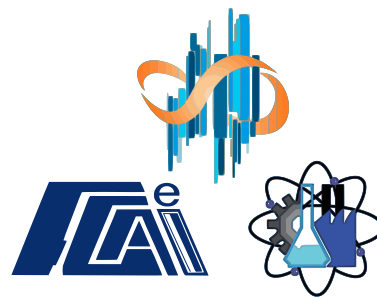




UNIVERSIDAD AUTÓNOMA DEL
ESTADO DE MORELOS



FACULTAD DE CONTADURÍA, ADMINISTRACIÓN E
INFORMÁTICA
SECRETARÍA DE INVESTIGACIÓN Y POSGRADO
MAESTRÍA EN OPTIMIZACIÓN Y COMPUTO APLICADO.

Imputación equivalente de una base de datos de fluidos geotérmicos

T E S I S

QUE PARA OBTENER EL TÍTULO DE:
MAESTRA EN OPTIMIZACIÓN Y COMPUTO APLICADO

PRESENTA:

Mariana Alelí Román Flores

DIRECTOR DE TESIS:
Lorena Díaz González

COASESOR DE TESIS:
Guillermo Santamaría Bonfil

Cuernavaca, Morelos

28 de octubre de 2019

*“La conformidad es el carcelero de la libertad y el enemigo del
crecimiento.”*
John F. Kennedy.

Esta tesis está dedicada:

A Dios por guiar mi camino.

A mis padres Adriana y Antonio, quienes con su amor, esfuerzo y paciencia me han ayudado a cumplir hoy una meta más.

A mis hermanos Valeria y Jafet por su cariño y apoyo incondicional durante este proceso.

A mis papitos Ma. de Jesús y Sebastian porque con sus oraciones, consejos y palabras de aliento hicieron de mi una mejor persona, por que siempre me acompañan en cada uno de mis sueños y metas.

A mi novio Miguel Ángel por su paciencia, motivación y apoyo en los momentos más complicados. Te quiero, amor.

A mis tíos Sergio, Gamaliel, Lorena y Ana Lilia quienes con su ayuda, cariño y apoyo me han alentado siempre a seguir adelante y sobre todo por ser mi inspiración a superarme.

Mariana A. Román Flores

Agradecimientos

En primer lugar doy infinitamente gracias a Dios por llenar mi vida de bendiciones.

Gracias a toda mi familia por brindarme su amor y comprensión. Por el apoyo en las incontables decisiones que eh tomado a lo largo de mi vida, unas buenas, otras malas. Por permitirme desenvolverme como ser humano. Son lo más bello que Dios ha puesto en mi camino y por quienes estoy inmensamente agradecida con él.

Mi más grande y profundo agradecimiento a la Dra. Lorena Díaz y al Dr. Guillermo Santamaría por brindarme la confianza para desarrollar este trabajo. Por compartir conmigo sus conocimientos y sobretodo, gracias por la paciencia orientación y el apoyo durante el proceso de investigación y redacción para culminar con éxito la meta propuesta.

No puedo dejar de agradecerte especialmente a ti Javier, mi compañero y gran amigo a quien estimo tanto y a quien le debo su apoyo incondicional. Por compartir conmigo tus conocimientos y por los ánimos que siempre me diste para seguir adelante.

Y por supuesto a mi querida Universidad y a todos los profesores, por permitirme concluir con esta etapa de mi vida.

Mariana A. Román Flores

Resumen

Los geotermómetros estadísticos se utilizan en la exploración de sistemas geotérmicos explotables. Su precisión depende de los datos que, debido al ruido de los sensores y las circunstancias experimentales, son escasos, caros e incompletos. En la práctica, solo se utilizan registros completos de elementos químicos para el modelado geotermométrico, las relaciones de restricción y la incertidumbre que se pueden modelar y la estimación de la temperatura de polarización. Por lo tanto, proponemos un método de imputación equivalente para completar una base de datos geotérmica basada en solutos a partir de pozos productivos reportados. El método completa registros que determinan el mejor modelo de imputación basado en ambos, la precisión en la predicción de los valores observados y la equivalencia estadística entre los datos imputados y observados. Los valores se estiman mediante algoritmos de imputación simples (punto de referencia, máquinas de vectores de soporte y redes neuronales artificiales) y múltiples (imputaciones multivariadas por ecuaciones encadenadas). Los resultados muestran que, la metodología propuesta produce datos equivalentes y precisos en la mayoría de los solutos probados con cambios insignificantes en la correlación. Esto también muestra que la precisión por sí sola no es suficiente para seleccionar el mejor método de imputación. Al combinar las pruebas de equivalencia con un análisis de precisión, los registros geotérmicos de la literatura se pueden completar de manera equivalente, aumentando sustancialmente la información disponible para el modelado geotérmico estadístico.

Abstract

Statistical geothermometers are used in the exploration of exploitable geothermal systems. Its accuracy depends on data which due to sensors noise and experimental circumstances is scarce, expensive, and incomplete. In practice only complete chemical elements records are used for geothermometric modeling, constraining relations and uncertainty which can be modeled, and biasing temperatures estimation. Therefore, we propose an equivalent imputation method for building a solute-based geothermal database from reported productive wells. The method completes records determining the best imputation model based on both, accuracy on predicting observed values and the statistical equivalence between imputed and observed data. Values are estimated by single (benchmark, Support Vector Machines, and Artificial Neural Networks) and multiple (Multivariate Imputations by Chained Equations) imputation algorithms. Results show that, the proposed methodology produces equivalent and accurate data in most of the tested solutes with negligible changes in correlation. These also show that accuracy on its own is not sufficient for selecting the best imputation method. By combining equivalence tests with an accuracy analysis, geothermal records from literature can be completed in an equivalent fashion substantially increasing the available information for statistical geothermal modeling.

Índice general

1. Introducción	1
1.1. Planteamiento del problema	4
1.2. Objetivo general	5
1.3. Objetivos específicos	5
1.4. Estructura de la tesis	6
2. Métodos de imputación de datos faltantes	7
2.1. Mecanismos de pérdida de datos	8
2.1.1. Faltan completamente al azar (MCAR, del inglés <i>Missing Completely At Random</i>)	9
2.1.2. Faltan al azar (MAR, del inglés <i>Missing At Random</i>)	10
2.1.3. No faltan al azar (MNAR, del inglés <i>Missing Not At Random</i>)	10
2.2. Imputación única	13
2.2.1. Media y mediana	14
2.2.2. Regresión estocástica	15
2.2.3. Máquinas de vectores de soporte	16
2.2.4. Redes Neuronales Artificiales	19
2.3. Imputación múltiple	20
2.3.1. Imputación multivariante por ecuaciones encadenadas	21
2.4. Criterios de evaluación de modelos de imputación	28
2.4.1. Medidas de error	29
Error cuadrático medio (RMSE, del inglés <i>Root Mean Square Error</i>)	30

Error absoluto medio (MAE, del inglés Mean Absolute Error)	30
Coeficiente U de Theil	31
2.4.2. Pruebas de equivalencia	31
3. Metodología	36
3.1. Estructura de la metodología aplicada	37
3.2. BD de fluidos geotérmicos	38
3.3. Normalización de datos	44
3.4. Proceso de imputación	44
3.4.1. Selección de predictores	45
3.5. Entrenamiento de modelos de imputación única y múltiple	49
3.5.1. Validación de modelos	50
3.5.2. Pruebas de equivalencia	51
3.5.3. Selección del modelo adecuado	51
3.5.4. Adición de variable imputada en la nueva BD . .	52
4. Resultados	53
4.1. Modelos de imputación	54
4.2. Validación de modelos de imputación	57
4.3. Pruebas de equivalencia de los modelos de imputación . .	60
4.4. Modelos de imputación seleccionados	64
4.5. Nueva base de datos geotérmicos mundial	67
4.5.1. Estadística descriptiva final	67
5. Conclusiones y trabajos futuros	68
Referencias	93

Índice de figuras

1.1. Proceso de iteración agua-roca	3
2.1. Esquema de pasos básicos en la imputación múltiple. . .	21
2.2. Ejemplo del flujo de valores de la media y varianza que de- terminan convergencia y no convergencia en las iteraciones del algoritmo MICE.	27
2.3. Escenarios de las pruebas de equivalencia	34
3.1. Diagrama de flujo de la metodología empleada para la imputación de la BDFG.	37
3.2. Distribución de campos geotérmicos	42
3.3. Mapa de calor de la matriz de correlación	46
3.4. Mapa de calor del indicador de casos disponibles	47
4.1. Distribución de datos observados e imputados de <i>Li</i> , <i>Mg</i> y <i>Ca</i>	55
4.2. Distribución de datos observados e imputados de <i>Cl</i> , <i>SO₄</i> y <i>HCO₃</i>	56

Índice de cuadros

3.1. Localidades geotérmicas mundiales y fuentes de literatura utilizadas para la creación de la base de datos geotérmica de fluidos geotérmicos.	39
3.2. Información estadística de temperatura y elementos químicos.	43
3.3. Matriz de predictores	49
4.1. Validación de modelos de imputación	58
4.2. Resultados de pruebas de equivalencia	62
4.3. Modelos seleccionados	65
4.4. Estadística final de la BDFG	67

Capítulo 1

Introducción

La geotermia es una fuente de energía amigable con el medio ambiente, sus recursos han sido explotados desde 1904 en la primera planta de energía geotérmica en Larderello, Italia [1]. Esta fuente de energía se basa en la explotación de reservorios geotérmicos, los cuales se forman en zonas con altas temperaturas a poca profundidad del nivel del suelo y acuíferos cubiertos por roca permeable que ocasionan vapor a altas presiones. Por medio de pozos geotérmicos, cuyas profundidades oscilan entre los 1000 a 2000 m de profundidad, se extrae el vapor a presión, el cual, hace girar una turbina conectada al generador de energía eléctrica, luego que el vapor ha pasado por la turbina, se enfría y convierte en líquido para reinyectarse al pozo y comenzar el ciclo nuevamente.

En la etapa de exploración de nuevos reservorios geotérmicos, la estimación de las temperaturas de fondo es uno de los procedimientos geoquímicos más importantes para evaluar su potencial energético. La estimación de este parámetro puede realizarse de dos maneras: (i) mediante la perforación de pozos someros de diámetro pequeño, que nos otorgan temperaturas con mayor precisión y exactitud, sin embargo, la realiza-

ción de esta actividad representa aproximadamente el 10 % del total de un pozo geotérmico de mayor diámetro (cuyo costo atendiendo a su profundidad puede oscilar entre 3 y 5 millones de dólares); y (ii) mediante herramientas geotermométricas que proporcionan temperaturas estimadas. Considerando el costo que representa sus mediciones in-situ con equipo especializado y aunado a las pérdidas económicas generadas por la perforación de pozos someros fallidos, las herramientas geotermométricas también conocidas como geotermómetros, constituyen el medio más económico y viable para inferir las temperaturas de fondo antes de realizar perforaciones.

Los geotermómetros, son ecuaciones creadas a partir de bases de datos con mediciones in-situ de temperatura y composiciones químicas de fluidos geotérmicos muestreados en las manifestaciones naturales sub-superficiales (p. ej. manantiales hidrotermales, fumarolas, etc.) y pozos productores.

En general, los geotermómetros se fundamentan teóricamente en los procesos geoquímicos y termodinámicos que dominan en los sistemas geotérmicos. Esto es, cuando los fluidos están en el interior del yacimiento, la ocurrencia de procesos geoquímicos de interacción agua-roca dan lugar a que algunos de los componentes químicos residentes en la roca se particionen favorablemente hacia el fluido durante un tiempo determinado, hasta alcanzar un estado de equilibrio químico (ver Fig.1.1). De este modo, la composición de los fluidos es gobernada por la solubilidad de algunos minerales, la cual a su vez, depende de las condiciones de presión y temperatura que prevalecen en el reservorio [2].

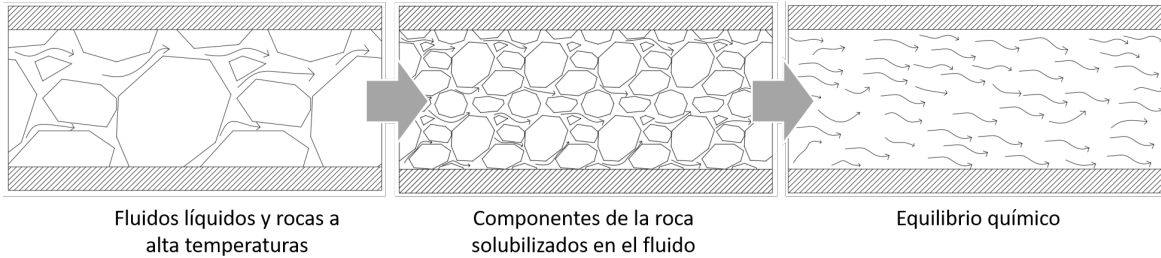


Figura 1.1: Proceso de iteración agua-roca

Estas herramientas se han utilizado en la industria geotérmica durante casi 60 años. Entre los geotermómetros reportados en la literatura podemos encontrar los geotermómetros catiónicos de Na/K , $K - Mg$, $Li - Mg$, $Na - Li$, $Na/K - Ca$, Na/K , Mg y $Na/K - Ca * Mg$. Sin embargo, uno de los principales inconvenientes de estas herramientas es que tienden a obtener una estimación sesgada de la temperatura de fondo, debido entre otras cosas, a que la cantidad de datos disponibles para su desarrollo es pequeña, incompleta y ruidosa [3]. Por ejemplo, [4] utilizó una red neuronal artificial de perceptrón multicapa entrenada con solo 20 datos de campos geotérmicos de Turquía lo cual limita su aplicación para otros campos geotérmicos. [5] entrenó una red neuronal con un algoritmo de retro propagación utilizando 39 datos de pozos geotérmicos de varios países, mismos que fueron utilizados para la validación de sus modelos lo que ocasiona un sesgo estadístico en la ecuación. Por otra parte [2] presentan tres nuevas ecuaciones implementando una red neuronal de dos capas utilizando una base de datos con 112 registros de campos geotérmicos de diferentes países, dos de ellos proveen mejores resultados cuando las temperaturas son mayores a 160° . Finalmente, [6] implementa una red neuronal con algoritmos genéticos para optimizar los pesos de las neuronas de capa oculta utilizando 324 datos recolectados de pozos geotérmicos de diversos pozos productores, sin embargo, en

este trabajo concluyen que se necesitan datos más confiables para tener un geotermómetro que represente mejor los recursos por debajo de los 160°C.

1.1. Planteamiento del problema

Ante tal escenario, [7] compiló una base de datos con 708 registros de fluidos geotérmicos y temperaturas de fondo medidas de pozos productores de diferentes países, derivados de mediciones de campo y laboratorio reportados en artículos publicados en revistas arbitradas y memorias de congresos internacionales con arbitraje. Desafortunadamente, la base de datos de fluidos geotérmicos (BDFG) presenta ausencia de datos en la mayoría de las variables, ya que no fueron reportadas por los autores de los trabajos consultados. Aunque representa una buena opción aumentar los datos disponibles mediante la recopilación de análisis químicos a partir de los estudios geotermométricos, o incluso mediante muestreos físicos (de manera más costosa), la información recopilada sobre elementos químicos sigue estando en la mayoría de los casos incompleta [8]. Alternativamente, este problema puede solucionarse mediante un proceso de *imputación de datos*, el cual consiste en el reemplazamiento de datos faltantes por valores calculados [9].

El presente trabajo se centra en la obtención de una nueva base de datos de fluidos geotérmicos (NBDFG) completa a partir de la base de datos compilada y proporcionada por [7] mediante el diseño e implementación de una *metodología de imputación equivalente* que considera técnicas de imputación única y múltiple fundamentada en la aplicación de medidas de error y pruebas de equivalencia.

1.2. Objetivo general

Con el fin de incrementar la información y obtener una nueva base de datos de fluidos geotérmicos completa y con ella contribuir al futuro desarrollo de una herramienta geotermométrica que mejore las estimaciones de la temperatura de fondo de reservorios geotérmicos, se implementará una metodología de imputación equivalente, basada en la comparación y selección de modelos de imputación única y imputación multivariable .

1.3. Objetivos específicos

1. Aplicar modelos de imputación:
 - **Única:** Media y mediana, regresión estocástica, máquinas de vectores de soporte y redes neuronales y
 - **Múltiple:** Imputación multivariada por ecuaciones encadenadas.
2. Validar los modelos usando las medidas de error RMSE, MAE y U de Theil's.
3. Realizar un análisis comparativo de distribuciones mediante pruebas de equivalencia.
4. Seleccionar el mejor modelo de imputación para cada variable con base en los resultados de validación y análisis estadístico de los modelos.
5. Completar los registros faltantes con los datos imputados por los modelos seleccionados

1.4. Estructura de la tesis

El resto del documento está organizado de la siguiente manera: En el capítulo 1 se presenta una breve introducción a la geotermia, así como el planteamiento del problema, objetivo general y los objetivos específicos de este proyecto. En el capítulo 2 se presentan los fundamentos teóricos de las técnicas de imputación que se implementarán en la metodología (Capítulo 3), su clasificación y aplicaciones en diversas áreas. El capítulo 3 describe la información contenida en la BDFG y la estructura de la metodología aplicada para este proyecto. El capítulo 4 muestra los resultados de la validación, pruebas de equivalencia, y análisis comparativo de los modelos. Finalmente, en el capítulo 5 se presentan las conclusiones de este trabajo, los trabajos futuros y las referencias.

Capítulo 2

Métodos de imputación de datos faltantes

Introducción

En este capítulo se analizan los fundamentos teóricos de los métodos de imputación única e imputación múltiple implementados en este trabajo de tesis, se describe la teoría en la que se sustentan y las áreas en que se han aplicado. Así como también, se describen los parámetros de error y el procedimiento de las pruebas estadísticas que se utilizarán para comparar y seleccionar el modelo con el mejor desempeño en la imputación de la BDFG.

2.1. Mecanismos de pérdida de datos

Los valores faltantes son un problema común en casi todos los estudios de investigación [10], [11], particularmente, en los de exploración de reservorios geotérmicos los registros de composiciones geoquímicas incompletas pueden ocurrir debido a las mediciones debajo de los límites de detección de los sensores o las técnicas utilizadas para el análisis [12], [13]. Por ejemplo, [14], [15] estiman las temperaturas de los reservorios de Dixie Valley y las aguas termales utilizando el método de geotermometría multicomponente, sin embargo, se informó que faltaban las concentraciones de aluminio (Al) y magnesio (Mg), posiblemente debido a su límite inferior de detección o a la contaminación de la muestra.

En [16] los autores propusieron un método basado en el análisis de componentes principales (PCA) y la geotermometría multicomponente para estimar las firmas térmicas de aguas subterráneas en Surprise Valley utilizando grandes conjuntos de datos de muestras de fluidos. Las muestras incompletas que comprenden concentraciones de boro (*B*), sodio (*Na*), potasio (*K*), calcio (*Ca*), magnesio (*Mg*), cloro (*Cl*), sulfato (*SO₄*), bicarbonato (*HCO₃*), óxido de silicio (*SiO₂*) o flúor (*F*), fueron removidas debido a su concentración debajo del límite inferior de detección de los sensores. En otros casos, los registros de composiciones geoquímicas están simplemente incompletos y los autores no determinan las razones de su pérdida. Por ejemplo, en [17], se realiza un metanálisis basado en PCA y pruebas de hipótesis estadísticas basadas en la química de la esfalerita de los depósitos de plomo (*Pb*) y zinc (*Zn*). Se identificaron elementos trazas tales como galio (*Ga*), germanio (*Ge*) e indio (*In*), entre otros, que presentan datos faltantes debido al límite inferior de detección y/o errores de medición.

La forma en que se manejan los datos faltantes depende de las razones de cómo o por qué faltan los datos. De acuerdo a [18], antes de que se lleve a cabo cualquier proceso de imputación, es muy importante identificar los mecanismos de datos faltantes para seleccionar el algoritmo de imputación más apropiado. Si entendemos la presencia de valores perdidos como un fenómeno probabilístico necesitamos un mecanismo estadístico que describa las leyes que rigen su aparición, y que capte las posibles relaciones entre la aparición de valores perdidos y los datos no observados en sí mismos.

Para poder definir estos mecanismos, denotamos Z como variable con datos faltantes, S como un conjunto de variables completas y r_z como una variable binaria que toma el valor de 1 si faltan los datos en Z y 0 en caso contrario.

Las categorías de la clasificación de [18] pueden expresarse mediante las siguientes declaraciones.

2.1.1. Faltan completamente al azar (MCAR, del inglés *Missing Completely At Random*)

Este mecanismo se presenta cuando la pérdida de datos en una variable es completamente aleatoria, es decir, la probabilidad de que falte un dato no depende de los valores observados o faltantes., MCAR es bastante común en los datos geoquímicos y puede representarse mediante la declaración 2.1.

$$Pr(r_z = 1|S, Z) = Pr(r_z = 1), \quad (2.1)$$

Donde la probabilidad de perder un valor en Z no depende de S o Z . Por ejemplo, un recipiente que contiene una muestra de concentraciones

químicas de una zona de estudio se rompe o se pierde por accidente, por lo que ya no se pueden medir las concentraciones.

2.1.2. Faltan al azar (MAR, del inglés *Missing At Random*)

La suposición de MAR surge si la probabilidad de pérdida de datos de una variable depende de los datos observados y no de los datos faltantes en sí. Este patrón ocurre con mucha menos frecuencia en la geoquímica, sin embargo, bajo este supuesto se desarrollan los métodos más eficientes [19]. MAR puede representarse mediante la declaración 2.2.

$$Pr(r_z = 1|S, Z) = Pr(r_z = 1|S), \quad (2.2)$$

Donde la probabilidad de perder un valor en Z depende de S (pero no de sí mismo), por lo tanto, su valor se puede estimar usando S. Por ejemplo, un sensor que falla ocasionalmente durante el proceso de obtención de datos debido a un corte de energía. En este ejemplo, las variables donde faltan datos no son la causa de los datos incompletos sino la correlación con una causa externa.

2.1.3. No faltan al azar (MNAR, del inglés *Missing Not At Random*)

MNAR ocurre cuando la probabilidad de que los datos faltantes en una variable se relaciona con la misma variable. Este mecanismo también es muy común en geoquímica particularmente debido a los límites de detección de los equipos de medición, sin embargo, las estimaciones de estos datos son muy difíciles de calcular debido a que para usar un análisis

apropiado se debe profundizar más en la razón de pérdida. MNAR incluye todos aquellos casos que no pueden ser clasificados como MCAR o MAR y se puede representar mediante la declaración 2.3.

$$Pr(r_z = 1|S, Z) = Pr(r_z = 1|Z), \quad (2.3)$$

Donde la probabilidad de tener datos faltantes en Z depende solo de sí misma. Por ejemplo, si el valor de concentración de un elemento químico está por debajo del límite de detección de un sensor, por tanto este no puede adquirir la información.

A diferencia de MAR, MCAR no puede predecirse solo a partir de las variables disponibles en la base de datos.

Generalmente, los datos contendrán poca información que nos permita decidir si los datos faltantes son MCAR, MAR o MNAR. En [17] los autores declaran específicamente que los valores de elementos incompletos faltan al azar (MAR). Vale la pena señalar que, es un requisito de cualquier método de imputación que el mecanismo de datos faltantes sea MAR. Desafortunadamente, las suposiciones para justificar el uso de un método de imputación son generalmente fuertes y no comprobables por completo [20].

Una vez que se verifica la factibilidad de la imputación, es decir, cuando se comprueba que los datos observados contienen información útil para predecir los valores faltantes, un procedimiento de imputación adecuado puede hacer uso de ella y obtener resultados precisos.

En el área geoquímica, un enfoque típico es considerar que los datos faltantes se producen debido a los límites de detección más bajos del sensor, que es un mecanismo MNAR [12]. En este caso, los valores de

los elementos químicos se estiman utilizando métodos de reemplazo (por ejemplo, una fracción del límite de detección o una fracción ajustada por parámetros de distribución lognormal), métodos de Estimación de máxima verosimilitud (MLE), Imputación aleatoria, Imputación múltiple (MI), Cosimulación combinada Modelo de Markov, por mencionar algunos [12], [13], [17]. En otros casos [14], [15], [21], [22] los valores de concentración de elementos desconocidos se estiman utilizando índices de saturación y equilibrio de minerales, o métodos estadísticos robustos [17]. Por ejemplo, en [21] se usa un método llamado *Fix-Al* para estimar los valores de *Al* faltantes al suponer que la concentración del elemento está limitada por el equilibrio termodinámico entre éste y su mineral portador. Mientras que en [14], [15], [22], las concentraciones desconocidas se calculan mediante la optimización numérica de la agrupación de índices de saturación de minerales cercanos a cero. Por otro lado, la mayoría de las metodologías de imputación asumen que los registros incompletos se deben a un mecanismo MAR [23]-[25]. En ambos casos, se puede llevar a cabo la imputación; sin embargo, en el caso del MNAR, debemos especificar qué registros están incompletos debido a la detección del límite inferior del sensor, lo que no es práctico y es difícil de demostrar utilizando solo la literatura. Por simplicidad aceptamos que los datos que faltan son de tipo MAR.

Por lo anterior, se considerarán seis técnicas de aplicación general basadas en modelos estadísticos para estimar los valores faltantes de la BDFG, las cuales se pueden clasificar en dos grupos: de imputación simple y múltiple, para lo cual la forma de estimar o predecir los valores perdidos diferenciará unas técnicas de otras. Según [26] un buen método de imputación debería:

- tomar en cuenta el proceso que originó los datos faltantes,

- preservar las relaciones entre las variables y
- preservar la incertidumbre entre esas relaciones.

En las siguientes secciones se detalla la aplicación de cada una de dichas técnicas.

2.2. Imputación única

Los métodos de imputación única se utilizan para estimar un valor por cada dato faltante a partir de los datos observados, generando un conjunto de datos completo. Para el presente trabajo se considera la imputación a través de los siguientes métodos: media, mediana, regresión estocástica (RE), máquinas de vectores de soporte (SVM, del inglés *Support Vector Machines*) y redes neuronales artificiales (RNA).

La imputación por la media, mediana y regresión estocástica han sido ampliamente utilizadas en distintas áreas, tales como, contaminación del ambiente, calidad del aire y medicina. Por ejemplo, en [20] evaluaron los métodos de imputación de la media, *hot deck* y maximización de expectativas (EM) en concentraciones de PM_{10} y concluyeron que el error de estos métodos es significativo cuando el porcentaje de datos faltantes es muy alto (e.g., 50 %). [27], [28] aplicaron la interpolación e imputación de la media en un conjunto de datos de concentraciones de PM_{10} , simularon diferentes porcentajes de datos faltantes y concluyeron que la media es el mejor método únicamente cuando el número de valores perdidos es pequeño. [29] compararon las técnicas de imputación única (e.g., la media, mediana y regresión lineal) y múltiple en conjuntos de datos simulados de calidad del aire y obtuvieron que las técnicas múltiples son mejores. [30] compararon el rendimiento de diversos métodos de imputación en un

conjunto de datos de calidad del aire y concluyeron que son superiores los métodos estadísticos multivariantes (e.g., la regresión estocástica) a los métodos univariados (e.g. mean). [31] compararon diversos métodos de imputación usando datos clínicos sobre el funcionamiento cognitivo de las personas, en donde la imputación de la media y la regresión estocástica obtuvieron el peor y el mejor rendimiento, respectivamente. En contraste, [32] demostraron que para el 30 % de datos faltantes estos dos métodos produjeron resultados similares. [33] aplicaron varios métodos de imputación estadística (e.g., mean, *hot-deck* e *imputación múltiple*), y técnicas de aprendizaje máquina (e.g., perceptron multicapa, mapas de auto organización y k-vecinos más cercanos) en un conjunto de datos de cáncer de mama, en donde los métodos de aprendizaje máquina presentaron mejores resultados.

2.2.1. Media y mediana

El método de sustitución por la media propuesto por primera vez por [34] es uno de los procedimientos de imputación más antiguo y más sencillo. Los datos faltantes se sustituyen con el promedio de los valores observados de la variable. El promedio se obtiene por la ecuación 2.4.

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} \quad (2.4)$$

donde n es el número de valores observados y x es cada valor observado de la variable. En su aplicación se asume que los datos faltantes son de tipo MCAR.

Por otro lado, la imputación de la mediana se realiza a partir del valor medio del vector que contiene los datos observados ordenados de

una variable incompleta, los valores faltantes de dicha variable se imputan a través de las ecuaciones 2.5 y 2.6.

Si la longitud del vector es par:

$$\tilde{x} = X\left[\frac{n+1}{2}\right] \quad (2.5)$$

Si la longitud del vector es impar:

$$\tilde{x} = \frac{X\left[\frac{n+1}{2}\right] + X\left[\frac{n}{2}\right]}{2} \quad (2.6)$$

donde X es el vector de datos observados ordenados y n la longitud de X .

2.2.2. Regresión estocástica

Un método un poco más robusto y popular es la regresión estocástica propuesto por [35], el cual incorpora la información de otras variables con la idea de producir mejores imputaciones. La imputación por regresión consiste en construir un modelo de regresión utilizando los valores observados de las variables que se correlacionen con la variable de interés y a partir del modelo ajustado se estiman los valores faltantes. La inclusión de tantas variables como sea posible tiende a hacer que el supuesto MAR sea mayormente aceptable [36]. La ecuación general de la regresión, para estimar los valores faltantes se representa mediante la ecuación 2.7.

$$\hat{Y} = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p + \varepsilon \quad (2.7)$$

donde β_0 y β_1, \dots, β_p son los coeficientes de \hat{Y} , x_1, \dots, x_p son los predictores y ε una variable con ruido aleatorio agregado a \hat{Y} . El término de error aleatorio es una variante normal con una media de cero y una des-

viación estándar igual al error estándar de la estimación de la ecuación de regresión [37].

La adición del error aleatorio es una estrategia utilizada para evitar que se subestime la varianza y se sobreestimen las correlaciones con la variable imputada. La imputación de la regresión estocástica es un importante paso adelante. En particular, conserva no solo los pesos de regresión, sino también la correlación entre las variables.

2.2.3. Máquinas de vectores de soporte

Las SVM tienen su origen en los trabajos de aprendizaje estadístico y fueron introducidas en los años 90 por Vapnik y colaboradores [38]. Aunque originalmente las SVM fueron pensadas para resolver problemas de clasificación, hoy en día, se utilizan para resolver otros tipos de problemas, como los de regresión para lo cual es muy común nombrarlas por el acrónimo SVR (del inglés Support Vector Regression). [39]. Este método está recibiendo cada vez más atención y se ha aplicado con éxito en una amplia gama de problemas, entre ellos la estimación de datos faltantes. Por ejemplo, en [40] presentan SVR como un método novedoso para la estimación de los valores faltantes en el perfil de expresión génica. Sus resultados muestran que la imputación por SVR tiene gran capacidad de predicción de los valores faltantes y es robusto contra el ruido.

El objetivo de SVR dado un conjunto de datos de entrenamiento $S = (x_1, y_1), \dots, (x_n, y_n)$, donde $x_i \in \mathbb{R}^d$ e $y_i \in \mathbb{R}$, es encontrar los parámetros $w = (w_1, \dots, w_d)$ que permitan definir un hiperplano como una función lineal que mejor se ajuste al conjunto de datos de entrenamiento:

$$f(x) = (w_1, x_1 + \dots + w_d, x_d) + b = \langle \mathbf{w}, \mathbf{x} \rangle + b \quad (2.8)$$

donde $w_i \in \mathbb{R} \forall_i = 1, \dots, d$ y $b \in \mathbb{R}$.

Para obtener un hiperplano óptimo, se debe cuantificar el error entre el valor real y la predicción de cada dato mediante una función de pérdida definida por la ecuación 2.9. El objetivo principal de esta función es anular los errores asociados con los puntos que caen dentro de lo que se denomina *margen blando*, el cual es un tubo formado a una distancia $\pm \varepsilon$ alrededor de la función lineal (2.8). De tal forma que todos los puntos de datos que queden sobre o fuera de la región definida por $\pm \varepsilon$ serán considerados *vectores de soporte*. En general, cuanto más grande es ε , menos vectores de soporte se necesitan. De esta forma, la resolución del problema dependería únicamente de los vectores de soporte y no de la dimensión de los datos de entrada.

$$L_\varepsilon(y, f(x)) = \begin{cases} 0 & \text{si } |y - f(x)| \leq \varepsilon \\ |y - f(x)| - \varepsilon & \text{en otro caso} \end{cases} \quad (2.9)$$

Es decir, sí la diferencia absoluta entre la predicción $f(x)$ y su valor real y_i es menor o igual que ε la función de pérdida se anula. En el caso contrario se definen dos variables de holgura ξ_i y ξ_i^* para cuantificar el error de predicción. Así, la variable $\xi_i > 0$ cuando $f(x) - y_i > \varepsilon$ y $\xi_i^* > 0$ cuando $y_i - f(x) > \varepsilon$. Por tanto, la función estándar de SVR se define como:

$$\min \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n (\xi_i + \xi_i^*) \quad (2.10)$$

donde w es la magnitud del vector o hiperplano, C es el parámetro de regularización que controla la compensación entre el margen ($\pm \varepsilon$) y el error de predicción (ξ_i). Un valor muy grande para la constante C

ayudaría a que la función predictora represente mejor el conjunto de datos. Por el contrario, un valor demasiado pequeño para C permitiría valores de ξ_i elevados, es decir, estaríamos admitiendo un número muy elevado de datos mal representados.

Por su parte, se presentan casos en los que los datos no tienen una tendencia lineal, es decir que en este tipo de datos jamás se encontrará un hiperplano óptimo, cuando se presenta esta situación el procedimiento es exactamente igual, la única diferencia es que antes de llevar a cabo el procedimiento anterior para buscar el hiperplano óptimo se implementa una función llamada *Kernel*:

$$\phi = \mathbb{R} \rightarrow F \quad (2.11)$$

El objetivo de la función Kernel denotada por ϕ es simplemente mapear el espacio de entrada \mathbb{R} (por ejemplo de 2 dimensiones) a uno de mayor dimensionalidad F (por ejemplo a 3 dimensiones) y aquí encontrar el hiperplano que mejor se ajuste a nuestro conjunto.

Para este caso la función de regresión que se busca es la siguiente:

$$f(x) = \langle w, \phi(x) \rangle + b \quad (2.12)$$

y en el caso de la función estándar de SVR se agrega la función ϕ , la cual quedaría de la siguiente manera:

$$\min \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n (\xi_i + \xi_i^*) \quad (2.13)$$

sujeto a

$$\begin{aligned} y_i - \langle w, \phi(x_i) \rangle &\leq \varepsilon + \xi_i & i = 1, \dots, n \\ \langle w, \phi(x_i) \rangle - y_i &\leq \varepsilon + \xi_i^* & i = 1, \dots, n \\ \xi_i \geq 0, \xi_i^* &\geq 0 & i = 1, \dots, n \end{aligned}$$

Existen diferentes funciones Kernel, sin embargo en este trabajo se implementa únicamente la función de Kernel lineal y la función de Kernel radial o (también llamado Kernel Gaussiano).

Kernel lineal:

$$K(x, x') = \langle x, x' \rangle \quad (2.14)$$

Kernel radial:

$$K(x, x') = \exp(-\gamma \|x - x'\|^2), \gamma > 0 \quad (2.15)$$

donde γ es un parámetro propio del kernel

El rendimiento del algoritmo depende de una buena configuración de los parámetros C y de los del Kernel. Por tanto, el problema de la selección óptima de parámetros se complica aún más por el hecho de que la complejidad del modelo depende de estos parámetros.

2.2.4. Redes Neuronales Artificiales

Por otro lado, las redes neuronales también han sido utilizadas en imputación de datos, por ejemplo, en tareas de clasificación de patrones [19], datos de plantas de energía industrial [41], entrevistas telefónicas y por computadora [42], censos de población [43] e incluso en inventarios de recursos nacionales [44]. A pesar de que se requiere entrenar varias redes

neuronales e implica un alto costo computacional, este enfoque mejora la calidad de una base de datos, ya que son adecuadas para estimar datos faltantes cuando existen relaciones no lineales entre las variables y se puede extender fácilmente para imputar variables discretas.

Las redes neuronales y el algoritmo de aprendizaje supervisado perceptrón multicapa (MLP, del inglés *Multi-Layer Perceptron*) que dado un conjunto de entrenamiento $(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(m)}, y^{(m)})$, donde $x^{(i)} \in \mathbb{R}^n$ es un elemento de entrenamiento y $y^{(i)} \in \mathbb{R}^o$ es la variable objetivo, puede entrenarse para aprender una función no lineal que se aproxima a $f : x \rightarrow y$. Como lo indica el nombre de este método, un MLP usa múltiples capas compuestas por unidades conocidas como neuronas. Consiste en una capa de entrada con $m + 1$ unidades donde cada unidad representa una característica para un ejemplo $\{x^{(i)} | x_1^{(i)}, \dots, x_n^{(i)}\}$, una o más capas ocultas, donde cada unidad aplica una suma lineal ponderada de los valores de salida de la capa anterior seguida de una función de activación, no lineal $a_j^{[l]} = g^{[l]}(\sum_k w_{jk}^{[l]} a_k^{[l-1]} + b_j^{[l]})$ y una capa de salida, con o unidades de salida, que reciben los valores de activación de la última capa oculta en la red y aplican una función lineal seguida de una función no lineal en el caso de un problema de clasificación o la función de identidad en el caso de un problema de regresión.

2.3. Imputación múltiple

A diferencia de las técnicas de imputación única, el procedimiento de imputación múltiple propuesto por [18], crea varios conjuntos de datos (indicados por m). Es decir, un valor faltante en el conjunto de datos original se reemplaza por m valores estimados en función de los valores observados de la variable objetivo y su relación con otras variables.

La figura 2.1 muestra los tres pasos principales de la imputación múltiple. El lado izquierdo de la imagen indica que el análisis comienza con un conjunto de datos observados e incompletos. En el primer paso se imputan $m = 3$ conjuntos de datos. Los tres conjuntos son idénticos en los datos observados, pero difieren en los valores imputados. En el segundo paso se analizan los resultados mediante alguna métrica estadística como el coeficiente de correlación. El último paso es combinar los conjuntos para obtener uno final, de acuerdo con el método descrito por [45] se obtiene la media de las m estimaciones y se imputa en el punto correspondiente.

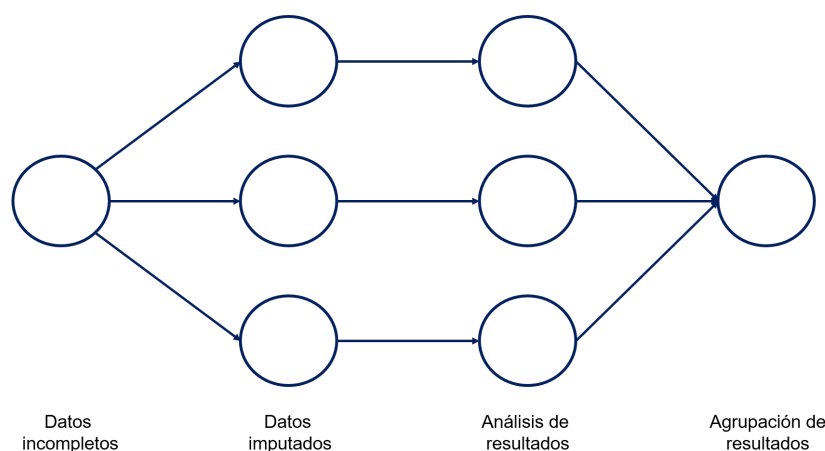


Figura 2.1: Esquema de pasos básicos en la imputación múltiple.

2.3.1. Imputación multivariante por ecuaciones encadenadas

La imputación multivariable por ecuaciones encadenadas (MICE) fue desarrollado por van Buuren y Groothuis-Oudshoorn [26], es una técnica particular de imputación múltiple que puede manejar datos tanto MAR como MNAR. Sin embargo, la imputación múltiple de datos tipo

MNAR requiere análisis adicionales que contribuye para generar mejores imputaciones. MICE ejecuta una serie de modelos de regresión, uno para cada variable incompleta. Esta técnica ha sido ampliamente utilizada para la imputación de datos faltantes de calidad energética [11], así como también datos de epidemiología [46], rasgos mamíferos [47], pronósticos médicos [48], pruebas de conocimiento y habilidades humanas [49], estudios en los que destacó por brindar un buen rendimiento.

El algoritmo MICE consiste en 4 pasos generales:

Algorithm 1 Imputación Multivariada por Ecuaciones Encadenadas

1. Los valores faltantes se imputan temporalmente con la media de los valores observados de la variable.
2. Para la primera variable incompleta (Y), los valores imputados en el paso 1 se eliminan nuevamente.
3. Genera un modelo de regresión (logístico, lineal, etc.) adecuado a Y , utilizando sus datos observados. En el modelo, Y es la variable dependiente y los predictores seleccionados son las variables independientes.
4. Los valores faltantes de Y se estiman con el modelo de regresión generado en el paso 3. Para los modelos posteriores donde Y es un predictor, se utilizarán sus valores observados e imputados.
5. Los pasos 2 a 4 se repiten para cada variable incompleta. Una vez que se imputen todas las variables se considerará una iteración. Una vez completado el número designado de iteraciones, se repite todo el proceso de imputación para generar m conjuntos de datos imputados.

Debido a que cada problema de la vida real es diferente, cada uno necesita adaptaciones especiales, que son la parte más importante en el procedimiento MICE, las principales se mencionan a continuación:

Modelo de imputación univariado

En MICE se especifica el método de regresión para cada variable incompleta de acuerdo a la escala de la misma, para variables numéricas pueden considerarse métodos como regresión lineal, regresión lineal bayesiana o coincidencia predictiva media (PMM, del inglés *Predictive Mean Matching*) y para variables categóricas métodos como regresión logística, regresión polinomial o análisis discriminante lineal. El modelo de regresión especificado toma un conjunto de variables completas como predictores correlacionadas con la variable objetivo. Para el presente trabajo se implementa MICE con el modelo de coincidencia predictiva media descrito a continuación.

Coincidencia predictiva media

La coincidencia predictiva media (PMM) es un esquema de imputación que combina algunos aspectos de los métodos de imputación paramétricos y no paramétricos [50]. PMM imputa los valores faltantes por medio de la selección aleatoria a partir de una submuestra de valores observados cercanos a su predicción, dichas predicciones se calculan mediante un modelo de regresión lineal.

La ventaja principal de utilizar este método es que, debido a que las imputaciones se basan en valores observados no se producirán imputaciones sin sentido (por ejemplo, concentraciones químicas negativas).

El algoritmo 2 describe los pasos para llevar a cabo el método PMM.

Algorithm 2 Coincidencia Predictiva Media

1. Mediante un modelo de regresión lineal especificado, se estima el valor faltante de la variable objetivo.
 2. Para cada estimación, se selecciona un pequeño conjunto de donantes candidatos (generalmente con 1, 3 o 10) de todos los valores completos cercanos al valor de la estimación.
 3. Se realiza un muestreo aleatorio entre los candidatos y se toma uno de ellos para reemplazar el valor faltante.
-

Hay varias formas de seleccionar al donante para el reemplazamiento de los datos faltantes. Para describir algunas se define y_i como un i -ésimo valor observado y \hat{y}_j como el valor predicho para el j -ésimo dato faltante.

- Define un umbral ϵ y toma todos los valores donde $y_i - \hat{y}_j \leq \epsilon$ como candidatos, luego selecciona aleatoriamente un donante para reemplazar el valor faltante.
- Toma el candidato más cercano como donante, es decir, el valor par el cual $y_i - \hat{y}_j$ es el mínimo.
- Encuentra los d candidatos más cercanos para los cuales $y_i - \hat{y}_j$ es el mínimo y selecciona uno aleatoriamente como donante.

Una desventaja obvia de este método, es la posible selección del mismo valor del donante muchas veces lo que ocasionaría un sesgo muy grande en las estimaciones. Es más probable tener este problema si la muestra es pequeña o tiene más datos faltantes que observados en una cierta región. Para disminuir las probabilidades de que esto ocurra se recomienda utilizar valores grandes para d como 3, 5 o 10.

Predictores del modelo univariado

Se debe especificar el conjunto de predictores que se utilizarán para cada variable incompleta. Para seleccionar adecuadamente las variables predictoras, Buuren y colaboradores [26] recomiendan incluir las variables que se utilizarían en los análisis si los datos estuvieran completos, las variables que satisfacen el supuesto de MAR (es decir, las que están relacionadas con la pérdida de datos), incluso versiones de variables transformadas, combinadas o recodificadas del conjunto de datos. Una vez que se identifican las variables anteriores, se inspeccionan dos parámetros de suma importancia, la correlación con la variable objetivo y el indicador de casos disponibles. Este último nos informa la cantidad de datos utilizables de los predictores, los que tienen demasiados valores faltantes no pueden incluirse en el modelo. De lo contrario, si la proporción de estos parámetros excede un umbral, las variables son aptas para incluirse en el modelo. Para verificar el valor adecuado de estos umbrales se requiere experimentar con múltiples valores.

Esquema de visitas

El esquema de visitas consiste en determinar el orden en el que las variables incompletas serán imputadas. En la práctica común, las imputaciones se realizan de izquierda a derecha (o viceversa). Teóricamente el orden es irrelevante, pero algunos esquemas son más eficientes que otros como el *monotónico* en el que las variables se ordenan de acuerdo con el número creciente de datos faltantes, ya que con dicho esquema se alcanza una convergencia casi inmediata [51] del algoritmo.

Convergencia

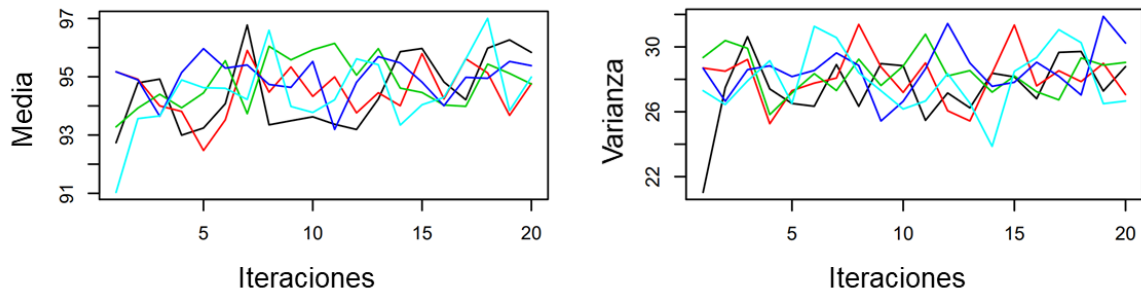
No existe un método claro para determinar si el algoritmo ha convergido. Lo que se hace a menudo es contrastar uno o más parámetros como la media y la varianza de las imputaciones en cada iteración. El trazado de estos valores da una buena idea de si la variabilidad entre imputaciones se ha estabilizado y si las estimaciones están libres de tendencia.

La figura 2.2a ejemplifica de forma visual la convergencia de una variable “ x ”, las líneas representan las variaciones en el valor de la media y varianza de las m imputaciones en cada iteración. En la figura 2.2a se aprecia que las líneas se mezclan muy bien entre sí desde el principio, en comparación con la figura 2.2b que ejemplifica una no convergencia, en la que se muestran que las líneas apenas se mezclan y se van lentamente a un estado estable.

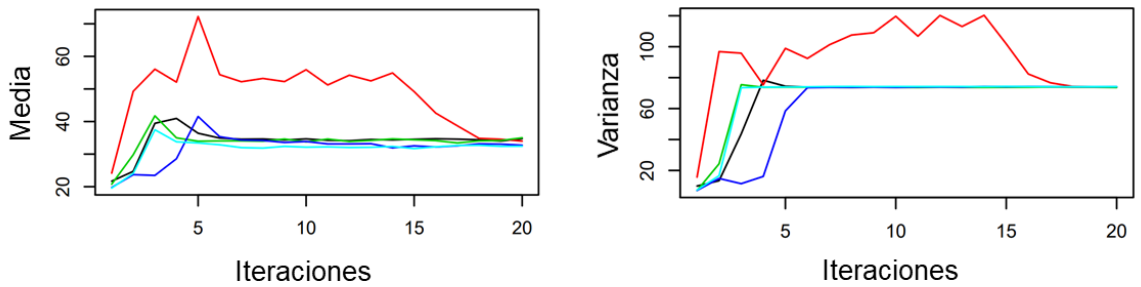
Una vez que se observe la convergencia, es recomendable calcular algunas iteraciones adicionales para evaluar la convergencia en tramos más largos.

Número de iteraciones

Sin duda, evaluar la convergencia del algoritmo ayudará en gran medida a determinar el número necesario de iteraciones. Sin embargo, también se debe de tomar en cuenta la cantidad de datos faltantes. Por ejemplo, para cantidades moderadas de 5 a 10 iteraciones es suficiente para producir buenos resultados. Para grandes cantidades la convergencia será más lenta por tanto será necesario incrementar el número de iteraciones.



(a) Convergencia



(b) No convergencia

Figura 2.2: Ejemplo del flujo de valores de la media y varianza que determinan convergencia y no convergencia en las iteraciones del algoritmo MICE.

Valor m

En particular, el tamaño del conjunto de datos y la cantidad de datos faltantes pueden ayudar a determinar cuántos conjuntos de datos imputados (m) generar. Para cantidades moderadas de 5 a 10 conjuntos de datos imputados serán suficientes. Sin embargo, para cantidades muy grandes, m puede aumentar hasta 40. Todo dependerá del caso. Por ejemplo, para imputar un conjunto con datos faltantes que tenga cientos de variables y miles registros, crear 40 conjuntos de datos imputados

puede ser poco práctico debido al costo computacional. Por el contrario, imputar un conjunto con datos faltantes con 20 variables y cientos de registros podría ejecutarse en minutos y, por lo tanto, crear 40 conjuntos de datos imputados sería bastante factible.

Una vez que se han imputado los datos, los m conjuntos de datos imputados están “completo” en el sentido de que no tienen datos faltantes el siguiente paso será ejecutar un análisis estándar (por ejemplo, medir la correlación entre las variables) en cada uno de los conjuntos de datos imputados y realizar las correlaciones a los modelos en caso de ser necesario. Finalmente, se deben combinar las estimaciones de cada conjunto de datos para obtener el resultado final.

La ecuación 2.16 propuesta por [45] indica como llevar a cabo dicha combinación.

$$\bar{x}_b = \frac{\sum_{k=1}^m x_b^{(k)}}{m} \quad (2.16)$$

Donde, se suman los valores estimados x_b de cada conjunto k , donde $k = 1, \dots, m$ y se divide entre el número total de conjuntos m para obtener un solo valor a imputar \bar{x}_b

2.4. Criterios de evaluación de modelos de imputación

Cuando se escogen diferentes modelos para la imputación de una misma variable como se pretende en este trabajo de tesis, se implementan diversos modelos y por tanto, diversas posibles soluciones. La pregunta que surge aquí es: ¿Cómo elegir o bajo qué criterios seleccionar un

modelo?. Una práctica frecuente que se presenta en la literatura [10], [11], [24], es evaluar la calidad de los modelos de imputación mediante funciones de pérdida que miden la diferencia entre los valores observados y sus predicciones correspondientes. Entonces, se podría argumentar que el problema puede reducirse a seleccionar el modelo de imputación que obtiene el error más pequeño. Sin embargo, para identificar el mejor método de imputación, no es suficiente medir el error interno, es decir el error entre los valores imputados y los datos observados que constituyen la base de datos de entrenamiento y validación. En consecuencia, además de las medidas de error, se propone aplicar pruebas de hipótesis estadísticas para garantizar un grado de similitud aceptable entre la distribución de los datos observados y la distribución de datos imputados que son realmente datos faltantes y que por lo tanto, no son parte de la base de datos de entrenamiento y validación. Ambos criterios de evaluación se describen a continuación.

2.4.1. Medidas de error

Las medidas de error son un criterio importante para probar los métodos propuestos y seleccionar el mejor modelo de imputación. Estas medidas surgen a partir de funciones de pérdida, que se aplican cuando se elimina artificialmente un subconjunto de valores del conjunto de prueba y después estos son estimados por un método de imputación. El principal objetivo de estas funciones es medir las diferencias entre la estimación dada por un modelo y su valor real, o bien para comparar las estimaciones de un modelo contra las de un modelo de referencia. En el presente trabajo se utilizarán las medidas de error clásicas, que se presentan a continuación con su formulación matemática correspondiente.

Error cuadrático medio (RMSE, del inglés Root Mean Square Error)

Esta medida surge de una función de pérdida cuadrática y se define como la raíz cuadrada de la media de los errores al cuadrado.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2} \quad (2.17)$$

Donde n es el número de muestras, y_i corresponde al i -ésimo valor real (observado) y \hat{y}_i la estimación de y_i . RMSE indica la exactitud del modelo, es decir, en que medida las estimaciones del modelo se acercan al valor real. La exactitud está relacionada con el sesgo de las estimaciones, cuanto menor es el sesgo más exacta es una estimación.

Error absoluto medio (MAE, del inglés Mean Absolute Error)

MAE surge de la función de pérdida del error absoluto y se define como el promedio de los errores, sin tener en cuenta su signo, es decir, el promedio de las diferencias absolutas.

$$MAE = \frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i| \quad (2.18)$$

Donde n es el número de muestras, y_i corresponde al i -ésimo valor real (observado) y \hat{y}_i la estimación de y_i . MA mide la precisión del modelo, lo que se refiere a la dispersión del conjunto de valores obtenidos, cuanto menor es la dispersión mayor es la precisión.

Aunque, las definiciones de RMSE y MAE son bastante parecidas, difieren en el hecho de que la primera mide la cercanía al valor real y la

segunda la frecuencia de resultados similares en distintas mediciones. Sin embargo, cuando los resultados se analizan utilizando ambas métricas, se debe tener en cuenta que cuanto mayor sea la diferencia entre ellas, mayor será la varianza en los errores individuales en la muestra, teniendo en cuenta que cuanto más pequeños sean sus valores, mejor será el modelo.

Coefficiente U de Theil

El parámetro U de Theil (también llamado coeficiente de diferencia r) es una medida de exactitud que indica la eficiencia de un modelo de predicción cuando se compara con otros. Se ha interpretado como la división del RMSE del modelo de referencia R entre el RMSE del modelo comparado C , como lo muestra la ecuación 2.19.

$$r = \frac{\sqrt{\sum_{i=1}^n (\hat{y}_i^R - y_i)^2}}{\sqrt{\sum_{i=1}^n (\hat{y}_i^C - y_i)^2}} \quad (2.19)$$

Donde n es el número de muestras y_i corresponde al i -ésimo valor real (observado) y \hat{y}_i la estimación de y_i . Cuando $r < 1$, indica que el error obtenido en el modelo de referencia es menor que el obtenido en el modelo comparado. Para $r = 1$, el modelo de referencia es igual que el modelo comparado. Por otro lado, si $r > 1$ el error del modelo de referencia es mayor que el del modelo comparado.

2.4.2. Pruebas de equivalencia

Anteriormente, la validación de un modelo de imputación se ha realizado mediante pruebas de significancia, donde la hipótesis de no diferencia es la hipótesis nula (H_0), en este caso, si la H_0 es aceptada quiere decir que el modelo es bueno. Por ejemplo, en [31], [46], se utilizan prue-

bas de hipótesis estadísticas como la prueba t , la prueba de McNemar, ANOVA y la prueba de la suma de rangos de Wilcoxon para verificar la similitud entre las muestras observadas y las imputadas, que se logra al aceptar la H_0 . Si bien esta idea es atractiva, las pruebas de hipótesis tradicionales no se consideran adecuadas para la validación de modelos de imputación. En estas pruebas de significancia convencionales, es importante tener en cuenta que si se acepta la H_0 no significa que se haya probado que sea verdadera, sólo que no se ha demostrado que sea falsa. En otras palabras, estas pruebas no pueden asegurar que las muestras son iguales o que provienen de la misma distribución [52]. Por lo anterior se propone probar la efectividad de las pruebas de equivalencia.

Las pruebas de equivalencia surgen originalmente en el campo farmacéutico [53]-[55], donde, se requiere que los medicamentos genéricos produzcan efectos semejantes a los de patente. De manera similar, para la validación de un modelo de imputación se requiere que éste produzca valores equivalentes con respecto a una muestra de valores medidos [56], [57]. Es decir, para mostrar la *equivalencia* entre dos muestras de datos, la diferencia de sus medias deben ser menor que una medida considerada de importancia mínima determinada por el investigador.

Un enfoque de prueba de equivalencia muy simple es el procedimiento de “dos pruebas t de un solo lado” (TOST, del inglés *Two-One Side Test*) en donde se especifica una *zona de indiferencia* definida por un límite de equivalencia superior ($+\Delta$) e inferior ($-\Delta$) para examinar si la diferencia de los datos observados son más grandes que dichos límites.

Los límites de equivalencia se pueden definir en puntajes brutos o en una diferencia estandarizada como $\Delta = d$ de Cohen. La d de Cohen es una medida del tamaño del efecto como diferencia de medias estandarizada. Es decir, cuántas desviaciones típicas de diferencia hay entre

los resultados de los dos conjuntos que se comparan. El cálculo de este parámetro se realiza mediante la ecuación 2.20

$$\Delta = d = \frac{\bar{x}_1 - \bar{x}_2}{\sigma} \quad (2.20)$$

Donde \bar{x}_1 es la media de los valores observados, \bar{x}_2 la media de los valores estimados y σ es la agrupación de las desviaciones estándar de ambas muestras calculada por la ecuación 2.21

$$\sigma = \sqrt{\frac{(n_1 - 1)DS_1^2 + (n_2 - 1)DS_2^2}{n_1 + n_2 - 2}} \quad (2.21)$$

Cuando el tamaño de ambas muestras es desigual, como en el caso de las muestras que se examinarán en este trabajo de tesis, debe realizarse la prueba de equivalencia que no supone varianzas iguales, la cual se basa en las dos pruebas t de la ecuación 2.22.

$$t_1 = \frac{\bar{x}_1 - \bar{x}_2 - (-\Delta)}{\sqrt{\frac{DS_1^2}{n_1} + \frac{DS_2^2}{n_2}}} \quad (2.22)$$

$$t_2 = \frac{\bar{x}_1 - \bar{x}_2 - (+\Delta)}{\sqrt{\frac{DS_1^2}{n_1} + \frac{DS_2^2}{n_2}}}$$

Donde n_1 y n_2 es el tamaño de las muestras de los valores observados y los valores estimados respectivamente, y por último DS_1^2 y DS_2^2 las desviaciones estándar de las muestra correspondientes.

En la prueba TOST, H_A significa que la diferencia de la media de dos conjuntos más un intervalo de confianza (por ejemplo el 95 %) cae dentro de la zona de indiferencia y está lo suficientemente cerca de cero, por lo que los dos conjuntos de datos se consideran equivalentes. Para esto, la

H_0 se descompone en dos:

$$\begin{aligned} H_{01} : t_1 &\leq -\Delta \\ H_{02} : t_2 &\geq +\Delta \end{aligned} \tag{2.23}$$

donde al rechazar H_{01} y H_{02} se afirma que la diferencia entre la media de los valores observados μ y la media de los valores estimados $\hat{\mu}$ es menor o igual a $+\Delta$ y mayor o igual que $-\Delta$ respectivamente.

Cuando se aplican las pruebas de equivalencia existe cuatro posibles resultados como lo ilustra la figura 2.3 en donde se representan la diferencia de las medias con cuadros negros, los intervalos de confianza con líneas horizontales alrededor de la diferencia de las medias y por ultimo los límites de equivalencia $+\Delta = 0.5$ y $-\Delta = -0.5$ indicadas por las líneas verticales punteadas.

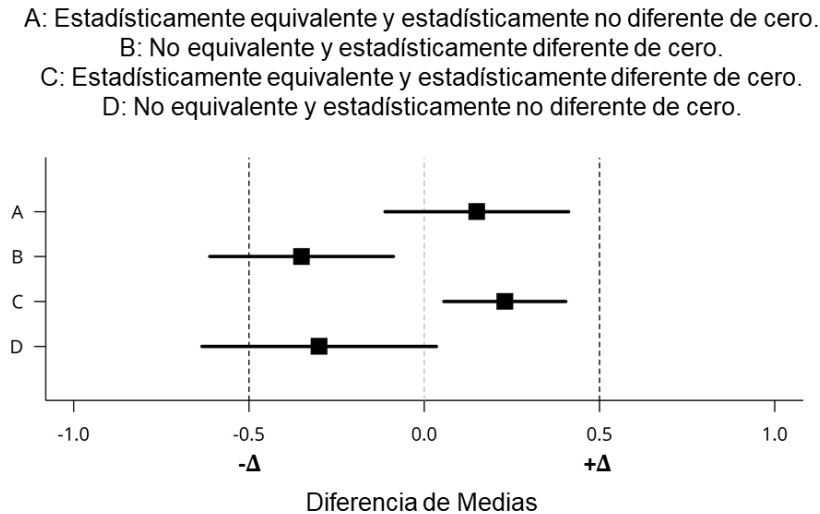


Figura 2.3: Escenarios de las pruebas de equivalencia

El escenario *A* concluye una equivalencia y no diferencia de cero entre dos muestras ya que las diferencias de sus medias e intervalos de confianza se encuentran entre los valores de $\pm\Delta$ e incluyen al cero. El escenario *B* representa no equivalencia ya que su intervalo de confianza alrededor de la diferencia de medias, supera el límite inferior de equivalencia y no incluye al cero. En el escenario *C* las muestras son equivalentes y estadísticamente diferentes de cero porque el intervalo de confianza se encuentra entre los valores $\pm\Delta$, pero excluyen al cero. Finalmente, en el escenario *D* se concluye indeterminación debido a que las muestras no son equivalentes y estadísticamente no diferente de cero debido a que el intervalo de confianza alrededor de la diferencia de medias, supera el límite inferior de equivalencia pero también incluye al cero.

Capítulo 3

Metodología

Los datos geotérmicos basados en solutos han sido generados dentro de diferentes lapsos de tiempo, por diferentes sectores y con objetivos distintos para su explotación. Estos datos se recopilan y emplean para construir una base de datos geoquímica y con ella, desarrollar modelos para mejorar la estimación de temperaturas de fondo de pozos geotérmicos. Desafortunadamente, estos registros recopilados conllevan el riesgo de estar incompletos debido a diversas razones, por lo que completarlos es primordial para explotar al máximo su uso.

En este capítulo se describe la metodología de imputación equivalente propuesta para lograr el completado de la BDFG. En ella se desarrolla la aplicación de las 6 técnicas de imputación propuestas, así como, el estudio comparativo de las mismas utilizando medidas de error convencionales y pruebas de equivalencia TOST.

3.1. Estructura de la metodología aplicada

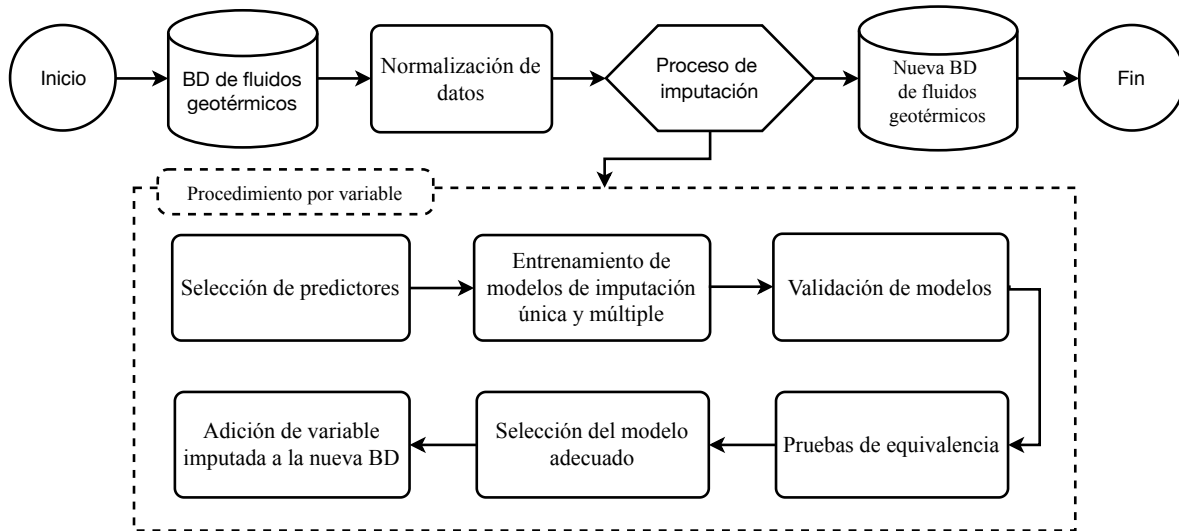


Figura 3.1: Diagrama de flujo de la metodología empleada para la imputación de la BDFG.

En la Fig.3.1 se muestra un esquema representativo de la metodología propuesta y desarrollada para la imputación de la BDFG, que consta de 4 fases principales: (i) lectura de la base de datos, la cual incluye entre otras variables, la composición química de fluidos geotérmicos y la temperatura de fondo medida en pozos productores; (ii) normalización de datos, en donde se realiza el escalado y normalización de datos; (iii) proceso de imputación, esta a su vez se subdivide en 6 fases, atendiendo particularmente variable por variable (incompleta) se seleccionan las variables que nos ayudarán de estimar los valores faltantes, se entrenan los modelos de imputación con la información disponible, los modelos se validan por medio de medidas de error aplicadas en un conjunto de prueba, una vez que se generan las imputaciones se realizan pruebas de

equivalencia a fin de evaluar su similitud con los datos observados, a partir de los resultados obtenidos en las medidas de error y pruebas de equivalencia, se selecciona el modelo que genere mejores resultados para la variable objetivo y se toman sus estimaciones para el completado de la misma; (iv) nueva base de datos, se refiere a la BDFG inicial imputada, es decir, completa.

En las siguientes secciones se explica de forma detallada la aplicación de esta metodología a la BDFG, los resultados obtenidos se presentan en el capítulo 4.

3.2. BD de fluidos geotérmicos

La creación de esta base de datos fue una actividad fundamental para el desarrollo de la tesis doctoral de [7]. Los datos de la BDFG fueron obtenidos de diversos artículos publicados en revistas y memorias de congresos internacionales arbitradas. En los artículos consultados se encontraron registros de aproximadamente 140 campos geotérmicos distribuidos en veinticinco países, entre ellos, Chile, China, Costa Rica, Djibouti, El Salvador, Etiopía, Alemania, Grecia, Hungría, Islandia, India, Indonesia, Italia, Japón, Santa Lucía, Dominica, México, Nueva Zelanda, Filipinas, Portugal, Rusia, Taiwán, Tailandia, Turquía y Estados Unidos.

En la Tabla 3.1 se reporta la fuente bibliográfica de donde fueron obtenidos dichos registros; el país donde se localizan los pozos geotérmicos productores muestreados; el número de registros por país (n); y el nombre de los campos geotérmicos; utilizados para la creación de la BDFG.

La distribución de los campos geotérmicos descritos en la tabla 3.1 se aprecia en la figura 3.2, cada campo se denota por símbolos cuadrados rellenos, diferenciados por color para cada país. Como se puede apreciar,

la mayor cantidad de fuentes termales ocurren en la unión de las placas tectónicas o lo que se denomina cinturón de fuego, debido a que es principalmente en estas zonas en donde el magma se almacena muy cerca de la superficie terrestre.

Cuadro 3.1: Localidades geotérmicas mundiales y fuentes de literatura utilizadas para la creación de la base de datos geotérmica de fluidos geotérmicos.

País	Campos geotérmicos	n	Referencias
Chile	El Tatio	11	[5], [58]
China	Yangb, Yangbajain y Tibet	12	[59]-[62]
Costa Rica	Miravalles y Guanacaste	47	[63]-[65]
Djibouti	Asal	4	[66]
El Salvador	Ahuachapan, Berlin, Chinameca, Las burras, Playón de Salitre y San Vicente	16	[5], [67]-[70]
Etiopía	Aluto-Langano	10	[59], [61]
Alemania	Hamburg	1	[61]
Grecia	Nisyros y Aghiasmata	4	[71], [72]
	Zunil y Tecuamburro	10	[61], [63], [73]
Hungría	Varosliger	1	[74]

Continúa en la siguiente página

País	Campos geotérmicos	n	Referencias
Islandia	Námafjall, Húsavík, Krafla, Nesjavellir, Reykjanes y Svartsengi	119	[61], [67], [74]-[84]
India	Puga	6	[61], [85]
Indonesia	Cisolok, Cisukarame, Citaman, Darajat, Kamojang, Salak, y Wayang Windu	7	[59], [86]
Italia	Larderello, Latera, Mofete, Northern Latium y Phlegrean	18	[87]-[89]
Japón	Beppu, Fushime, Hatchobaru, Kyushu, Matsukawa, Otake, Sumikawa, Takigami y Uenotai	46	[5], [61], [67], [90]-[95]
Santa Lucia	Qualibou Caldera	1	[96]
Dominica	Dominica	2	[88]
México	Cerro Prieto, Los Azules, Primavera y Las Tres Vírgenes	94	[58], [63], [67], [75], [83], [90], [97]-[102]

Continúa en la siguiente página

País	Campos geotérmicos	n	Referencias
Nueva Zelanda	Ngawha, Rotorua, Korako, Waiotapu, Kawerau, Rotokawa, Orakei Tauhara, Wairakei y	85	[5], [58], [61], [63], [64], [67], [74], [75], [90], [103]-[105]
Philippines	Tongonan, Bao, Okoy Valley	5	[59], [63]
Portugal	Sao Miguel y Azores	26	[106]
Rusia	Kamchatka	1	[74]
Taiwan	Chingshu y Tuchang	2	[75]
Tailandia	San Kampaeng	4	[59], [61], [107]
Turkia	Canakkale-Tuzla, Aydin-Germencik y Denizli-Kizildere	68	[5], [6], [74], [108]-[115]
Estados Unidos	California, Idaho, Nevada, Nuevo México y Utah	108	[5], [60], [61], [63], [74], [88], [90], [116]-[121]

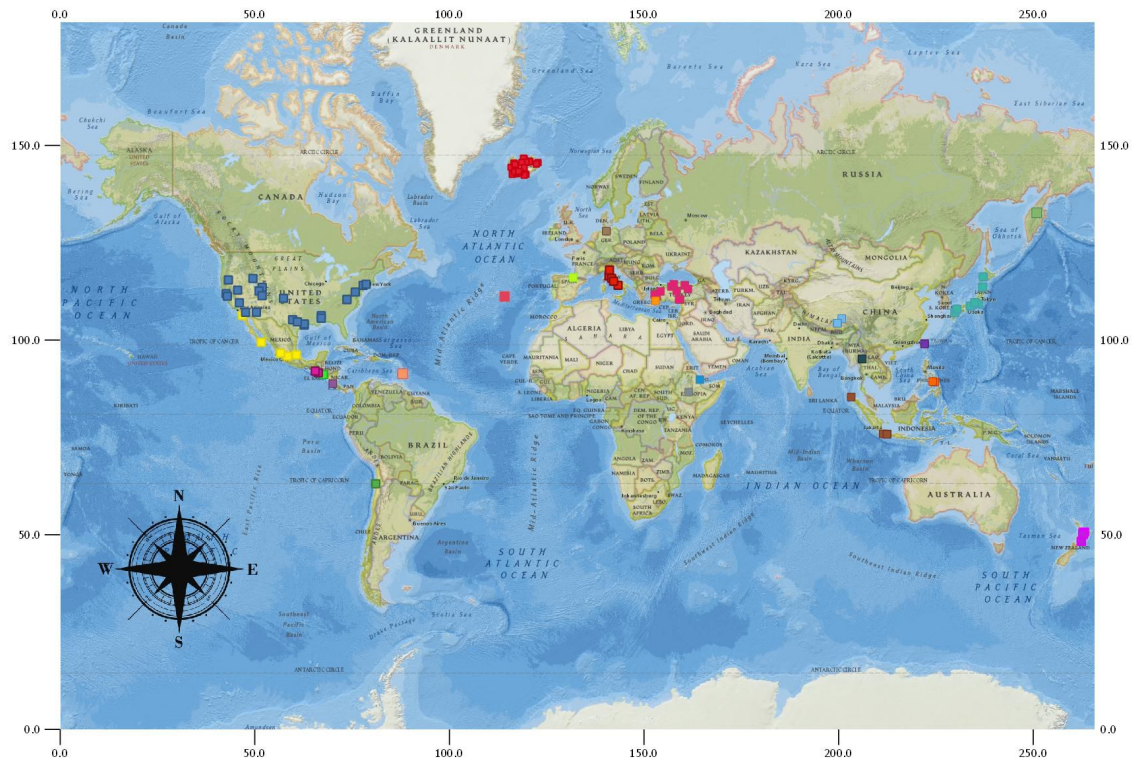


Figura 3.2: Distribución de campos geotérmicos

El formato de la BDFG consta de 9 columnas que incluyen los datos de *Temperatura* medida (en $^{\circ}C$) en pozos de producción y las composiciones de fluidos geotérmicos de 8 elementos químicos (en mg/L) Li , Na , K , Mg , Ca , Cl , SO_4 y HCO_3 y 708 filas que representan los registros reportados de las composiciones de fluidos muestreados.

La Tabla 3.2 presenta las estadísticas descriptivas básicas de la BDFG por variable recopilada, a saber: los valores mínimo (mín) y máximo (máx), media (\bar{x}) y mediana (\tilde{x}), desviación estándar (Sd), asimetría (γ_1), curtosis (γ_2) y el porcentaje de datos faltantes. Como se puede observar, *Temperatura*, Na y K tienen un porcentaje de pérdida de 0%, es decir, no tienen datos faltantes, esto debido, a que se estableció

como condición que se compilarían únicamente aquellos registros que reportaran al menos estas tres variables, dado que los geotermómetros basados en la relación Na/K han demostrado proporcionar temperaturas más confiables y consistentes en los estudios de exploración [2].

Cuadro 3.2: Información estadística de temperatura y elementos químicos.

Variable	Mín	Máx	\bar{x}	Sd	γ_1	γ_2	Pérdida %
<i>Temperatura</i>	59	359	216	69	-0.35	2	0 %
<i>Na</i>	22	565579	11472	520014	5	42	0 %
<i>K</i>	0.5	66473	1583	6555	8	76	0 %
<i>Li</i>	0.02	215	14	24	7	60	67 %
<i>Mg</i>	0.001	3920	115	512	5	30	16 %
<i>Ca</i>	0.06	55600	2303	7685	3	16	6 %
<i>Cl</i>	2	524690	6918	28523	12	205	22 %
<i>SO₄</i>	0.6	2500	140	247	4	27	27 %
<i>HCO₃</i>	0.01	3074	350	566	2	8	58 %

Las concentraciones de composición se muestran en mg/L , mientras que la temperatura se muestra en $^{\circ}C$.

En contraste, el porcentaje de datos faltantes para el resto de las variables varió de 6 % para *Ca* al 67 % para *Li*. De los 708 registros de la BDFG, 150 registros (21 %) tienen datos observados en todas las variables.

Los datos faltantes se imputaron utilizando seis técnicas diferentes, explotando la información disponible. Para ello se hizo la lectura de la BDFG de un archivo de entrada en formato de hoja de cálculo a un dataframe en la plataforma de lenguaje de programación R.

3.3. Normalización de datos

En el desarrollo de geotermómetros, escalar cationes y concentraciones de aniones utilizando la función logarítmica es una práctica común [122]. Esto también se sugiere en la comunidad de imputación [122] como una forma de inicializar correctamente los datos de entrada para los algoritmos de imputación. En este trabajo se realizó un escalado aplicando el logaritmo natural. Además, para abordar las diferencias entre escalas de características (por ejemplo, temperatura frente a concentraciones químicas), se llevó a cabo un procedimiento de escala adicional para ajustar los valores de todas las variables a un rango específico. A este procedimiento se le conoce comúnmente como *normalización*. Cabe mencionar que hay diversas formas de normalización, sin embargo, en este trabajo se utiliza la *normalización basada en la unidad*, la cual, escala todos los valores a un rango entre 0 y 1 mediante la siguiente ecuación:

$$X' = \frac{X - x_{min}}{x_{max} - x_{min}} \quad (3.1)$$

donde X representa la variable con los valores iniciales, x_{min} y x_{max} el valor mínimo y máximo de la variable X respectivamente.

3.4. Proceso de imputación

El proceso de imputación de datos se aplicó para cada variable incompleta de la BDFG, por lo que, los pasos descritos a continuación se refieren al procedimiento de imputación de una variable en particular a la cual llamaremos *variable objetivo*.

3.4.1. Selección de predictores

Una vez que los datos se prepararon adecuadamente, el siguiente paso fue obtener los datos disponibles que se utilizaron más adelante para generar los modelos de imputación. Cada modelo se construyó utilizando diferentes conjuntos de datos dependiendo su enfoque, ya sea univariado o multivariado. El univariado se refiere, a que el modelo utiliza únicamente los datos disponibles de una variable para imputar los valores faltantes de la misma, como es el caso de la media y la mediana, por lo que, para generar ambos modelos, se extrajeron los datos disponibles de la variable objetivo. Por ejemplo, para la imputación de Li se extrajeron sus 250 datos disponibles.

El enfoque multivariado se refiere a que los modelos utilizan dos o más variables (predictores) para estimar los datos faltantes de la variable objetivo. Sin embargo, estos a su vez pueden o no admitir datos faltantes en su ejecución. En la regresión estocástica, máquinas de vectores de regresión y redes neuronales no se admiten datos faltantes debido a que estos necesitan datos para generar una salida, de lo contrario, se ocasionan errores de compilación. Además, hacer que la suposición de que los datos perdidos son de tipo MAR sea más aceptable, implica que el número de predictores debe elegirse lo más grande posible [123]. Por tanto, para la construcción de estos modelos es necesario considerar solo las variables completas *Temperatura*, *Na* y *K* para incluirlas como predictores.

En contraste, MICE permite considerar variables incompletas en la construcción del modelo, debido a que (como lo describimos en la subsección 2.3.1), el algoritmo comienza con una imputación simple de la media y con esto las variables con datos faltantes se completan para la

estimación de la variable objetivo. Por tanto, la selección de predictores conlleva un análisis más extenso, como se describe a continuación.

Para especificar el conjunto de predictores de la variable objetivo, en primer lugar calculamos las correlaciones de Pearson utilizando los datos disponibles de cada par de variables de la BDFG. En la Fig.3.3 se muestra la matriz de correlación en forma de mapa de calor. Su representación gráfica es básicamente una línea recta diagonal en los ejes cartesianos en los que las abscisas son las variables y los coeficientes están representados por colores, los cuadros azul marino y rojo indican una correlación de 1 y -1 respectivamente, mientras los colores se van aclarando, los valores van disminuyendo hasta llegar al blanco que indica correlación 0.

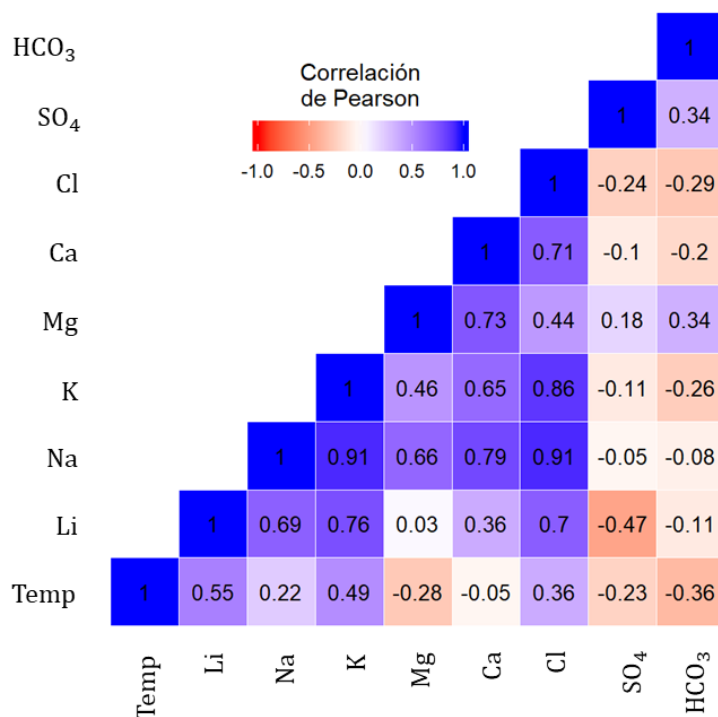


Figura 3.3: Mapa de calor de la matriz de correlación

Por ejemplo, podemos observar que *Li* tiene una correlación por arriba de 0.5 con *Temperatura* (~ 0.5), *Na* (~ 0.6), *K* (~ 0.7) y *Cl* (~ 0.7), así como una correlación muy cercana a 0 con *Ca* (~ 0.3), *SO₄* (~ -0.4), *HCO₃* (~ -0.1), y *Mg* (~ 0).

En segundo lugar, por medio del *indicador de casos disponibles*, se mide cuántos registros con datos faltantes en la variable objetivo tienen valores observados en el predictor. La proporción será baja si faltan tanto en la variable objetivo como la candidata a predictor en los mismos registros. Si es así, el predictor contiene poca información para imputar la variable objetivo, y podría descartarse para el modelo.

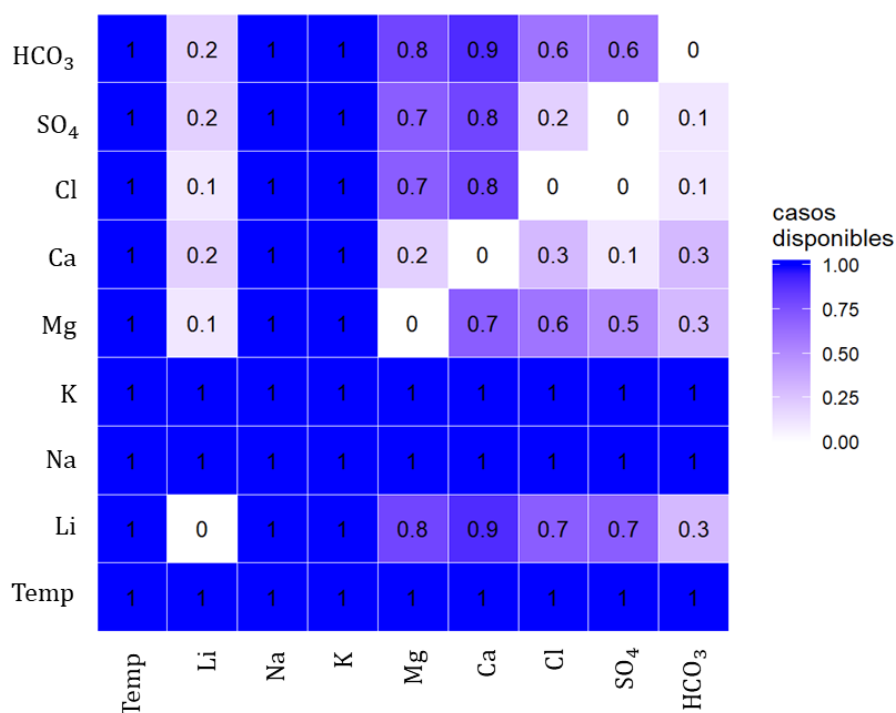


Figura 3.4: Mapa de calor del indicador de casos disponibles

En la Fig.3.4 se muestra el mapa de calor del indicador de casos disponible por cada par de variables (medido en porcentaje%). Las filas representan las variables a imputar (omitiendo *Temperatura*, *Na*, *K*) y

las columnas las variables candidatas a predictor, los cuadros de color azul fuerte, indican que la variable candidato esta informada al 100% cuando la variable objetivo tiene datos faltantes en los mismos registros, los cuadros de color blanco indican que la variable candidato está informada en un 0% es decir, tiene gran cantidad de datos faltantes en los mismos registros que la variable objetivo. Notese que la matriz del indicador de casos disponibles es asimétrica, por ejemplo, se puede observar que cuando Li (fila) tiene datos faltantes, las variables candidatas *Temperatura*, *Na*, *K*, *Ca* y *Mg* tienen un porcentaje de datos observados mayor al 75%. mientras que *Cl*, *SO₄*, *HCO₃* tienen un porcentaje $\leq 50\%$ de datos observados.

Ahora bien, una vez que conocemos los valores de ambos parámetros, buscamos establecer un umbral (para cada uno) que defina una correlación y un indicador de casos disponibles aceptable entre la variable objetivo y las variables candidatas. La finalidad de esto fue crear un conjunto de predictores lo suficientemente grande para aprovechar toda la información disponible y mejorar las estimaciones, cuidando que no se generara un sesgo debido a los valores faltantes o a la poca relación entre las variables.

En la Tabla 3.3 se identifican los valores para los umbrales de correlación (r) y porcentaje de datos observados (%) por variable objetivo, así como, los predictores (indicados con 1) que se incluyen en su modelo MICE.

Por ejemplo, para seleccionar el conjunto de predictores de Li se estableció un umbral de correlación ≥ 0.3 (absoluto) y un porcentaje de datos observados $\geq 70\%$. De acuerdo a la información mostrada en la Fig. 3.3 y 3.4 las variables que superan ambos umbrales son *Temperatura*, *Na*, *Ca*, *Cl*, por tanto, estas se definen como el conjunto de predictores en el

modelo de MICE para la variable Li .

Cuadro 3.3: Matriz de predictores

	Umbrales		Predictores								
	r	%	<i>Temperatura</i>	<i>Li</i>	<i>Na</i>	<i>K</i>	<i>Mg</i>	<i>Ca</i>	<i>Cl</i>	<i>SO₄</i>	<i>HCO₃</i>
<i>Li</i>	0.3	70	1	0	1	1	0	1	1	0	0
<i>Mg</i>	0.2	90	1	0	1	1	0	0	0	0	0
<i>Ca</i>	0.05	30	1	0	1	1	0	0	1	0	0
<i>Cl</i>	0.05	10	1	1	1	1	1	1	0	0	0
<i>SO₄</i>	0.05	10	1	1	1	1	1	1	1	0	1
<i>HCO₃</i>	0.05	20	1	0	1	1	1	1	1	1	0

r indica el coeficiente de correlación de Pearson y % indica el porcentaje de datos observados.

Para especificar los umbrales de cada variable objetivo, se realizaron pruebas con diferentes valores y al final se eligió la combinación que generaba el conjunto de predictores más conveniente para cada una. Observe que la matriz predictora no necesariamente es simétrica. Por ejemplo Ca es predictor de Li pero Li no es predictor de Ca . Por otro lado, es importante mencionar que en los predictores de las variables SO_4 y HCO_3 se encontraron valores muy pequeños de correlación y porcentaje de datos observados. A pesar de esto, fueron seleccionados los que obtuvieron los valores más altos en comparación con las demás.

3.5. Entrenamiento de modelos de imputación única y múltiple

A partir de la selección de predictores, los conjuntos de datos generados se dividieron de manera aleatoria en entrenamiento y prueba. Utilizando el conjunto de entrenamiento, y para obtener el mejor desempeño de cada uno de los algoritmos de imputación, se realizó el entrenamiento

y sintonización de sus parámetros:

Para el caso de SVR, los parámetros sintonizados fueron $\epsilon \in \{0, 0.1, 0.2, \dots, 1\}$, el parámetro de Costo $C \in \{0.01, 0.11, \dots, 10\}$, y en el caso del kernel radial $\gamma \in \{0.01, 0.11, \dots, 10\}$.

Para el caso de la RNA, se utilizó una arquitectura con dos capas ocultas de 15 unidades por capa; se eligió ReLu como la función de activación para las unidades de capas ocultas. La capa de salida contiene una sola unidad que usa la función de identidad $g(x) = x$ como su función de activación. Para el aprendizaje de parámetros, se utilizó el algoritmo de optimización LM - BFGS (un algoritmo que estima la matriz de Hesse utilizando una cantidad limitada de memoria). Para evitar el sobreajuste se utilizó un término de regularización $\alpha=0.01$.

En el caso de MICE, se utilizaron 10 iteraciones para la imputación de cada valor faltante; el número de conjuntos $m = 10$; como secuencia de visitas se usó un criterio monotónico ascendente; y como método de regresión se utilizó el algoritmo *pmm*.

3.5.1. Validación de modelos

Una vez que se generaron los modelos con el conjunto de entrenamiento, los valores de la variable objetivo en el conjunto de entrenamiento se eliminan en forma temporal, posteriormente, estos valores se estiman mediante los modelos entrenados y finalmente, se obtiene la diferencia entre los valores estimados y los reales mediante las funciones de pérdida 2.17, 2.18 y 2.19. Los valores obtenidos se desescalan y desnormalizan mediante sus funciones inversas para analizarlos en su escala normal. El conocimiento de estos parámetros nos ayudó a evaluar si las imputaciones son sobrestimadas o subestimadas con respecto a los datos originales

y reconfigurar los parámetros de los modelos si se considera necesario.

3.5.2. Pruebas de equivalencia

En esta etapa de la metodología, los modelos resultantes se utilizaron para obtener el conjunto de datos imputados de la variable objetivo de la BDFG.

Con la finalidad de evaluar la similitud entre los valores observados e imputados, se aplicaron las pruebas de equivalencia TOST. Para esto, se definieron los intervalos de equivalencia por medio de la diferencia estandarizada d de Cohen probando valores en un rango de $\Delta_{L,U} \in 0.1, 0.125, 0.15, \dots, 5$, se estableció $\alpha=0.5$ y un intervalo de confianza $IC = 95\%$.

3.5.3. Selección del modelo adecuado

Seleccionar un método de imputación adecuado es una decisión de gran importancia, ya que para un conjunto de datos determinado, algunas técnicas de imputación podrían dar mejores aproximaciones a los valores verdaderos que otras. Hay que tomar en cuenta que muchas veces la técnica de imputación seleccionada puede ser adecuada para algunas variables pero, no para otras. Entonces, al seleccionar el modelo adecuado para cada variable incompleta de la BDFG se toman en cuenta dos criterios importantes: en primer lugar, que la distribución de los valores imputados por el modelo sean equivalentes a los observados, dentro de los límites $\Delta_{L,U}$ establecidos; y en segundo lugar, tomando en cuenta únicamente los modelos que cumplen lo anterior, se busca el modelo que obtenga el error más pequeño de acuerdo a los parámetros $RMSE$ y MAE .

3.5.4. Adición de variable imputada en la nueva BD

En esta etapa los valores obtenidos por el modelo resultante se utilizaron para la imputación de la variable objetivo. Finalmente, una vez completada la variable objetivo es integrada a la NBDFG. Este proceso se aplica de manera iterativa a todas las variables incompletas.

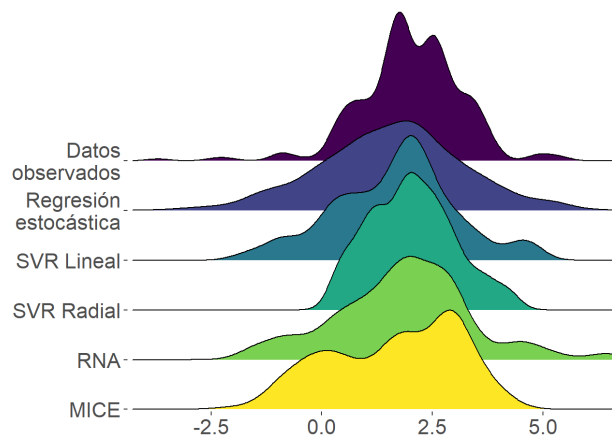
Capítulo 4

Resultados

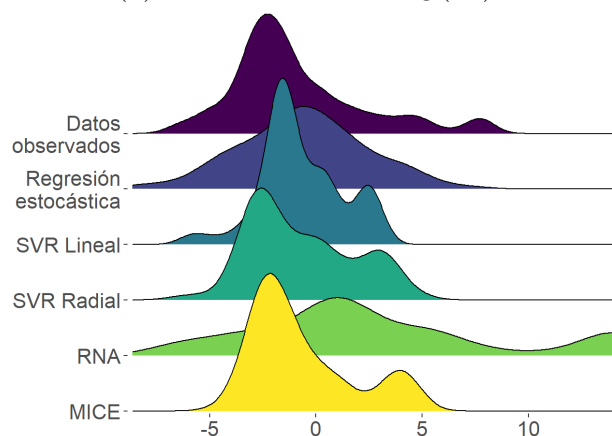
En este capítulo se presentan los resultados de las etapas más relevantes de la metodología de imputación aplicada, tales como: (i) comparación de las distribuciones entre los valores observados y los datos imputados, (ii) resultados de la validación de todos los modelos de imputación mediante tres parámetros de error convencionales: coeficiente de diferencias (U de Theil), RMSE y MAE; (iii) resultados de las pruebas de equivalencia TOST; y, (iv) modelos de imputación seleccionados a partir de una combinación de las pruebas anteriores.

4.1. Modelos de imputación

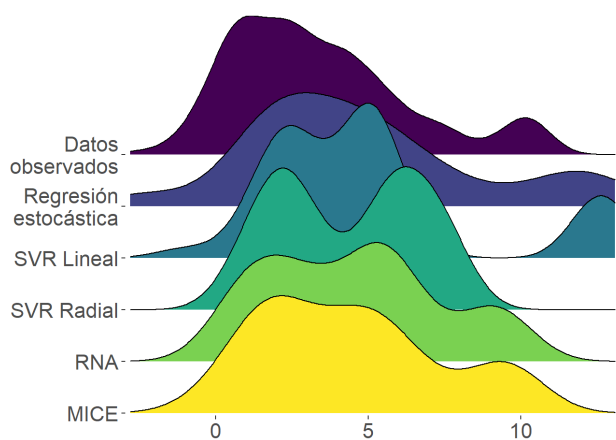
Los datos faltantes se imputaron utilizando seis técnicas diferentes. En las figuras ?? y ?? se presentan tanto las distribuciones de las concentraciones observadas como las concentraciones generadas por las 5 técnicas de imputación aplicadas (Regresión estocástica, SVR Lineal, SVR Radial, RNA y MICE) para cada una de las 3 variables objetivo separadas por cationes –a) $\log(Li)$, b) $\log(Mg)$, c) $\log(Ca)$ – y aniones –a) $\log(Cl)$, b) $\log(SO_4)$ y c) $\log(HCO_3)$ –. En donde, se puede apreciar que las distribuciones de los datos observados no son normales, a pesar de la transformación logarítmica aplicada. También se observa que cada una de las técnicas de imputación generaron diversas distribuciones de datos imputados que, en forma gráfica y cualitativa, no es posible identificar cual es el mejor modelo para imputar los datos faltantes en cada una de las 6 variables. Finalmente, cabe mencionar que el número de datos faltantes es diferente en cada variable objetivo, por ejemplo, Li tiene un alto porcentaje, 67%, lo que significa que aproximadamente había 475 valores faltantes, los cuales después de haberse imputado por diferentes técnicas, son presentados gráficamente en la distribución de los datos imputados. En contraste, hay otras variables con pocos valores faltantes, como Ca que solo tiene un 6%, esto es, 43 registros incompletos, que después de haberse imputado, son presentados en forma gráfica. Lo mismo ocurre con las otras variables objetivo, son presentadas gráficamente los conjuntos de datos que fueron imputados dependiendo de sus porcentajes de datos faltantes, esto es, Mg (16%), Cl (22%), SO_4 (27%) y HCO_3 (58%).



(a) Concentraciones $\log(Li)$



(b) Concentraciones $\log(Mg)$



(c) Concentraciones $\log(Ca)$

Figura 4.1: Distribución de datos observados e imputados de Li , Mg y Ca

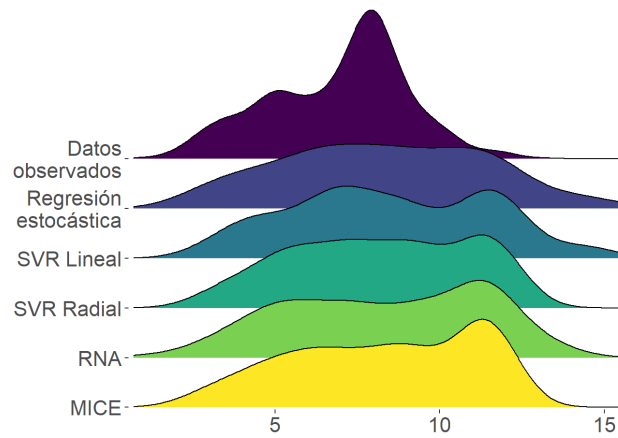
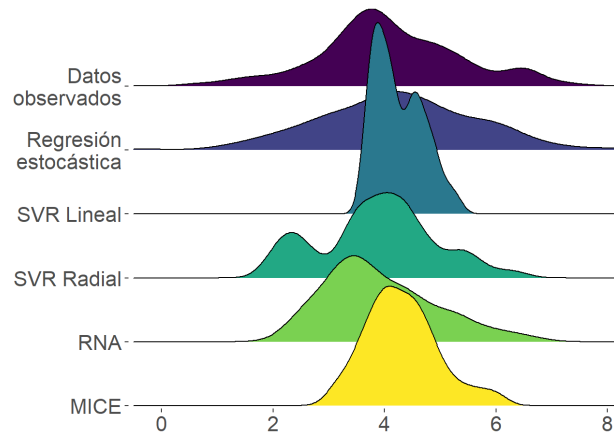
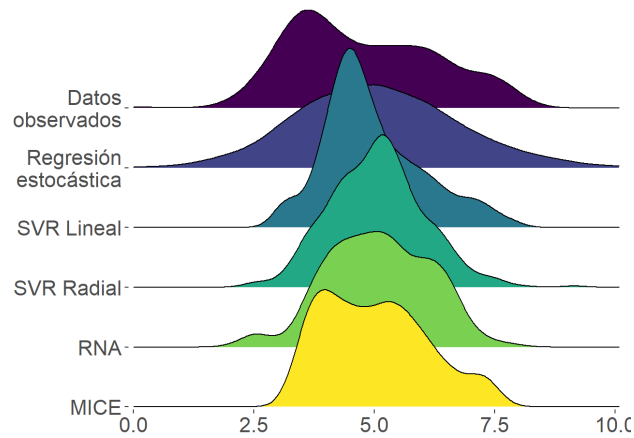
(a) Concentraciones $\log(Cl)$ (b) Concentraciones $\log(SO_4)$ (c) Concentraciones $\log(HCO_3)$

Figura 4.2: Distribución de datos observados e imputados de Cl , SO_4 y HCO_3

4.2. Validación de modelos de imputación

La evaluación del desempeño de las técnicas de imputación se realizó usando el 20% de los datos excluidos para este proceso y que no fueron usados para el entrenamiento. Una vez entrenados los modelos, fueron aplicados al conjunto de prueba para obtener un conjunto de datos imputados. Los errores entre los datos observados y los datos imputados, fueron calculados mediante los parámetros de error: el coeficiente de diferencias U de Theil, RMSE y MAE. El parámetro U de Theil, también llamado coeficiente de diferencia, es una medida de exactitud de la predicción de un modelo de referencia comparado con otros modelos.

En la tabla 4.1 se presentan los errores obtenidos por las técnicas de imputación –media, mediana, Regresión estocástica, SVR Lineal, SVR Radial, RNA y MICE– para cada una de las 6 variables –*Li*, *Mg*, *Ca*, *Cl*, *SO₄* y *HCO₃⁻*. Para interpretar los resultados de U de Theil, es importante destacar que, los errores de los modelos de referencia se muestran en las filas y, los errores de los modelos comparados se encuentran en las columnas. Si el error es menor que 1, significa que el error del modelo de referencia es menor que el otro modelo comparado y viceversa. Si el valor es igual a 1 significa que el error es igual entre el modelo de referencia y el modelo a comparar. Por otro lado, es conveniente destacar que los valores de los parámetros de RMSE y MAE se expresan en las mismas unidades de concentración de las variables, esto es, en *mg/L*.

En cada una de las variables, los mejores modelos se identifican con letras negritas.

Cuadro 4.1: Validación de modelos de imputación

Método	Coeficiente de diferencias (U de Theil)							RMSE	MAE
	Media	Median	RE	SVR L	SVR R	RNA	MICE		
Li									
Media	1	0.99	1.46	1.53	1.48	1.70	1.33	40	17
Mediana	1.00	1	1.47	1.54	1.49	1.71	1.33	41	17
RE	0.68	0.67	1	1.04	1.01	1.16	0.90	27	15
SVR L	0.65	0.64	0.95	1	0.97	1.11	0.86	26	13
SVR R	0.67	0.66	0.98	1.03	1	1.14	0.89	27	10
RNA	0.58	0.58	0.85	0.89	0.87	1	0.77	23	9
MICE	0.75	0.74	1.09	1.15	1.11	1.28	1	30	12
Mg									
Media	1	0.99	1.21	0.97	2.15	1.99	1.44	496	110
Mediana	1.01	1	1.21	0.97	2.15	1.99	1.44	496	110
RE	0.82	0.82	1	0.80	1.78	1.64	1.19	409	92
SVR L	1.02	1.02	1.24	1	2.21	2.04	1.48	509	129
SVR R	0.46	0.46	0.56	0.45	1	0.92	0.67	229	48
RNA	0.50	0.50	0.60	0.48	1.08	1	0.72	249	49
MICE	0.69	0.69	0.83	0.67	1.49	1.37	1	243	61
Ca									
Media	1	0.99	1.18	1.22	4.17	4.84	1.61	7737	2073
Mediana	1.01	1	1.18	1.22	4.17	4.84	1.61	7740	2071
RE	0.84	0.84	1	1.02	3.51	4.07	1.35	6515	1795
SVR L	0.81	0.81	0.97	1	3.41	3.96	1.31	6327	1653
SVR R	0.23	0.23	0.28	0.29	1	1.16	0.38	1853	434
RNA	0.20	0.20	0.24	0.25	0.86	1	0.33	1597	338
MICE	0.62	0.62	0.73	0.75	2.59	3.00	1	4 801	1212
Cl									
Media	1	1.01	0.74	0.49	1.06	0.99	1.33	52960	9859
Mediana	0.99	1	0.74	0.48	1.06	0.99	1.32	52853	9726
RE	1.33	1.33	1	0.65	1.42	1.32	1.77	70690	11176
SVR L	2.03	2.04	1.52	1	2.17	2.03	2.71	108016	11594
SVR R	0.93	0.93	0.70	0.45	1	0.93	1.24	49622	5905
RNA	1.01	1.01	0.75	0.49	1.07	1	1.33	53186	7348
MICE	0.75	0.75	0.56	0.36	0.80	0.74	1	39776	5465
SO₄									
Media	1	0.99	0.91	0.98	1.01	1.03	1.07	289	100
Mediana	1.00	1	0.91	0.99	1.02	1.04	1.08	291	99
RE	1.09	1.08	1	1.08	1.11	1.13	1.17	317	148
SVR L	1.01	1.00	0.92	1	1.02	1.04	1.08	292	104
SVR R	0.98	0.97	0.89	0.97	1	1.01	1.05	284	96
RNA	0.96	0.95	0.88	0.95	0.98	1	1.03	279	88
MICE	0.93	0.92	0.84	0.92	0.94	0.96	1	269	97
HCO₃									
Media	1	0.98	0.88	1.03	1.15	1.51	1.48	648	366
Mediana	1.01	1	0.90	1.05	1.17	1.54	1.48	660	365
RE	1.12	1.10	1	1.16	1.30	1.69	1.66	729	452
SVR L	0.96	0.94	0.85	1	1.11	1.45	1.42	623	319
SVR R	0.86	0.84	0.76	0.89	1	1.30	1.28	560	293
RNA	0.66	0.64	0.58	0.68	0.76	1	0.98	429	240
MICE	0.67	0.66	0.59	0.70	0.78	1.01	1	437	228

En el primer caso de la tabla 4.1 se presenta la variable Li , en donde, el menor error fue obtenido por el modelo de RNA, que le lleva una gran ventaja principalmente a los métodos de media y mediana con un coeficiente de diferencias de 0.58, seguido de los métodos MICE (0.77), Regresión estocástica (0.85), SVR Radial (0.87), y SVR Lineal (0.89). Los resultados de RMSE y MAE también indican que el mejor modelo es RNA con los errores más pequeños.

Para Mg , el modelo que obtuvo el menor error de acuerdo a la U de Theil, RMSE y MAE corresponde a SVR Radial, el cual en orden ascendente obtuvo mejores resultados que los modelos de SVR Lineal (0.45), Media y Mediana (0.46), Regresión estocástica (0.56), y RNA (0.92), con el cual solo obtuvo una ligera ventaja.

En el caso de la variable Ca , según el coeficiente de U de Theil, el modelo de RNA fue el mejor, principalmente comparado con media y mediana (0.2), Regresión estocástica (0.24) y SVR Lineal (0.25), MICE (0.33); y finalmente, comparado con SVR Radial (0.86), con el cual RNA presentó una ventaja más ligera. Esto es, para la Ca , los mejores modelos fueron RNA y SVR Radial. Estos resultados de U de Theil son totalmente concordante con los obtenidos por los parámetros de RMSE y MAE.

Según el coeficiente de U de Theil, para imputar la variable de Cl , el mejor modelo fue MICE. A los métodos que más ventaja les lleva MICE, en orden de mayor a menor ventaja, son SVR Lineal (0.36), Regresión estocástica (0.56), RNA (0.74), Media y Mediana (ambas con 0.75) y finalmente, SVR Radial (0.8), que quedó muy cerca de MICE. Los parámetros de U de Theil, RMSE, MAE son completamente concordantes en la identificación de los mejores (MICE y SVR Radial) y peores (SVR Lineal y regresión estocástica) modelos. Sin embargo, los resultados de MAE indican una pequeña diferencia en el orden de RNA

y media-mediana.

Continuando con la discusión de los resultados, para la variable SO_4 , según el coeficiente de diferencias U de Theil en orden ascendente, MICE presenta mayores ventajas para los modelos Regresión estocástica (0.84), SVR lineal y Mediana (ambos con 0.92), Media (0.93), SVR Radial (0.94) y RNA (0.96). Como se puede apreciar en esta variable, la competencia entre todos los modelos está muy cerrada. El parámetro RMSE también indica que el mejor modelo es MICE (con un valor de 269). No obstante, según MAE el mejor modelo es RNA (con un valor de 88). Aunque los tres parámetros coinciden que el peor modelo fue regresión estocástica.

Finalmente, para la variable HCO_3 , según el parámetro U de Theil, el mejor modelo es RNA, el cual en orden ascendente presenta mayores ventajas para los modelos Regresión estocástica (0.58), Mediana (0.64), Media (0.66), SVR Lineal (0.68), SVR Radial (0.76), MICE (0.98). El parámetro RMSE coincide con U de Theil, indicando que RNA fue el mejor modelo. Sin embargo, según el parámetro MAE, el mejor modelo fue MICE. Nuevamente, los tres parámetros coinciden que el peor modelo fue regresión estocástica.

4.3. Pruebas de equivalencia de los modelos de imputación

Las pruebas de equivalencia se utilizaron para evaluar la presencia o ausencia de una similitud considerable entre las distribuciones de los datos observados y los imputados de cada variable. El análisis se realizó entorno a dos rangos de valores: *acceptables* los cuales comprenden intervalos de equivalencia entre $(0, 0.3]$, y *no acceptables* valores entre $(0.3, 0.5]$.

Es importante señalar que usamos límites simétricos alrededor de cero para todas las pruebas de equivalencia. En la tabla 4.2 se reportan todos aquellos algoritmos en dónde se obtuvieron imputaciones estadísticamente equivalentes y no estadísticamente diferentes a los datos observados. A continuación se describe cada una de las columnas presentadas en la tabla 4.2: (i) el método de imputación utilizado; (ii) los límites d Cohen (i.e. diferencia estandarizada de las medias) representan el porcentaje de diferencia de ambas distribuciones; (iii) los límites en valores crudos son los valores en mg/L que representan el porcentaje (d de Cohen); (iv) la diferencia de las medias representa (disimilitud entre los conjuntos de datos observados e imputados de la variable); (v) los rangos de valores 90 TOST IC los cuales representan el intervalo de confianza (IC) del 90 % alrededor de la diferencia de las medias; y, vi) por ultimo, se presenta el respectivo valor de significancia (p) de cada prueba. Cabe señalar que, la nomenclatura $* p \leq 0.05$ y $** p \leq 0.01$ es usada para denotar la significancia estadística de las pruebas.

De acuerdo a [54] para concluir equivalencia entre dos conjuntos se deben rechazar H_{01} y H_{02} (Ver ecuación 2.23), es decir, aceptar la H_A . En este sentido, lo que se busca es que la diferencia de las medias de los conjuntos, más los valores del IC, estén contenidas dentro de los límites de los valores crudos (zona de indiferencia).

Adicionalmente, es importante mencionar que, cuando nos referimos a *valores equivalentes* o a una *conclusión de equivalencia*, significa que la diferencia entre los grupos es menor de lo que se considera significativo y estadísticamente cae dentro del intervalo indicado por los límites de equivalencia.

Cuadro 4.2: Resultados de pruebas de equivalencia

Método	Límites d Cohen	Límites valores crudos	Diferencia de medias	90 TOST IC	<i>p</i>
Li					
SVR Lineal	± 0.175	± 4.1	0.9	[-2.1; 4]	*
RE	± 0.2	± 4.7	-2.8	[-2.1; 4]	*
SVR Radial	± 0.3	± 5.9	1.6	[-1.1; 4.3]	**
MICE	± 0.3	± 5.9	2.4	[-0.2; 5.1]	*
Mg					
RE	± 0.4	± 150	90	[49; 131]	*
SVR Radial	± 0.4	± 145	108	[73; 142]	*
MICE	± 0.4	± 145	105	[70; 139]	*
Cl					
SVR Radial	± 0.5	± 45420	-49821	[-66481; -33160]	0.66
MICE	± 0.5	± 18644	-25901	[-32087; -19716]	0.97
Ca					
RNA	± 0.275	± 1659	538	[-519; 1596]	*
MICE	± 0.275	± 1794	-28	[-1409; 1352]	*
SO₄					
RE	± 0.3	± 102	-43	[-96; 8]	*
RNA	± 0.4	± 78	51	[28; 75]	*
MICE	± 0.4	± 73	46	[26; 66]	*
HCO₃					
SVR Radial	± 0.275	± 153	75	[5; 145]	*
MICE	± 0.275	± 131	69	[7; 131]	*
RNA	± 0.3	± 138	75	[14; 135]	*

En el primer caso, se presentan los resultados de las imputaciones de *Li*, en donde, se concluye equivalencia para: SVR Lineal en los límites $\Delta_{L,U} = \pm 0,175$ mostrando que la diferencia de sus medias 0.9 más su IC es menor que sus límites en valores crudos, es decir, $-2,1 \geq -4,1$ y $4 \leq 4,1$ lo cual significa que sus valores se encuentran dentro de la zona

de indiferencia; de igual manera en RE donde en los límites $\Delta_{L,U} = \pm 0,2$ se cumple que la diferencia de sus medias 2.8 más su IC $-2.1 \geq -4.7$ y $4 \leq 4.7$ (donde ± 4.7 son sus límites en valores crudos); SVR Radial con los límites $\Delta_{L,U} = \pm 0,3$ que, dada la diferencia de sus medias 1.6 más su IC cumplen la condición $-1.1 \geq -5.9$ y $4.3 \leq 5.9$; por último, MICE concluye una equivalencia en los límites $\Delta_{L,U} = \pm 0,3$ se obtuvo una diferencia de medias una diferencia de medias de 2.4 más su IC que cumplen $-0.2 \geq -5.9$ y $5.1 \leq 5.9$.

Para *Mg* los algoritmos que concluyeron equivalencia fueron: RE con límites $\Delta_{L,U} = \pm 0,4$ demostrando que la diferencia de sus medias 90 más su IC es menor que sus límites en valores crudos, es decir, $49 \geq -150$ y $131 \leq 150$ lo cual significa que sus valores se encuentran dentro de la zona de indiferencia; SVR Radial donde en los límites $\Delta_{L,U} = \pm 0,4$ se cumple que la diferencia de sus medias 108 más su IC $73 \geq -145$ y $142 \leq 145$; por último, MICE concluye una equivalencia en los límites $\Delta_{L,U} = \pm 0,4$ se obtuvo una diferencia de medias una diferencia de medias de 105 más su IC que cumplen $70 \geq -145$ y $139 \leq 145$.

En contraste con lo anterior, para la variable *Cl* ningún algoritmo de imputación obtuvo valores equivalentes a los observados. Sin embargo, se reportan aquellos que se acercaron más en los límites máximos (± 0.5), por ejemplo, en SVR Radial con una diferencia de medias de -49821 más un IC se rechazó la H_A de equivalencia debido a que $-66481 \leq -45420$ - $33160 \geq 45420$. Por otro lado, MICE con una diferencia en sus medias de -25901 aceptó las $H_{01,02}$ donde $-1409 \leq -18644$ $1352 \geq 18644$. En otras palabras, los valores de ambos algoritmos se encuentran fuera de la zona de indiferencia.

Continuando con la descripción de los resultados de las pruebas de equivalencia, para *Ca*, se aprecia que el algoritmo de RNA obtuvo valores

equivalentes a los observados, en los límites $\Delta_{L,U} = \pm 0,275$, con una diferencia de medias de 538 más su IC del 90 % cumplen que $-519 \geq -1659$ y $1596 \leq 1659$; de igual manera MICE con límites $\Delta_{L,U} = \pm 0,275$ y dada la diferencia de sus medias 1.6 más su IC, cumplen la condición $-1409 \geq -1794$ y $1352 \leq 1794$.

En el caso de SO_4 , según las pruebas de equivalencia los algoritmos que obtienen valores equivalentes son: RE en los límites $\Delta_{L,U} = \pm 0,3$ demostrando que la diferencia de sus medias -43 más su IC se concentran dentro de los límites en valores crudos, es decir, $-96 \geq -102$ y $8 \leq 102$; RNA con los límites $\Delta_{L,U} = \pm 0,4$, una diferencia en sus medias de 51 más su IC cumplen la condición $28 \geq -78$ y $75 \leq 78$; por último, MICE concluye una equivalencia en los límites $\Delta_{L,U} = \pm 0,4$ se obtuvo una diferencia de medias una diferencia de medias de 46 más su IC que cumplen $26 \geq -73$ y $66 \leq 73$.

Finalmente, para HCO_3 se observa que el algoritmo de SVR Radial obtuvo valores equivalentes a los observados, en los límites $\Delta_{L,U} = \pm 0,275$, con una diferencia de medias de 75 más su IC donde se cumple que $5 \geq -153$ y $145 \leq 153$; RNA con los límites $\Delta_{L,U} = \pm 0,275$, una diferencia en sus medias de 69 más su IC cumplen la condición $7 \geq -131$ y $131 \leq 13$; por último, MICE con límites $\Delta_{L,U} = \pm 0,3$ y dada la diferencia de sus medias 75 más su IC, se acepta que $14 \geq -138$ y $135 \leq 138$.

4.4. Modelos de imputación seleccionados

En la tabla 4.3 se resumen los modelos que de acuerdo a los criterios de selección –pruebas de equivalencia y parámetros de error RMSE, MAE y U de Theil– proporcionaron las mejores aproximaciones a los valores observados. En este sentido lo que se busca, por variable, es obtener el

modelo con menor error (RMSE/MAE/U.Theil) que además satisfaga el menor intervalo de equivalencia aceptable.

Cuadro 4.3: Modelos seleccionados

Variable	Model	Errores		TOST parámetros		
		RMSE	MAE	Límites d Cohen	Límites valores crudos	Diferencia de medias
Li	SVR Linear	26	13	± 0.175	± 4.1	0.9
Mg	SVR Radial	229	48	± 0.4	± 145	108
Ca	RNA	1597	338	± 0.275	± 1659	538
SO₄	RE	317	148	± 0.3	± 102	-43
HCO₃	MICE	429	240	± 0.275	± 131	69

Para la variable *Li*, los modelos que concluyeron equivalencia entre los valores imputados y los observados fueron SVR Lineal ($d=\pm 0.175$), RE ($d=\pm 0.2$), SVR Radial ($d=\pm 0.3$) y MICE($d=\pm 0.3$). Sin embargo, de los cuatro modelos mencionados anteriormente, el algoritmo que se aproximó mejor a los valores reales de acuerdo a los parámetros de error, fue **SVR Lineal** con un RMSE = 26 y MAE = 13. Por tal motivo, para imputar los valores faltantes de dicha variable se tomaron en cuenta las estimaciones de este modelo. Es importante mencionar que, a pesar de que el algoritmo de RNA obtuvo el mejor resultado en la etapa de validación, este no obtuvo estimaciones equivalentes por lo que se descartó en la selección del modelo para *Li*.

En el caso de *Mg*, ambos criterios de selección señalan que los mejores resultados se obtuvieron mediante el modelo de **SVR Radial**. Esto debido a que, a pesar de que los algoritmos de RE, MICE concluyeron equivalencia en el mismo intervalo ($d=\pm 0.4$) que SVR Radial, este último fue el que obtuvo el valor mínimo en los parámetros de error con un RMSE=229 y MAE=48. Por tanto, se utilizaron las estimaciones de

dicho modelo para la imputación de los valores faltantes de *Mg*.

Por otro lado, para la imputación de la variable *Ca* se utilizó el algoritmo **RNA**, ya que los dos criterios de selección apuntan que es el algoritmo que produjo los mejores resultados para la estimación de los valores faltantes de dicha variable. A pesar de que concluyó equivalencia en el mismo límite que MICE ($d=0.275$), las RNA obtuvieron los valores más pequeños en los parámetros de error con RMSE=1,597 y MAE=338.

En contraste, para la variable *Cl* el algoritmo que se aproximó más a los valores reales del conjunto de prueba fue MICE con un RMSE=39,776 y MAE=5,465. Desafortunadamente, al aplicarse las pruebas estadísticas **ningún algoritmo concluyó equivalencia** entre las estimaciones y los valores observados de *Cl*, aún cuando el límite se amplió hasta ± 0.5 , permitiendo gran margen de disimilitud.

En el caso de *SO₄* los algoritmos de MICE y RNA obtuvieron un valor de: RMSE=269, MAE=97; y RMSE=279, MAE=88, respetivamente. Esto decir, más pequeño que RE con un RMSE=317 y MAE=148. Sin embargo, este ultimo concluyó equivalencia en un intervalo menor al de los otros dos. Por tanto las estimaciones proporcionadas por **RE** fueron imputadas en los valores faltantes de *SO₄*.

Por ultimo, para *HCO₃* los modelos que concluyeron equivalencia entre los valores imputados y los observados fueron SVR Radial y MICE (ambos con $d=\pm 0.275$), y RNA ($d=\pm 0.3$). Sin embargo, a pesar de que el algoritmo de RNA se aproximó mejor a los valores reales de acuerdo a los parámetros de error, **MICE** concluyó equivalencia en un intervalo más pequeño. Por tanto, para imputar los valores faltantes de dicha variable se tomaron en cuenta las estimaciones de este modelo.

4.5. Nueva base de datos geotérmicos mundial

4.5.1. Estadística descriptiva final

Una vez que se realizó el análisis para determinar el algoritmo adecuado para la imputación de cada variable incompleta de la BDFG, cada una de ellas fue completada e incorporada a la nueva base de datos de fluidos geotérmicos (NBDFG), con excepción de *Cl*, debido a que por el momento ningún algoritmo produjo resultados satisfactorios para su imputación. En la tabla 4.4 se muestra la estadística final obtenida de la NBDFG. Como se puede observar, las variables que no presentaron valores faltantes quedaron de la misma manera ya que no sufrieron modificaciones.

Cuadro 4.4: Estadística final de la BDFG

Variable	Mín	Máx	\bar{x}	<i>Sd</i>	γ_1	γ_2
<i>Temperatura</i>	59	359	216	69	-0.35	2
<i>Na</i>	22	565579	11472	520014	5	42
<i>K</i>	0.5	66473	1583	6555	8	76
<i>Li</i>	0.025	215	13.5	23	6	56
<i>Mg</i>	0.001	3920	96	471	4	28
<i>Ca</i>	0.06	55600	2304	7548	3	17
<i>SO₄</i>	0.6	2500	124	219	4	26
<i>HCO₃</i>	1	3074	309	462	2	7

Capítulo 5

Conclusiones y trabajos futuros

Con la finalidad de contribuir al desarrollo de la geotermometría de solutos, el presente trabajo partió de la base de datos geoquímicos BDFG que fue previamente desarrollada por [7], la cual contiene 708 registros de fluidos geotérmicos y temperaturas de fondo de pozos geotérmicos productores de diferentes partes del mundo. Dicha base de datos contiene 9 variables entre ellas *Temperatura*, *Na*, *K*, *Li*, *Mg*, *Ca*, *Cl*, *SO₄* y *HCO₃*. Desafortunadamente, 6 de estas variables presentaban los siguientes porcentajes de datos faltantes: *Li* (67 %), *Mg* (16 %), *Ca* (6 %), *Cl* (22 %), *SO₄* (27 %) y *HCO₃* (58 %). Como se puede observar, únicamente la *Temperatura*, *Na* y *K* se encontraban completas. Se destaca que, del total de 708 registros de la BDFG, únicamente 150 (21 %) estaban completos en todas las variables. Esto claramente sesgaría realizar un análisis de casos completos.

Por lo tanto, en el presente trabajo, se propuso resolver el problema de los datos faltantes en esta base de datos inicial, mediante la implemen-

tación de una *metodología de imputación equivalente*, constituida por 5 técnicas de imputación única –media, mediana, regresión estocástica, máquinas de vectores de soporte y redes neuronales–; y la técnica de imputación múltiple, MICE. Antes de aplicar estas técnicas de imputación, se validó que un proceso de imputación fuera posible a través de la identificación del mecanismo de pérdida de información de tipo MAR.

En forma previa al entrenamiento de los modelos, para cada una de las 6 variables Li , Mg , Ca , Cl , SO_4 y HCO_3 se seleccionaron las variables predictoras y se obtuvieron los conjuntos individuales de datos observados, los cuales se dividieron de manera aleatoria en dos subconjuntos, 80% para el entrenamiento y 20% para la validación de los modelos. De acuerdo a los parámetros de validación RMSE, MAE y U de Theil, la técnica de RNA obtuvo los menores errores de aproximación para las variables Li , Ca y HCO_3 . Mientras que, SVR Radial fue el mejor modelo para la aproximación de la variable de Mg . Por último, MICE obtuvo el menor error para las variables Cl y SO_4 .

Sin embargo, las pruebas de equivalencia permiten evaluar la capacidad de generalización de los modelos en forma externa al entrenamiento, esto es, evaluaron la presencia o ausencia de una similitud considerable entre la distribución de los datos observados y los imputados de cada variable objetivo. Según las pruebas de equivalencia, los modelos que obtuvieron valores estadísticamente equivalentes dentro de determinados límites d Cohen, son los siguientes: (i) SVR Lineal únicamente para la variable Li ; (ii) Regresión estocástica para Li , Mg y SO_4 ; (iii) SVR Radial para Li , Mg , Cl y HCO_3 ; (iv) MICE para todas las variables objetivo; (vi) RNA para Ca , SO_4 y HCO_3 .

Finalmente, después de tomar en cuenta tanto los criterios de validación –RMSE, MAE y U de Theil– y las pruebas de equivalencia, se

seleccionó el mejor modelo para cada variable objetivo: (i) SVR Lineal para *Li*; (ii) SVR Radial para *Mg*; (iii) RNA para *Ca*; (iv) Regresión estocástica para *SO₄*; (v) MICE para *HCO₃*. Como se puede apreciar, para cada variable objetivo resultó seleccionada una técnica de imputación diferente, lo cual indica la heterogeneidad del modelado requerido para cada variable.

Aún cuando es notorio que, desafortunadamente, para *Cl*, ningún algoritmo concluyó equivalencia entre los valores estimados y los valores observados, para el resto de las variables si fue posible. En particular el caso del *Li* es importante, ya que muchos geotermómetros estadísticos dependen de él para la estimación de la temperatura de fondo. Por otra parte, el hecho de que los métodos estudiados no fueran capaces de completar la variable *Cl* de forma equivalente es importante, ya que este es un anión sumamente importante en la composición química de los fluidos geotérmicos. Por lo tanto, en un trabajo futuro se realizará un pre-procesamiento de filtrado de registros por balance cargas a la base de datos inicial antes de aplicar las técnicas de imputación.

Referencias

- [1] A. K. Sani, R. M. Singh, T. Amis e I. Cavarretta, “A review on the performance of geothermal energy pile foundation, its design process and applications”, *Renewable and Sustainable Energy Reviews*, vol. 106, págs. 54-78, 2019, ISSN: 1364-0321.
- [2] L. Díaz-González, E. Santoyo y J. Reyes-Reyes, “Tres nuevos geotermómetros mejorados de Na/K usando herramientas computacionales and geoquimiométricas: aplicación a la predicción de temperaturas de sistemas geotérmicos”, *Revista mexicana de ciencias geológicas*, vol. 25, n.º 3, págs. 465-482, 2008.
- [3] C. G. García-López, K. Pandarinath y E. Santoyo, “Solute and gas geothermometry of geothermal wells: A geochemometrics study for evaluating the effectiveness of geothermometers to predict deep reservoir temperatures”, *International Geology Review*, vol. 56, n.º 16, págs. 2015-2049, 2014, ISSN: 19382839. DOI: 10.1080/00206814.2014.984352.
- [4] A. Ferhat Bayram, “Application of an artificial neural network model to a Na–K geothermometer”, *Journal of Volcanology and Geothermal Research*, vol. 112, n.º 1-4, págs. 75-81, 2001, ISSN: 03770273. DOI: 10.1016/S0377-0273(01)00235-9.

- [5] I. Can, "A new improved Na/K geothermometer by artificial neural networks", *Geothermics*, vol. 31, n.º 6, págs. 751-760, 2002, ISSN: 0375-6505. DOI: [https://doi.org/10.1016/S0375-6505\(02\)00044-5](https://doi.org/10.1016/S0375-6505(02)00044-5). dirección: <http://www.sciencedirect.com/science/article/pii/S0375650502000445>.
- [6] U. Serpen, "Hydrogeological investigations on Balçova geothermal system in Turkey", *Geothermics*, vol. 33, n.º 3, págs. 309-335, 2004, ISSN: 0375-6505. DOI: <https://doi.org/10.1016/j.geothermics.2003.08.011>. dirección: <http://www.sciencedirect.com/science/article/pii/S037565050300107X>.
- [7] L. Díaz-González, "Desarrollo de nuevas herramientas estadísticas and geotermométricas para la industria geotérmica: Universidad Nacional Autónoma de México, Postgrado en Ingeniería (Energía)", Tesis doct., Tesis de Doctorado, 2008.
- [8] L. Shevenell, R. Penfield, R. Zehner, G. Johnson y M. Coolbaugh, "National Geothermal Data System Geochemical Data for Exploration", vol. 36, n.º 1983, 2012, ISSN: 01935933.
- [9] G. E. Batista y M. C. Monard, "An analysis of four missing data treatment methods for supervised learning", *Applied artificial intelligence*, vol. 17, n.º 5-6, págs. 519-533, 2003.
- [10] B. L. Dickson y A. M. Giblin, "An evaluation of methods for imputation of missing trace element data in groundwaters", *Geochemistry: Exploration, Environment, Analysis*, vol. 7, n.º 2, págs. 173-178, 2007.
- [11] C. C. Turrado, F. S. Lasheras, J. L. Calvo-Rollé y A. J. Piñón-Pazos, "A new missing data imputation algorithm applied to elec-

- trical data loggers”, *Sensors (Switzerland)*, vol. 15, n.º 12, págs. 31 069-31 082, 2015, ISSN: 14248220. DOI: 10.3390/s151229842.
- [12] J. Palarea-Albaladejo, J. A. Martín-Fernández y A. Buccianti, “Compositional methods for estimating elemental concentrations below the limit of detection in practice using R”, *Journal of Geochemical Exploration*, vol. 141, págs. 71-77, 2014, ISSN: 03756742. DOI: 10.1016/j.gexplo.2013.09.003. dirección: <http://dx.doi.org/10.1016/j.gexplo.2013.09.003>.
- [13] B. Ghane y O. Asghari, “Accuracy evaluation of different statistical and geostatistical censored data imputation approaches (Case study: Sari Gunay gold deposit)”, *Int. J. Min. & Geo-Eng*, vol. 50, n.º 1, págs. 49-60, 2016.
- [14] L. Peiffer, C. Wanner, N. Spycher, E. L. Sonnenthal, B. M. Kennedy y J. Iovenitti, “Optimized multicomponent vs. classical geothermometry: Insights from modeling studies at the Dixie Valley geothermal area”, *Geothermics*, vol. 51, págs. 154-169, 2014, ISSN: 03756505. DOI: 10.1016/j.geothermics.2013.12.002. dirección: <http://dx.doi.org/10.1016/j.geothermics.2013.12.002>.
- [15] N. Spycher, L. Peiffer, E. L. Sonnenthal, G. Saldi, M. H. Reed y B. M. Kennedy, “Integrated multicomponent solute geothermometry”, *Geothermics*, vol. 51, págs. 113-123, 2014, ISSN: 03756505. DOI: 10.1016/j.geothermics.2013.10.012. dirección: <http://dx.doi.org/10.1016/j.geothermics.2013.10.012>.
- [16] A. P. Fowler, N. Spycher, R. A. Zierenberg y C. A. Cantwell, “Identification of blind geothermal resources in Surprise Valley, CA, using publicly available groundwater well water quality data”,

- Applied Geochemistry*, vol. 80, págs. 24-48, 2017, ISSN: 18729134. DOI: 10.1016/j.apgeochem.2017.03.001. dirección: <http://dx.doi.org/10.1016/j.apgeochem.2017.03.001>.
- [17] M. Frenzel, T. Hirsch y J. Gutzmer, “Gallium, germanium, indium, and other trace and minor elements in sphalerite as a function of deposit type - A meta-analysis”, *Ore Geology Reviews*, vol. 76, págs. 52-78, 2016, ISSN: 01691368. DOI: 10.1016/j.oregeorev.2015.12.017. dirección: <http://dx.doi.org/10.1016/j.oregeorev.2015.12.017>.
- [18] D. B. Rubin, “Multiple imputations in sample surveys - A phenomenological Bayesian approach to nonresponse”, *Proceedings of the Survey Research Methods Section of the American Statistical Association*, págs. 20-34, 1978. dirección: https://www2.amstat.org/sections/srms/Proceedings/papers/1978%7B%5C_%7D004.pdf.
- [19] P. J. García-Laencina, J. L. Sancho-Gómez y A. R. Figueiras-Vidal, “Pattern classification with missing data: A review”, *Neural Computing and Applications*, vol. 19, n.º 2, págs. 263-282, 2010, ISSN: 09410643. DOI: 10.1007/s00521-009-0295-6. arXiv: [arXiv:1011.1669v3](https://arxiv.org/abs/1011.1669v3).
- [20] N. Razak, Y. Z. Zubairi y R. M. Yunus, “Imputing missing values in modelling the PM10 concentrations”, *Sains Malaysiana*, vol. 43, n.º 10, págs. 1599-1607, 2014.
- [21] Z. Pang y M. Reed, “Theoretical chemical thermometry on geothermal waters: problems and methods”, *Geochimica et Cosmochimica Acta*, vol. 62, n.º 6, págs. 1083-1091, 1998, ISSN: 00167037. DOI: 10.1016/S0016-7037(98)00037-4.

- [22] D. L. Siler, Y. Zhang, N. F. Spycher, P. F. Dobson, J. S. McClain, E. Gasperikova, R. A. Zierenberg, P. Schiffman, C. Ferguson, A. Fowler y C. Cantwell, “Play-fairway analysis for geothermal resources and exploration risk in the Modoc Plateau region”, *Geothermics*, vol. 69, n.º November 2016, págs. 15-33, 2017, ISSN: 03756505. DOI: 10.1016/j.geothermics.2017.04.003.
- [23] Q. Shang, Z. Yang, S. Gao y D. Tan, “An Imputation Method for Missing Traffic Data Based on FCM Optimized by PSO-SVR”, *Journal of Advanced Transportation*, vol. 2018, págs. 1-21, 2018, ISSN: 0197-6729. DOI: 10.1155/2018/2935248.
- [24] I. B. Aydilek y A. Arslan, “A hybrid method for imputation of missing values using optimized fuzzy c -means with support vector regression and a genetic algorithm”, *Information Sciences*, vol. 233, págs. 25-35, 2013, ISSN: 0020-0255. DOI: 10.1016/j.ins.2013.01.021. dirección: <http://dx.doi.org/10.1016/j.ins.2013.01.021>.
- [25] M. Gómez-Carracedo, J. Andrade, P. López-Mahía, S. Muniategui y D. Prada, “A practical comparison of single and multiple imputation methods to handle complex missing data in air quality datasets”, *Chemometrics and Intelligent Laboratory Systems*, vol. 134, págs. 23-33, 2014.
- [26] S. van Buuren y C. G. M. Oudshoorn, *Multivariate Imputation by Chained Equations : Mice V1.0 User’s manual*. TNO, 2000. dirección: <https://www.narcis.nl/publication/RecordID/oai:tudelft.nl:uuid:55f1a228-7982-4ea3-8841-96571562b900>.
- [27] M. N. Norazian, Y. A. Shukri, R. N. Azam y A. M. M. Al Bakri, “Estimation of missing values in air pollution data using single

- imputation techniques”, *ScienceAsia*, vol. 34, n.º 3, págs. 341-345, 2008.
- [28] N. M. Noor, M. M. A. B. Abdullah, A. S. Yahaya y N. A. Ramli, “Comparison of linear interpolation method and mean method to replace the missing values in environmental data set”, *Small*, vol. 5, pág. 10, 2015.
- [29] A. S. Yahaya y F. Ahmad, “International Journal of Applied Science and Technology Vol. 1 No. 6; November 2011”, vol. 1, n.º 6, págs. 278-285, 2011.
- [30] H. Junninen, H. Niska, K. Tuppurainen y J. Ruuskanen, “Methods for imputation of missing values in air quality data sets”, vol. 38, págs. 2895-2907, 2004. DOI: 10.1016/j.atmosenv.2004.02.026.
- [31] J. M. Engels y P. Diehr, “Imputation of missing longitudinal data : a comparison of methods”, vol. 56, págs. 968-976, 2003. DOI: 10.1016/S0895-4356(03)00170-7.
- [32] F. M. Shrive, H. Stuart, H. Quan y W. A. Ghali, “Dealing with missing data in a multi-question depression scale : a comparison of imputation methods”, vol. 10, págs. 1-10, 2006. DOI: 10.1186/1471-2288-6-57.
- [33] M. Jerez, I. Molina, P. J. Garcı, E. Alba, N. Ribelles, L. Franco y M. Martı, “Artificial Intelligence in Medicine Missing data imputation using statistical and machine learning methods in a real breast cancer problem”, *Elsevier*, vol. 50, págs. 105-115, 2010. DOI: 10.1016/j.artmed.2010.05.002.

- [34] S. S. Wilks, “Moments and distributions of estimates of population parameters from fragmentary samples”, *The Annals of Mathematical Statistics*, vol. 3, n.º 3, págs. 163-195, 1932.
- [35] S. F. Buck, “A method of estimation of missing values in multivariate data suitable for use with an electronic computer”, *Journal of the Royal Statistical Society. Series B (Methodological)*, págs. 302-306, 1960.
- [36] P. D. Allison, *Missing data*. Sage publications, 2001, vol. 136.
- [37] D. A. Newman, “Longitudinal Modeling With Randomly and Systematically Missing Data : A Simulation of Ad Hoc, Maximum Likelihood, and Multiple Imputation Techniques”, *Organizational Research Methods*, vol. 6, n.º 3, págs. 328-362, 2003. DOI: 10.1177/1094428103254673.
- [38] C. Cortes y V. Vapnik, “Support-vector networks”, *Machine Learning*, vol. 20, n.º 3, págs. 273-297, 1995, ISSN: 1573-0565. DOI: 10.1007/BF00994018. dirección: <https://doi.org/10.1007/BF00994018>.
- [39] H. Drucker, C. J. Burges, L. Kaufman, A. J. Smola y V. Vapnik, “Support vector regression machines”, en *Advances in neural information processing systems*, 1997, págs. 155-161.
- [40] X. Wang, A. Li, Z. Jiang y H. Feng, “Missing value estimation for DNA microarray gene expression data scheme”, vol. 10, págs. 1-10, 2006. DOI: 10.1186/1471-2105-7-32.
- [41] F. V. Nelwamondo, S. Mohamed y T. Marwala, “Missing data: A comparison of neural network and expectation maximization techniques”, *Current Science*, págs. 1514-1521, 2007.

- [42] E. L. Silva-Ramírez, R. Pino-Mejías, M. López-Coello y M. D. Cubiles-de-la-Vega, “Missing value imputation on missing completely at random data using multilayer perceptrons”, *Neural Networks*, vol. 24, n.º 1, págs. 121-129, 2011, ISSN: 08936080. DOI: 10.1016/j.neunet.2010.09.008. dirección: <http://dx.doi.org/10.1016/j.neunet.2010.09.008>.
- [43] S. Nordbotten, *Neural network imputation applied to the Norwegian 1990 population census data*, 1996. dirección: <https://brage.bibsys.no/xmlui/handle/11250/178156>.
- [44] T. Maiti, C. P. Miller y P. K. Mukhopadhyay, “Neural network imputation: An experience with the National Resources Inventory Survey”, *Journal of Agricultural, Biological, and Environmental Statistics*, vol. 13, n.º 3, págs. 255-269, 2008, ISSN: 10857117. DOI: 10.1198/108571108X337394.
- [45] D. B. Rubin, “Multiple imputation for nonresponse in surveys”, 1987.
- [46] A. D. Shah, J. W. Bartlett, J. Carpenter, O. Nicholas y H. Hemingway, “Comparison of random forest and parametric imputation models for imputing missing data using MICE: A CALIBER study”, *American Journal of Epidemiology*, vol. 179, n.º 6, págs. 764-774, 2014, ISSN: 14766256. DOI: 10.1093/aje/kwt312.
- [47] C. Penone, A. D. Davidson, K. T. Shoemaker, M. Di Marco, C. Rondinini, T. M. Brooks, B. E. Young, C. H. Graham y G. C. Costa, “Imputation of missing data in life-history trait datasets: Which approach performs the best?”, *Methods in Ecology and Evolution*, vol. 5, n.º 9, págs. 1-10, 2014, ISSN: 2041210X. DOI: 10.1111/2041-210X.12232.

- [48] I. Eekhout, H. C. De Vet, J. W. Twisk, J. P. Brand, M. R. De Boer y M. W. Heymans, “Missing data in a multi-item instrument were best handled by multiple imputation at the item score level”, *Journal of Clinical Epidemiology*, vol. 67, n.º 3, págs. 335-342, 2014, ISSN: 08954356. DOI: 10.1016/j.jclinepi.2013.09.009. dirección: <http://dx.doi.org/10.1016/j.jclinepi.2013.09.009>.
- [49] C. Ordóñez Galán, F. Sánchez Lasheras, F. J. de Cos Juez y A. Bernardo Sánchez, “Missing data imputation of questionnaires by means of genetic algorithms with different fitness functions”, *Journal of Computational and Applied Mathematics*, vol. 311, págs. 704-717, 2017, ISSN: 03770427. DOI: 10.1016/j.cam.2016.08.012. dirección: <http://dx.doi.org/10.1016/j.cam.2016.08.012>.
- [50] Y. Ding y A. Ross, “A comparison of imputation methods for handling missing scores in biometric fusion”, *Pattern Recognition*, vol. 45, n.º 3, págs. 919-933, 2012.
- [51] S. v. Buuren y K. Groothuis-Oudshoorn, “mice: Multivariate imputation by chained equations in R”, *Journal of statistical software*, págs. 1-68, 2010.
- [52] C. Ialongo, “The logic of equivalence testing and its use in laboratory medicine”, vol. 27, n.º 1, págs. 5-13, 2017.
- [53] V. Luzar-Stiffler y C. Stiffler, “Equivalence Testing the Easy Way”, *Journal of Computing and Information Technology*, vol. 10, n.º 3, pág. 233, 2002, ISSN: 1330-1136. DOI: 10.2498/cit.2002.03.12. dirección: <http://ieeexplore.ieee.org/document/1024659/%20http://cit.srce.unizg.hr/index.php/CIT/article/view/1487>.

- [54] D. Lakens, A. M. Scheel y P. M. Isager, “Equivalence testing for psychological research: A tutorial”, *Advances in Methods and Practices in Psychological Science*, vol. 1, n.º 2, págs. 259-269, 2018.
- [55] B. A. Rabe, S. Day, M. H. Fiero y M. L. Bell, “Missing data handling in non-inferiority and equivalence trials: A systematic review”, *Pharmaceutical Statistics*, n.º August 2017, págs. 477-488, 2018, ISSN: 15391612. DOI: 10.1002/pst.1867.
- [56] A. P. Robinson y R. E. Froese, “Model validation using equivalence tests”, *Ecological Modelling*, vol. 176, n.º 3-4, págs. 349-358, 2004.
- [57] J. J. Dolado, D. Rodriguez, M. Harman, W. B. Langdon y F. Sarro, “Evaluation of estimation models using the Minimum Interval of Equivalence”, *Applied Soft Computing Journal*, vol. 49, págs. 956-967, 2016, ISSN: 15684946. DOI: 10.1016/j.asoc.2016.03.026. dirección: <http://dx.doi.org/10.1016/j.asoc.2016.03.026>.
- [58] A. J. (J. Ellis y W. A. J. Mahon, *Chemistry and geothermal systems*, English. New York : Academic Press, 1977, Includes bibliographies and index, ISBN: 0122374509.
- [59] M. P. Hochstein, “Assessment and modelling of geothermal reservoirs (small utilization schemes)”, *Geothermics*, vol. 17, n.º 1, págs. 15-49, 1988, ISSN: 0375-6505. DOI: [https://doi.org/10.1016/0375-6505\(88\)90004-1](https://doi.org/10.1016/0375-6505(88)90004-1). dirección: <http://www.sciencedirect.com/science/article/pii/0375650588900041>.

- [60] C. O. Grigsby, J. W. Tester, P. E. Trujillo y D. A. Counce, "Rock-water interactions in the Fenton Hill, new Mexico, hot dry rock geothermal systems I. fluid mixing and chemical geothermometry", *Geothermics*, vol. 18, n.º 5, págs. 629-656, 1989, ISSN: 0375-6505. DOI: [https://doi.org/10.1016/0375-6505\(89\)90098-9](https://doi.org/10.1016/0375-6505(89)90098-9). dirección: <http://www.sciencedirect.com/science/article/pii/0375650589900989>.
- [61] M. Kühn, *Reactive flow modeling of hydrothermal systems*. Springer Science Business Media, 2004, vol. 103.
- [62] Q. Guo, Y. Wang y W. Liu, "Major hydrogeochemical processes in the two reservoirs of the Yangbajing geothermal field, Tibet, China", *Journal of Volcanology and Geothermal Research*, vol. 166, n.º 3, págs. 255-268, 2007, ISSN: 0377-0273. DOI: <https://doi.org/10.1016/j.jvolgeores.2007.08.004>. dirección: <http://www.sciencedirect.com/science/article/pii/S0377027307002569>.
- [63] W. F. Giggenbach, "Geothermal solute equilibria. Derivation of Na-K-Mg-Ca geoindicators", *Geochimica et Cosmochimica Acta*, vol. 52, n.º 12, págs. 2749-2765, 1988, ISSN: 0016-7037. DOI: [https://doi.org/10.1016/0016-7037\(88\)90143-3](https://doi.org/10.1016/0016-7037(88)90143-3). dirección: <http://www.sciencedirect.com/science/article/pii/0016703788901433>.
- [64] W. F. Giggenbach y R. C. Soto, "Isotopic and chemical composition of water and steam discharges from volcanic-magmatic-hydrothermal systems of the Guanacaste Geothermal Province, Costa Rica", *Applied Geochemistry*, vol. 7, n.º 4, págs. 309-332, 1992, ISSN: 0883-2927. DOI: <https://doi.org/10.1016/0883->

- 2927(92)90022-U. dirección: <http://www.sciencedirect.com/science/article/pii/088329279290022U>.
- [65] A. Yock, “Chemical and isotopic studies in the Miravalles geothermal field”, The United Nations University, Geothermal Training Programme, inf. téc. 17, 1998, págs. 461-499.
- [66] F. D’Amore, D. Giusti y A. Abdallah, “Geochemistry of the high-salinity geothermal field of Asal, Republic of Djibouti, Africa”, *Geothermics*, vol. 27, n.º 2, págs. 197-210, 1998, ISSN: 0375-6505. DOI: [https://doi.org/10.1016/S0375-6505\(97\)10009-8](https://doi.org/10.1016/S0375-6505(97)10009-8). dirección: <http://www.sciencedirect.com/science/article/pii/S0375650597100098>.
- [67] D. Nieva y R. Nieva, “Developments in geothermal energy in Mexico—part twelve. A cationic geothermometer for prospecting of geothermal resources”, *Heat recovery systems and CHP*, vol. 7, n.º 3, págs. 243-258, 1987.
- [68] T. Campos, “Geothermal resources of el salvador. Preliminary assessment”, *Geothermics*, vol. 17, n.º 2, págs. 319-332, 1988, ISSN: 0375-6505. DOI: [https://doi.org/10.1016/0375-6505\(88\)90061-2](https://doi.org/10.1016/0375-6505(88)90061-2). dirección: <http://www.sciencedirect.com/science/article/pii/0375650588900612>.
- [69] A. Aiuppa, M. L. Carapezza y F. Parello, “Fluid geochemistry of the San Vicente geothermal field (El Salvador)”, *Geothermics*, vol. 26, n.º 1, págs. 83-97, 1997.
- [70] F. Damore y J. T. Mejia, “Chemical and physical reservoir parameters at initial conditions in Berlingeothermal field, El Salvador: a first assessment”, *Geothermics*, vol. 28, n.º 1, págs. 45-73,

- 1999, ISSN: 0375-6505. DOI: [https://doi.org/10.1016/S0375-6505\(98\)00044-3](https://doi.org/10.1016/S0375-6505(98)00044-3). dirección: <http://www.sciencedirect.com/science/article/pii/S0375650598000443>.
- [71] T. Kavouridis, D. Kuris, C. Leonis, V. Liberopoulou, J. Leontiadis, C. Panichi, G. L. Ruffa y A. Caprai, “Isotope and chemical studies for a geothermal assessment of the island of Nisyros (Greece)”, *Geothermics*, vol. 28, n.º 2, págs. 219-239, 1999, ISSN: 0375-6505. DOI: [https://doi.org/10.1016/S0375-6505\(99\)00005-X](https://doi.org/10.1016/S0375-6505(99)00005-X). dirección: <http://www.sciencedirect.com/science/article/pii/S037565059900005X>.
- [72] E. Dotsika, I. Leontiadis, D. Poutoukis, R. Cioni y B. Raco, “Fluid geochemistry of the Chios geothermal area, Chios Island, Greece”, *Journal of Volcanology and Geothermal Research*, vol. 154, n.º 3, págs. 237-250, 2006, ISSN: 0377-0273. DOI: <https://doi.org/10.1016/j.jvolgeores.2006.02.013>. dirección: <http://www.sciencedirect.com/science/article/pii/S0377027306001041>.
- [73] S. J. Goff, F. Goff y C. J. Janik, “Tecuamburro Volcano, Guatemala: exploration geothermal gradient drilling and results”, *Geothermics*, vol. 21, n.º 4, págs. 483-502, 1992, ISSN: 0375-6505. DOI: [https://doi.org/10.1016/0375-6505\(92\)90003-R](https://doi.org/10.1016/0375-6505(92)90003-R). dirección: <http://www.sciencedirect.com/science/article/pii/S037565059290003R>.
- [74] R. Fournier y R. Potter, “Magnesium correction to the Na-K-Ca chemical geothermometer”, *Geochimica et Cosmochimica Acta*, vol. 43, n.º 9, págs. 1543-1550, 1979, ISSN: 0016-7037. DOI: [https://doi.org/10.1016/0016-7037\(79\)90147-9](https://doi.org/10.1016/0016-7037(79)90147-9). direc-

- ción: [http://www.sciencedirect.com/science/article/pii/0016703779901479](http://www.sciencedirect.com/science/article/pii/S0016703779901479).
- [75] R. O. FOURNIER, “Water geothermometers applied to geothermal energy”, en *Applications of Geochemistry in Geothermal Reservoir Development*, F. D’Amore, ed. Rome: UNITAR/UNDP Centre on Small Energy Resources, 1981, págs. 37-69.
- [76] S. Arnórsson, E. Gunnlaugsson y H. Svavarsson, “The chemistry of geothermal waters in Iceland. II. Mineral equilibria and independent variables controlling water compositions”, *Geochimica et Cosmochimica Acta*, vol. 47, n.º 3, págs. 547-566, 1983.
- [77] S. Arnórsson, E. Gunnlaugsson y H. Svavarsson, “The chemistry of geothermal waters in Iceland. III. Chemical geothermometry in geothermal investigations”, *Geochimica et Cosmochimica Acta*, vol. 47, n.º 3, págs. 567-577, 1983.
- [78] S. Arnórsson y E. Gunnlaugsson, “New gas geothermometers for geothermal exploration—calibration and application”, *Geochimica et Cosmochimica Acta*, vol. 49, n.º 6, págs. 1307-1325, 1985.
- [79] S. Arnórsson, “Geothermal systems in Iceland: structure and conceptual models—I. High-temperature areas”, *Geothermics*, vol. 24, n.º 5-6, págs. 561-602, 1995.
- [80] J. Mungania, “Borehole geology of well RN-9, Reykjanes, SW-Iceland”, UNU Geothermal Training Programme, inf. téc. 12, 1993, pág. 38.
- [81] Z. Ping y H. Ármannsson, “Gas geothermometry in selected Icelandic geothermal fields with comparative examples from Kenya”, *Geothermics*, vol. 25, n.º 3, págs. 307-347, 1996, ISSN: 0375-6505.

- DOI: [https://doi.org/10.1016/0375-6505\(96\)00006-5](https://doi.org/10.1016/0375-6505(96)00006-5). dirección: <http://www.sciencedirect.com/science/article/pii/S0375650596000065>.
- [82] S. Zhanxue, “Geothermometry and chemical equilibria of geothermal fluids from Hveragerdi, SW-Iceland, and selected hot springs Jiangxi province, SE-China”, The United Nations University, Geothermal Training Programme, inf. téc. 14, 1998, págs. 373-402.
- [83] M. Lippmann, A. Truesdell y G. Frye, “The Cerros Prieto and Salton Sea geothermal fields-are they really alike”, en *Proceedings of the 24th Workshop on Geothermal Reservoir Engineering, Stanford University, Stanford, California*, 1999, págs. 25-27.
- [84] A. Stefánsson y S. Arnórsson, “Feldspar saturation state in natural waters”, *Geochimica et Cosmochimica Acta*, vol. 64, n.º 15, págs. 2567-2584, 2000, ISSN: 0016-7037. DOI: [https://doi.org/10.1016/S0016-7037\(00\)00392-6](https://doi.org/10.1016/S0016-7037(00)00392-6). dirección: <http://www.sciencedirect.com/science/article/pii/S0016703700003926>.
- [85] B. Moon y P. Dharam, “Geothermal energy in India. Present status and future prospects”, *Geothermics*, vol. 17, n.º 2, págs. 439-449, 1988, ISSN: 0375-6505. DOI: [https://doi.org/10.1016/0375-6505\(88\)90073-9](https://doi.org/10.1016/0375-6505(88)90073-9). dirección: <http://www.sciencedirect.com/science/article/pii/S0375650588900739>.
- [86] *The chemistry of geothermal fluids in Indonesia and their relationship to water and vapour dominated systems*, vol. 28, Proceedings World Geothermal Congress, 2000.
- [87] G. Gianelli y G. Scandiffio, “The Latera geothermal system (Italy): Chemical composition of the geothermal fluid and hypotheses on

- its origin”, *Geothermics*, vol. 18, n.º 3, págs. 447-463, 1989, ISSN: 0375-6505. DOI: [https://doi.org/10.1016/0375-6505\(89\)90068-0](https://doi.org/10.1016/0375-6505(89)90068-0). dirección: <http://www.sciencedirect.com/science/article/pii/0375650589900680>.
- [88] G. Michard, D. Grimaud, F. D’Amore y R. Fancelli, “Influence of mobile ion concentrations on the chemical composition of geothermal waters in granitic areas. Example of hot springs from piemonte (Italy)”, *Geothermics*, vol. 18, n.º 5, págs. 729-741, 1989, ISSN: 0375-6505. DOI: [https://doi.org/10.1016/0375-6505\(89\)90103-X](https://doi.org/10.1016/0375-6505(89)90103-X). dirección: <http://www.sciencedirect.com/science/article/pii/037565058990103X>.
- [89] G. Caprarelli, M. Tsutsumi y B. Turi, “Chemical and isotopic signatures of the basement rocks from the Campi Flegrei geothermal field (Naples, southern Italy): inferences about the origin and evolution of its hydrothermal fluids”, *Journal of Volcanology and Geothermal Research*, vol. 76, n.º 1, págs. 63-82, 1997, ISSN: 0377-0273. DOI: [https://doi.org/10.1016/S0377-0273\(96\)00072-8](https://doi.org/10.1016/S0377-0273(96)00072-8). dirección: <http://www.sciencedirect.com/science/article/pii/S0377027396000728>.
- [90] Y. K. Kharaka y R. H. Mariner, “Chemical Geothermometers and Their Application to Formation Waters from Sedimentary Basins”, en *Thermal History of Sedimentary Basins*, N. D. Naeser y T. H. McCulloh, eds. Springer New York, 1989, págs. 99-117.
- [91] T. Noda y K. Shimada, “Water mixing model calculation for evaluation of deep geothermal water”, *Geothermics*, vol. 22, n.º 3, págs. 165-180, 1993.

- [92] K. Ariki, H. Kato, A. Ueda y M. Bamba, “Characteristics and management of the Sumikawa geothermal reservoir, northeastern Japan”, *Geothermics*, vol. 29, n.º 2, págs. 171-189, 2000.
- [93] S. Furuya, M. Aoki, H. Gotoh y T. Takenaka, “Takigami geothermal system, northeastern Kyushu, Japan”, *Geothermics*, vol. 29, n.º 2, págs. 191-211, 2000, ISSN: 0375-6505. DOI: [https://doi.org/10.1016/S0375-6505\(99\)00059-0](https://doi.org/10.1016/S0375-6505(99)00059-0). dirección: <http://www.sciencedirect.com/science/article/pii/S0375650599000590>.
- [94] H. Okada, Y. Yasuda, M. Yagi y K. Kai, “Geology and fluid chemistry of the Fushime geothermal field, Kyushu, Japan”, *Geothermics*, vol. 29, n.º 2, págs. 279-311, 2000.
- [95] N. Takeno, “Thermal and geochemical structure of the Uenotai geothermal system, Japan”, *Geothermics*, vol. 29, n.º 2, págs. 257-277, 2000, ISSN: 0375-6505. DOI: [https://doi.org/10.1016/S0375-6505\(99\)00062-0](https://doi.org/10.1016/S0375-6505(99)00062-0). dirección: <http://www.sciencedirect.com/science/article/pii/S0375650599000620>.
- [96] K. Wohletz, G. Heiken, M. Ander, F. Goff, F.-D. Vuataz y G. Wadge, “The Qualibou caldera, St. Lucia, West Indies”, *Journal of Volcanology and Geothermal Research*, vol. 27, n.º 1, págs. 77-115, 1986, ISSN: 0377-0273. DOI: [https://doi.org/10.1016/0377-0273\(86\)90081-8](https://doi.org/10.1016/0377-0273(86)90081-8). dirección: <http://www.sciencedirect.com/science/article/pii/0377027386900818>.
- [97] *Geoquímica hidrotermal del campo geotérmico de Cerro Prieto*, Segundo Simposio sobre el Campo Geotérmico de Cerro Prieto, Baja California, 1979, págs. 209-213.

- [98] A. Maimoni, "Minerals recovery from salton sea geothermal brines: a literature review and proposed cementation process", *Geothermics*, vol. 11, n.º 4, págs. 239-258, 1982, ISSN: 0375-6505. DOI: [https://doi.org/10.1016/0375-6505\(82\)90031-1](https://doi.org/10.1016/0375-6505(82)90031-1). dirección: <http://www.sciencedirect.com/science/article/pii/S0375650582900311>.
- [99] *Initial chemical and reservoir conditions at Los Azufres wellhead power plant startup*, Proceedings 10th Workshop on Geothermal Reservoir Engineering, Stanford University, ene. de 1985.
- [100] S. Villa Merlo, M. Chacón Franco y G. Medina Orozco, "Utilización de la relación atómica Na⁺/K⁺ para identificar zonas de mayor actividad hidrotermal en el campo geotérmico de la Primavera, Jalisco", *Geotermia Revista Mexicana de Geoenergía*, vol. 3, págs. 241-254, 1987.
- [101] M. P. Ochoa y A. E. Meza de Luna, "Determinaciones isotópicas de Deuterio, Oxígeno-18 and Carbono-13 en fluidos del campo geotérmico de Los Azufres, Mich.", Bachelor thesis, Instituto Tecnológico de la Laguna, México, Torreón, Coahuila, México, 1989, pág. 90.
- [102] S. P. Verma, K. Pandarinath, E. Santoyo, E. González-Partida, I. S. Torres-Alvarado y E. Tello-Hinojosa, "Fluid chemistry and temperatures prior to exploitation at the Las Tres Vírgenes geothermal field, Mexico", *Geothermics*, vol. 35, n.º 2, págs. 156-180, 2006, ISSN: 0375-6505. DOI: <https://doi.org/10.1016/j.geothermics.2006.02.002>. dirección: <http://www.sciencedirect.com/science/article/pii/S0375650506000162>.

- [103] W. A. J. Mahon y J. B. Finlayson., “Chemistry of the Broadlands geothermal area New Zealand”, *American Journal of Science*, vol. 272, n.º 1, págs. 48-68, 1972. DOI: 10.2475/ajs.272.1.48.
- [104] R. Henley y M. Stewart, “Chemical and isotopic changes in the hydrology of the tauhara geothermal field due to exploitation at wairakei”, *Journal of Volcanology and Geothermal Research*, vol. 15, n.º 4, págs. 285-314, 1983, ISSN: 0377-0273. DOI: [https://doi.org/10.1016/0377-0273\(83\)90104-X](https://doi.org/10.1016/0377-0273(83)90104-X). dirección: <http://www.sciencedirect.com/science/article/pii/S037702738390104X>.
- [105] *Geochemistry and the Exploration of the Ngawha Geothermal System, New Zealand*, The 12th Workshop on Geothermal Reservoir Engineering, Stanford University, ene. de 1987.
- [106] M. Carvalho, V. Forjaz y C. Almeida, “Chemical composition of deep hydrothermal fluids in the Ribeira Grande geothermal field (São Miguel, Azores)”, *Journal of Volcanology and Geothermal Research*, vol. 156, n.º 1, págs. 116-134, 2006, Volcanic geology of the Azores Islands, ISSN: 0377-0273. DOI: <https://doi.org/10.1016/j.jvolgeores.2006.03.015>. dirección: <http://www.sciencedirect.com/science/article/pii/S0377027306001302>.
- [107] S. Praserdvigai, “Geothermal development in Thailand”, *Geothermics*, vol. 15, n.º 5, págs. 565-582, 1986, ISSN: 0375-6505. DOI: [https://doi.org/10.1016/0375-6505\(86\)90066-0](https://doi.org/10.1016/0375-6505(86)90066-0). dirección: <http://www.sciencedirect.com/science/article/pii/S0375650586900660>.

- [108] K. Karul, "Geothermal activity in Turkey", *Geothermics*, vol. 17, n.º 2, págs. 557-564, 1988, ISSN: 0375-6505. DOI: [https://doi.org/10.1016/0375-6505\(88\)90086-7](https://doi.org/10.1016/0375-6505(88)90086-7). dirección: <http://www.sciencedirect.com/science/article/pii/0375650588900867>.
- [109] S. Arnórsson, "Geothermal systems in Iceland: Structure and conceptual models—II. Low-temperature areas", *Geothermics*, vol. 24, n.º 5-6, págs. 603-629, 1995.
- [110] A. GOKGOZ, "Geochemistry of the Kizildere-Tekkehamam-Buldand-Pamukkale geothermal fields, Turkey", *Geothermal Training in Iceland 1998*, págs. 115-156, 1998. dirección: <https://ci.nii.ac.jp/naid/20001587358/en/>.
- [111] H. Mutlu y N. Güleç, "Hydrogeochemical outline of thermal waters and geothermometry applications in Anatolia (Turkey)", *Journal of Volcanology and Geothermal Research*, vol. 85, n.º 1-4, págs. 495-515, 1998.
- [112] Ü. Gemici y G. Tarcan, "Hydrogeochemistry of the Simav geothermal field, western Anatolia, Turkey", *Journal of Volcanology and Geothermal Research*, vol. 116, n.º 3, págs. 215-233, 2002, ISSN: 0377-0273. DOI: [https://doi.org/10.1016/S0377-0273\(02\)00217-2](https://doi.org/10.1016/S0377-0273(02)00217-2). dirección: <http://www.sciencedirect.com/science/article/pii/S0377027302002172>.
- [113] G. Tarcan y Ü. Gemici, "Water geochemistry of the Seferihisar geothermal area, İzmir, Turkey", *Journal of Volcanology and Geothermal Research*, vol. 126, n.º 3, págs. 225-242, 2003, ISSN: 0377-0273. DOI: [https://doi.org/10.1016/S0377-0273\(03\)00149-5](https://doi.org/10.1016/S0377-0273(03)00149-5). dirección: <http://www.sciencedirect.com/science/article/pii/S0377027303001495>.

- [114] Ş. Şimşek, N. andıldırım y A. Gülgör, “Developmental and environmental effects of the Kızıldere geothermal power project, Turkey”, *Geothermics*, vol. 34, n.º 2, págs. 234-251, 2005, Environmental Aspects of Geothermal Development, ISSN: 0375-6505. DOI: <https://doi.org/10.1016/j.geothermics.2004.12.005>. dirección: <http://www.sciencedirect.com/science/article/pii/S0375650505000180>.
- [115] G. Tarcan, “Mineral saturation and scaling tendencies of waters discharged from wells (¿150 şC) in geothermal areas of Turkey”, *Journal of Volcanology and Geothermal Research*, vol. 142, n.º 3, págs. 263-283, 2005, ISSN: 0377-0273. DOI: <https://doi.org/10.1016/j.jvolgeores.2004.11.007>. dirección: <http://www.sciencedirect.com/science/article/pii/S0377027304003762>.
- [116] R. M. Capuano y D. R. Cole, “Fluid-mineral equilibria in a hydrothermal system, Roosevelt hot springs, Utah”, *Geochimica et Cosmochimica Acta*, vol. 46, n.º 8, págs. 1353-1364, 1982, ISSN: 0016-7037. DOI: [https://doi.org/10.1016/0016-7037\(82\)90271-X](https://doi.org/10.1016/0016-7037(82)90271-X). dirección: <http://www.sciencedirect.com/science/article/pii/001670378290271X>.
- [117] O. D. Christensen, R. A. Capuano y J. N. Moore, “Trace-element distribution in an active hydrothermal system, Roosevelt hot springs thermal area, Utah”, *Journal of Volcanology and Geothermal Research*, vol. 16, n.º 1, págs. 99-129, 1983, ISSN: 0377-0273. DOI: [https://doi.org/10.1016/0377-0273\(83\)90086-0](https://doi.org/10.1016/0377-0273(83)90086-0). dirección: <http://www.sciencedirect.com/science/article/pii/S0377027383900860>.

- [118] K. Murray, M. Jonas y C. Lopez, “Geochemical exploration of the Calistoga geothermal resource area, Napa Valley, California”, *Geothermal Resources Council Transactions*, vol. 9, n.º 1, págs. 339-344, 1985.
- [119] R. O. Fournier, “Double-diffusive convection in geothermal systems: the salton sea, California, geothermal system as a likely candidate”, *Geothermics*, vol. 19, n.º 6, págs. 481-496, 1990, ISSN: 0375-6505. DOI: [https://doi.org/10.1016/0375-6505\(90\)90001-R](https://doi.org/10.1016/0375-6505(90)90001-R). dirección: <http://www.sciencedirect.com/science/article/pii/037565059090001R>.
- [120] J. N. Moore, M. C. Adams, T. L. Sperry, K. K. Bloomfield y R. Kunzman, “Preliminary results of geochemical monitoring and tracer tests at the Cove Fort-Sulphurdale geothermal system, Utah”, en *Proc., The 25th Workshop on Geothermal Reservoir Engineering, Stanford University, Stanford, CA*, 2000.
- [121] J. L. Palandri y M. H. Reed, “Reconstruction of in situ composition of sedimentary formation waters”, *Geochimica et Cosmochimica Acta*, vol. 65, n.º 11, págs. 1741-1767, 2001, ISSN: 0016-7037. DOI: [https://doi.org/10.1016/S0016-7037\(01\)00555-5](https://doi.org/10.1016/S0016-7037(01)00555-5). dirección: <http://www.sciencedirect.com/science/article/pii/S0016703701005555>.
- [122] O. M. Espinoza-Ojeda y E. Santoyo, “A new empirical method based on log-transformation regressions for the estimation of static formation temperatures of geothermal, petroleum and permafrost boreholes”, *Journal of Geophysics and Engineering*, vol. 13, n.º 4, págs. 559-596, ago. de 2016. DOI: 10.1088/1742-2132/13/4/559. dirección: [http://iopscience.iop.org/1742-2140/13/4/559%](http://iopscience.iop.org/1742-2140/13/4/559%20)

20<http://stacks.iop.org/1742-2140/13/i=4/a=559?key=crossref.48448ffbd35e16e8bb2e269f8327359f>.

- [123] S. Buuren y K. Groothuis-Oudshoorn, “mice : Multivariate Imputation by Chained Equations in R”, *J. Stat. Softw.*, vol. 45, n.º 3, 2011, ISSN: 1548-7660. DOI: 10.18637/jss.v045.i03. dirección: <http://www.jstatsoft.org/v45/i03/>.



UNIVERSIDAD AUTÓNOMA DEL
ESTADO DE MORELOS



FACULTAD DE CONTADURÍA, ADMINISTRACIÓN E INFORMÁTICA

Cuernavaca, Morelos a 03 de septiembre de 2019.

DRA. LAURA PATRICIA CEBALLOS GILES
DIRECTORA DE LA FCAeI
PRESENTE

En mi carácter de director de Tesis, titulada **Imputación equivalente de una base de datos de fluidos geotérmicos**, que presenta la estudiante **Mariana Alelhi Román Flores** con matrícula **10010403**, para obtener el grado de **Maestro en Optimización y Computo Aplicado**, cumple con los lineamientos teóricos, metodológicos y de investigación requeridos en el Reglamento de Titulación de Posgrado en la UAEM.

Con base en los argumentos precedentes me permito otorgar la **LIBERACIÓN DE TESIS** por lo que de mi parte no existe inconveniente para que el estudiante continúe con los trámites que esta Secretaría de Investigación y Posgrado tenga establecidos para obtener el grado mencionado.

Atentamente
Por una humanidad culta
Una universidad de excelencia

Dra. Lorena Díaz González
Profesor- investigador
Centro de Investigación en Ciencias



UNIVERSIDAD AUTÓNOMA DEL
ESTADO DE MORELOS



FACULTAD DE CONTADURÍA, ADMINISTRACIÓN E INFORMÁTICA

Cuernavaca, Morelos a 21 de Agosto del 2019.

DRA. LAURA PATRICIA CEBALLOS GILES
DIRECTORA DE LA FCAeI
PRESENTE

En mi carácter de revisor de Tesis, hago de su conocimiento que he leído con interés la tesis para obtener el grado de Maestría en Optimización y Cómputo Aplicado, de la estudiante Mariana Alelhi Román Flores, con matrícula 10010403, con el título **Imputación equivalente de una base de datos de fluidos geotérmicos** por lo cual, me permito informarle que después de una revisión cuidadosa de dicha tesis, concluyo que el trabajo se caracteriza por el establecimiento de objetivos académicos pertinentes y una metodología adecuada para su logro. Además construye una estructura coherente y bien documentada, por lo cual considero que los resultados obtenidos contribuyen al conocimiento del tema tratado.

Con base en los argumentos precedentes me permito expresar mi **VOTO APROBATORIO** por lo que de mi parte no existe inconveniente para que la estudiante continúe con los trámites que esta Secretaría de Investigación y Posgrado tenga establecidos para obtener el grado mencionado.

Atentamente,

Dr. Outmane Oubram
Profesor- investigador
Facultad de Ciencias Químicas e Ingeniería

Por una humanidad culta
Una universidad de excelencia



UNIVERSIDAD AUTÓNOMA DEL
ESTADO DE MORELOS



FACULTAD DE CONTADURÍA, ADMINISTRACIÓN E INFORMÁTICA


Cuernavaca, Morelos a 21 de Agosto del 2019.

DRA. LAURA PATRICIA CEBALLOS GILES
DIRECTORA DE LA FCAeI
PRESENTE

En mi carácter de revisor de Tesis, hago de su conocimiento que he leído con interés la tesis para obtener el grado de Maestra en Optimización y Cómputo Aplicado, de la estudiante Mariana Alelhi Román Flores, con matrícula 10010403, con el título **Imputación equivalente de una base de datos de fluidos geotérmicos** por lo cual, me permito informarle que después de una revisión cuidadosa de dicha tesis, concluyo que el trabajo se caracteriza por el establecimiento de objetivos académicos pertinentes y una metodología adecuada para su logro. Además construye una estructura coherente y bien documentada, por lo cual considero que los resultados obtenidos contribuyen al conocimiento del tema tratado.

Con base en los argumentos precedentes me permito expresar mi **VOTO APROBATORIO** por lo que de mi parte no existe inconveniente para que la estudiante continúe con los trámites que esta Secretaría de Investigación y Posgrado tenga establecidos para obtener el grado mencionado.

Atentamente,



Dr. José Alberto Hernández Aguilar
Profesor- investigador
FCAeI - UAEM

Por una humanidad culta
Una universidad de excelencia



UNIVERSIDAD AUTÓNOMA DEL
ESTADO DE MORELOS



FACULTAD DE CONTADURÍA, ADMINISTRACIÓN E INFORMÁTICA

Cuernavaca, Morelos a 21 de Agosto del 2019.

DRA. LAURA PATRICIA CEBALLOS GILES
DIRECTORA DE LA FCAeI
PRESENTE

En mi carácter de revisor de Tesis, hago de su conocimiento que he leído con interés la tesis para obtener el grado de Maestra en Optimización y Cómputo Aplicado, de la estudiante Mariana Alehí Román Flores, con matrícula 10010403, con el título **Imputación equivalente de una base de datos de fluidos geotérmicos** por lo cual, me permito informarle que después de una revisión cuidadosa de dicha tesis, concluyo que el trabajo se caracteriza por el establecimiento de objetivos académicos pertinentes y una metodología adecuada para su logro. Además construye una estructura coherente y bien documentada, por lo cual considero que los resultados obtenidos contribuyen al conocimiento del tema tratado.

Con base en los argumentos precedentes me permito expresar mi **VOTO APROBATORIO** por lo que de mi parte no existe inconveniente para que la estudiante continúe con los trámites que esta Secretaría de Investigación y Posgrado tenga establecidos para obtener el grado mencionado.

Atentamente,

Dr. Luis Manuel Gaggero Sager
Profesor- investigador
CIICAP-UAEM

Por una humanidad culta
Una universidad de excelencia



UNIVERSIDAD AUTÓNOMA DEL
ESTADO DE MORELOS



FACULTAD DE CONTADURÍA, ADMINISTRACIÓN E INFORMÁTICA

Cuernavaca, Morelos a 21 de Agosto del 2019.

DRA. LAURA PATRICIA CEBALLOS GILES
DIRECTORA DE LA FCAeI
PRESENTE

En mi carácter de revisor de Tesis, hago de su conocimiento que he leído con interés la tesis para obtener el grado de Maestra en Optimización y Cómputo Aplicado, de la estudiante Mariana Alelhi Román Flores, con matrícula 10010403, con el título **Imputación equivalente de una base de datos de fluidos geotérmicos** por lo cual, me permito informarle que después de una revisión cuidadosa de dicha tesis, concluyo que el trabajo se caracteriza por el establecimiento de objetivos académicos pertinentes y una metodología adecuada para su logro. Además construye una estructura coherente y bien documentada, por lo cual considero que los resultados obtenidos contribuyen al conocimiento del tema tratado.

Con base en los argumentos precedentes me permito expresar mi **VOTO APROBATORIO** por lo que de mi parte no existe inconveniente para que la estudiante continúe con los trámites que esta Secretaría de Investigación y Posgrado tenga establecidos para obtener el grado mencionado.

Atentamente,

Dr. Guillermo Santamaría Bonfil
Profesor- investigador
CONACYT-INEEL

Por una humanidad culta
Una universidad de excelencia