bayesianos (ABB).

3.2.7.1. Aplicación de la técnica de imputación

Variable pH in situ Se realizarán 2 métodos de imputación por la distribución de los datos, siendo éstos los más adecuados por la naturaleza de la variable.

- Máxima Verosimilitud - Algoritmo EM Inicializando valores con $\mu = 5$ y $\sigma^2 = 0.1$ se obtuvo con 4 iteraciones en la siguiente salida:

```
 > imputacion.em<-em.norm(x)
6.199167 0.160051
6.239139 0.1125215
6.240471 0.1108822
6.240516 0.1108275
```

  Por lo que el algoritmo converge y obtiene $\mu = 6.240516$ y $\sigma^2 = 0.1108275$.
- Imputación por un valor aleatorio (VADE) Se recuerda que este procedimiento se basa en la distribución de la variable, y en R se utiliza la siguiente sintaxis:

```
> x<-ph.in.situ
> N<-length(x)
> Xobs<-x[!is.na(x)]
> Xmis<-x[is.na(x)]
> sust<-rnorm(length(Xmis),mean=mean(Xobs))
> sust
[1] 6.348273 5.233151
```

  Por lo que se tiene que $pHinsitu_{aus1} = 6.348273$ y $pHinsitu_{aus2} = 5.233151$

Variable Dureza Se realizará la imputación por la fórmula que se definió en la información auxiliar:

Dureza (mg/l de $CaCO_3$)= $2.50[Ca^{++}] + 4.16[Mg^{++}]$

A continuación se muestra parte de la salida del programa ejecutado en R; donde se muestra en la columna *pred* los datos predichos que son muy similares a los datos reales

que están en la siguiente columna *Dureza* y la última columna representa el valor absoluto de la diferencia del predicho y del valor real.

```
> pred<-2.5*Ca+4.16*Mg
> Dureza1<-data.frame(pred,Dureza,abs(pred-Dureza))
> Dureza1
     pred       Dureza        abs.pred...Dureza.
22 864.0436     858.31             5.7336
23 876.6634     870.04             6.6234
24 842.2400     837.05             5.1900
25 887.6634     881.94             5.7234
26 820.7096     814.93             5.7796
27 737.2312     732.00             5.2312
28 752.9478         NA                 NA
29 744.6496     740.00             4.6496
30 985.1782     978.00             7.1782
31 752.8774     748.00             4.8774
32 734.6214     730.00             4.6214
33 747.4652     744.00             3.4652
34 740.0366     736.00             4.0366
35 749.8214     748.00             1.8214
36 748.5024     744.00             4.5024
37 705.9576     704.00             1.9576
38 733.8378     728.00             5.8378
39 855.7476     852.00             3.7476
40 740.8524     736.00             4.8524
41 708.3718     704.00             4.3718
42 722.6316     718.00             4.6316
43 742.6150     738.00             4.6150
44 712.4774         NA                 NA
45 676.8704     676.00             0.8704
46 713.2984     708.40             4.8984
47 711.0836     705.20             5.8836
48 669.9048     666.80             3.1048
49 708.1020     703.60             4.5020
50 722.1190     717.60             4.5190
```

```
51 747.4060     742.40            5.0060
52 723.2870     718.80            4.4870
```

Por lo que se tiene que $Dureza_{aus1} = 752.9478$ y $Dureza_{aus2} = 712.4774$

Variable Alcalinidad Se realizarán 3 métodos de imputación combinados, siendo éstos los más adecuados por la naturaleza de la variable y por las covariables completas que se tienen.

- Método de Regresión Múltiple: Se analizó en la información auxiliar que la variable alcalinidad puede estar relacionada proporcionalmente con el aumento de Na (Sodio), Mg (Magnesio), K (Potasio) y Ca (Calcio).
  Se tiene el siguiente modelo de regresión encontrado en base a las variables completas:

```
> Alcalinidad_imp

Call:
lm(formula = Alcalinidad ~ Na + Ca + K + Mg)

Coefficients:
(Intercept)           Na           Ca            K           Mg
    73.1255       0.6785       1.3033      -1.7555       2.3623
```

Por lo que se puede imputar el valor donde están completas todas las variables es el caso 36 (ver Tabla 21),esto es:

pH	Ca	Na	Mg	Cl	SO4	T	Dureza	pH in situ	Alcalinidad	Fluor	K
6.0	137.76	158.0	97.14	18.96	435.0	25.4	744.00	5.69	NA	0.84	12.80

Por lo que usando el modelo de regresión se obtiene el valor de la observación 36:

```
> caso36<-Alcalinidad_imp$coefficients[[1]]+Na[36]*Alcalinidad_imp$coefficients[[2]]+
Ca[36]*Alcalinidad_imp$coefficients[[3]]+K[36]*Alcalinidad_imp$coefficients[[4]]+
Mg[36]*Alcalinidad_imp$coefficients[[5]]
> caso36
[1] 566.8728
```

Se imputa esta observación y se procede a implementar los dos métodos de imputación a continuación.

- Máxima Verosimilitud - Algoritmo EM: Incializando valores con $\mu = 500$ y $\sigma^2 = 10$ se obtuvo con 3 iteraciones la siguiente salida:

```
> x<-Alcalinidad0
> imputacion.em<-em.norm(x)
479.857 7674.111
479.5213 7794.971
479.5157 7796.983
```

Por lo que el algoritmo converge y obtiene $\mu = 479.5157$ y $\sigma^2 = 7796.983$.

- Imputación por un valor aleatorio (VADE): Se recuerda que este procedimiento se basa en la distribución de la variable se obtiene con el código siguiente:

```
> N<-length(x)
> Xobs<-x[!is.na(x)]
> Xmis<-x[is.na(x)]
> sust<-rnorm(length(Xmis),mean=mean(Xobs))
> sust
[1] 477.5086
```

Por lo que se tiene que $Alcalinidad_{aus1} = 477.5086$

Variable Temperatura (T) Se realizarán 2 métodos de imputación, la variable presenta un solo valor ausente y se usarán técnicas deterministas básicas tales como vecino más cercano y la mediana; también una técnica estocástica como la imputación por Métodos Bayesianos (ABB).

- Vecino más cercano: Para utilizar esta técnica de imputación se asigna distancias a cada una de las observaciones, en base a los valores de la variable completa que tenga más relación con la *Temperatura*. La variable elegida para este procedimiento es el *pH* y se utilizará la distancia euclidiana.

```
> dato<-data.frame(T,pH)
> dato
> imputacion.ejemplo<-SeqKNN(dato,10)
> imputacion.ejemplo
```

Por lo que se obtiene $T_{aus1} = 25.1$.

- Imputación por la mediana: Este método es simple de utilizar y útil con pocos datos faltantes.

```
> Temp<-T
> median(na.omit(Temp))
> median(na.omit(Temp))
[1] 25.5
> Temp[22]<-median(na.omit(Temp))
```

El valor de la mediana imputando el caso 22 es de 25.5.

- Imputación por Métodos Bayesianos (ABB): Se utilizará para el dato faltante un interesante método con parámetros robustos, se obtuvo con la aplicación de este método la siguiente salida:

```
> Temp1<-T
> ImputacionBB(Temp1,7)
Mu 1   25.435
U 1   0.03111201
Mu 2   25.435
U 2   0.03111201
Mu 3   25.42667
U 3   0.03116309
Mu 4   25.435
U 4   0.03111201
Mu 5   25.43167
U 5   0.03111577
Mu 6   25.43667
U 6   0.03111846
Mu 7   25.43833
U 7   0.03113046
mu_IM 25.43405
Varianza(mu_IM) 0.03114018
```

En este caso el valor de media una vez imputado el caso 22 es de 25.43405

3.2.7.2. Validación de la Técnica de Imputación

Esta fase de la guía metodológica es de mucha importancia pues ayudará a decidir cual de las técnicas de imputación aplicada es la más efectiva, mediante la comparación del vector de parámetros observado y el o los vectores obtenidos luego de la imputación deseando que estos no difieran significativamente.

Tabla 23. Efectos de la Imputación en la variable pH in situ

Estimadores	Datos Incompletos	Datos Completados por Máxima Verosimilitud - Algoritmo EM	Datos completados por Imputación Aleatoria (VADE)
n	58	60	60
Media	6.240517	6.240517	6.225524
Mediana	6.2	6.2	6.2
Varianza	0.1127699	0.1089472	0.1261152
Desviación Estándar	0.3358123	0.3300715	0.355127
Error Estándar	0.04409432	0.04261205	0.04584669
Coeficiente de Asimetría	0.2513334	0.2558546	-0.04989591
Curtosis	-0.1585441	-0.05712382	0.2564878
Rango	1.5	1.5	1.866849
Mínimo	5.6	5.6	5.233151
Máximo	7.1	7.1	7.1
Percentil 25	6.0025	6.0075	6
Percentil 50	6.2	6.2	6.2
Percentil 75	6.485	6.455	6.455

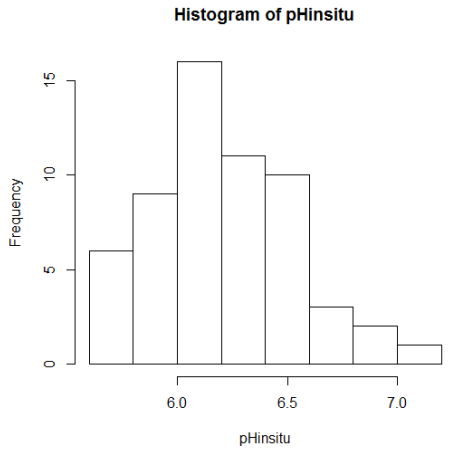

Figura 3.6: Histograma de la variable pH in situ sin Imputar

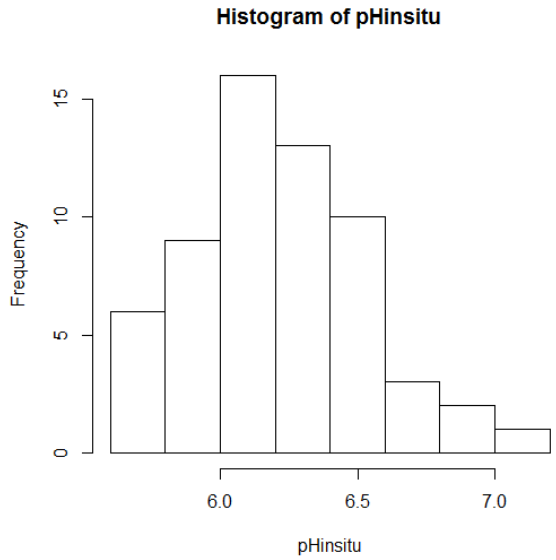

Figura 3.7: Histograma de la variable pH in situ imputada usando Máxima Verosimilitud - Algoritmo EM

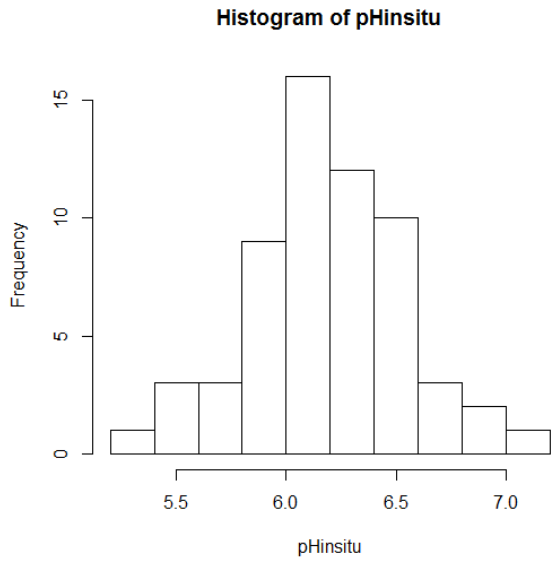

Figura 3.8: Histograma de la variable pH in situ usando Imputación Aleatoria (VADE)

Al comparar uno a uno los elementos del vector de parámetros estimados de la Tabla 23, para cada uno de los métodos mediante los cuales se imputó la base de datos, es posible determinar que las diferencias son menores al hacer uso de las técnicas de Máxima Verosimilitud-Algoritmo EM, por lo que en este caso ésta sería la técnica más indicada sin

embargo, es importante mencionar que las diferencia entre ambas técnicas son pequeñas y esto se debe a que gracias a la implementación de la guía metodológica ambas técnicas son adecuadas, por lo que se podrá imputar haciendo uso de cualquiera de las dos, pero siempre es recomendable usar una estocástica como es la imputación aleatoria para este caso.

En los gráficos 3.6, 3.7 y 3.8 se observa que se preserva la distribución de la variable original tras la aplicación de ambas técnicas de imputación, por lo que, este análisis no resulta concluyente, sin embargo ayuda a confirmar lo planteado en el análisis del vector de parámetros.

Tabla 24. Efectos de la Imputación en la variable Dureza

Estimadores	Datos Incompletos	Datos completados por Modelo Químico
n	58	60
Media	730.0655	730.1538
Mediana	724.9	724.9
Varianza	7795.315	7545.177
Desviación Estándar	88.29108	86.86298
Error Estándar	11.59319	11.21396
Coeficiente de Asimetría	-2.635892	-2.678612
Curtosis	17.45917	18.12123
Rango	756	756
Mínimo	222	222
Máximo	978	978
Percentil 25	706.55	707.25
Percentil 50	724.9	724.9
Percentil 75	743.6	744

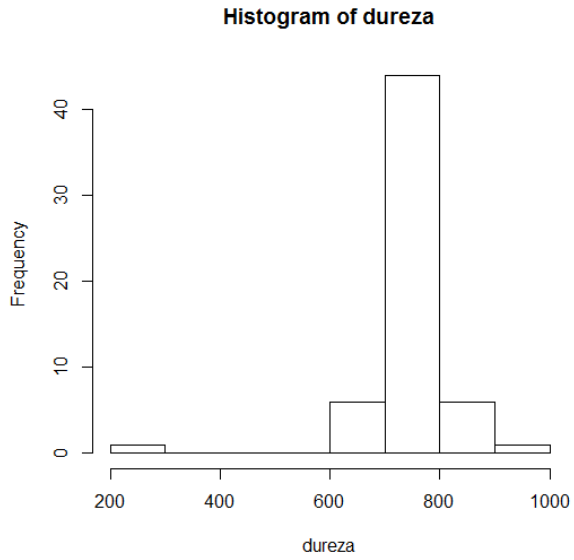

Figura 3.9: Histograma de la variable Dureza sin Imputar

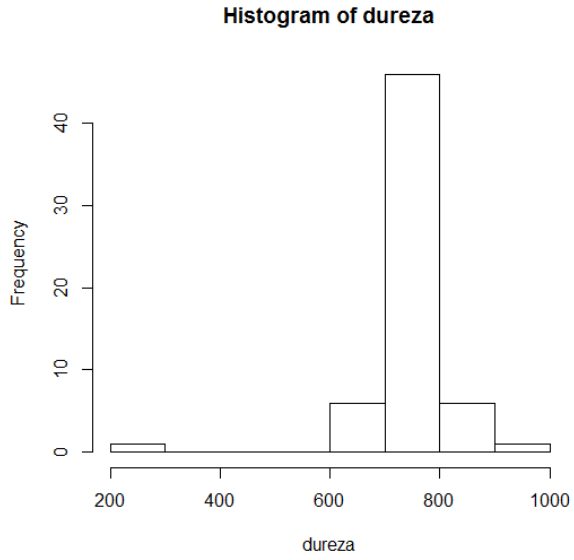

Figura 3.10: Histograma de la variable Dureza imputada por Modelo Químico

En en el caso de la variable *Dureza* se ejemplifica que existen diferentes ramas de imputación en las cuales tiene mucho peso la información auxiliar, por lo que, en este caso ya existía un modelo probado a priori sobre los coeficientes que relacionaban a las variables en estudio y al aplicar este modelo para la imputación de los datos se obtienen valores cercanos al resto de valores observados para las variables, por lo que el sesgo es

pequeño y los vectores de parámetros se asemejan (Tabla 24), lo cual permite establecer que este es un buen método para imputar esta variable, ya que al considerar otros métodos de imputación que subestiman la información auxiliar y se incrementa el sesgo, por lo que esta sería una decisión errónea al momento de seleccionar la forma en la que se imputarán los datos.

Tabla 25. Efectos de la Imputación en la variable Alcalinidad

Estimadores	Datos Incompletos	Datos Completados por Regresión Múltiple - Máxima Verosimilitud - Algoritmo EM	Datos Completados por Regresión Múltiple - Imputación Aleatoria (VADE)
n	58	60	60
Media	478.0095	479.5156	479.4822
Mediana	508.175	508.175	508.175
Varianza	7934.407	7797.017	7797.085
Desviación Estándar	89.07529	88.30072	88.3011
Error Estándar	11.69616	11.39957	11.39962
Coeficiente de Asimetría	-1.180375	-1.204657	-1.203524
Curtosis	1.319932	1.421967	1.420067
Rango	419.57	419.57	419.57
Mínimo	195.87	195.87	195.87
Máximo	615.44	615.44	615.44
Percentil 25	417.645	419.475	419.475
Percentil 50	508.175	508.175	508.175
Percentil 75	527.3375	528.3625	528.3625

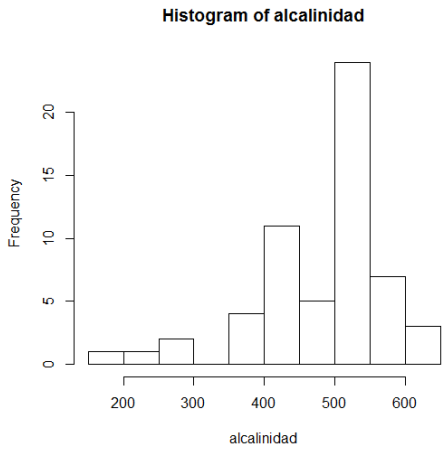

Figura 3.11: Histograma de la variable Alcalinidad sin Imputar

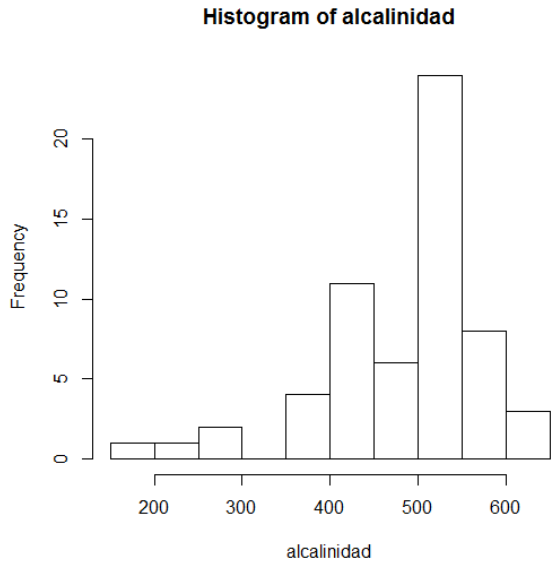

Figura 3.12: Histograma de la variable Alcalinidad Completados por Regresión Múltiple - Máxima Verosimilitud - Algoritmo EM

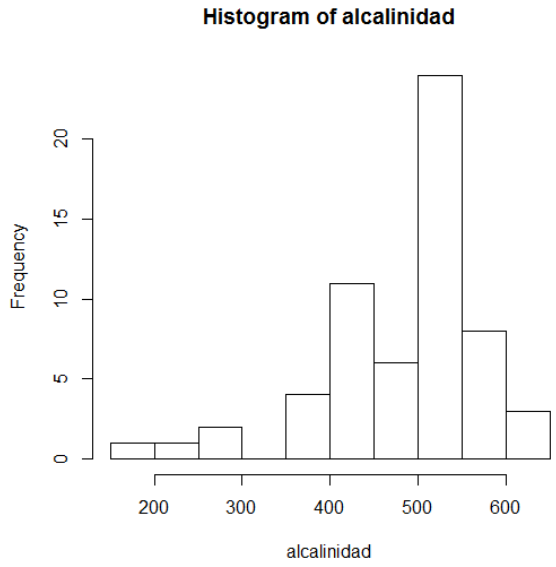

Figura 3.13: Histograma de la variable Alcalinidad Completados por Regresión Múltiple - Imputación Aleatoria (VADE)

En la Tabla 25 se observa que ambas técnicas producen vectores de parámetros casi iguales, por lo que ambas técnicas son adecuadas para la imputación de esta variable, sin embargo es importante hacer notar que siempre que se este dudando entre dos métodos

y uno de ellos sea determinístico y el otro estocástico como en este caso, se recomienda emplear el método estocástico, pues si llegara a darse el caso en que la variable incremente su número de omisiones estos métodos son más eficientes en la reducción del sesgo en las estimaciones, gracias a su componente aleatorio.

En los gráficos 3.11, 3.12 y 3.13 se observa que se preserva la distribución de la variable original tras la aplicación de ambas técnicas de imputación por lo que este análisis no resulta concluyente, sin embargo ayuda a confirmar lo planteado en el análisis del vector de parámetros.

Tabla 26. Efectos de la Imputación en la variable Temperatura

Estimadores	Datos Incompletos	Datos Completados por el Vecino más cercano	Datos Completados por la Mediana	Datos Completados por ABB (Bootstrap)
n	59	60	60	60
Media	25.4339	25.42833	25.435	25.4339
Mediana	25.5	25.5	25.5	25.5
Varianza	1.898831	1.868506	1.86672	1.866648
Desviación Estándar	1.377981	1.366933	1.36628	1.366253
Error Estándar	0.1793978	0.1764703	0.176386	0.1763825
Coeficiente de Asimetría	-0.4308177	-0.4222255	-0.4369921	-0.4346461
Curtosis	3.787561	3.885928	3.907438	3.906575
Rango	8.2	8.2	8.2	8.2
Mínimo	21.3	21.3	21.3	21.3
Máximo	29.5	29.5	29.5	29.5
Percentil 25	25.15	25.075	25.225	25.225
Percentil 50	25.5	25.5	25.5	25.5
Percentil 75	25.8	25.8	25.8	25.8

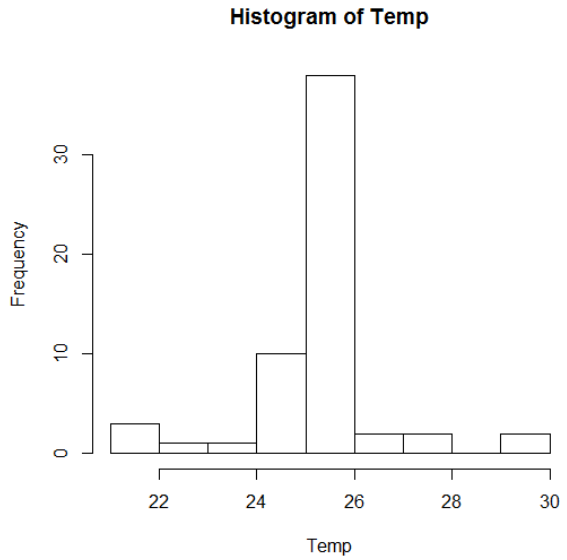

Figura 3.14: Histograma de la variable Temperatura sin Imputar

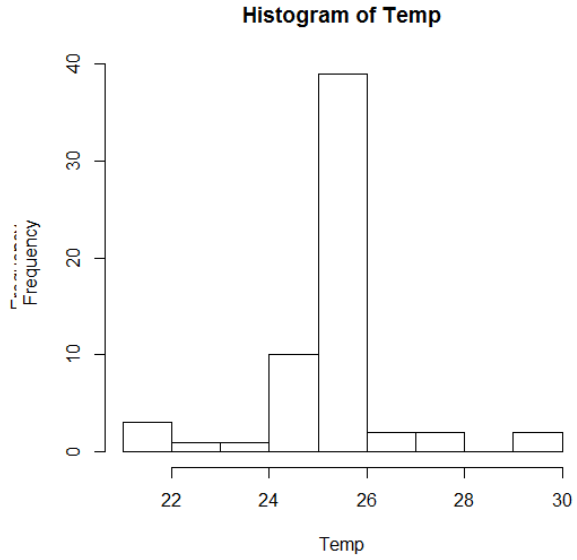

Figura 3.15: Histograma de la variable Temperatura Completados por el Vecino más cercano

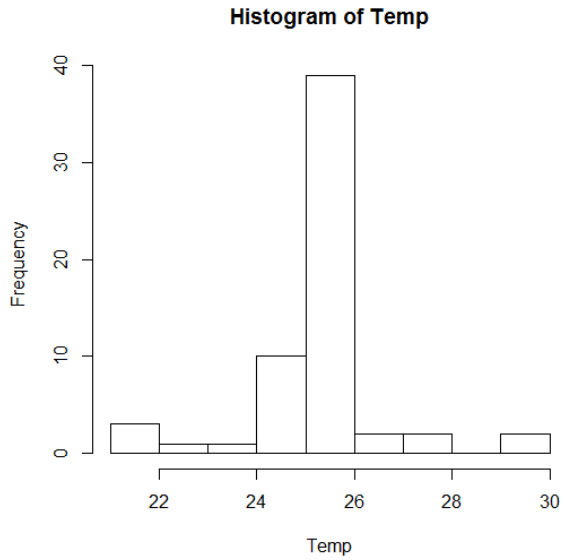

Figura 3.16: Histograma de la variable Temperatura Completados por la mediana

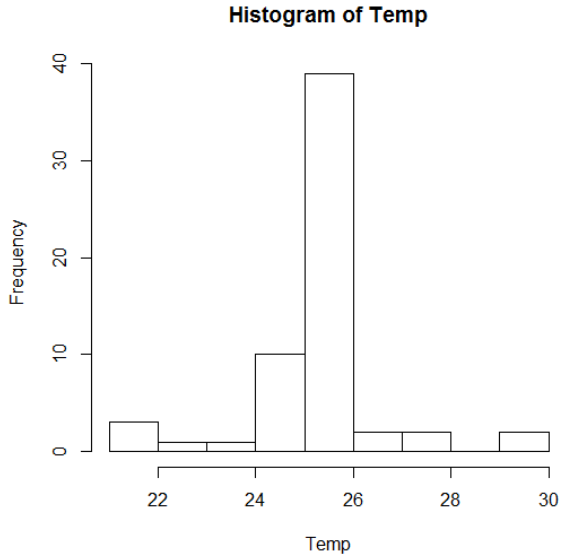

Figura 3.17: Histograma de la variable Temperatura Completados por ABB (Bootstrap)

En este caso para la *Temperatura* (Tabla 26) se observa que al igual que en la alcalinidad los diferentes métodos sugeridos tras la implementación de la guía metodológica conducen a resultados plausibles y que los tres obtienen vectores de parámetros similares, por lo que nuevamente la recomendación es la de elegir un método estocástico entre las sugerencias planteadas tras la consecución del método de selección.

Es posible observar que en la aplicación de las diferentes técnicas los resultados son satisfactorios, ya que las diferencias entre los valores observados y los esperados son muy pequeñas, además al observar los histogramas es posible notar que luego de la aplicación de la imputación se preserva la distribución de las variables, por lo que los resultados no conllevan a sesgos significativos.

Es importante hacer notar que las técnicas se seleccionaron haciendo uso de la guía metodológica y los resultados reflejan que al seguir estos lineamientos es posible determinar técnicas que conlleven a resultados plausibles, por lo que la guía cumple con su objetivo y es una herramienta a utilizar en la búsqueda de una buena técnica de imputación.

Conclusiones y Recomendaciones

Con base en nuestra investigación es posible concluir que:

- Se lograron los objetivos planteados al inicio pues fue posible generar un conjunto de pasos lógicos para la selección de una técnica adecuada de imputación. Luego del estudio exhaustivo de la bibliografía disponible fue posible definir y caracterizar las técnicas de imputación de datos, con el fin de identificando así sus propiedades, funcionalidad, fortalezas y debilidades.
- Se logró elaborar códigos en R específicos para cada una de las técnicas estudiadas los cuales serán de utilidad para cualquier investigador que consulte nuestro trabajo.
- Mediante la implementación práctica de la guía metodológica fue posible imputar de manera eficiente la base de datos proporcionada por el SNET, para aquellas variables que son útiles para los investigadores del área de tal forma que a posteriori puedan realizar con ella cualquier análisis estadístico clásico que ellos requieran aplicar.
- La utilidad primordial de nuestro trabajo reside en el logro plasmado en la implementación práctica en la que se puso de manifiesto que tras la aplicación de la guía metodológica a un conjunto de datos real fue posible obtener un conjunto de variables completas, cuyos vectores de parámetros son muy cercanos a los valores de los parámetros de los valores observados, por lo que se comprobó la eficiencia de la guía la cual es una herramienta muy útil en cualquier área de investigación.
- Se recomienda al lector interesado en aplicar una técnica específica de imputación, indagar a profundidad sobre esta ya que las técnicas son muy extensas y en esta tesis se han expuesto únicamente generalidades de las mismas.
- Se recomienda dar continuidad a la investigación de pruebas que proporciones información más específica y fehaciente sobre la identificación de los mecanismos de pérdida que pueden presentar las bases de datos.
- Se recomienda un análisis previo exhaustivo que permita al investigador conocer a profundidad su base de datos de tal forma que comprenda a la perfección la relación

existente entre las variables de su estudio y de esta forma pueda hacer uso eficiente y explotar al máximo la información auxiliar que pueda rodear a la variable de interés.

Anexos

Códigos Capítulo I

1.3.1.1. Prueba t de Student para contrastar el mecanismo de pérdida de información (MCAR)

```
library(BaylorEdPsych)
library(mvnmle)
data(EndersTable1_1)
### Univariante
ejemplo<-data.frame(EndersTable1_1$IQ,EndersTable1_1$JP)
LittleMCAR(EndersTable1_1)
prueba<-LittleMCAR(ejemplo)
prueba$data$DataSet1
prueba$data$DataSet2
t.test(prueba$data$DataSet1[1],prueba$data$DataSet2[1])
```

1.3.1.2. Prueba de Little MCAR

```
ejemplo1<-LittleMCAR(EndersTable1_1)
#Patrones de datos
ejemplo1$data

gmean <- mlest(EndersTable1_1)$muhat
gmean
gcov <- mlest(EndersTable1_1)$sigmahat
gcov

#media solo con IQ
mean(ejemplo1$data$DataSet4)
#contribución
d2_1=2*t(91.5-100.00)%*%solve(189.60)%*%(91.50-100)
d2_1
```

```
## media con IQ y WB
mean(ejemplo1$data$DataSet2)
d2_2=8*t(matrix(c(87.75,9.13),nrow=2)-matrix(c(100.00,10.27),nrow=2))%*%
solve(matrix(c(189.60,12.21,12.21,11.04),nrow=2))%*%(matrix(c(87.75,9.13),
nrow=2)-matrix(c(100.00,10.27),nrow=2))
d2_2

## media con IQ y JP
mean(ejemplo1$data$DataSet3)
d2_3=1*t(matrix(c(108,10),nrow=2)-matrix(c(100.00,10.23),nrow=2))%*%
solve(matrix(c(189.60,22.31,22.31,8.68),nrow=2))%*%(matrix(c(108,10),nrow=2)-
matrix(c(100.00,10.23),nrow=2))
d2_3

## media con IQ , JP y WB
mean(ejemplo1$data$DataSet1)
d2_4=9*t(matrix(c(111.89,11.89,11.44),nrow=3)-matrix(c(100.00,10.23,10.27),nrow=3))%*%
solve(matrix(c(189.60,22.31,12.21,22.31,8.68,5.61,12.21,5.61,11.04),nrow=3))%*%
(matrix(c(111.89,11.89,11.44),nrow=3)-matrix(c(100.00,10.23,10.27),nrow=3))
d2_4
d2_1+d2_2+d2_3+d2_4

LittleMCAR(EndersTable1_1)
```

1.4.2.2. Imputación Haciendo uso de la Media

```
#Incondicional
#Tenemos los datos de una serie
remesas <- read.csv(file="remesas1.csv",head=TRUE,sep=",")
n=length(remesas[[1]])
n
remesas[[1]]
media<-mean(na.omit(remesas[[1]]))
media
b<-is.na(remesas[[1]])
b
for(i in 1:n) {if(b[i]==TRUE) remesas[[1]][i]<-media}
remesas

## Imputacion condicional
#Tenemos los datos de una serie
remesas <- read.csv(file="remesas1.csv",head=TRUE,sep=",")
```

```
#Visualizamos los datos como grupos
remesa_st=ts(remesas,start=1991,frequency=12)
remesa_st
#insertamos como vector el grupo con datos faltante
a2007<-c(269.0,320.2,310.3,338.0,NA,324.6,NA,281.6,323.8,283.5,351.1,270.5)
media_2007<-mean(na.omit(a2007))
media_2007
```

1.4.2.3. Imputación usando la mediana

```
#Usando los datos de las remesas del año 2007
median(na.omit(a2007))

#1.4.2.4. Imputación por Regresión
#Función de Imputación por Regresión Simple
#La segunda columna es la variable dependiente y es la que tiene los datos faltantes
ImputacionReg<-function(x) {
        n=length(x[[2]])
        nx<- na.omit(x)
           nx
        reg<-lm(nx[[2]]~nx[[1]])
          b<-is.na(x[[2]])
       for (i in 1:n) {if(b[i]==TRUE) x[i,2]<-reg$coefficients[[1]]+x[i,1]*reg$coefficients[[2]]}
       cat("Modelo de Regresión: y=",reg$coefficients[[1]],"+",reg$coefficients[[2]],"*x \n")
          x
   }
peso<-c(82,75,70,68,44,NA,72,85,95,70,75,59,69,68,75,70,NA,57,63,80,NA,54,54)
estatura<-c(185,185,180,178,159,172,176,183,185,179,186,169,176,176,174,177,170,
161,170,190,185,162,165)
datos<-data.frame(estatura,peso)
datos
ImputacionReg(datos)
```

1.4.2.5. Imputación por series de tiempo

```
#Tenemos los datos de una serie
remesas <- read.csv(file="remesas1.csv",head=TRUE,sep=",")
remesas
```

```
remesas[[1]][0:197]

remesa_st=ts(remesas[[1]][0:197],start=1991,frequency=12)
remesa_st
pred.hw2=HoltWinters(remesa_st)

#Observamos el mejor valor del alpha
pred.hw2$alpha

#Observamos el mejor valor del beta
pred.hw2$beta

#Observamos el mejor valor del gamma
pred.hw2$gamma

#Graficamos los datos con al predicción del suavizado
#exponencial triple
plot(pred.hw2)

#Observamos los valores predichos
fitted(pred.hw2)

# Sumas de cuadrados error para el Suavizado Exponencial
# Triple
pred.hw2$SSE

#MSE
MSE_ET<-pred.hw2$SSE/197;MSE_ET

#Tipo de Estacionalidad
pred.hw2$seasonal

#Predicción hasta agosto de 2007
predict(pred.hw2,n.ahead=3)
```

1.4.2.6. Imputación usando el vecino más cercano

```
###Función nnmiss extraída de la librería seqKnn
nnmiss<-function (x, xmiss, ismiss, K)
{
    xd <- as.matrix(scale(x, xmiss, FALSE)[, !ismiss])
    dd <- drop(xd^2 %*% rep(1, ncol(xd)))
```

```
    od <- order(dd)[seq(K)]
    od <- od[!is.na(od)]
    K <- length(od)
    distance <- dd[od]
    s <- sum(1/(distance + 1e-15))
    weight <- (1/(distance + 1e-15))/s
    xmiss[ismiss] <- drop(weight %*% x[od, ismiss, drop = FALSE])
    xmiss
}

###Función SeqKNN extraída de la librería seqKnn
SeqKNN<-function (data, k)
{
    x <- as.matrix(data)
    N <- dim(x)
    p <- N[2]
    N <- N[1]
    nas <- is.na(drop(x %*% rep(1, p)))
    xcomplete <- x[!nas, ]
    xbad <- x[nas, , drop = FALSE]
    missing <- c()
    for (i in seq(nrow(xbad))) {
        missing[i] <- sum(is.na(xbad[i, ]))
    }
    missingorder <- order(missing)
    xnas <- is.na(xbad)
    xbadhat <- xbad
    cat(nrow(xbad), fill = TRUE)
    for (i in seq(nrow(xbad))) {
        j <- order(missingorder[i])
        xinas <- xnas[missingorder[i], ]
        xbadhat[missingorder[i], ] <- nnmiss(xcomplete, xbad[missingorder[i],
            ], xinas, K = k)
        xcomplete <- rbind(xcomplete, xbadhat[missingorder[i],
            ])
    }
    x[nas, ] <- xbadhat
    x
}


dato<-matrix(c(3,4,5,2,3,1,4,6,2,6,3,4,
5,NA,3,1,4,1,3,5,4,2,6,3,4,1,NA,3,6,4),nrow=15,ncol=2)
dato
imputacion.ejemplo<-SeqKNN(dato,10)
imputacion.ejemplo
```

1.4.2.8. Algoritmo EM(Expectation Maximization)

```
em.norm <- function(Y){
          Yobs <- Y[!is.na(Y)]
          Ymis <- Y[is.na(Y)]
          n <- length(c(Yobs, Ymis))
          r <- length(Yobs)
 # Valores Iniciales
      mut <- mean(Yobs) #*
      sit <- var(Yobs)*(r-1)/r  #**
 # Definimos la Función de Máxima Verosimilitud
      ll <- function(y, mu, sigma2, n){
      -.5*n*log(2*pi*sigma2)-.5*sum((y-mu)^2)/(sigma2)
         }
#Se Calcula la Función de Verosimilitud con los Valores iniciales
     lltm1 <- ll(Yobs, mut, sit, n)
     repeat{
# Paso-E (Estimación)
     EY <- sum(Yobs) + (n-r)*mut
     EY2 <- sum(Yobs^2) + (n-r)*(mut^2 + sit)
# Paso-M (Maximización)
     mut1 <- EY / n
     sit1 <- EY2 / n - mut1^2
# Se Actualiza los  valores de los Parámetros
     mut <- mut1
     sit <- sit1
#Se calcula la Función de Máxima Verosimilitud con los Valores Iniciales
    llt <- ll(Yobs, mut, sit, n)
#Se despliega los parámetros de Máxima Verosimilitud Actuales
     cat(mut, sit, "\n")
#Se detiene la convergencia según el error prescrito
     if ( abs(lltm1 - llt) < 0.001) break
       lltm1 <- llt
     }
# Se llenan los valores perdidos con el parámetro mu encontrado.
       b<-is.na(Y)
       for (i in 1:n){if(b[i]==TRUE) Y[i]<-mut}
       Y
}

x<-c(5.226416,4.246948,5.719032,5.062461,6.621925,NA,4.720129,NA,4.199522,
5.607660,NA,4.401418,5.273086,4.567361,4.205563,4.353384,3.852736,NA,6.328042,NA)

imputacion.em<-em.norm(x)
imputacion.em
```

1.4.3.1. Imputación por redes Neuronales

```
# Datos para entrenar la red
x1<-c(0,1,0,2,4,2,1,2,5,4,4)
x2<-c(1,1,2,3,2,5,3,4,1,4,4)
x3<-c(0,0,1,1,0,1,1,1,0,1,0)
datos<-data.frame(x1,x2,x3)
datos

library(nnet)
# Creamos la red 2-1
and.nn<-nnet(x3~x1+x2, datos, size=0, decay=1e-4, linout=T,
skip=T, maxit=1000, Hess=T)
# observamos los resultados con los pesos de los nodos
summary(and.nn)

# Ya que este es un proceso estocástico los valores pueden
#cambiar en cada ejecución
# pero el número de conexiones con este código no cambiará
# por lo tanto siempre
# con estos datos nos resultaran 3 pesos con una red neuronal

# Datos para probar la red
x1<-c(0,2,1,0,4)
x2<-c(0,0,5,3,5)
x3<-c(NA,NA,NA,NA,NA)
pruebadatos<-data.frame(x1,x2,x3)
### con los datos de prueba creamos predicciones y vemos la
# efectividad de la red
predicciones=predict(and.nn,newdata=pruebadatos)
predicciones

# Se puede observar que la red trabaja con una efectividad
#del 100% pero necesitamos la variable
# sea binaria por lo que hacemos una aproximación sin decimales
round(predicciones, 0)
```

1.4.3.2. Imputación por Regresión Aleatoria o Estocástica

```
#La segunda columna es la variable dependiente y es la que tiene
# los datos faltantes
```

```
ImputacionRegA<-function(x) {
        n=length(x[[2]])
        nx<- na.omit(x)
           nx
        reg<-lm(nx[[2]]~nx[[1]])
        residuos.est<-stdres(reg)
          b<-is.na(x[[2]])
       for (i in 1:n) {if(b[i]==TRUE){
            residuo<-sample(residuos.est,1)
            x[i,2]<-reg$coefficients[[1]]+x[i,1]*reg$coefficients[[2]]+residuo[[1]]}}
       cat("Modelo de Regresión: y=",reg$coefficients[[1]],"+",reg$coefficients[[2]],"*x \n")
          x
   }
peso<-c(82,75,70,68,44,NA,72,85,95,70,75,59,69,68,75,
   70,NA,57,63,80,NA,54,54)
estatura<-c(185,185,180,178,159,172,176,183,185,179,186,169,
   176,176,174,177,170,161,170,190,185,162,165)
datos<-data.frame(estatura,peso)
datos
ImputacionRegA(datos)
```

1.4.3.3. Imputación Aleatoria de un Caso Seleccionado

```
#VADU
x<-c(5.226416,4.246948,5.719032,5.062461,6.621925,NA,
4.720129,NA,4.199522,5.607660,NA,4.401418,5.273086,4.567361,
4.205563,4.353384,
3.852736,NA,6.328042,NA)
N<-length(x)
Xobs<-x[!is.na(x)]
Xmis<-x[is.na(x)]
n<-length(Xmis)
i=min(Xobs)
f=max(Xobs)
muestra<- seq(from = i, to = f, by = 0.0001)
vade<-sample(muestra,n)
sust<-vade
length(vade)
#Sustituir valores
b<-is.na(x)
j=1
for (i in 1:N){
```

```
     if(b[i]==TRUE){
     x[i]<-sust[j]
     j=j+1}
     }
x  # vector completo

#VADE
N<-length(x)
x<-c(5.226416,4.246948,5.719032,5.062461,6.621925,NA,4.720129,
NA,4.199522,5.607660,NA,4.401418,5.273086,4.567361,
4.205563,4.353384,3.852736,NA,6.328042,NA)
Xobs<-x[!is.na(x)]
Xmis<-x[is.na(x)]
sust<-rnorm(length(Xmis),mean=mean(Xobs))
sust
#Sustituir valores
b<-is.na(x)
j=1
for (i in 1:N){
     if(b[i]==TRUE){
     x[i]<-sust[j]
     j=j+1}
     }
x  # vector completo
```

1.4.3.4. Bootstrap, Imputación por Métodos Bayesianos (ABB)

```
#     Approximate Bayesian Bootstrap      #
library(LaplacesDemon)

ImputacionBB<-function(Y,m){
      m=m
      Yobs<-Y[!is.na(Y)]
      Ymis<-Y[is.na(Y)]
      k <- length(Yobs)
      r <- length(Ymis)
      n <- length(c(Yobs, Ymis))
      bootstrap<-ABB(Y,m)
      mu_estimada<-c(rep(0,m))
      boot<-c(rep(0,r))
      U<-c(rep(0,m))
       for (i in 1:m){
```

```
           sumobs<-0
           sumboot<-0
           mu_estimada[i]<-(1/n)*(sum(Yobs)+sum(bootstrap[[i]]))
           cat("Mu",i," ",mu_estimada[i], "\n")
           boot<-bootstrap[[i]]
            for (j in 1:k){sumobs<-sumobs+(Yobs[j]-mu_estimada[i])^2}
            for (l in 1:r){sumboot<-sumboot+(boot[l]-mu_estimada[i])^2}
           U[i]<-(1/(n*(n-1)))*(sumobs+sumboot)
           cat("U",i," ",U[i], "\n")
            }
        mu_IM=mean(mu_estimada)
        cat("mu_IM",mu_IM, "\n")
       v_estimada<-mean(U)+((m+1)/m)*var(mu_estimada)
        cat("Varianza(mu_IM)",v_estimada, "\n")

        }
Y<-c(4,5,3,7,4,1,8,4,1,NA,4,6,1,NA,6,NA,7,2,7,2,1)
ImputacionBB(Y,7)
```

Códigos Capítulo III

3.2.1. Base de datos con valores faltantes

```
base_jab<-read.csv(file="jabali.csv",head=TRUE,sep=";")
attach(base_jab)
base_jab0<-data.frame(pH,Ca,Na,Mg,Cl,SO4,T,Dureza,ph.in.situ,Alcalinidad,Fluor,K)
base_jab0  #ver datos
```

3.2.2. Identificación del patrón de pérdida de datos

```
library("colorspace")
library(VIM)
incvars<-c("pH","Ca","Na","Mg","Cl","SO4","T","Dureza","ph.in.situ","Alcalinidad","Fluor","K")
aggr(base_jab0[, incvars], numbers=TRUE, prop = c(TRUE, FALSE))
```

3.2.4.1. Mecanismo de pérdida de datos

```
library(BaylorEdPsych)
library(mvnmle)
LittleMCAR(base_jab0)
```

3.2.7. Selección de la técnica de imputación

```
ph_insitu<-lm(ph.in.situ~Na+Cl+SO4+Fluor)
ph_insitu
ph_insitu$coefficients[[1]]

n<-length(ph.in.situ)
estim<-c(rep(0,n))
for (i in 1:n){ estim[i]<-ph_insitu$coefficients[[1]]+Na[i]*ph_insitu$coefficients[[2]]+
Cl[i]*ph_insitu$coefficients[[3]]+SO4[i]*ph_insitu$coefficients[[4]]+
Fluor[i]*ph_insitu$coefficients[[5]]}

comp<-data.frame(estim,ph.in.situ,ph.in.situ-comp)
#Comparación de los estimados y reales
na.omit(comp)
mean(abs(na.omit(ph.in.situ-estim)))
#media de las diferencias de los estimados y reales

hist(ph.in.situ)
qqnorm(ph.in.situ) #grafico de normalidad
qqline(ph.in.situ)
shapiro.test(ph.in.situ) #test de normalidad

# Imputar la variable Dureza
pred<-2.5*Ca+4.16*Mg
Dureza1<-data.frame(pred,Dureza,abs(pred-Dureza))
Dureza1

# Imputar la variable Alcalinidad
#Modelo de regresion Alcalinidad
Alcalinidad_imp<-lm(Alcalinidad~Na+Ca+K+Mg)
Alcalinidad_imp

comp<-data.frame(estim1,Alcalinidad,Alcalinidad-estim1)
 #Comparación de los estimados y reales
na.omit(comp)
mean(abs(na.omit(Alcalinidad-estim1)))
 #media de las diferencias de los estimados y reales

alcalinidad1<-c(411.7, 505, 517, 504, 491, 549, 555, 566, 553, 562, 426,
409.74, 403.7, 383.75, 416.73, 411.74, 509.32, 507.03,
503.76, 589.62, 493.3, 545.81, 527.5, 552, 505.8, 600.68,
526.85, 522.01, 615.44, 611.82, 514.44, 556.79, 514.64, 520.12,
515.64, 530.95, 525.83, 515.56, 519.41, 532.87, 513.21,
523.37, 511.38, 415.17, 420.39, 472.63, 449.49, 473.05, 448.4,
469.68, 407.88, 368.26, 377.91, 377.14)
```

```
hist(alcalinidad1)  #histograma
qqnorm(alcalinidad1) #grafico de normalidad
qqline(alcalinidad1) #recta de normalidad

caso36<-Alcalinidad_imp$coefficients[[1]]+Na[36]*Alcalinidad_imp$coefficients[[2]]+
Ca[36]*Alcalinidad_imp$coefficients[[3]]+K[36]*Alcalinidad_imp$coefficients[[4]]+
Mg[36]*Alcalinidad_imp$coefficients[[5]]

Alcalinidad0<-Alcalinidad
Alcalinidad0[36]<-caso36
Alcalinidad0

#Alcalinidad Usando Algoritmo EM
x<-Alcalinidad0
imputacion.em<-em.norm(x)

#Alcalinidad Usando VADE
x<-Alcalinidad0
x
N<-length(x)
Xobs<-x[!is.na(x)]
Xmis<-x[is.na(x)]
sust<-rnorm(length(Xmis),mean=mean(Xobs))
sust
#Sustituir valores
b<-is.na(x)
j=1
for (i in 1:N){
     if(b[i]==TRUE){
     x[i]<-sust[j]
     j=j+1}
     }
x  # vecotr completo


##Imputación de la temperatura por el vecino mas cercano:

dato<-data.frame(T,pH)
dato
imputacion.ejemplo<-SeqKNN(dato,10)
imputacion.ejemplo

## Imputación de la temperatura por la mediana:
Temp<-T
median(na.omit(Temp))
```

```
Temp[22]<-median(na.omit(Temp))
Temp


Temp1<-T
ImputacionBB(Temp1,7)
```

3.2.7.2. Validación de la Técnica de Imputación

```
Asimetria=function(x) {
m3=mean((x-mean(x))^3)
skew=m3/(sd(x)^3)
skew}

Curtosis=function(x) {
m4=mean((x-mean(x))^4)
curt=m4/(sd(x)^4)-3
curt}

Resumen<-function(x){
          cat("n",length(x),"\n")
          cat("Media",mean(x),"\n")
          cat("Mediana",median(x),"\n")
          cat("Varianza",var(x),"\n")
          cat("Desviación Estándar",sd(x),"\n")
          cat("Error Estándar",sd(x)/sqrt(length(x)),"\n")
          cat("Coeficiente de Asimetría",Asimetria(x),"\n")
          cat("Curtosis",Curtosis(x),"\n")
          cat("Rango",max(x)-min(x),"\n")
          cat("Mínimo",min(x),"\n")
          cat("Máximo",max(x),"\n")
          cat("Percentil 25",quantile(x,.25),"\n")
          cat("Percentil 50",quantile(x,.5),"\n")
          cat("Percentil 75",quantile(x,.75),"\n")
              }


#### pH in situ sin imputar
pHinsitu<-ph.in.situ
hist(pHinsitu)
Resumen(na.omit(pHinsitu))
```

```
#### pH in situ usando Máxima verosimilitud - Algoritmo EM
pHinsitu<-ph.in.situ
pHinsitu[7]=pHinsitu[22]<-6.240516
pHinsitu
hist(pHinsitu)
Resumen(pHinsitu)

#### pH in situ usando Imputación Aleatoria VADE
pHinsitu<-ph.in.situ
pHinsitu[7]<-6.348273
pHinsitu[22]<-5.233151
pHinsitu
hist(pHinsitu)
Resumen(pHinsitu)

####Dureza sin Imputar
dureza<-Dureza
dureza
hist(dureza)
Resumen(na.omit(dureza))


####Dureza con imputacion por el modelo químico
dureza<-Dureza
dureza[28]<-752.9478
dureza[44]<-712.4774
dureza
hist(dureza)
Resumen(dureza)


##### Alcalinidad sin Imputar
alcalinidad<-Alcalinidad
alcalinidad
hist(alcalinidad)
Resumen(na.omit(alcalinidad))

##### Alcalinidad Usando regresión Multiple -
# Máxima Verosimilitud -EM
alcalinidad<-Alcalinidad
alcalinidad[13]<-479.5157
alcalinidad[36]<-566.8728
alcalinidad
hist(alcalinidad)
Resumen(alcalinidad)
```

```
##### Alcalinidad Usando regresión Multiple -
# Imputación Aleatoria VADE
alcalinidad<-Alcalinidad
alcalinidad[13]<-477.5086
alcalinidad[36]<-566.8728
alcalinidad
hist(alcalinidad)
Resumen(alcalinidad)

##### Temperatura sin Imputar
Temp<-T
Temp
hist(Temp)
Resumen(na.omit(Temp))

##### Temperatura usando el Vecino más cercano
Temp<-T
Temp[22]<-25.1
hist(Temp)
Resumen(Temp)

##### Temperatura usando la mediana
Temp<-T
Temp[22]<-25.5
hist(Temp)
Resumen(Temp)

##### Temperatura usando ABB
Temp<-T
Temp[22]<-25.43405
hist(Temp)
Resumen(Temp)
```

Bibliografía

[1] Acuña, E. y Rodríguez, C. *The treatment of missing values and its effect in the classifier accuracy: Four different methods to deal with missing values* , S.p.i. 2-3p. (2009)

[2] Alfaro, Rosario y Alfaro, Rafael *Aplicación de algunos métodos de relleno a series anuales de lluvia de diferentes regiones de Costa Rica . International Journal of Tropical Biology and Conservation, Top. Meteor. Oceanogr*, 7(1):1-20, San José. (2000)

[3] Capa Santos Holger y Pacheco Toscano Adriana. *Alternativas a los problemas de la imputación clásica de datos. Una aplicación al sistema nacional interconectado del Ecuador.*,CIMAC Y TCÍA. LTDA. Quito. (2006)

[4] Craig K. Enders *Applied Missing Data Analysis*, THE GUILFORD PRESS, New York. (2010)

[5] Dempster, N. M. and Laird, D. B. Rubin. *Journal of the Royal Statistical Society.*,Series B (Methodological), Vol. 39, No. 1. pp. 1-38. (1977)

[6] Doménico, Patrick A. y Schwartz, Franklin. *W. Physical and Chemical Hydrogeology.* John Wiley & Sons, Inc. USA. (1990)

[7] Donado Garzon, Leonardo David. *"Hidrogeoquímica"Hidrogeología.* Aplicaciones y casos de estudio latinoamericanos. Colombia, v.1 , p.1 - 538. (1999)

[8] Driscoll, Fletcher G. *Grondeater and Wells.* Johnson Division. Second Edition. USA. (1986)

[9] Droesbeke, J.-J. and Lavallee, P. *La non-réponse dans les enquêtes.*Methodologica, nº 4, pp. 1-39. (1996)

[10] Graham, J. W, Hofer, S. M. & MacKinnon, D. P, *Maximizing the usefulness of data obtained with planned missing value patterns: An application of maximum likelihood procedures. Multivariate Behavioral Research*, 31(2), 197-218. (1996)

[11] Goicochea, P. *Imputación basada en árboles de clasificación*. Eustat. (2002)

[12] Infante Saba, Ortega José y Cedeño Fernando. *Estimación de datos faltantes en estaciones meteorológicas de Venezuela vía un modelo de redes neuronales.*,Revista de Climatología Vol. 8 (0208): 51-70 ISSN 1578-8768. Caracas.(2008)

[13] Juárez Alonso, Carlos Alberto *Fusión de Datos. Imputación y Validación.*, Tesis Doctoral. Universitat Politécnica de Cataluya, Barcelona. (2004)

[14] Kim, J. *"A Note on Approximate Bayesian Bootstrap Imputation."*,Biometrika. Vol. 89, No. 2, Pp. 470-477. (2002)

[15] Little, R. J. A. & Rubin, D. B., *"Statistical analysis with missing data."*Biometrika. Vol. 89, No. 2, Pp. 470-477. New York: John Wiley & Sons. (1987)

[16] Little, R.J Rubin, D.B. *Statistical Analysis with Missing Data*, Statistical Analysis with Missing Data. Second Edition. New Jersey: John Wiley & Sons. (2002)

[17] Marco R. Steenbergen. *Maximum Likelihood Programming in R*, University of North Carolina. (2006)

[18] Marí, Gonzalo. *Metodología de Imputación de datos Faltantes y cálculo de la varianciа de las estimaciones*, Instituto Nacional de Estadística y Censos. Programa MECOVI-Argentina. Buenos Aires. (2001)

[19] Matthai, A. *."Estimation of Parameters f rom Incomplete Data with Application to Design of Sample Surveys ."*, Sankhya 2, 11,145-152. (1951)

[20] McKnight, P. E, McKnight, K. M, Sidani, S, & Figueredo, A. J. *Missing data: A gentle introduction*. New York, NY: Guilford Press. (2007)

[21] Navarro Pastor, José Blas. *Aplicación de Redes Neuronales Artificiales Al tratamiento de Datos Incompletos.*, Tesis Doctoral. Universidad Autónoma de Barcelona, Barcelona. (1998)

[22] Pacheco, A. J. y Capa, H. *Tratamiento Estadístico a la Pérdida e Inconsistencia de Datos del Módulo de Registro Histórico del Sistema de Manejo de Energía del Ecuador del Centro Nacional de Control de Energía – CENACE.*,Escuela Politécnica Nacional. Quito. (2009)

[23] Parzen, Michael, Stuart Lipsitz, and Garrett Fitzmaurice. *"A Note on Reducing the Bias of the Approximate Bayesian Bootstrap Imputation Variance Estimator."*, Biometrika. Vol. 92, No. 4, Pp. 971-974. (2005)

[24] Platek, R. *Metodología y tratamiento de la no-respuesta*, seminario internacional de estadística en Euskadi. Eustat. (1986)

[25] Romero Rojas, Jairo Alberto. *Acuiquímica*. Escuela Colombiana de Ingeniería. Santafé de Bogotá. (1996)

[26] Rubin, D.B. *Multiple Imputation for Nonresponse in Surveys.* Wiley, New York. (1987)

[27] Uribe, Iván Amón. *Guía Metodológica para la selección de técnicas de depuración de datos.*Universidad Nacional de Colombia. Medellín. (2010)

[28] Otero García, Deborah. *Imputación de datos faltantes en un Sistema de Información sobre Conductas de Riesgo.*Tesis de Maestría, Universidades de Santiago de Compostela A. Coruña. (2011)

[29] Useche Lelly y Mesa Dulce. *Una Introducción a la imputación de Valores Perdidos.* Terra Nueva Etapa, año/vol. XXII, número 031, pp. 127-151. Universidad Central de Venezuela. Caracas. (2006)

[30] Wilks,S. S. *"Moments and Distributions of Estimates of Population Parameters from Fragmentary Samples."* Annals of Mathematical Statistics, 3. 163-195. (1932)