UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

POSGRADO EN CIENCIA E INGENIERÍA DE LA COMPUTACIÓN

Metodologías de interpretación automática de datos geofísicos multidimensionales

T E S I S

QUE PARA OPTAR POR EL GRADO DE:
MAESTRO EN CIENCIA E INGENIERÍA DE LA COMPUTACIÓN

P R E S E N T A:
MANUEL ORTIZ OSIO

Director de Tesis:
Dr. Erik Molino Minero Re
IIMAS, UNAM

Ciudad Universitaria, CD. MX. Octubre, 2022

Agradecimientos

Al Consejo Nacional de Ciencia y Tecnología por el apoyo financiero durante el periodo que comprendió el desarrollo de este trabajo.

A mi director de tesis, el Dr. Erik Molino Minero Re, por su gran apoyo, guía y confianza.

Al departamento de geofísica de la Facultad de Ingeniería de la UNAM por los datos brindados, producto de las prácticas de campo de las materias de *Prospección gravimétrica y magnetométrica* y *Prospección eléctrica*; así también por su ayuda con la interpretación de los mismos.

A mis profesores del IIMAS, por su guía y su entrega en cada clase.

Al Ing. Alejandro García, al Dr. Andrés Tejero Andrade, al Dr. Esteban Hernández y a Yosselin Angeles por sus comentarios e impresiones sobre la interpretación de las imágenes.

Resumen

Se presenta una metodología basada en algoritmos de aprendizaje computacional cuyo objetivo es realizar interpretaciones automáticas de datos geofísicos multidimensionales.

Se implementó la base de un preprocesamiento *adaptativo*, apoyada en la transformada de ondícula y en algoritmos de detección de anomalías, con objetivo de filtrar los datos de entrada. COn esto se busca optimizar el desempeño de los algoritmos de aprendizaje computacional, teniendo como entrada datos con calidad mejorada.

El agrupamiento, sector central de la metodología planteada, usa los algoritmos de k-medias, k-medianas, k-medoides y mapas auto-organizados para realizar la interpretación automatizada, encontrando los patrones existentes en las variables de las bases de datos usadas.

En este trabajo se usaron datos provenientes de levantamientos magnetométricos terrestres, cuyas variables son los resultados de filtros de realce de anomalías típicos, así como datos de tomografías de resistividad eléctrica en dos dimensiones usando dos arreglos electródicos.

La metodología propuesta es capaz de fungir como intérprete automático, cuyos resultados son correlacionables con información *in situ* y con la experiencia de intérpretes en el área de geofísica.

Abstract

A methodology based on machine learning algorithms, whose objective is to perform automatic interpretations of multidimensional geophysical data, is presented.

An adaptive preprocessing based on the wavelet transform and anomaly detection algorithms was implemented to filter the input data, to improve the performance of the machine learning algorithms by having input data with improved quality.

Clustering, the core of the proposed methodology, uses the k-means, k-medians, k-medoids and self-organized maps algorithms to carry out the automated interpretation, finding the patterns in the variables of the databases used.

In this work, data from ground magnetometric surveys were used, whose variables are the results of typical anomaly enhancement filters, as well as data from electrical resistivity tomography in two dimensions using two electrode arrays.

The proposed methodology is capable of acting as an automatic interpreter, whose results can be correlated with information in situ, as well as with the experience of geophysics interpreters.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

1.1. Motivación

De acuerdo con el portal del servicio geológico mexicano (SGM, 2017), la geofísica es la aplicación de los principios y prácticas de la física para la resolución de problemas relacionados con la Tierra, midiendo de modo indirecto las variaciones de las magnitudes físicas en el espacio y en el tiempo. La geofísica tiene como uno de sus objetivos analizar múltiples imágenes en un dominio definido, se busca realizar una interpretación de las estructuras que son fuente de las variables registradas.

Se han desarrollado herramientas para reducir incertidumbre en la interpretación de datos geofísicos multidimensionales, dentro de los que destacan:

1. la inversión conjunta, donde analíticamente se resuelven de forma simultánea las ecuaciones de los campos físicos involucrados, algunos detalles se pueden consultar en el trabajo de Gallardo and Meju (2004);

2. aprendizaje computacional, donde se busca realizar una clasificación geológica con datos de levantamientos regionales, algunos detalles se pueden consultar en el trabajo de Carneiro et al. (2012).

El punto 1 resuelve el problema desde un punto de vista físico, buscando determinar las propiedades físicas que producen la respuesta del medio durante la toma

de los datos. Esta técnica puede ser muy inestable, obteniendo resultados considerablemente diferentes ante pequeños cambios en los parámetros. El punto 2 resuelve el problema desde un punto de vista matemático, buscando patrones en las variables registradas asociándolas a una litología conocida. Esta técnica funciona bien para la clasificación geológica, sin embargo no siempre es posible disponer de una base de datos con etiquetas asociadas para efectuar un entrenamiento supervizado.

1.2. Planteamiento del problema

Las interpretaciones son cualitativas, por lo que las conclusiones dependen de la experiencia del intérprete y de su capacidad para analizar datos multidimensionales, este proceso puede ser repetitivo y complicado.

1.3. Trabajo previo

Se han utilizado técnicas de aprendizaje computacional para automatizar la búsqueda de patrones en imágenes satelitales (Bedini (2009), Ghimire et al. (2010), Goncalves et al. (2008), Huang and Shibuya (2020), Waske and Braun (2009), Yu et al. (2012)). Las herramientas más usadas son *bosques aleatorios*, *máquinas de soporte vectorial* y *mapas auto-organizados*. La escala de estos estudios es muy grande, las estructuras buscadas, generalmente estructuras geológicas regionales, tienen dimensiones de kilómetros y son muy sensibles a la posición (georreferencia de estas estructuras).

Basados en los trabajos anteriores, otros autores han mezclado datos de imágenes satelitales con levantamientos aeromagnéticos (Bachri et al. (2020), Carneiro et al. (2012), Chen et al. (2020), Costa et al. (2019), Cracknell et al. (2014), Cracknell and Reading (2013), Harris and Grunsky (2015), Kuhn et al. (2018), Kuhn et al. (2019), Nathan et al. (2020)), aplicando de forma general las mismas herramientas de aprendizaje computacional. Su principal objetivo es entrenar un sistema de forma supervisada para que pueda predecir litología.

1.4. Objetivo general

Proponer una metodología para la agrupación de datos geofísicos multidimensionales de pequeña escala, con base en algoritmos de aprendizaje no supervizado. Esto con la finalidad de tener herramientas computacionales que sean de apoyo para la reducción en la complejidad y en la disminución en los tiempos de interpretación, es decir, la implementación de un intérprete automático.

1.5. Objetivos específicos

- Implementar una biblioteca con la capacidad de pre-procesar datos y aplicar herramientas de aprendizaje y agrupamiento a datos geofísicos.

- Implementar la base de un pre-procesamiento *adaptativo* apoyada en algoritmos de detección de anomalías y en la transformada de ondícula

1.6. Organización de la tesis

En el capítulo 2 se revisan las bases en las que se fundamenta la implementación de los algoritmos. En el capítulo 3 se muestra la metodología propuesta, funcionando para magnetometría y tomografías de resistividad eléctrica en dos dimensiones, bases de datos cuyas características se revisan en el capítulo 4. Finalmente en el capítulo 5 se muestran resultados de la aplicación de la metodología planteada, exhibiendo imágenes obtenidas de la interpretación automática.

La base de datos a la que se tiene acceso para este trabajo contiene información de levantamientos geofísicos que involucran datos de componente total de campo magnético y datos de resistividad eléctrica obtenida de dos arreglos electródicos. Estos datos provienen de levantamientos realizados en prácticas escolares de las materias de prospección eléctrica y prospección gravimétrica y magnetométrica de la carrera de ingeniería geofísica, impartida en la Facultad de Ingeniería de la *UNAM*. El acceso a los datos es otorgado por el Ing. Alejandro García Serrano, el Dr. Andrés

Tejero Andrade y el M.C. David Escobedo Zenil.

El código implementado, así como ejemplos de uso, se pueden consultar en el siguiente enlace https://github.com/CecilRamza/Interprete_geofisica.

Capítulo 2

Marco teórico

2.1. Introducción

En este capítulo se definen de forma breve los aspectos teóricos usados para el desarrollo de este trabajo.

Comienzo con dar las bases del pre-procesamiento aplicado, tema que involucra la transformada de ondícula y los algoritmos de detección de anomalías. La implementación propuesta requiere que a algunas bases de datos se les apliquen transformaciones, temas que se tocan más adelante, así como una técnica de inmersión (*embedding*) de datos.

El desarrollo continua con los algoritmos de reducción de dimensionalidad y con las transformaciones basadas en el escalamiento de los datos y de su histograma, usadas para el realce de imágenes. Estas herramientas se aplican para preparar los datos de entrada al proceso de aprendizaje computacional.

Tocando la materia de aprendizaje computacional, se muestran las bases de los métodos basados en agrupamiento, así como las bases de los mapas auto-organizados.

Finalmente se introducen los temas de prospección geofísica, exponiendo las características generales de los métodos y del procesamiento al que se somete cada

conjunto de datos. Se busca también exponer el estado del arte en materia de aprendizaje computacional aplicado a este tipo de datos.

2.2. Transformada de ondícula

De acuerdo con la teoría del *análisis de Fourier*, una señal puede ser representada como la suma infinita de una serie de senos y cosenos (Oppenheim et al., 1998). La desventaja en este tipo de análisis es que la información en frecuencia y en tiempo no se ven a la vez: en el dominio del tiempo no se tiene información de la frecuencia y en el dominio de la frecuencia no se tiene información del tiempo.

La transformada de ondícula es una de las soluciones para obtener información en tiempo y en frecuencia al mismo tiempo. Este método se basa en la aplicación de una ondícula escalable que se desplaza a lo largo de la señal, este proceso se repite cambiando las características de traslación y escala de la ondícula, resultando en un conjunto de representaciones en tiempo-frecuencia de la señal a distinta resolución, denominadas como representaciones en tiempo-escala (Valens, 1999).

2.2.1. Ondículas

Una ondícula, u onduleta, es una señal oscilatoria de corta duración que puede generar familias de ondículas ortogonales (Edwards, 1991). Las propiedades más importantes de las ondículas son las condiciones de admisibilidad y regularidad.

Las funciones que satisfacen las condiciones de admisibilidad:

$$\int \frac{|\Psi(\omega)|^2}{|\omega|} d\omega < \infty \tag{2.1}$$

se pueden usar para analizar y reconstruir una señal sin perder información (Valens, 1999), donde $\Psi(\omega)$ es la transformada de Fourier de $\psi(t)$. Esta condición implica que la transformada de Fourier de $\psi(t)$ se anula en la frecuencia cero.

Esto indica que el espectro de las ondículas debe tener la forma de un filtro pasa-banda. Además, en la frecuencia cero el valor promedio de la ondícula debe ser nulo y en el dominio temporal necesita ser una función oscilatoria (Valens, 1999).

Refiriéndonos a la transformada discreta de ondícula, la base ortogonal de una ondícula es un conjunto de funciones definidas como (Edwards, 1991):

$$\psi(t) = \sum_{k=0}^{M-1} c_k \psi(2t - k). \tag{2.2}$$

Si una ondícula puede verse como un filtro pasa-banda, entonces una serie de ondículas *dilatadas* es equivalente a un banco de filtros pasa-banda (Valens, 1999). Sin embargo, para cubrir todo el espectro usando únicamente *dilataciones* sería necesario un número infinito de ondículas, lo cual puede ser poco práctico para fines discretos. Para resolver este problema se agrega una *función de escalamiento* cuyo espectro tiene la forma de un filtro pasa-bajas.

La elección de la ondícula a usar dependerá del tipo de señal a analizar, de la información que se quiera obtener de ella. Un criterio para seleccionar una ondícula consiste en elegir aquella cuya forma sea la más parecida a la señal estudiada; otro método se basa en realizar pruebas con diferentes ondículas, seleccionando aquella con la que obtenga el mejor resultado (GISI, 2018).

A la ondícula con la que se aplicará la transformación se le llama *ondícula madre*. A esta se le aplicarán modificaciones que están en función de parámetros de escalamiento y de desplazamiento. El escalamiento consiste en *alargar* o *comprimir* la ondícula, mientras que el desplazamiento radica en el recorrido de la ondícula a lo largo de la señal. La *ondícula madre* $\psi_{a,b}(t)$ se define entonces como:

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}} \psi \left(\frac{t - b}{a} \right) \tag{2.3}$$

donde a es el escalamiento y b es el desplazamiento.

2.2.2. Transformada contínua de ondícula

La transformada contínua de ondícula (CWT por sus siglas en inglés) se define como la integral del producto de una señal $x(t)$ con una *ondícula madre* (Mallat, 1999):

$$CWT(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t)\psi\left(\frac{t-b}{a}\right) dt \tag{2.4}$$

Esta integral correlaciona la variación de x en un vecindario b, cuyo tamaño es proporcional a a, en otras palabras correlaciona la señal de interés con una ondícula a cierta escala. Los *coeficientes de ondícula* indican la relación que existe entre la señal de interés y la *ondícula madre*.

2.2.3. Transformada discreta de ondícula

La transformada discreta de ondícula (DWT por sus siglas en inglés) se obtiene al discretizar el escalamiento y el desplazamiento (GISI, 2018):

$$\begin{aligned} a &= 2^{-j} \\ b &= k2^{-j} \end{aligned} \tag{2.5}$$

con j y k enteros. La *ondícula madre* de forma discreta se define entonces como:

$$\psi_{j,k}(t) = 2^{j/2}\psi(2^j t - k) \tag{2.6}$$

por lo que la DWT queda como:

$$DWT_{j,k} = 2^{j/2} \int_{-\infty}^{\infty} x(t)\psi(2^j t - k)dt \tag{2.7}$$

La recostrucción de la señal está en función de la ondícula y de la función de escalamiento:

$$x(t) = \sum_k \sum_j c_{j,k}\phi(t) + \sum_k \sum_j d_{j,k}\psi(t) \tag{2.8}$$

donde $c_{j,k}$ son los coeficientes de aproximación, asociados con la función de escalamiento, y $d_{j,k}$ son los coeficientes de detalle, asociados con la ondícula. Estos coeficientes otorgan información sobre las características de la señal analizada. Con su manipulación es posible atenuar o eliminar componentes no deseadas de la señal original.

La aplicación de la *DWT* condujo a la implementación de *bancos de filtros*, que son un conjunto de filtros que constan de un filtro pasa-bajas (función de escalamiento) y de un filtro pasa-altas (ondícula). Como producto de la *DWT*, aplicando de forma recursiva el par de filtros a la señal salida del filtro pasa-bajas, se obtiene una descomposición de la señal a distintos niveles.

2.3. Detección de anomalías

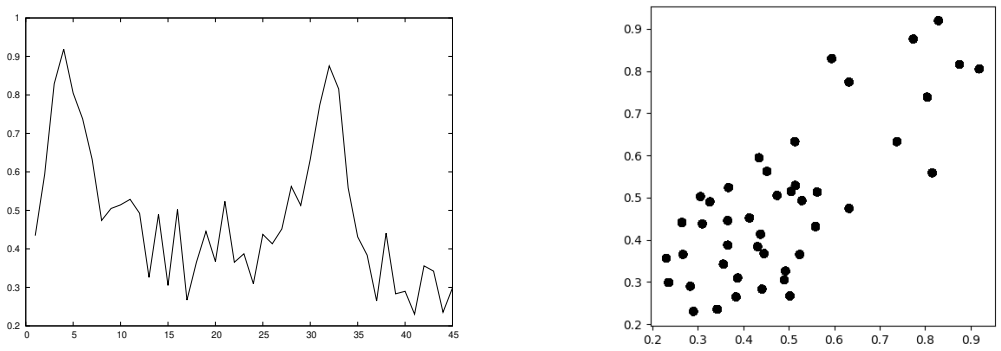
Una anomalía es una instancia que no se asemeja a la mayoría de las observaciones. La práctica tradicional con respecto a las anomalías es asociarlas a ruido o error en las observaciones, sin embargo también pueden ser indicio de un cambio de importancia en el objeto de estudio (Montiel et al., 2021).

Al proceso de identificación de observaciones potencialmente anómalas se le conoce como *detección de anomalías* (Markou and Singh, 2003). Los algoritmos de *detección de anomalías* caen dentro de una rama de la *inteligencia artificial* (*IA*) conocida como *aprendizaje no supervisado* (sección 2.6).

A continuación se abordarán de forma general los algoritmos aplicados en este trabajo: los *bosques de aislamiento* y el *factor local de valor atípico*, así como una técnica de *incrustación*, o *inmersión* (*embedding* en inglés), usada como etapa anterior a los algoritmos antes mencionados.

2.3.1. Inmersión

Una *incrustación* o *inmersión* es un espacio de dimensiones bajas que puede traducirse en vectores de dimensiones altas (Google, 2020).



(a) . (b) .

Figura 2.1: Ejemplo de aplicación de la técnica de inmersión empleada. (a) serie de datos original, (b) inmersión aplicada, los ejes son la variable dependiente del dato i (eje x) y del dato $i+1$ (eje y).

Un ejemplo de esta técnica, consiste en la representación de una serie de tiempo (siendo equivalente a una serie de datos equiespaciados) en un espacio bidimensional definido por la variable dependiente de cada dato y de su k vecino más cercano. Con esta metodología es posible representar una serie de datos en forma de un gráfico de dispersión en dos dimensiones, cuyos ejes son la variable dependiente del dato i, la variable dependiente del dato $i+1$, hasta el dato $i+k$; espacio donde es posible aplicar algoritmos de *detección de anomalías* para estimar vectores atípicos de la serie de datos de interés. La figura 2.1 muestra un ejemplo de la aplicación de este algoritmo para $k=1$.

2.3.2. Bosques de aislamiento

El algoritmo de *bosques de aislamiento* (*IF* por sus siglas en inglés) cuantifica el *nivel de anomalía* de un vector[1] en función de qué tan difícil es aislarlo del resto de los vectores (Liu et al., 2008).

Este algoritmo consiste en trazar hiperplanos en el espacio de atributos, definido por las n-dimensiones de la base de datos, cada hiperplano es perpendicular a alguno de los ejes de este espacio, elegido de forma aleatoria divide el espacio en dos subre-

[1]Entiéndase como una observación n-dimensional.

giones. Si el vector de interés es el único existente en una subregión este se encontrará aislado de los demás, en caso contrario se repetirá el proceso de aislamiento en la subregión en donde este se encuentre. El número de iteraciones necesarias para aislar un vector es una medida del *nivel de anomalía*. El método de *IF* es considerado un algoritmo global, ya que compara el número de iteraciones para cada vector con el del resto (Montiel et al., 2021).

Algunas propiedades de los *IF* son (Liu et al., 2008):

1. submuestreo, funcionan muy bien cuando el número de muestras es pequeño;
2. inmersión, ocurre cuando vectores típicos se encuentran cercanos a vectores anómalos, el submuestreo es una opción para evitarlo;
3. enmascaramiento, ocurre cuando existen muchos valores atípicos, complicando el proceso de isolación, de la misma forma el submuestreo es una opción para evitar este fenómeno;
4. vectores de múltiples dimensiones, esta característica vuelve menos eficientes a los métodos basados en distancia (Verleysen and François, 2005), siendo también sensibles los *IF*;
5. únicamente vectores típicos, los *IF* funcionan bien incluso si la base de datos no contiene vectores anómalos.

El *nivel de anomalía* se calcula de la siguiente manera:

$$s(x, m) = 2^{\frac{-E(h(x))}{c(m)}} \tag{2.9}$$

donde:

$$c(m) = \begin{cases} 2H(m-1) - \frac{2(m-1)}{n} & m > 2 \\ 1 & m = 2 \\ 0 & \text{en otro caso} \end{cases} \tag{2.10}$$

con n igual al tamaño de la base de datos, m igual al tamaño del conjunto muestreado y H es el número armónico[2]. $c(m)$ representa el promedio de $h(x)$ dado m y $E(h(x))$ es el valor promedio de $h(x)$ para un conjunto de árboles de aislamiento.

Finalmente, dado un vector x:

- si s es cercano a 1, entonces x probablemente sea una anomalía,
- si s es menor a 0,5, entonces x probablemente sea típico.

2.3.3. Factor local de valor atípico

El algoritmo del *factor local de valor atípico* (LOF por sus siglas en inglés) es un algoritmo que se considera local, tiene la premisa de que dos vectores cercanos suelen tener la misma densidad respecto a su vecindad (Montiel et al., 2021).

Breunig et al. (2000) definen la *distancia de alcanzabilidad* como:

$$\text{distancia-alcanzabilidad}_k(A, B) = \text{máx}\left(k\text{-distancia}(B), d(A, B)\right) \tag{2.11}$$

donde k-distancia(B) es la distancia del vector B a su k-ésimo vecino más cercano. La *densidad de alcanzabilidad local* de un vector A se define como:

$$lrd_k(A) = 1/\left(\frac{\sum_{B \in N_k(A)} \text{distancia-alcanzabilidad}_k(A, B)}{|N_k(A)|}\right) \tag{2.12}$$

donde $N_k(A)$ es el conjunto de los k vecinos más cercanos. Finalmente, el factor LOF se expresa de la siguiente forma:

$$LOF_k(A) = \frac{\sum_{B \in N_k(A)} lrd_k(B)}{|N_k(A)| lrd_k(A)} \tag{2.13}$$

que es la comparación de las *densidades de alcanzabilidad local*. Surgen entonces tres posibles casos:

[2]Puede ser estimado como $H(i) = \ln(i) + \gamma$, donde $\gamma = 0{,}5772156649$.

- $LOF(K) \sim 1$, densidades similares como vecinos;
- $LOF(K) < 1$, punto interior (*inlier*); y
- $LOF(K) > 1$, punto exterior (*outlier*).

Este algoritmo, al igual que los bosques de aislamiento (sección 2.3.2), es suceptible a la *maldición de la dimensionalidad* (Verleysen and François, 2005), por lo que se obtienen mejores resultados en espacios de dimensiones bajas.

2.4. Reducción de dimensionalidad

La *reducción de dimensionalidad* es la transformación de datos multidimensionales en una representación de baja dimensionalidad que mantiene lo más posible sus características generales. La importancia de estas herramientas radica en que ayudan a mitigar el efecto de la *maldición de la dimensionalidad* (Jimenez and Landgrebe, 1998). La *reducción de la dimensionalidad* facilita la clasificación, visualización y compresión de datos multidimensionales (Van Der Maaten et al., 2009).

Los algoritmos de *reducción de dimensionalidad* pueden clasificarse en técnicas lineales y en técnicas no lineales, donde estas últimas resaltan sobre las primeras en que pueden manejar datos con estructuras complejas. Un ejemplo de técnica lineal es el *análisis de componentes principales*, mientras que *isomap* es un ejemplo de técnica no lineal.

Un criterio para decidir si se reduce la dimensionalidad de un conjunto de variables es calculando el grado de correlación que tienen estas. Una de las técnicas más usadas es el *índice de correlación de Pearson*.

2.4.1. Índice de correlación de Pearson

El *índice de correlación de Pearson* es una medida de dependencia lineal entre dos variables (de Oviedo, 2009).

Sea un par de variables (X, Y), el *coeficiente de correlación de Pearson* $r_{X,Y}$ es:

$$r_{X,Y} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y} \tag{2.14}$$

donde $\sigma_{X,Y}$ es la covarianza de (X, Y) y σ_X y σ_Y son la desviación estándar de X y Y respectivamente. El índice varía dentro del intervalo $[-1, 1]$, donde 1 es una correlación lineal perfecta, -1 es una correlación negativa perfecta y 0 es la inexistencia de una correlación lineal.

2.4.2. Análisis de componentes principales

El *análisis de componentes principales* (*PCA* por sus siglas en inglés), es una técnica ampliamente usada para la *reducción de dimensionalidad*, compresión de información, extracción de características y visualización (Jolliffe, 2005). El *PCA* puede definirse como la proyección ortogonal de los datos en un espacio de menor dimensionalidad, en la que se maximiza la varianza de estos datos proyectados (Bishop and Nasrabadi, 2006).

Sea un conjunto de datos x_n con $n = 1, ..., N$ con dimensionalidad D, y sea M las dimensiones del espacio al que se desea proyectar sabiendo que $M < D$. Buscamos maximizar la varianza proyectada de:

$$u^T S u$$

donde u es un vector unitario y S es la matriz de covarianza:

$$S = \frac{1}{N} \sum_{n=1}^{N} (x_n - \bar{x})(X_n - \bar{x})^T \tag{2.15}$$

Necesitamos agregar una restricción para evitar que $||u_1|| \to \infty$, agregamos un multiplicador de Lagrange λ_1:

$$u_1^T S u_1 + \lambda_1 \left(1 - u_1^T u_1\right) \tag{2.16}$$

Haciendo la derivada con respecto a u_1 igual a cero, un punto estacionario será:

$$\begin{aligned} Su_1 &= \lambda_1 u_1 \\ u_1^T Su_1 &= \lambda_1 \end{aligned} \tag{2.17}$$

por lo que u_1 debe ser un vector propio de S. La varianza será máxima cuando u_1 sea igual al vector propio con el valor propio más grande, A este vector propio se le conoce como *primer componente principal* (Bishop and Nasrabadi, 2006).

Podemos definir *componentes principales* adicionales eligiendo diferentes direcciones ortogonales que maximicen la varianza proyectada.

2.4.3. Isomap

Isomap es un método no lineal de *reducción de dimensionalidad*. El algoritmo consiste en lo siguiente (Tenenbaum et al., 2000):

1. determinar los vecinos de cada vector,
   - dado un radio, y
   - dados los k vecinos más cercanos;
2. construcción del gráfico de vecindad,
   - un punto se conecta con otro si es uno de los k vecinos más próximos del otro, y
   - la longitud de la arista es igual a su distancia euclidiana;
3. computar el camino más corto entre dos nodos;
4. computar la inmersión (*embedding*) a una dimensión menor usando el *escalamiento multidimensional* (*MDS* por sus siglas en inglés).

Uno de los principales problemas de este algoritmo es que es muy sensible al número elegido de vecinos más cercanos, pudiendo producir errores en la matriz de distancias (Balasubramanian and Schwartz, 2002).

2.5. Realce de imágenes

El realce de una imagen se obtiene al procesarla de tal forma que se obtenga otra imagen en la que alguna o algunas de sus características sean *más convenientes* de acuerdo a una aplicación específica (Gonzalez, 2009), es decir los procesos están orientados al problema a resolver. Algunas transformaciones básicas son:

- transformación logarítmica,
- procesamiento del histograma.

2.5.1. Transformación logarítmica

La forma general de una transformación logarítmica es:

$$s = c \log (1 + r) \tag{2.18}$$

donde c es una constante y se asume que $r \geq 0$. Esta transformación mapea un rango angosto de valores a un rango más ancho a la salida, resultando en una expansión de los niveles en la imagen resultante.

2.5.2. Procesamiento de histogramas

La manipulación de histogramas tiene por objetivo realzar o mejorar la calidad de una imagen (Gonzalez, 2009). El histograma de una imagen es la representación gráfica de la distribución que existe de sus niveles con el número de pixeles o puntos que definen la imagen, es decir el histograma representa la frecuencia relativa de ocurrencia de cada nivel.

La ecualización del histograma es una técnica que ayuda a mejorar el contraste de una imagen al volver su histograma más uniforme. Como consecuencia, una imagen ecualizada permite que ciertos detalles sean visibles en regiones oscuras o brillantes (Gonzalez, 2009). La ecualización se computa de la siguiente forma:

1. cálculo de la frecuencia acumulada del nivel en cuestión dividida entre el número total de pixeles o puntos,

2. multiplicación del valor obtenido en el punto anterior por el valor del nivel máximo de la imagen.

aplicándose de forma recursiva para cada nivel de la imagen a procesar.

2.6. Aprendizaje computacional

El *aprendizaje computacional*, *aprendizaje máquina* o *aprendizaje automático* (*ML* por sus siglas en inglés), es una disciplina que se enfoca en dos preguntas relacionadas entre sí: ¿cómo se puede elaborar o construir sistemas computacionales que puedan mejorar con la experiencia de forma automática? y ¿cuáles son las leyes teóricas que rigen todo sistema de aprendizaje? (Mitchell, 1997).

Los algoritmos de *ML* se pueden clasificar en:

- aprendizaje supervisado, donde el producto del algoritmo es una función que correlaciona las entradas y salidas esperadas en el sistema, y
- aprendizaje no supervisado, donde el modelado se crea a partir de las entradas al sistema.

En las siguientes secciones se describirán de forma somera los algoritmos de aprendizaje no supervisado aplicados en este trabajo: k-medias, k-medianas, k-medoides y mapas auto-organizados.

2.6.1. Métodos de agrupamiento

El aprendizaje no supervisado, dentro del contexto de métodos basados en distancia, se centra en herramientas de agrupamiento. Estos algoritmos usan la distancia como regla para tomar decisiones.

Ante la ausencia de una variable objetivo explícita, se asume que la distancia codifica indirectamente el objetivo del entrenamiento, por lo que buscamos encontrar grupos o cúmulos como elementos de salida (Mitchell, 1997).

k-medias y k-medianas

Estas herramientas no tienen una metodología para encontrar el mínimo global, por lo que es necesario recurrir a un algoritmo heurístico (Flach, 2012).

El algoritmo itera agrupando los datos usando una métrica de distancia (generalmente la *distancia euclidiana*) para decidir cuál es el centroide más cercano a cada vector. El número de centroides será igual al número de grupos en los que se desee agrupar la base de datos. Se calcula la nueva posición de cada centroide con base en el valor de la media o el valor de la mediana (según el método, k-medias o k-medianas respectivamente) de todos los vectores que pertenezcan al centroide en cuestión.

En estos algoritmos no hay garantía de que los puntos de convergencia sean el mínimo global ni de qué tan lejos se encuentra este.

En la práctica es recomendable ejecutar el algoritmo un cierto número de veces, de modo que se pueda elegir la mejor distribución de centroides. Existen varias métricas que ayudan a decidir el número de grupos existentes en la base de datos estudiada, algunas de estas son los índices de *Davies-Bouldin* y de *Silhouette*.

Índice de Davies-Bouldin Este índice fue introducido por David L. Davies y Donald W. Bouldin en 1979 como una métrica para evaluar algoritmos de agrupamiento (Halkidi et al., 2001). Sea $R_{i,j}$ una métrica de qué tan bueno es un agrupamiento; debe considerar la separación entre los grupos ($M_{i,j}$), que es deseablemente grande, y también considerar la dispersión S dentro de cada grupo, que es deseablemente baja. Una solución para $R_{i,j}$ es:

$$R_{i,j} = \frac{S_i + S_j}{M_{i,j}} \tag{2.19}$$

a partir de la cual se puede definir el *índice de Davies-Bouldin* (DB) para un número N de grupos:

$$D_i \equiv \max_{j \neq i} R_{i,j} \tag{2.20}$$

$$DB = \frac{1}{N} \sum_{i=1}^{N} D_i \tag{2.21}$$

Este índice es siempre positivo, un valor bajo equivale a que la agrupación es mejor.

Índice de *Silhouette* Es un método de validación usado en algoritmos de agrupamiento basado en qué tan similar es un vector a su centroide en relación con el resto de los centroides. El rango de este índice es $[-1, 1]$, donde un valor alto (positivo) implica que los vectores están bien asignados a su propio centroide y mal asignados al resto de ellos (Rousseeuw, 1987). Sea entonces a_i una representación de lo bien que el vector i está asignado a su centroide:

$$a_i = \frac{1}{|C_I| - 1} \sum_{j \in C_I, i \neq j} d(i, j) \tag{2.22}$$

donde $d(i, j)$ es la distancia entre los vectores i y j dentro del centroide C_I. a_i puede verse como una medida de qué tan bien está asignado i a su cúmulo C_I. Por otro lado:

$$b_i = \min_{k \neq i} \frac{1}{|C_K|} \sum_{j \in C_K} d(i, j) \tag{2.23}$$

el centroide con valor b_i más pequeño es el centroide vecino de i. El *índice de Silhouette* es entonces:

$$s_i = \frac{b_i - a_i}{\max\{a_i, b_i\}} \tag{2.24}$$

k-medoides

Es una adaptación del algoritmo descrito en la sección anterior, en el que cada centroide estará definido necesariamente como un vector existente en la base de datos

estudiada. Este algoritmo requiere analizar la base de datos por parejas de vectores, lo que puede ser muy costoso para bases de datos grandes.

En lugar de minimizar la *distancia euclidiana*, como lo harían los algoritmos de k-medias y k-medianas[3], se busca minimizar la suma de *disimilitudes* entre pares de vectores (Flach, 2012), usando por ejemplo la *distancia Manhattan*.

2.6.2. Mapas auto-organizados

Los *mapas auto-organizados* (*SOM* por sus siglas en inglés) son un tipo de red neuronal artificial que fueron introducidos por *Teuvo Kohonen* en 1986. Los *SOM* son una técnica de aprendizaje no supervisado usada para representar datos de alta dimensionalidad a una dimensionalidad más baja (generalmente bidimensional) preservando la estructura topológica de estos (Kohonen, 1990). Esta herramienta es usada para encontrar estructuras y patrones en bases de datos, así también como técnica de reducción de dimensionalidad ayudando a la visualización de bases de datos complejas (Temprano, 2020).

Un mapa consiste de componentes denominados *nodos* o *neuronas* que están dispuestos en una malla que es típicamente rectangular o hexagonal. El número de *neuronas* y la topología del mapa dependerán del objetivo de estudio. A cada *neurona* se le asigna un vector de pesos que representa la posición de esta en el espacio de características, el entrenamiento consiste en ajustar los pesos de cada *neurona* de forma en que se minimice el *error de cuantización* y el error *topológico*. Los pesos de cada *neurona* pueden inicializarse de forma aleatoria o usando las dos componentes principales obtenidas de la ejecución de un *PCA* (2.4.2). La *neurona* cuyo vector de pesos sea el más similar a un vector de entrada se le denomina *neurona ganadora* (*BMU* por sus siglas en inglés). Los pesos de la *BMU* y de su vecindad se ajustan de acuerdo al vector de entrada. La fórmula de actualización de los pesos $W_v(s)$ de una *neurona* v es:

[3]Estos algoritmos no están restringidos a trabajar usando la *distancia euclidiana*, sin embargo es la más usada.

$$W_v(s+1) = W_v(s) + \theta(u, v, s)\alpha(s)(D(t) - W_v(s)) \tag{2.25}$$

donde s es el la iteración, t es el índice de la muestra del entrenamiento, u es el índice de la *BMU* para el vector de entrada $D(t)$, $\alpha(s)$ es el factor de aprendizaje (que decrece monotónicamente respecto a s) y $\theta(u, v, s)$ es la función de vecindad en términos de la *BMU*. Se espera que la función de vecindad sea más pequeña con cada iteración, generalmente se usa una *función Gaussiana* o una *ondícula de Ricker* (también conocida como *Mexican-hat*).

Una forma de visualizar un *SOM* es mediante la *distancia euclidiana* que hay entre los pesos de las neuronas dentro de su vecindario más próximo, a esta imagen se le nombra *matriz de distancias unificada* (*U-matrix*) (Ultsch, 1990). Mediante esta representación se pueden visualizar las regiones en donde se encuentran más concentradas las neuronas, pudiendo ser consideradas como *cúmulos*, así como las regiones que definen los límites entre dichos grupos.

Finalmente, los errores se definen como:

- error de cuantización, es el promedio de las distancias entre cada vector usado para entrenar y su respectiva *BMU*;
- error topológico, también conocido como error topográfico, es el promedio de la relación de vecindad que existe entre la *BMU* de cada vector usado para entrenar y de su *segunda neurona ganadora*, es una medida de *suavidad* usada para evaluar la regularidad del mapa.

2.7. Prospección geofísica

Dentro del sentido más amplio, la *geofísica* es la aplicación de la física para estudiar el interior de la Tierra, desde la superficie hasta el núcleo interior (Reynolds, 2011), aunque a veces también se refiere al estudio de la luna, de otros planetas y del espacio que los separa, así como de la atmósfera.

La *geofísica aplicada*, según Sheriff (2002), es la aplicación e interpretación de mediciones de propiedades físicas de la Tierra para determinar las condiciones del subsuelo, usualmente con objetivos económicos o con aplicaciones ingenieriles, así como sus implicaciones con su ambiente local. Algunas de sus aplicaciones pueden ser la exploración de agua subterranea, de minerales u otros recursos económicos, el mapeo de vestigios arqueológicos o para localizar tuberias o cables enterrados.

Los *métodos geofísicos* responden a los contrastes de las propiedades físicas de los materiales que se encuentran en el subsuelo. La geofísica aplicada ofrece una gran cantidad de herramientas que, usadas correctamente y en las situaciones adecuadas, brindarán información muy útil.

Ningún método geofísico concluye con una solución única a una situación particular (Reynolds, 2011). Es de suma importancia que los *datos geofísicos* sean interpretados dentro de un marco que esté delimitado por las características físicas de la zona estudiada. De forma prácticamente unánime se apoya la idea de que datos geofísicos complementarios, producto de la aplicación de múltiples métodos, facilita la detección del objetivo ayudando a caracterizar la región de estudio (Butler et al., 2006).

Un concepto de suma importancia es el de *anomalía*, según Sheriff (2002) puede definirse como:

- una desviación de uniformidad en las propiedades físicas, como una perturbación en un campo fuera de lo normal, de lo uniforme o de lo predecible;
- el resultado de un valor observado menos su respectivo valor teórico;
- una porción de un levantamiento geofísico que es diferente en apariencia del levantamiento en general;
- una desviación que es de interés para la exploración, como una característica que puede ser asociada a un objetivo económico.

De forma simple, de acuerdo con Carrasco et al. (2021), puede decirse en pocas palabras que una anomalía sucede cuando los datos corregidos no siguen la respuesta

esperada para un modelo ideal que se asume para la Tierra.

En las siguientes subsecciones se describirán brevemente las técnicas de prospección magnetométrica y de prospección eléctrica, métodos de los que provienen las bases de datos que son usados en esta tesis, así como un resúmen del estado del arte de la aplicación de técnicas de aprendizaje computacional a datos geofísicos.

2.7.1. Magnetometría

El método magnetométrico es una técnica de exploración geofísica que consiste en medir las variaciones del campo magnético terrestre para inferir las estructuras magnetizables presentes en el subsuelo (Blakely, 1996). Los levantamientos magnetométricos consisten en realizar mediciones de campo magnético sobre la superficie terrestre con dispositivos conocidos como *magnetómetros*. El registro del campo magnético se atribuye a un punto en superficie, pudiendo entonces generar perfiles (curvas) o mallas (imágenes) a partir de la adquisición en estaciones con ubicaciones controladas.

Los métodos gravimétricos y magnetométricos tienen mucho en común, sin embargo tienen diferencias sustanciales, por ejemplo un mapa de gravedad usualmente está dominado por efectos regionales, mientras que un mapa de campo magnético generalmente muestra una multitud de anomaías locales. La magnetometría es la técnica de prospección geofísica más versátil, sin embargo carece de unicidad en la interpretación de sus datos (Murray et al., 1990).

El campo magnético terrestre se compone principalmente por tres elementos:

- el campo principal, de origen interno y variación lenta;
- un campo externo de menor magnitud que el campo principal y con variación temporal importante; y
- anomalías locales superficiales de pequeña magnitud que son prácticamente constantes en el tiempo, son el objetivo de la prospección magnetométrica.

Los *magnetómetros de campo total* miden la magnitud del campo magnético, incluyendo las contribuciones de los campos internos y externos. Para estudiar el efecto de las variaciones locales es necesario eliminar los efectos producidos por el campo magnético principal y los campos magnéticos externos, al campo resultante se le conoce como *anomalía de campo total*:

$$\Delta T = F_{obs} - \delta f - F_{IGRF} \tag{2.26}$$

donde F_{obs} es la lectura del *magnetómetro*, δf es la contribución del campo magnético externo y F_{IGRF} es la contribución del campo magnético principal.

Para remover la influencia magnética que tiene el sol sobre la ionósfera (campo externo) se debe monitorear al campo magnético de forma continua en una localidad fija. Para remover la influencia del campo geomagnético (campo principal) se calcula su valor en una posición y momento específicos mediante un modelo matemático definido por el *campo geomagnético de referencia internacional* (*IGRF* por sus siglas en inglés) (IAGA, 2019).

Existen otros instrumentos similares a los *magnetómetros* en los que se puede medir de forma simultánea la magnitud del campo magnético en dos posiciones espaciales diferentes, generalmente en sentido vertical, a estos aparatos se les conoce como *gradiómetros*.

En cuanto a la interpretación, comunmente se hace uso de filtros de *realce de anomalías* descritos en la literatura, ejemplos de ellos son:

- **reducción al polo**, su objetivo es cambiar matemáticamente el ángulo de inclinación del campo geomagnético, de forma que las anomalías se visualicen como producidas por una magnetización vertical; su función de transferencia es:

$$\mathcal{F}\{\Psi_r\} = \frac{|k|^2}{a_1 k_x^2 + a_2 k_y^2 + i|k|(b_1 k_x + b_2 k_y)} \tag{2.27}$$

- **gradiente vertical**, su objetivo es definir la tasa de cambio del campo magnético en sentido vertical, a partir de las lecturas de un gradiómetro se calcula como:

$$GV = \frac{\Delta T_{inf} - \Delta T_{sup}}{\Delta z} \tag{2.28}$$

- **gradiente horizontal**, se define como el módulo de las derivadas en dirección x y en dirección y:

$$GH = \sqrt{\left(\frac{\partial \Delta T}{\partial x}\right)^2 + \left(\frac{\partial \Delta T}{\partial y}\right)^2} \tag{2.29}$$

- **señal analítica**, que puede verse como una combinación de los gradientes vertical y horizontal, se calcula mediante la *transformada de Hilbert* de ΔT (ecuación 2.26);

- **derivada inclinada (*tilt derivative*)**, usada para mapear estructuras someras:

$$TDR = \tan^{-1}\left(\frac{GV}{GH}\right) \tag{2.30}$$

- **gradiente horizontal de la derivada inclinada**, equivalente al gradiente horizontal:

$$GH_{TDR} = \sqrt{\left(\frac{dTDR}{dx}\right)^2 + \left(\frac{dTDR}{dy}\right)^2} \tag{2.31}$$

La interpretación consiste entonces en analizar las imágenes obtenidas de la aplicación de los filtros anteriores. Debido a las características complejas de los mapas magnéticos, regularmente se realizan interpretaciones cualitativas. Algunos autores llaman a la interpretación como un arte fino (Murray et al., 1990). Un intérprete experimentado puede encontrar anomalías observando *patrones* en los mapas magnetométricos. La interpretación se realiza cotejando los distintos mapas de realce, comparando con modelos directos, o inclusive realizando una inversión de datos (Grant, 1972).

En cuanto a la geometría de los levantamientos, pueden interpretarse como perfiles individuales (curvas en dos dimensiones, magnitud del campo vs. distancia) o

como mapas (imágenes en tres dimensiones, dos coordenadas espaciales vs. magnitud del campo en escala de color). Los mapas se generan a partir de la interpolación de las estaciones levantadas, esperando que inicialmente se tenga una buena densidad de datos.

Por otro lado, con el fin de controlar la calidad de los datos, durante la adquisición se toman las siguientes medidas:

- De forma automática el equipo realiza el promedio de varias lecturas tomadas de forma consecutiva (técnica denominada apilamiento o *stacking*), a partir de estas lecturas se calcula la *calidad* del dato; es buena práctica el mantener la calidad lo más cercana al valor máximo (comunmente es 99).
- Aunado al punto anterior, pueden adquirirse manualmente un número n de lecturas en el mismo lugar con el fin de aplicar un apilamiento adicional.
- Los operadores deben estar libres de objetos metálicos que puedan introducir ruido a las lecturas de campo magnético.
- Deben evitarse fuentes de ruido externo, tales como lineas de tensión, rejas, tuberias, etc.

Casos de estudio, donde se puede comprobar la utilidad de esta técnica en la arqueología, pueden consultarse en los trabajos de Fassbinder (2017), Black and Johnston (1962), Gaffney (2008) y Juárez et al. (2017).

2.7.2. Geoeléctrica

El propósito de los levantamientos geoeléctricos (también conocidos como métodos de corriente continua) es determinar la distribución de *resistividades eléctricas* del subsuelo mediante mediciones efectuadas en superficie (Loke, 2004). La resistividad de un material depende de varios efectos, como puede ser el contenido mineralógico, porosidad, grado de saturación de agua, etc. Los métodos dentro de esta prospección se pueden clasificar en métodos de campo natural o métodos de fuente controlada, siendo estos últimos en los que se hará énfasis.

La ecuación que relaciona la resistividad eléctrica ρ con la diferencia de potencial $\Delta\phi$ entre dos puntos, producida por el *flujo* de una corriente eléctrica estacionaria I es:

$$\rho = k\frac{\Delta\phi}{I} \tag{2.32}$$

con k definido como:

$$k = \frac{2\pi}{\frac{1}{r_{AM}} - \frac{1}{r_{BM}} - \frac{1}{r_{AN}} + \frac{1}{r_{BN}}} \tag{2.33}$$

donde r es la distancia entre un par dado de *electrodos*. Un *electrodo* es una pieza de material conductor que es usado para hacer contacto con la superficie terrestre (Sheriff, 2002). Estos *electrodos* pueden tomar dos roles:

- *electrodo de corriente*, que se usará para generar el campo eléctrico controlado; y
- *electrodo de potencial*, usado para registrar el potencial primario generado por el campo eléctrico generado.

Al par de electrodos usados para *introducir* corriente eléctrica se les etiqueta como A y B, mientras que al par de electrodos usados para medir la diferencia de potencial se les etiqueta como M y N[4].

Existen muchas formas de crear arreglos usando un dispositivo de cuatro *electrodos*, conocidos mejor como *arreglos electródicos*, cada uno con una sensibilidad diferente ante cambios en las distribuciones de las propiedades eléctricas del subsuelo. Dos de los arreglos electródicos más comunes son:

- **dipolo-dipolo** axial, arreglo donde un dipolo envía corriente eléctrica y otro mide la diferencia de potencial, la separación entre pares de electrodos es mayor que la distancia entre los electrodos de cada pareja (figura 2.2);

[4]A veces también se les llama C_1 y C_2 a los electrodos de corriente y P_1 y P_2 a los electrodos de potencial.

Figura 2.2: Arreglo electródico dipolo-dipolo axial.

Figura 2.3: Arreglo electródico Wenner-Schlumberger axial.

- **Wenner-Schlumberger** axial, arreglo donde el dipolo de potencial se situa entre el dipolo de corriente, la distancia entre los electrodos puede ser la misma o la apertura del dipolo de corriente puede ser n veces mayor que la apertura del dipolo de potencial, donde n es un número entero (figura 2.3).

Una de las metodologías más prácticas en la prospección geoeléctrica es la *tomografía de resistividad eléctrica en 2 dimensiones* (*TRE2D*), donde se obtiene resolución tanto horizontal como a profundidad desplazando el arreglo electródico a través de una línea e incrementando la distancia entre los dipolos. Los resultados obtenidos con dos arreglos electródicos distintos pueden no ser iguales, ya que cada arreglo presenta una sensibilidad diferente, para mayor información se puede consultar el trabajo de Okpoli (2013). La geometría de los puntos de adquisición de una *TRE2D* (también llamados *puntos de atribución*) tienden a presentar la forma de un triángulo o de un trapecio invertido (figura 2.4); a cada *fila* de puntos, aquellos que tienen la misma pseudo-profundidad, se le conoce como *nivel*.

Los valores registrados en un levantamiento geoeléctrico tienden a tener un rango dinámico muy amplio, llegando a veces a saltos de 3 o más órdenes de magnitud. Esta característica se debe tanto al efecto superficial del contacto electrodo-suelo como a los contrastes de resistividad entre estructuras geoeléctricamente[5] muy diferentes.

Comunmente la interpretación de una *TRE2D* se realiza usando una técnica denominada como *inversión*, que consiste en encontrar el modelo geoeléctrico del

[5]Refiriéndome a las propiedades eléctricas de los materiales presentes en el subsuelo.

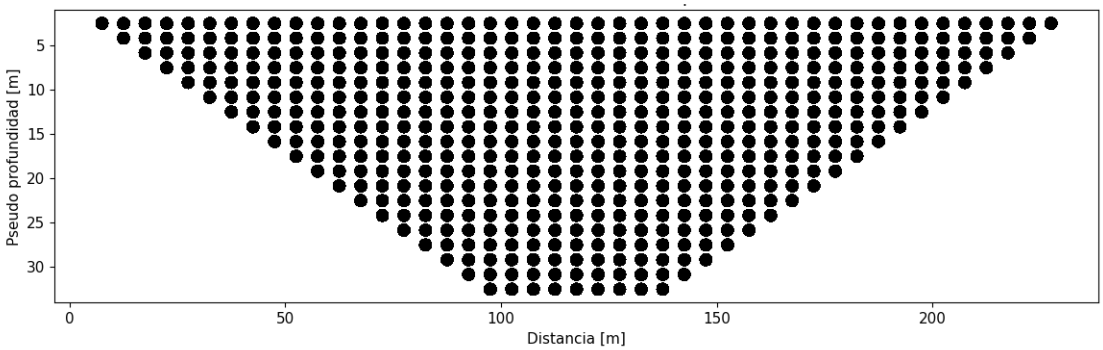


Figura 2.4: Ejemplo de puntos de atribución en una *TRE2D*, nótese la pérdida de densidad de datos a profundidad.

subsuelo (distribución de resistividades) que mejor explique los datos adquiridos. Es un proceso iterativo que se basa en la obtención de un número finito de modelos directos calculados analíticamente (Reynolds, 2011).

Respecto al control de calidad de los datos, durante la adquisición se toman las siguientes medidas:

- De forma automática el equipo realiza el promedio de varias lecturas tomadas de forma consecutiva (técnica denominada apilamiento o *stacking*), a partir de estas lecturas se calcula un error; es buena práctica mantener el error lo más bajo posible[6].
- Para mantener el error bajo, además de incrementar el número de lecturas para el apilamiento, se disminuye la resistencia de contacto entre los electrodos y el medio; esto se logra humedeciendo esta interfaz con una solución conductora.

Se pueden consultar casos de estudio, donde queda comprobada la utilidad de este técnica, en los siguientes trabajos:

- geohidrología: Redhaounia et al. (2016), Márquez et al. (2001);
- geotecnia: Lamsal et al. (2020), Arango-Galván et al. (2011).

[6]Esto dependerá de las características físicas del suelo y del grado de acoplamiento del dispositivo.

2.7.3. Inteligencia artificial aplicada a datos geofísicos

Como se ha descrito anteriormente, la geofísica tiene como uno de sus objetivos analizar múltiples imágenes que representan la distribución de variables físicas en un dominio definido. El objetivo es realizar una interpretación de las estructuras que son fuente de las variables registradas.

Las interpretaciones son cualitativas, realizadas sobre imágenes que son producto de la aplicación de herramientas matemáticas con el objetivo de realzar *anomalías* o de aproximar la distribución real de la variable en el subsuelo. Las conclusiones realizadas a partir de estas interpretaciones dependen enteramente de la experiencia del intérprete y de su capacidad de analizar de forma visual datos multidimensionales. Este proceso es repetitivo y complicado, las estructuras de interés presentan firmas diferentes de acuerdo con la física involucrada y de la escala de los levantamientos, por lo que el intérprete debe evaluar la existencia de esos patrones sujetos a distintos ambientes de estudio.

Se han utilizado técnicas de aprendizaje computacional para automatizar la búsqueda de patrones en imágenes satelitales (Bedini (2009), Ghimire et al. (2010), Goncalves et al. (2008), Huang and Shibuya (2020), Waske and Braun (2009), Yu et al. (2012)). Las herramientas más usadas son *bosques aleatorios*, *máquinas de soporte vectorial* y *mapas auto-organizados*. La escala de estos estudios es muy grande, las estructuras buscadas, generalmente estructuras geológicas regionales, tienen dimensiones de kilómetros y son muy sensibles a la posición (georreferencia de estas estructuras).

Basados en los trabajos anteriores, otros autores han mezclado datos de imágenes satelitales con levantamientos aeromagnéticos (Bachri et al. (2020), Carneiro et al. (2012), Chen et al. (2020), Costa et al. (2019), Cracknell et al. (2014), Cracknell and Reading (2013), Harris and Grunsky (2015), Kuhn et al. (2018), Kuhn et al. (2019), Nathan et al. (2020)), aplicando de forma general las mismas herramientas de aprendizaje computacional. Su principal objetivo es entrenar un sistema de forma supervisada para que pueda predecir litología, con vista para ser aplicado a zonas de

difícil acceso para trabajo geológico de campo.

Hay muchos trabajos en donde se aplican técnicas de aprendizaje computacional a datos sísmicos, entre ellos el trabajo de Molino-Minero-Re et al. (2018) donde se aplican *mapas auto-organizados* para clasificar datos sísmicos a partir de algunos de sus atributos.

Por otro lado existen algunas paqueterías que ofrecen un pre-procesamiento, por ejemplo filtrado e interpolación de datos, usando técnicas de aprendizaje computacional (Uieda, 2018).

Capítulo 3

Materiales y métodos

3.1. Introducción

En el capítulo anterior se revisaron las generalidades sobre la teoría en la que se sustentan los métodos usados en este trabajo. La figura 3.1 muestra el flujo de la metodología para esta tesis.

Este desarrollo se divide en cuatro etapas principales:

1. Pre-procesamiento *típico*: En este punto se aplica el procesamiento común aplicado a cada tipo de base de datos (capítulo 2.7), desde correcciones hasta realce de anomalías o inversión de datos.

2. Pre-procesamiento *adaptativo*: En esta etapa se aplican tanto el filtrado realizado en el dominio de la *DWT*, así como el retiro automático de vectores atípicos usando *IF* y *LOF*. Se reciben los datos crudos con un formato específico (sección 4), teniendo a la salida un arreglo con los datos filtrados conservando el mismo formato de entrada.

3. Entrenamiento no supervisado: Aquí se aplican los algoritmos de entrenamiento: *k-medias*, *k-medianas* y *SOM*. La entrada son bases de datos estructuradas, producto de un proceso intermedio de preparación, a la salida se obtienen los centroides o neuronas con sus parámetros y pesos respectivos.

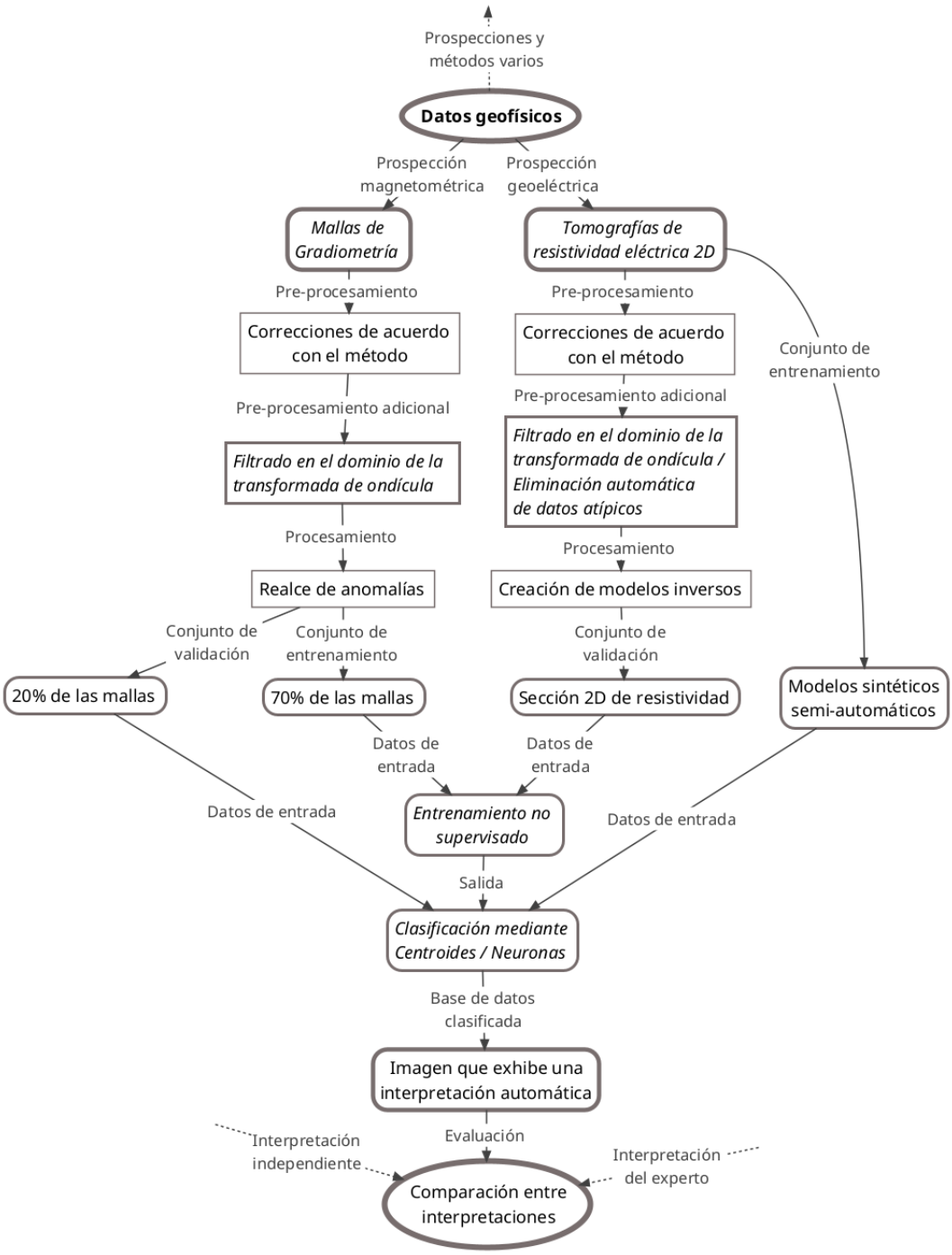


Figura 3.1: Esquema del procesamiento global realizado.

4. Elaboración de la imagen del conjunto de datos agrupado: creada a partir de aplicar el algoritmo de aprendizaje en un conjunto de datos desconocido por el sistema diseñado, obteniendo así una imagen que exhibe una interpretación asistida por computadora.

En las siguientes secciones se describirá el procedimiento de cada una de las etapas mencionadas anteriormente.

3.2. Pre-procesamiento *adaptativo*

Las características particulares de cada base de datos condiciona que la metodología de filtrado sea diferente para cada caso.

La adquisición de datos de *TRE2D* tradicional se realiza de dos formas posibles: en diagonal o por nivel, resultando en pérdida de densidad de datos a profundidad (sección 2.7.2). La metodología implementada es un filtrado unidimensional aplicado a cada *nivel* de la *TRE2D*.

Los datos de una malla de magnetometría muestran una estructura bidimensional (sección 2.7.1). Por esta razón la metodología empleada para el pre-proceso *adaptativo* se basa en el filtrado de imágenes.

Los resultados obtenidos de este pre-procesamiento se pueden consultar en los anexos A y B.

3.2.1. Filtrado de datos de *TRE2D*

El flujo del procesamiento aplicado se muestra en la figura 3.2. Las etapas del desarrollo implementado pueden resumirse en dos bloques principales: inicialización de los algoritmos de filtrado y la aplicación de estos.

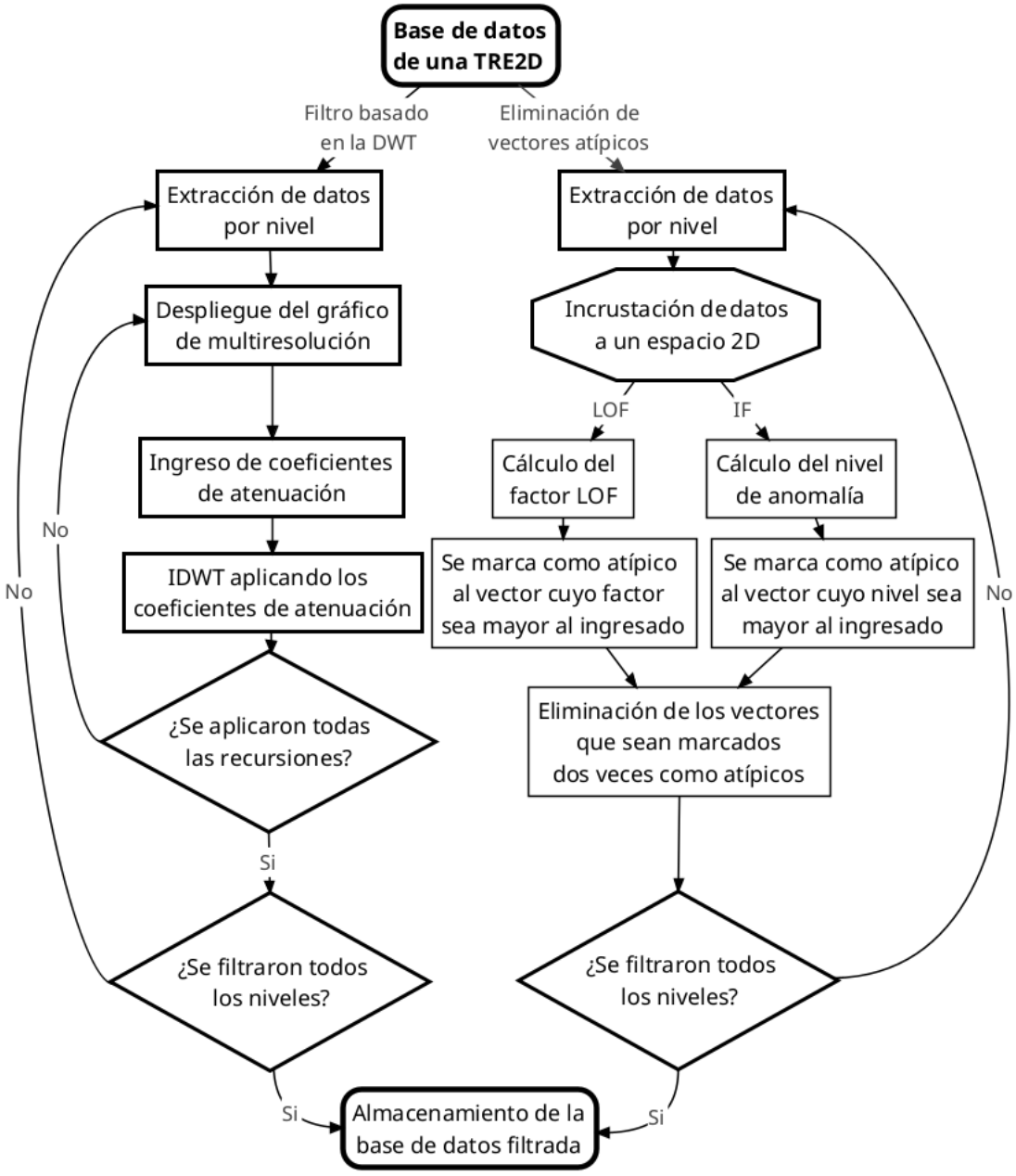


Figura 3.2: Esquema del proceso de filtrado aplicado a los datos de *TRE2D*.

Inicialización de los algoritmos

Los datos de entrada deben tener una configuración matricial con dimensiones $n \times m$, donde n es el número de vectores y m es el número de características de cada vector. Es necesario que las primeras dos columnas correspondan a las siguientes variables:

1. Columna 1: el *nivel* del *punto de atribución*. Se comienza con 1 para el *nivel* más somero, se espera que los vectores estén ordenados por nivel.

2. Columna 2: el cadenamiento del punto de atribución (coordenada en x).

Se espera que el archivo de entrada tenga un encabezado con títulos indistintos y con valores que se encuentren separados por comas.

Análisis en multiresolución 1D

Esta etapa usa la biblioteca *PyWavelets*, implementada por Lee et al. (2019). Los parámetros de entrada son:

1. Ondícula a usar: el identificador de la ondícula para la DWT y la $IDWT$. Estos identificadores son los implementados en la biblioteca *PyWavelets*.

2. Número de descomposiciones: son los niveles en los que se analizará la señal.

3. Número de recursiones: número de veces que se filtrará cada nivel de la *TRE2D*.

Ondícula: Puede usarse cualquier ondícula incorporada en la biblioteca *PyWavelets*, quedando como elección del usuario la familia y el orden de esta.

Número de descomposiciones: Es la cantidad de niveles de análisis, para cada uno de estos el usuario eligirá un factor de atenuación para la respectiva componente de detalle, buscando con esto disminuir la amplitud de componentes no deseadas a distinta escala.

Número de recursiones: Como resultado del filtrado se espera una señal que siga una tendencia deseada o esperada, quedando a criterio del analista si se ha conseguido dicha característica. Con el fin de aumentar la versatilidad del sistema se implementó la opción de aplicar un proceso recursivo, donde la señal se filtra n veces.

El resumen del proceso es el siguiente:

1. Extracción de los datos por nivel y aislamiento de la variable a filtrar. De forma que se obtenga una señal unidimensional lista para entrar al proceso de la *DWT*.

2. Despliegue del gráfico de descomposición. Aplicada la *DWT*, de acuerdo al número de niveles deseado, se muestra en pantalla un gráfico con las componentes de aproximación y detalle para cada uno. Es aquí donde el usuario a su consideración elige los factores de atenuación para cada componente de detalle.

3. Aplicación del filtro. Aquí se multiplica cada coeficiente de las componentes de detalle por su respectivo factor ingresado en el paso anterior, para entonces aplicar la *IDWT*.

4. Se evalua si se han llevado a cabo todas las recursiones del filtrado. De no ser así se entra nuevamente en el ciclo (regreso al punto 2).

5. Se evalua si se han filtrado todos los niveles de la *TRE2D*, de no ser así se regresa al punto 1.

6. Almacenamiento de la base de datos. Se sustituye la columna de la variable original por los vectores filtrados, conservando a la salida el mismo formato de entrada.

A continuación se presenta un ejemplo de aplicación del filtro:

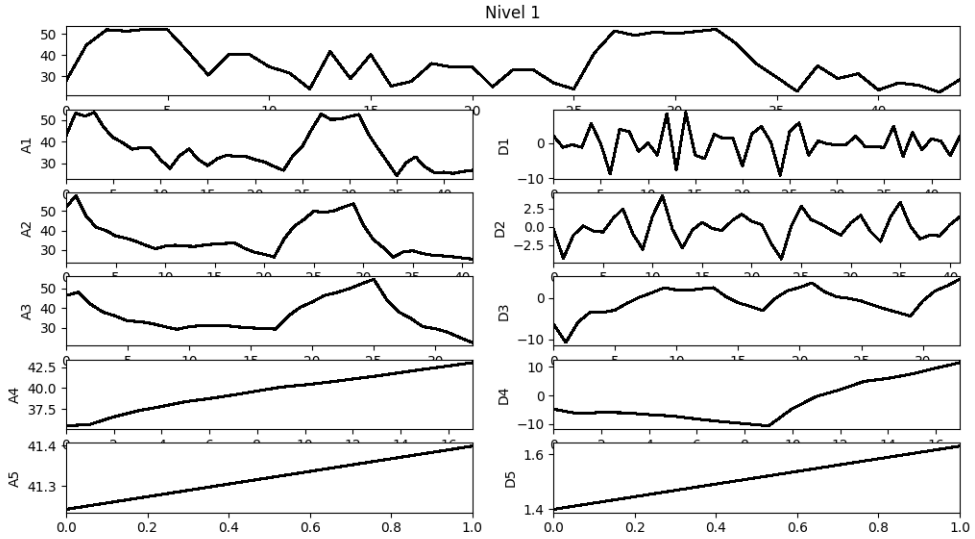


Figura 3.3: Descomposición de 5 niveles aplicada a un nivel de una *TRE2D*.

1. Parámetros de entrada:
   - Variable a filtrar: diferencia de potencial entre el dipolo de recepción.
   - Ondícula: *Daubechies* de segundo orden.
   - Descomposiciones: 5.
   - Recursiones: 1 y 2 recursiones con fines de comparación.
2. El gráfico de descomposición para el primer nivel de una *TRE2D* es el mostrado en la figura 3.3.
3. Los factores de atenuación elegidos son: `[0,0.5,1,1,1]`.

   Los resultados del filtrado se muestran en la figura 3.4.

Detección de vectores atípicos

Esta etapa usa la técnica de inmersión descrita en la sección 2.3.1, así como las herramientas de detección de anomalías revisadas en la sección 2.3. La

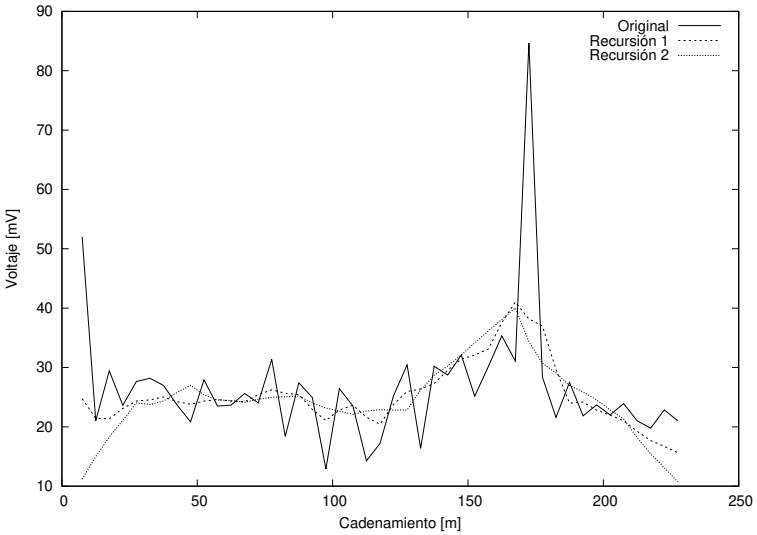


Figura 3.4: Comparación de filtrado usando diferente número de recursiones, para un nivel de una *TRE2D* usando la ondícula de *Daubechies* de segundo orden: 5 niveles de descomposición y factores de atenuación de $[0,0{,}5,1,1,1]$

.

inmersión tiene por objetivo evaluar el comportamiento de la señal espacial considerando los cambios que presenta en un punto siguiente.

Los algoritmos *IF* y *LOF* aplicados son los implementados en la biblioteca *Scikit-learn* (Pedregosa et al., 2011).

Los parámetros de entrada son:

- Aplicación de logaritmo base 10 a la variable de interés: variable booleana para decidir la aplicación de una transformación logarítmica (sección 2.5.1) al conjunto de datos.
- Número de vecinos para el método *LOF*.
- Umbral de selección para valor atípico.
- Método: identificador *LOF* o *IF*.

Aplicación de una transformación logarítmica: Como se revisó en la subsección 2.7.2, estas bases de datos tienden a tener valores extremos en resistividad, por lo que la aplicación de un algoritmo de detección de anomalías puede dar resultados poco satisfactorios. Por esta razón se agrega la opción de aplicar un escalamiento logarítmico a la variable de interés.

Número de vecinos: El método *LOF* requiere de un tamaño de vecindad para el cálculo del *factor de anomalía*.

Umbral de selección: Es un valor límite para decidir si un vector es considerado como atípico o no.

Método: Donde la selección puede ser *IF* o *LOF*.

El resumen del proceso es el siguiente:

a) Extracción de los datos por nivel y aislamiento de la variable a filtrar. De forma que se obtenga una señal unidimensional lista para entrar al algoritmo de inmersión.

b) Inmersión de datos usando un punto vecino.

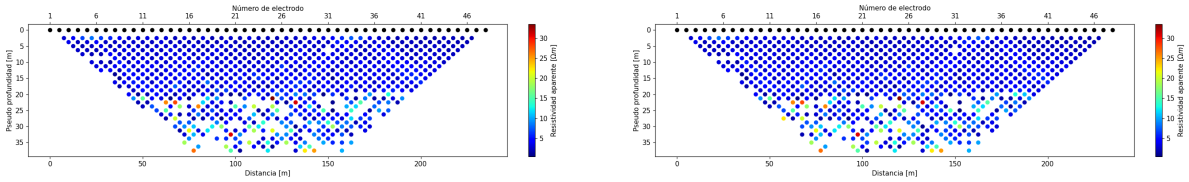
c) Para el algoritmo *LOF*:

- Cálculo del *factor de valor atípico* para cada vector.
- Se marca como vector atípico aquel cuyo factor calculado esté por encima del valor *umbral* ingresado por el usuario.

d) Para el algoritmo *IF*:

- Cálculo del *nivel de anomalía* para cada vector.
- Se marca como vector atípico aquel cuya puntuación sea mayor que el *umbral* ingresado por el usuario.

e) Se procede a la eliminación de vectores atípicos, considerando la técnica de inmersión usada se eliminarán los vectores que hayan sido marcados dos veces como atípicos, exceptuando las fronteras.



(a) Pseudosección filtrada usando *LOF* (b) Pseudosección filtrada usando *IF*

Figura 3.5: Comparación de la respuesta de los algoritmos *LOF* e *IF*.

f) Almacenamiento de la base de datos. Se eliminan de la base de datos original los vectores considerados como atípicos, conservando a la salida el mismo formato de entrada.

A continuación se presenta un ejemplo de aplicación del algoritmo:

a) Parámetros de entrada:

- Columna a analizar: resistividad aparente.
- Aplicación de logaritmo base 10: verdadero.
- Número de vecinos: 3.
- Umbral: 0,9.
- Método: *IF* y *LOF* para fines comparativos.

La figura 3.5 muestra la salida al aplicar los pasos anteriores.

3.2.2. Filtrado de datos de mallas de magnetometría

El flujo del procesamiento aplicado se muestra en la figura 3.6. El algoritmo tiene dos bloques principales, a partir de los cuales se generan tres posibles opciones de filtrado. Estos dos bloques son:

1. Filtro de una ventana. En este caso se extraen los datos de una ventana, porción de la malla original, a la que se le aplicará el proceso de filtrado.

2. Filtro de la malla completa. En este caso se aplicará el proceso de filtrado a toda la malla.

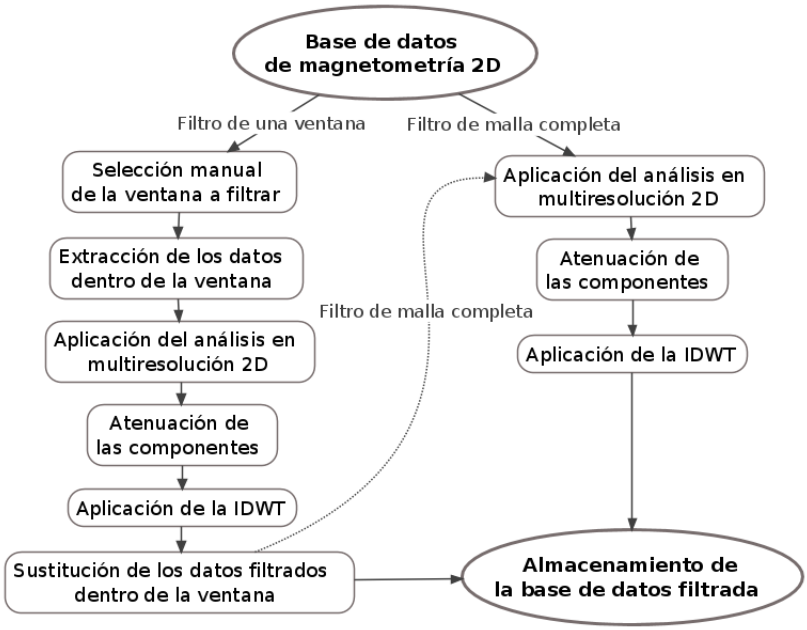


Figura 3.6: Esquema del proceso de filtrado aplicado a los datos de magnetometría 2D.

Las tres opciones de filtrado son entonces:

- Filtro de una ventana. Cuyo propósito es eliminar algún artefacto no deseado.
- Filtro de malla completa. Cuyo propósito es atenuar componentes no deseadas en toda la malla.
- Filtro de una ventana seguido del filtro de malla completa. Cuyo propósito es eliminar artefactos no deseados localizados en porciones específicas de la malla, para entonces atenuar componentes no deseadas en la malla completa considerando ahora el efecto de borde creado por el filtrado de la ventana.

Todas las opciones de filtrado implementadas usan la biblioteca *PyWavelets*, implementada por Lee et al. (2019).

Inicialización de los algoritmos

Los datos de entrada deben tener una configuración matricial con dimensiones $n \times m$, donde n es el número de vectores y m es el número de características. Se espera que las dos primeras columnas sean las coordenadas X, Y de cada estación de la malla. El archivo de entrada necesita un encabezado de títulos indistintos y requiere que sus valores se encuentren separados por comas.

Filtro de una ventana

Los parámetros de entrada son:

1. Columnas a filtrar: lista cuyos elementos especifican las variables que se desean filtrar. Todas serán filtradas de la misma forma.

2. Resolución_x y resolución_y: número de estaciones por linea y número de lineas, respectivamente. Se usan para interpolar los datos y generar una malla equiespaciada.

3. Ondícula: el identificador de la ondícula a usar para la *DWT2D*. Estos identificadores son de las ondículas incorporadas en la biblioteca *PyWavelets*.

4. Número de descomposiciones: son los niveles en los que se analizará la señal de acuerdo con la *DWT2D*.

5. Factores de atenuación: lista con los factores de atenuación para cada nivel de la *DWT2D*. El primer valor corresponde al peso de la componente de aproximación, los siguientes corresponden a las componentes de detalle para cada nivel de descomposición.

Columnas a filtrar: Como se mencionó en la subsección 2.7.1, las variables almacenadas en las bases de datos corresponden a los valores de campo magnético en dos sensores y la derivada que corresponde con la orientación de estos. Cada una de estas variables responde al mismo fenómeno físico, por lo que es intuitivo aplicar la misma técnica de filtrado a cada variable.

Factores de resolución: El *análisis en multiresolución 2D* aplicado está basado en las herramientas diseñadas para el filtrado de imágenes, por lo que a la entrada se espera una matriz de datos equiespaciados en ambas direcciones. Los factores solicitados juegan el papel de resolución espacial para la interpolación de los datos. Se recomienda entonces que estos valores sean iguales al número de estaciones registradas por línea y al número de líneas levantadas, de acuerdo a la orientación del levantamiento.

Ondícula Puede usarse cualquier ondícula incorporada la la biblioteca *PyWavelets*, quedando a elección del usuario la familia y el orden de esta.

Número de descomposiciones: Es la cantidad de niveles de análisis. Cada componente de detalle, así como la última componente de aproximación, es multiplicada por un factor de atenuación, buscando con esto disminuir la amplitud de componentes no deseadas a distinta escala.

Factores de atenuación: Son una lista cuyos valores son el peso por el que se multiplicará cada componente, comenzando por la última componente de aproximación seguida de las componentes de detalle en orden ascendente.

El resumen del proceso es el siguiente:

1. Interpolación de las variables solicitadas (columnas a filtrar) en una rejilla con coordenadas equiespaciadas.

2. Despliegue de la representación gráfica de la última variable ingresada solicitando dos coordenadas: las esquinas superior izquierda e inferior derecha de la ventana que se desea filtrar. Estas coordenadas se ingresan mediante dos *clicks* en la imagen desplegada.

3. Aislamiento de los datos dentro de la ventana definida y cambio de dimensiones de esta a la potencia de dos más cercana.

4. Aplicación de la *DWT2D* a cada variable.

5. Multiplicación de cada componente por su respectivo factor de atenuación.

6. Obtención de la *IDWT2D*.

7. Cambio del tamaño de la ventana a sus dimensiones originales y sustitución de sus nuevos valores.

8. Reestructuración matricial de la base de datos, donde las primeras dos columnas son las coordenadas, seguidas de las variables ya filtradas.

A continuación se presenta un ejemplo de aplicación del filtro:

1. Parámetros de entrada:

   - Columnas: campo magnético en sensor inferior, campo magnético en sensor superior, gradiente vertical.
   - Resolución en x: 10.
   - Resolución en y: 20.
   - Ondícula: *Daubechies* de cuarto orden.
   - Componentes: 3.
   - Pesos: `[0,0,0.1,0.3]`.

2. El despliegue de la representación gráfica del gradiente vertical se muestra en la figura 3.7, así como la ventana elegida.

La figura 3.8 muestra el resultado de la aplicación del filtro.

Filtro de malla completa

Los parámetros de entrada son los mismos que los del filtro de una ventana, presentando las mismas características.

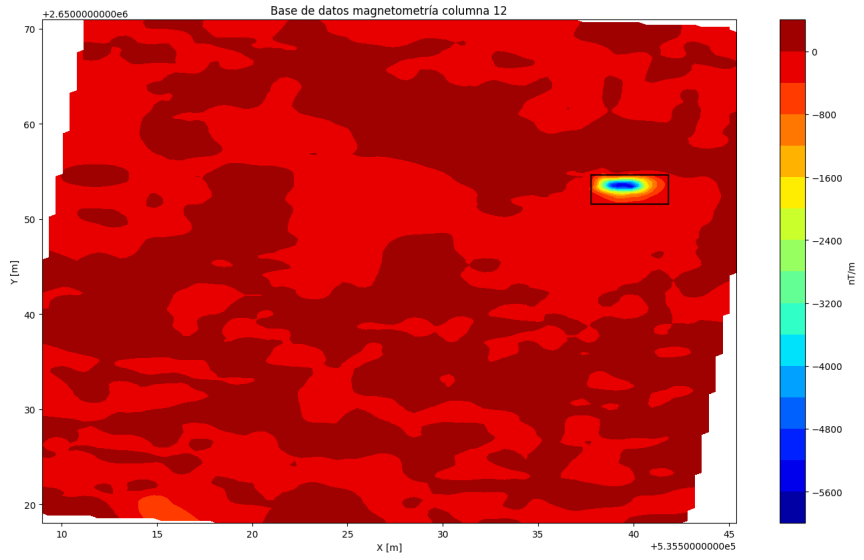


Figura 3.7: Gradiente vertical de una malla de magnetometría, se encierra en un rectángulo la ventana a filtrar.

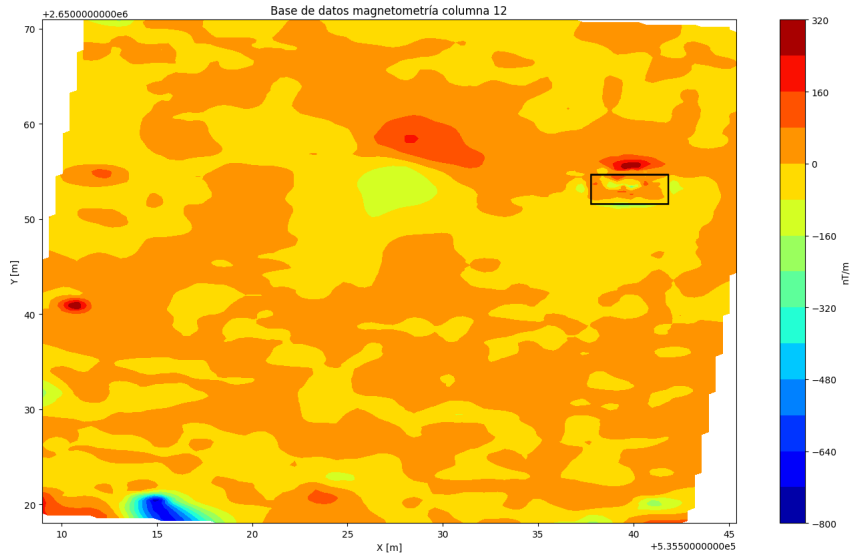


Figura 3.8: Gradiente vertical de una malla de magnetometría, se encierra en un rectángulo la ventana filtrada.

El resumen del proceso es el siguiente:

1. Interpolación de las variables solicitadas (columnas a filtrar) en una rejilla con coordenadas equiespaciadas.
2. Cambio de tamaño de la malla a la potencia de dos más cercana.
3. Aplicación de la *DWT2D* a cada variable solicitada.
4. Multiplicación de cada componente por su respectivo factor de atenuación.
5. Obtención de la *IDWT2D*.
6. Cambio del tamaño de la malla a sus dimensiones originales.
7. Reestructuración matricial de la base de datos, donde las primeras dos columnas son las coordenadas, seguidas de las variables ya filtradas.
8. Almacenamiento de la base de datos.

A continuación se presenta un ejemplo de aplicación del filtro:

1. Parámetros de entrada:
   - Columnas: campo magnético en sensor inferior, campo magnético en sensor superior, gradiente vertical.
   - Resolución en x: 10.
   - Resolución en y: 20.
   - Ondícula: *Daubechies* de cuarto orden.
   - Componentes: 3.
   - Pesos: `[1,0.1,0.5,0.7]`.

La figura 3.9 muestra el resultado de la aplicación del filtro.

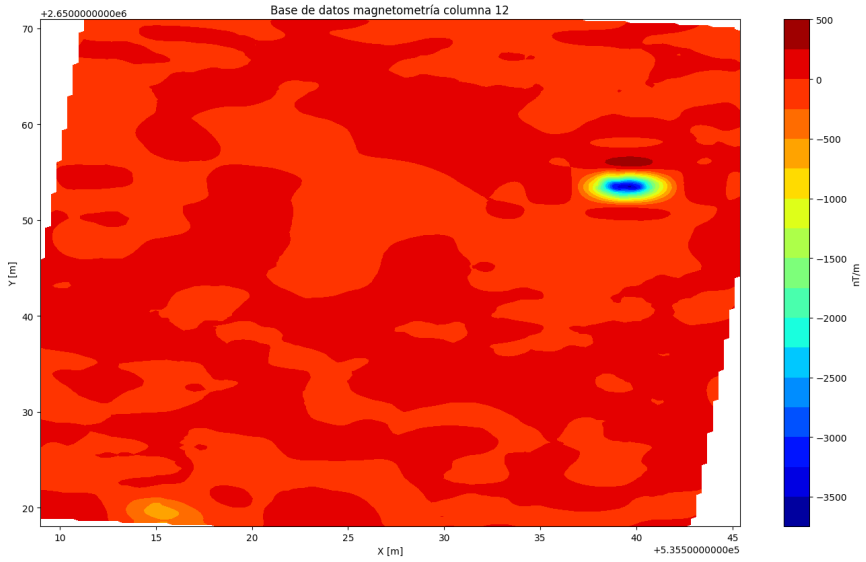


Figura 3.9: Gradiente vertical filtrado de una malla completa de magnetometría.

Filtro de una ventana y de malla completa

Este filtro consiste en la aplicación consecutiva del filtro de una ventana seguido del filtro de malla completa. A continuación se presenta un ejemplo de aplicación:

1. Parámetros de entrada:
    - Columnas: campo magnético en sensor inferior, campo magnético en sensor superior, gradiente vertical.
    - Resolución en x: 10.
    - Resolución en y: 20.
    - Ondícula: *Daubechies* de cuarto orden.
    - Componentes: 3.
    - Pesos del primer filtro: `[0,0,0.1,0.3]`.
    - Pesos del segundo filtro: `[1,0.1,0.5,0.7]`.

La figura 3.10 muestra el resultado de la aplicación del filtro.

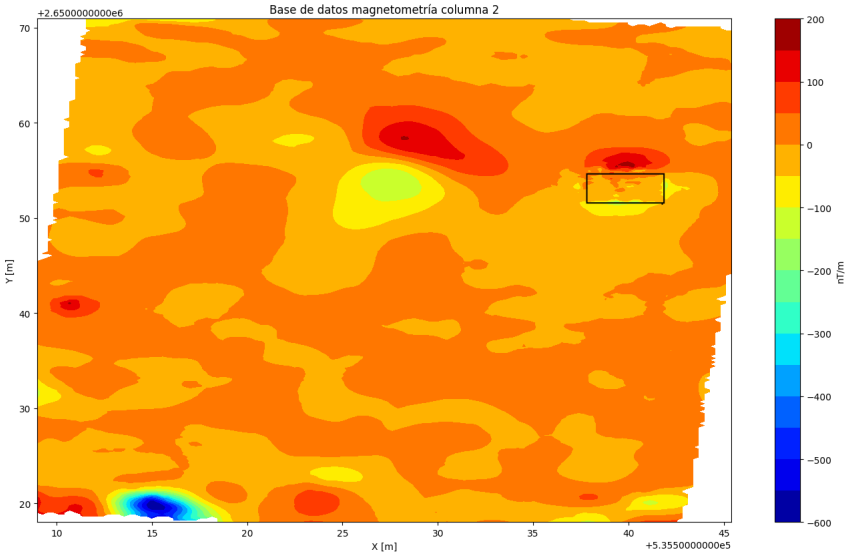


Figura 3.10: Gradiente vertical filtrado de una malla de magnetometría usando el filtro completo (ventana y malla completa).

3.3. Aprendizaje computacional

Como se mencionó en la sección 2.7, las bases de datos analizadas en este trabajo carecen de información sobre las etiquetas o clases a las que pertenece cada dato, por lo que la estrategia a seguir se basa en las herramientas de aprendizaje computacional enfocadas a entrenamiento no supervisado. Se implementaron dos aproximaciones:

- métodos de agrupamiento: k-medias y k-medianas,
- redes neuronales: mapas auto-organizados.

La figura 3.11 muestra de forma esquemática el proceso de aprendizaje computacional empleado en este trabajo. De forma general, el procedimiento completo se puede dividir en dos etapas: etapa de entrenamiento y etapa de agrupamiento, requiriendo un pre-procesado para escalar las variables y opcionalmente reducir la dimensionalidad de la base de datos.

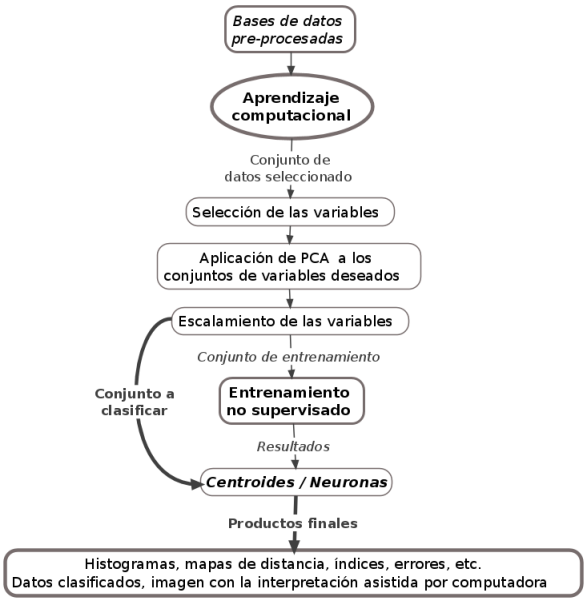


Figura 3.11: Esquema del proceso de aprendizaje computacional usado.

3.3.1. Pre-procesamiento

Las fases involucradas dentro de esta etapa son: inicialización de los algoritmos, reducción de dimensionalidad de conjuntos de variables ingresados por el usuario, y el escalamiento de las bases de datos. A continuación se describirá cada una de estas etapas.

Inicialización

El algoritmo implementado requiere del nombre de los archivos en donde se encuentran las bases de datos para entrenar, cada archivo ingresado se añade al conjunto de entrenamiento. Así mismo, como parámetro de entrada se encuentran los identificadores de las variables que se usarán tanto para entrenar como para agrupar los vectores de entrada, dicho identificador es el número de columna que almacena la variable de interés.

Reducción de dimensionalidad

En esta etapa se aplica *PCA* (sección 2.4.2) para reducir la dimensionalidad de conjuntos de variables ingresados por el usuario. Estos conjuntos se ingresan como una lista de listas, donde cada lista contiene los números de columna que se desean reducir de forma independiente a las listas restantes. Esta implementación para reducción de dimensionalidad se basa en la necesidad de *resumir* la información de variables que están estrechamente ligadas por el mismo fenómeno físico, por ejemplo:

- para magnetometría, la lectura o los filtros aplicados a los datos del sensor superior y del sensor inferior;
- para geoeléctrica, los valores de resistividad de los arreglos electródicos recíprocos;
- para ambos conjuntos de datos, los valores de una variable de los n vecinos más cercanos a cada vector, considerando sus coordenadas espaciales;

así como para evitar el efecto de la *maldición de la dimensionalidad* (Verleysen and François, 2005).

El algoritmo de *PCA* aplicado es el implementado en la biblioteca *Scikit-learn* (Pedregosa et al., 2011).

Escalamiento

Esta etapa busca llevar a los datos a un dominio en donde las variaciones de las variables de las bases de datos se encuentren comprendidas en un intervalo controlado. Los métodos empleados son: *StandardScaler*, *MinMaxScaler* y *MaxAbsScaler*, implementados en la biblioteca *Scikit-learn*; que respectivamente estandarizan, transforman a un dominio comprendido entre $[-1, 1]$ y escalan los valores entre $[0, 1]$.

3.3.2. Entrenamiento y clasificación

Esta etapa recibe los datos para aplicarlos al proceso de entrenamiento o agrupamiento. Cada aproximación (métodos de agrupamiento y redes neuronales) genera

distintos gráficos a la salida (errores, histogramas y mapas), que dependerán del método seleccionado.

Métodos de agrupamiento

Las herramientas empleadas en esta aproximación son:

- k-medias, usando la implementación de *Scikit-learn*; y
- k-medianas, usando la implementación de *PyClustering* (Novikov, 2019).

Se busca que estos algoritmos encuentren conjuntos similares de vectores con base en la distancia euclidiana del espacio multidimensional. Estos conjuntos se verán influenciados por la densidad y distribución de los vectores en dicho espacio, por lo que para evitar distribuciones de centroides sesgadas se implementaron dos alernativas al momento de entrenar al sistema:

- Aplicar una transformación a los datos, pudiendo ecualizar el histograma o aplicar el logaritmo base 10 a los datos (sección 2.5). Con esto se busca evitar la existencia de valores extremos que podrían comportarse como atípicos.
- Realizar el proceso de entrenamiento a los datos sin transformar, seguido de la supresión de los vectores que pertenezcan a la clase más poblada, posterior a esto se vuelve a entrenar al sistema. Con esto se busca ganar *resolución* al momento de agrupar, ya que se eliminará el efecto de la clase más poblada.

Como resultado de la clasificación se espera obtener una nueva base de datos con las siguientes variables:

- coordenadas espaciales x y y; y
- etiqueta de acuerdo con el centroide más cercano a cada vector, otorgando un identificador numérico y un color a cada grupo;

entonces, resultante del agrupamiento se obtiene una imagen cuya tercera dimensión es una etiqueta. Este resultado muestra, de forma general, las estructuras presentes en la base de datos analizada, pudiendo apreciar la distribución de los

vectores semejantes. Por lo tanto, esta imagen es un modelo simplificado de la distribución de las posibles estructuras en el subsuelo, siendo entonces una interpretación de los datos ingresados.

Los elementos resultantes de este proceso son:

- La etiqueta a la que pertenece cada vector de la base de datos a agrupar, aunada a las coordenadas espaciales que le corresponden.
- Histograma del uso de centroides, donde se muestra el número de vectores de la base de datos a clasificar que *caen* dentro del radio de cada centroide generado durante el entrenamiento.
- Malla que consta de la representación gráfica de los centroides usando esferas, cada una con las coordenadas de un centroide proyectado a un espacio tridimensional usando la herramienta *isomap* (sección 2.4.3) de *scikit-learn*, usando dos vecinos; los colores de cada esfera son independientes y elegidos automáticamente, correspondiendo unívocamente a los colores de la imagen de la base de datos agrupada. Esta malla se crea usando la biblioteca *PyVista* (Sullivan and Kaszynski, 2019).

Redes neuronales

La herramienta empleada para esta aproximación son los mapas auto-organizados, usando la implementación *minisom* de Vettigli (2018).

Se busca que este algoritmo distribuya los vectores de acuerdo al grado de similitud que guardan entre sí. Esta distribución cambiará de acuerdo a la topología de la red elegida, pudiendo ser rectangular o hexagonal, así como de la inicialización delos pesos de cada neurona. El mapa resultante servirá para visualizar la distribución de las posibles clases que existen en las bases de datos.

Para determinar los conjuntos de neuronas que son similares entre sí se usaron algoritmos de agrupamiento, pudiendo elegir entre k-medias, k-medianas y k-medoides. Encontrando entonces una etiqueta adicional que indica el *cúmulo* de neuronas al

que pertenece cada neurona individual, que a su vez es una etiqueta para el vector que la tenga como *BMU*.

Como se revisó en la sección 2.6.2, el mapa resultante de este algoritmo dependerá de la inicialización del vector de pesos de cada neurona, por lo que con cada ejecución del algoritmo se obtendrán mapas diferentes. Para evitar que la clasificación de un vector esté sesgada por el efecto aleatorio de la inicialización, se propuso implementar un *consenso* entre n redes entrenadas. Considerando esto se agregan dos etiquetas más a cada vector:

1. El color promedio de las n *BMU* de cada una de las n redes.
2. Considerando el punto anterior, se entrena un nuevo *SOM* para clasificar los nuevos colores de cada vector, llegando entonces a una versión más simplificada y fácil de interpretar.

Finalmente, como resultado de la clasificación se obtiene una nueva base de datos con las siguientes variables:

- coordenadas espaciales x y y;
- etiquetas que indican la *BMU* de los n mapas entrenados;
- etiquetas que indican el subconjunto de neuronas de los n mapas entrenados;
- etiqueta que indica el *consenso* de los n mapas; y
- etiqueta que indica el color de la *BMU* del mapa entrenado con los colores resultantes del *consenso* de los n mapas iniciales.

Los elementos resultantes de este proceso son:

- Las etiquetas a las que pertenece cada vector de la base de datos a clasificar, aunadas a las coordenadas espaciales que le corresponden.
- Mapa de aciertos de la red (*hit-map*).
- Mapa de distancias de la red (*u-matrix*).

- Gráfico con la evolución de los errores de cuantización y topológico.
- Malla cuyas celdas representan las neuronas de la red. El color de cada celda corresponde al subconjunto al que pertenece, siendo su saturación y brillo diferente para cada una, todo elegido automáticamente, correspondiendo unívocamente a los colores de la imagen de la base de datos agrupada.

Capítulo 4

Bases de datos

4.1. Introducción

En la sección 2.7 se mencionaron los aspectos teóricos involucrados en la adquisición de datos de mallas de magnetometría y de *TRE2D*. En este capítulo se describirán las características que tienen las bases de datos que fueron usadas en este trabajo.

Los datos fueron proporcionadas por el departamento de geofísica de la Facultad de Ingeniería de la Universidad Nacional Autónoma de México, conteniendo información de la componente total de campo magnético registrada por un gradiómetro e información de resistividad aparente registrada por un resistivímetro.

4.2. Mallas de magnetometría

Esta base de datos consta de dos zonas de estudio con objetivos arqueológicos similares:

1. *Xalasco*, Tlaxcala; y

2. *La Ferrería*, Durango.

Tabla 4.1: Número de vectores de las mallas de magnetometría levantadas en *La Ferrería*, Durango.

Malla	**Número de puntos**
$2F$	27608
$3F$	108535
$4F$	190345
$5F$	60755
$6F$	71666
$7F$	264130
Total de vectores:	723039

A pesar de que estas zonas exhiben características geomagnéticas diferentes (como la magnitud del campo geomagnético), las estructuras objetivo no presentan una dependencia fuerte a estas, es decir se pueden identificar estructuras similares independientemente si el levantamiento en cuestión se encuentra en Tlaxcala o en Durango, debido a que las anomalías están definidas por el contraste en la magnitud del campo, más que por su amplitud en sí (sección 2.7.1).

Se cuenta con seis mallas para cada zona de estudio (12 mallas en total), ninguno de sus puntos coincide espacialmente. El equipo usado fue un gradiómetro, por lo que se cuentan con datos de dos sensores separados verticalmente por 1 $[m]$. Se cuidó que la adquisición contara con el debido control de calidad, sin embargo algunas mallas de *La Ferrería* presentan picos de amplitud en ambos sensores, efecto que se filtró usando las metodologías descritas en la sección 3.2.2.

La tabla 4.1 muestra el número de puntos para cada malla filtrada y procesada de los datos levantados en la zona arqueológica de *La Ferrería*, mientras que la tabla 4.2 muestra lo propio para la zona arqueológica de *Xalasco*.

Tabla 4.2: Número de vectores de las mallas de magnetometría levantadas en *Xalasco*, Tlaxcala.

Malla	**Número de puntos**
$1X$	177227
$2X$	118142
$3X$	76439
$4X$	83075
$5X$	75465
$6X$	163421
Total de vectores:	693769

Todas las mallas fueron corregidas debidamente:

1. por IGRF, de acuerdo con la fecha de adquisición y ubicación de cada punto; y
2. por variación diurna, con una base magnetométrica ubicada cerca de cada malla y aproximando a un modelo lineal.

Antes de procesar cada malla, aplicando *filtros de realce de anomalías*, se aplica el proceso de filtrado descrito en la sección 3.2.2. Los parámetros usados para cada malla se muestran en la tabla 4.3. Todas las configuraciones de filtrado aplicado tienen en común los siguientes parámetros:

- Columnas: campo magnético en sensor inferior, campo magnético en sensor superior y el gradiente vertical.
- Ondícula: *Daubechies* de cuarto orden.
- Componentes: 3.

El lector es invitado a consultar el anexo A para más detalles sobre este proceso.

Se propone que las variables explicativas de cada punto sean aquellas que se usan generalmente para interpretar un levantamiento de magnetometría 2D (sección 2.7.1), además del valor de los n vecinos espacialmente más cercanos. Estas últimas

Tabla 4.3: Parámetros usados para el filtrado de las mallas de magnetometría. FM es el filtro de malla completa y FV es el filtro de ventana.

Malla	**Tipo de filtro**	**Pesos del primer filtro**	**Pesos del segundo filtro**
2F	FM	[1,0.1,0.5,0.7]	N/A
3F	FM	[1,0,0.5,0.7]	N/A
4F	FV-FM	[0,0,0.1,0.3]	[1,0.1,0.5,0.9]
5F	FV-FM	[0,0,0.1,0.3]	[1,0.1,0.5,0.9]
6F	FV-FM	[0,0,0.1,0.3]	[1,0.1,0.5,0.7]
7F	FM	[1,0.1,0.5,0.9]	N/A
1X	FM	[1,0.1,0.5,1]	N/A
2X	FM	[1,0.1,0.5,1]	N/A
3X	FV-FM	[0,0,0.1,0.3]	[1,0.1,0.5,1]
4X	FM	[1,0.3,0.7,1]	N/A
5X	FV-FM	[0,0,0.1,0.5]	[1,0.1,0.5,0.8]
6X	FM	[1,0.3,0.5,0.8]	N/A

variables tienen como finalidad agregar sensibilidad a contrastes o discontinuidades en los valores de cada variable producto del *realce de anomalías*.

La lista siguiente contiene las variables de estas bases de datos, la cual servirá de referencia para mencionar el orden de estas al analizar las matrices de dispersión:

1. coordenadas *UTM* que definen a cada punto espacialmente, no usadas para entrenar ni para agrupar;

2. componente total de campo magnético registrada por los dos sensores, de forma independiente;

3. gradiente horizontal;

4. reducción al polo;

5. señal analítica;

6. derivada inclinada;
7. gradiente horizontal de la derivada inclinada; y
8. gradiente vertical registrado por el gradiómetro.

Los puntos 2 al 7 de la lista anterior se aplican a los datos adquiridos por el sensor superior y a los datos adquiridos por el sensor inferior, seguidos de los valores de los n puntos vecinos más cercanos de acuerdo con sus coordenadas *UTM*.

En total, la base de datos correspondiente a las mallas de magnetometría consta de 1416808 vectores con 65 variables explicativas, considerando 4 vecinos para cada variable de la lista anterior. Con el fin de reducir el número de dimensiones de la base de datos, se propone aplicar reducción de dimensionalidad usando *PCA* (sección 2.4.2) sobre los 4 vecinos comunes a cada variable, este proceso se detalla en la sección 5.2.

Para visualizar la correlación entre las variables explicativas se analizaron las matrices de dispersión (figura 4.1) de las siguientes configuraciones considerando el 10 % de los vectores, elegidos aleatoriamente:

1. **Histograma ecualizado y sin considerar vecinos** (**a**): el orden de las variables, numerados según la lista anterior, es: 2, 8, 4, 5, 3, 6, 7. Los histogramas de cada variable se encuentran bien distribuidos, reflejando gráficos de dispersión muy uniformes sin marcar dependencias claras, excepto para las variables de campo total y de reducción al polo, para sensor superior y sensor inferior, las cuales muestran una ligera dependencia lineal.
2. **Histograma ecualizado y considerando 4 vecinos** (**b**): el orden de las variables, numerados según la lista anterior y colocando las componentes principales de los 4 vecinos más cercanos es: 2, 8, 4, 5, 3, 6, 7, primer componente de 2, primer componente 8, primer componente de 4, primer componente de 5, primer componente de 3, primer componente de 6, primer componente de 7 y segunda componente de 7. Los histogramas de cada variable se encuentran bien distribuidos, similar a los del punto anterior, sin embargo se observa una

clara dependencia lineal entre la componente principal de los vecinos con su respectiva variable.

3. **Histograma sin ecualizar y sin considerar vecinos** (**c**): el orden de las variables es el mismo que en el punto 1, se marcan las mismas dependencias lineales a pesar de que en este caso no se aplique una normalización.

4. **Histograma sin ecualizar y considerando 4 vecinos** (**d**): el orden de las variables es el mismo que en el punto 2, no existen muchas diferencias en cuanto al grado de correlación de las variables respecto a los casos anteriores.

Los *índices de correlación de Pearson* para las variables de campo total, reducción al polo, así como la primer componente principal de los 4 vecinos respectivos a cada variable, es superior al 0,95, comparando entre las variables del sensor superior con las del sensor inferior. En otro caso, el valor del índice es inferior a 0,7, indistintamente si se ecualiza el histograma o no.

4.3. *TRE2D*

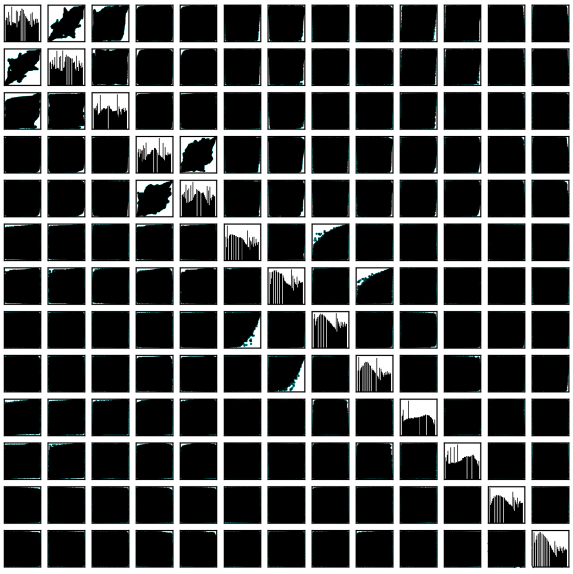
Esta base de datos consta de dos zonas de estudio con objetivos diferentes:

1. *Tamazulápam*, Oaxaca;

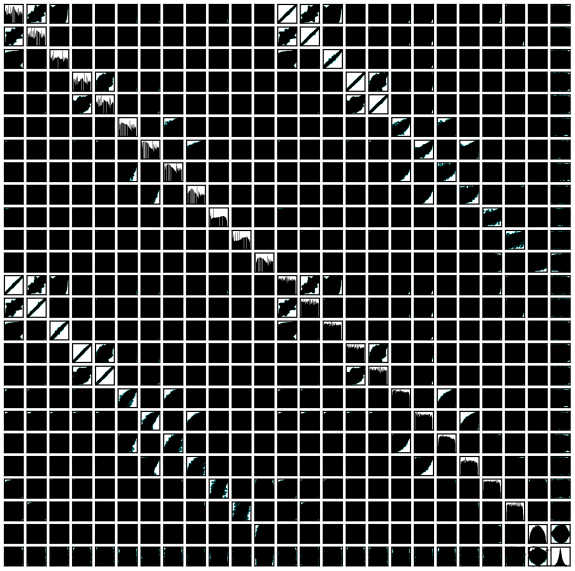
2. *Calpulalpan*, Tlaxcala.

Ambas zonas con objetivo geohidrológico, exhibendo características diferentes respecto a las estructuras buscadas y respecto a los contrastes y valores de resistividad.

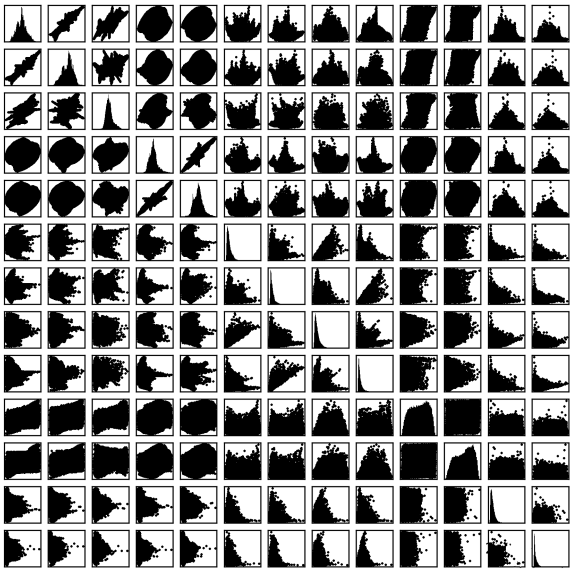
Los datos usados para interpretar levantamientos geoeléctricos son los resultantes del procedimiento conocido como *inversión* (sección 2.7.2), por lo que se usarán estos resultados para entrenar los algoritmos de aprendizaje computacional. La inversión de datos se realizó utilizando la biblioteca *ResIPy* (Blanchy et al., 2020), cuyos datos de entrada son las versiones filtradas de los datos adquiridos, usando los algoritmos



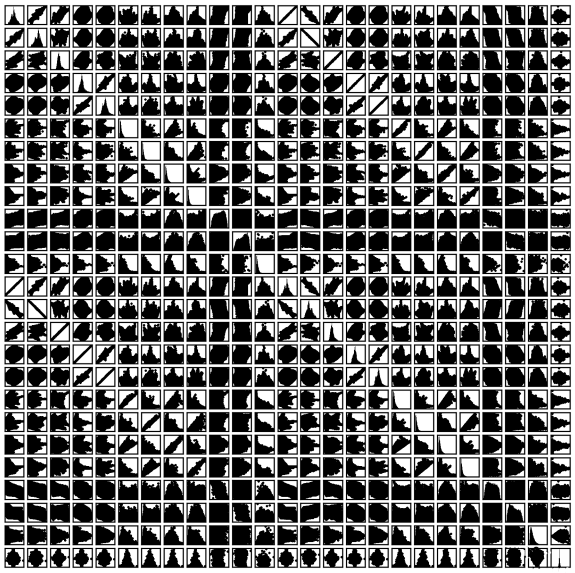
(a) Histograma ecualizado y sin considerar vecinos.



(b) Histograma ecualizado y considerando vecinos.



(c) Histograma sin ecualizar y sin considerar vecinos.



(d) Histograma sin ecualizar y considerando vecinos.

Figura 4.1: Gráficos de dispersión de las variables de la base de datos de magnetometría.

Tabla 4.4: Número de vectores de las bases de datos de las *TRE2D*.

Zona de estudio	**Número de puntos (entrenamiento)**	**Número de puntos (clasificación)**
Tamazulápam	194540	2101
Calpulalpan	17120	418
Total de vectores	211660	2519

detallados en la sección 3.2.1.

En todas las localidades se aplicaron dos arreglos electródicos, obteniendo entonces dos secciones geoeléctricas espacialmente coincidentes, pudiendo entonces construir una base de datos que contiene dos valores de resistividad para cada punto con coordenadas (x, z) dentro de la *TRE2D*. Se propone agregar como variable explicativa los valores de resistividad de los n vecinos espaciamente más cercanos, buscando con esto agregar sensibilidad a los contrastes y discontinuidades en la distribución de resistividades.

Debido a la escasa cantidad de datos en cada zona estudiada se propone la generación de modelos sintéticos para complementar las bases de datos. Estos modelos pueden realizarse de forma semiautomática usando la biblioteca *ResIPy*, considerando distintas distribuciones de resistividad y contrastes de esta de acuerdo a cada zona de estudio.

La tabla 4.4 muestra el número de puntos que le corresponden a cada base de datos. Los vectores usados para el entrenamiento fueron los modelos sintéticos.

Antes de invertir los datos de *TRE2D* adquiridos en campo, se aplicó el proceso de filtrado descrito en la sección 3.2.1. Todas las configuraciones de filtrado aplicado tienen en comun los siguientes parámetros:

1. Análisis en multiresolución 1D:
    - diferencia de potencial V_{MN} como variable a filtrar,
    - ondícula *Daubechies* de segundo orden,
    - 5 descomposiciones,
    - 1 recursión.
2. detección de vectores atípicos:
    - resistividad aparente ρ_a como variable a analizar,
    - aplicación de logaritmo base 10,
    - *IF* como método,
    - 0,9 de *umbral*.

Para más detalles del proceso de filtrado y de la generación de los modelos sintéticos se puede consultar el anexo B. Los parámetros de inversión fueron los definidos por defecto por la paquetería *ResIPy*.

Las variables de esta base de datos son:

1. coordenadas (x, z) que definen a cada punto espacialmente dentro de la *TRE2D*, no usadas para entrenar ni para agrupar; y
2. los valores de resistividad de los 2 arreglos electródicos.

Para el punto 2, cada valor de resistividad puede estar seguido de los valores respectivos de los n vecinos espacialmente más cercanos a cada vector. Por lo que la base de datos puede constar de 6 variables explicativas considerando 3 vecinos. Con el fin de reducir el número de dimensiones de la base de datos, se propone aplicar reducción de dimensionalidad usando *PCA* (sección 2.4.2) sobre los 3 vecinos comunes a cada variable, este proceso se detalla en la sección 5.3.

Para visualizar la correlación entre las variables explicativas se analizaron las matrices de dispersión (figura 4.2) de las siguientes configuraciones considerando el 50 % de los vectores:

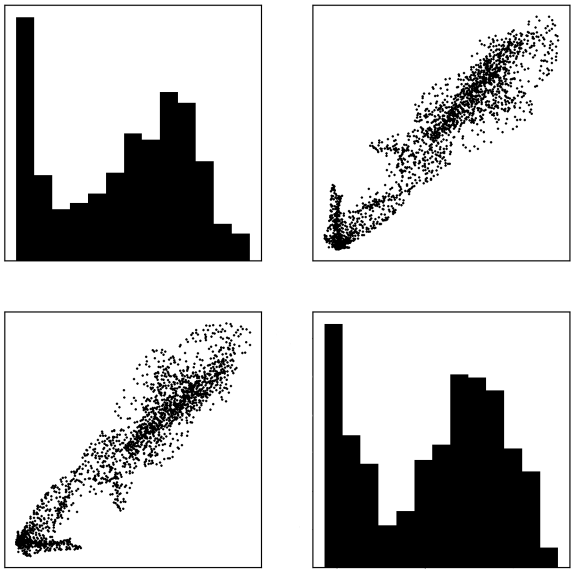
1. **Con transformación logarítmica y sin considerar vecinos** (**a**): donde la primer variable es la resistividad del arreglo Wenner-Schlumberger y la segunda es la resistividad del arreglo dipolo-dipolo, presentando una ligera dependencia lineal entre ambas.

2. **Con transformación logarítmica y considerando 3 vecinos** (**b**): donde las variables son las mismas que el punto anterior y agregando una componente principal de *PCA* para los vecinos espaciales de cada variable, se marca una tendencia lineal fuerte en estas últimas.

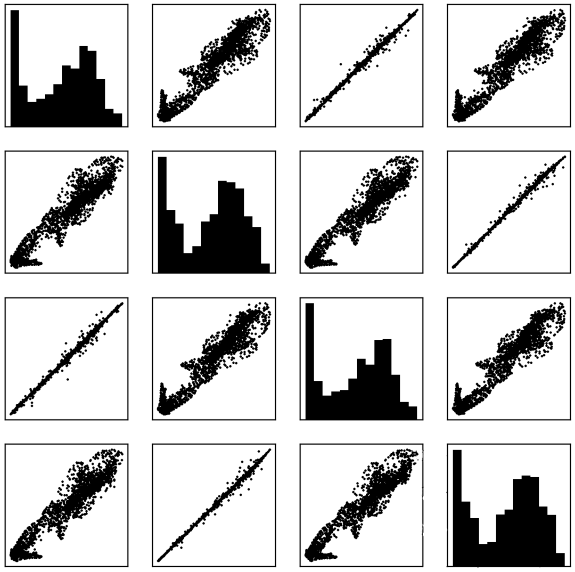
3. **Sin transformar y sin considerar vecinos** (**c**): el orden de las variables es el mismo que en el punto 1, no se aprecia una dependencia lineal clara.

4. **Sin transformar y considerando 3 vecinos** (**d**): el orden de las variables es el mismo que en el punto 2 apreciando poca dependencia lineal, similar al punto anterior.

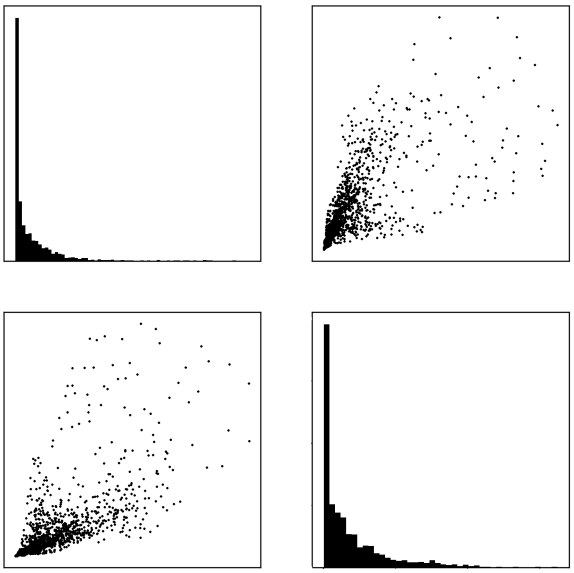
Los *índices de correlación de Pearson* para las variables sin transformar es de 0,7, mientras que transformadas alcanzan un valor de 0,9. El valor del coeficiente para la primer componente de los vecinos es superior al 0,9.



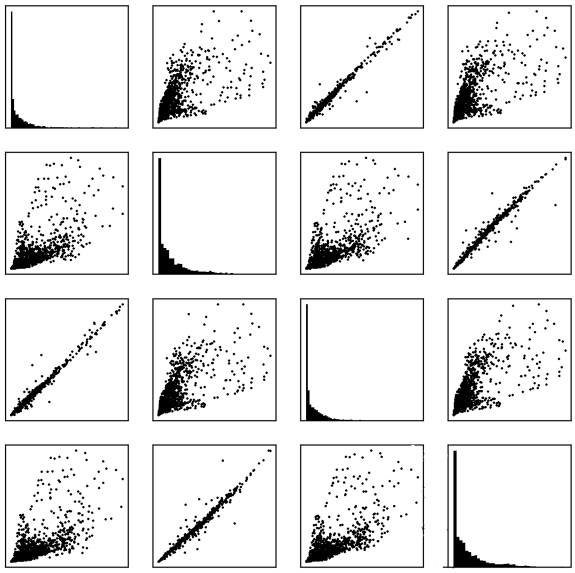
(a) Con transformación logarítmica y sin considerar vecinos.



(b) Con transformación logarítmica y considerando vecinos.



(c) Sin transformación logarítmica y sin considerar vecinos.



(d) Sin transformación logarítmica y considerando vecinos.

Figura 4.2: Gráficos de dispersión de las variables de la base de datos de *TRE2D*.

Capítulo 5

Resultados

5.1. Introducción

En el capítulo 3 se describieron los métodos implementados para cumplir con los objetivos de esta tesis, mismos que serán aplicados a las bases de datos descritas en el capítulo 4.

A la entrada de los algoritmos se reciben los datos geofísicos procesados, al nivel en el que sean interpretados por un experto en el área. A la salida de las herramientas implementadas se exhiben imágenes que muestran una versión simplificada de las posibles estructuras que generan las anomalías geofísicas, esto a partir de los patrones y similitudes que se encuentran en las bases de datos.

En las siguientes secciones se mostrarán los parámetros usados para obtener los resultados de la interpretación automática a partir de la aplicación de la metodología propuesta, tanto a los datos de magnetometría como a los datos geoeléctricos.

5.2. Mallas de magnetometría

Retomando lo mencionado en la sección 2.7.1, los datos de levantamientos magnetométricos se interpretan con base en los contrastes en la amplitud del campo

magnético y de sus *atributos*[1]. Generalmente a partir de análisis visuales se concluye sobre las características de las posibles estructuras presentes en el subsuelo. Se proponen entonces dos puntos de partida diferentes para efectuar el entrenamiento y agrupamiento de estos datos:

- ecualizando el histograma, y
- no ecualizando el histograma;

lo anterior para cada variable de la base de datos, ya que este *pre-proceso* puede ayudar a mejorar el contraste en una imagen (sección 2.5.2). Por otro lado, se desconoce de antemano los niveles de importancia que tiene cada variable para definir los patrones que pueden caracterizar una anomalía geofísica, por lo que se proponen los siguientes casos de conjuntos de variables a evaluar:

- considerando únicamente las variables del sensor inferior (contiene mayor información de las estructuras más someras, comparado con el sensor superior) y del gradiente vertical;
- mismas variables del punto anterior junto a las variables de los respectivos 4 vecinos espacialmente más cercanos, mismos que serán reducidos mediante *PCA*;
- considerando las variables del sensor inferior y del sensor superior, además del gradiente vertical;
- mismas variables del punto anterior junto a las variables de los respectivos 4 vecinos espacialmente más cercanos, mismos que serán reducidos mediante *PCA*;
- considerando las variables del sensor inferior, del sensor superior y el gradiente vertical, reduciendo mediante *PCA* cada pareja de variables con una correlación lineal fuerte[2];

[1] Refiriéndome a los filtros de realce de anomalías.

[2] Por ejemplo: las magnitudes del campo magnético y los filtros de realce de anomalías, para cada par de *sensor superior y sensor inferior.*

- mismas variables que el punto anterior junto a las variables de los respectivos 4 vecinos espacialmente más cercanos, mismos que serán reducidos mediante *PCA*; y
- considerando todas las 65 variables y reduciéndolas mediante *PCA*.

Uniendo las dos listas anteriores resulta en 14 combinaciones a evaluar. Los parámetros fijos en cada metodología de entrenamiento son:

- La selección de las mallas usadas para el entrenamiento, así como la malla a agrupar[3], se realiza aleatoriamente. Para que el conjunto de entrenamiento represente cerca del 80 % del número de datos involucrados, y que el restante 20 % corresponda al conjunto de clasificación, se seleccionan 5 mallas para conformar el conjunto de entrenamiento.
- Si aplica, el número de componentes principales resultante del empleo de *PCA* se elige de forma automática condicionando que dichas componentes representen el 80 % de varianza, esto con el fin de reducir de forma sustancial la cantidad de variables involucradas sin perder demasiada información.
- El escalamiento aplicado a cada variable fue *minmax* (sección 3.3.1), ya que los datos analizados pueden tomar valores positivos y negativos en proporciones similares.

5.2.1. Métodos de agrupamiento

Con la finalidad de elegir el mejor número de centroides que representa a las bases de datos se realizó una evaluación basada en los índices de *Davies-Bouldin* y de *Silhouette*. El cálculo de los índices se efectúa 10 veces, cambiando de forma aleatoria el conjunto de entrenamiento y el conjunto a agrupar.

La tabla 5.1 muestra los centroides con los que se obtuvieron los mejores índices promedio de las 10 ejecuciones, para todas las configuraciones de entrada.

[3]La malla a agrupar se elige aleatoriamente únicamente para aplicar la evaluación usando los índices de *Davies-Bouldin* y de *Silhouette*

Tabla 5.1: Mejores índices promedio para cada configuración de entrada de los datos de magnetometría.

Configuración	**Número de centroides**	**Índice Davies-Bouldin**	**Índice Silhouette**
Ecualizado y sensor inferior	2 - 3	0,96 - 1,03	0,39 - 0,23
Ecualizado y sensor inferior con vecinos	3 - 4	1,16 - 1,44	0,13 - 0,11
Ecualizado y ambos sensores	2 - 3	1,52	0,26
Ecualizado y ambos sensores con vecinos	3 - 4	1,44 - 1,53	0,26 - 0,22
Ecualizado y ambos sensores reducidos	3 - 4	1,18 - 2,10	0,11 - 0,00
Ecualizado y ambos sensores reducidos con vecinos	2 - 3	1,17 - 1,55	0,27 - 0,17
Ecualizado y usando todas las variables	2 - 3	1,01 - 1,02	0,38
No ecualizado y sensor inferior	2 - 3	0,83 - 0,85	0,39 - 0,36
No ecualizado y sensor inferior con vecinos	2 - 3	0,91 - 1,5	0,34 - 0,18
No ecualizado y ambos sensores	3 - 4	0,89 - 0,98	0,42 - 0,3
No ecualizado y ambos sensores con vecinos	3 - 4	1,19 - 1,44	0,29 - 0,28
No ecualizado y ambos sensores reducidos	3 - 4	2,09 - 1,24	0,24 - 0,27
No ecualizado y ambos sensores reducidos con vecinos	3 - 4	2,29 - 1,38	0,24 - 0,12
No ecualizado y usando todas las variables	3 - 4	2,29 - 1,38	0,24 - 0,12

Tabla 5.2: Mejores resultados de los parámetros de los *SOM* para los datos de magnetometría, obtenidos a partir de ensayos experimentales.

Variable	**Errores**	**u-matrix**	**hit-map**
Dimensiones	Mayores a 3×3	Mayores a 4×4	Menores a 10×10
Épocas		7000 iteraciones	
Vecindad	50 %	30 %	30 %
Aprendizaje	0,1	0,1	0,1

5.2.2. *SOM*

Para esta metodología se realizó un análisis cualitativo que involucró los siguientes puntos:

- los errores de cuantización y topológico,
- la estructura de la *u-matrix*, y
- la distribución del *hit-map*.

con las siguientes variables sujetas a la validación:

- dimensiones del mapa auto-organizado,
- número de épocas,
- coeficiente de vecindad, y
- tasa de aprendizaje.

La tabla 5.2 muestra los mejores resultados a partir de ensayos realizados. A continuación se realizará una descripción de cada uno:

1. Dimensiones:
    - respecto a los errores, mapas con dimensiones mayores a 3×3 tienden a disminuir de forma considerable sus errores, esto debido a que hay una mayor cantidad de neuronas para caracterizar la base de datos;

- respecto a la *u-matrix*, debido a las características de la base de datos, los grupos de neuronas tienden a situarse en los extremos del mapa, dimensiones mayores a 4×4 resultan en redes en las que es más clara la definición de grupos;
- respecto al *hit-map*, de la mano con la *u-matrix*, las neuronas más usadas tienden a localizarse cerca de los extremos, mapas con dimensiones menores a 10×10 tienden a definir de mejor forma las estructuras anómalas generalizando de mejor forma los patrones existentes en la base de datos.

2. Épocas: usar más de 7000 iteraciones tiende a aumentar el error topológico de manera considerable, pudiendo alcanzar valores mayores al 50 %. Con 7000 iteraciones las curvas de error de cuantización y topológico se situan por debajo de 0,2 y del 40 % respectivamente, de esta forma los mapas resultantes tienden a estar mejor estructurados.

3. Vecindad:

   - respecto a los errores, comenzar a actualizar a partir del 50 % de la red los disminuye de forma considerable;
   - respecto a la *u-matrix* y al *hit-map*, con la finalidad de distribuir de forma más homogenea los grupos de neuronas se puede ajustar el radio de vecindad al 30 %.

4. Aprendizaje: usar una taza de aprendizaje de 0,1 disminuye los errores de cuantización y topológico, comparado con el uso de una taza más alta. Además mejora la distribución de las neuronas de acuerdo con la *u-matrix* y el *hit-map*.

Es necesario mencionar que con *mejorar la distribución de las neuronas*, *mapas mejor estructurados* y *distribución más homogénea de los grupos de neuronas*, me refiero a que las neuronas más usadas no tiendan a localizarse en los extremos del mapa. Esta peculiaridad de que durante el entrenamiento se separen al extremo los conjuntos de neuronas depende en gran medida de las características del conjunto de entrenamiento.

5.2.3. Interpretación

Para elegir la mejor imágen se realizó una evaluación efectuada por intérpretes con experiencia en el método. Los aspectos evaluados fueron:

- nivel de detalle: visualización de elementos que puedan definir zonas de transición entre estructuras;
- definición de anomalías: las estructuras conformadas por los elementos del modelo generado presentan geometrías delimitadas por formas delgadas, así como patrones ordenados no aleatorios;
- paleta de colores: los colores del modelo generado presentan un contraste que facilita la visualización;
- estructura de los grupos o neuronas: los centroides o el mapa auto-organizado, según sea el caso, presenta una estrutura homogénea y es posible diferenciar los elementos de las estructuras regionales de los elementos que conforman la anomalía principal; y
- correlación con excavaciones: de existir evidencia, que las estructuras halladas en la zona arqueológica tengan correlación directa con las estructuras en la imagen analizada.

A continuación se muestran las imágenes que exhiben los resultados más representativos.

Malla 1X

La figura 5.1 muestra un agrupamiento aplicado a la malla 1X, los parámetros usados fueron:

- entrenamiento y agrupación usando los datos del sensor inferior sin considerar vecinos y sin ecualizar el histograma;
- *SOM* rectangular de 6×6, obtenida a partir de cinco redes de 6×6 con agrupación de neuronas usando k-medoides en 3 grupos;

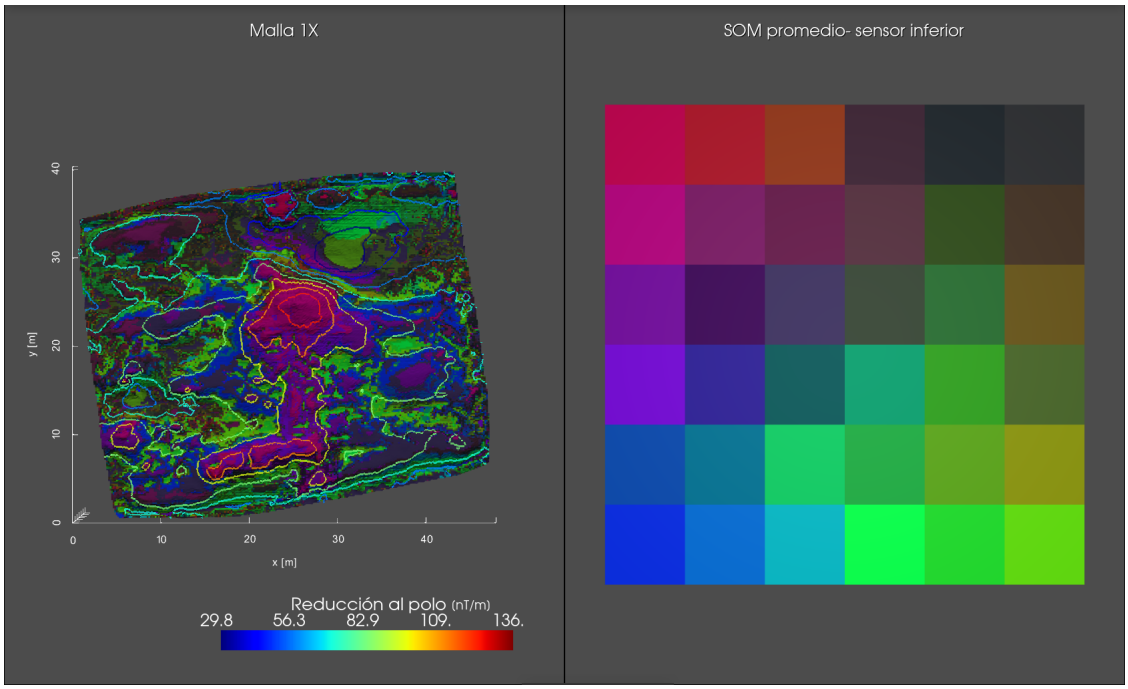


Figura 5.1: Agrupamiento aplicado a la malla 1X usando *SOM* rectangular.

- el error topológico se situó entre el 20 % al 30 %, mientras que el error de cuantización se mantuvo cerca de 0,1; y
- el porcentaje de uso de las redes fue superior al 80 %.

Respecto a la validación cualitativa:

- Nivel de detalle: se aprecian zonas de transición que delimitan de forma coherente las estructuras principales, definidas por los colores azules, morados y verdes.
- Definición de anomalías: el agrupamiento resultante preserva geometrías finas, marcando discontinuidades que pueden ser descritas como *agujeros*, presenta geometrías con patrones estructurados. Se visualiza la diferencia entre el efecto regional y la anomalía principal.
- Paleta de colores: los colores principales son azul, magenta y verde, la distribución de los tonos produce una imagen con suficiente contraste para ser interpretada.

- Estructura de las neuronas: los colores agrupados por el *SOM* se encuentran bien distribuidos, pudiendo definir cuatro grupos principales.

- Correlación con exavaciones: de acuerdo con Juárez et al. (2017) se realizó una excavación en la zona central de la malla 1X, encontrando una estructura antropogénica que corresponde a un suelo fragmentado con orientación N-S, concordando con la estructura descrita por los tonos magenta de la figura 5.1.

Dada la forma que tiene el mapa de reducción al polo, se puede atribuir que los tonos magentas a azules corresponden a la anomalía principal, que de acuerdo con el *SOM* se trata de elementos muy diferentes a las estructuras descritas por los tonos verdes. Se observa que la metodología de agrupamiento empleada distingue dos estructuras centrales, distinguidas por cambios de color del magenta al azul. Existe correlación entre la descripción realizada por Juárez et al. (2017) y la geometría de los elementos en tonos magenta, mostrando ciertas discontinuidades que se pueden aludir a la destrucción del suelo producto de la actividad humana.

A partir de la imagen obtenida se puede inferir una estructura que puede tener relación con la descrita en el párrafo anterior, definida por los elementos con tonos azules que tienden a ser la transición a los elementos de tonos verdes desde los tonos magenta. Esta estructura adicional presenta una geometría con bordes definidos y ángulos cercanamente rectos.

Esta metodología sobresale de los métodos de agrupación (k-medias y k-medianas) ya que las imágenes resultantes presentan detalles que estas últimas herramientas no son capaces de discriminar. Se usaron los datos del sensor inferior ya que es más sensible a las estructuras someras. Ecualizar el histograma no parece tener un impacto importante, así como agregar la información de los vecinos más cercanos. Con esta configuración se evita el uso de variables con correlación lineal fuerte.

Finalmente, una interpretación del *SOM* es:

- tonos magenta: la anomalía magnetométrica principal, correspondiendo a las estructuras más someras de acuerdo con los resultados arqueológicos, presenta

características muy diferentes a los elementos en tonos verdes;

- tonos azules: estructuras secundarias asociadas a una transición de las anomalías principales;
- tonos verdes: elementos que caracterizan las estructuras más profundas dentro del levantamiento, dada la diferencia con los elementos en tonos morados y al análisis de la imagen de reducción al polo; y
- tonos cafes y grises: elementos con menos presencia en la imagen agrupada, presentan tonos oscuros debido a que hubo mucha *discrepancia* en los colores de las cinco redes entrenadas, es posible que correspondan a vectores anómalos.

Malla 2X

La figura 5.2 muestra un agrupamiento aplicado a la malla 2X, los parámetros usados fueron:

- entrenamiento y agrupación usando los datos del sensor inferior sin considerar vecinos y sin ecualizar el histograma;
- *SOM* hexagonal de 6×6, obtenida a partir de cinco redes de 6×6 con agrupación de neuronas usando k-medoides en 3 grupos;
- el error topológico se situó entre el 40 % al 50 %, mientras que el error de cuantización se mantuvo entre 0,1 y 0,15; y
- el porcentaje de uso de la red fue superior al 90 %.

Respecto a la validación cualitativa:

- Nivel de detalle: se aprecian zonas de transición que delimitan las estructuras principales, las variaciones de los tonos verdes dan la sensación de incertidumbre en la continuidad de las estructuras conformadas por estos tonos.
- Definición de anomalías: el agrupamiento resultante preserva geometrías finas, marcando estructuras en el norte y centro de la zona estudiada, así como estructuras aisladas al este y suroeste. Los elementos en tonos morados representan el efecto regional.

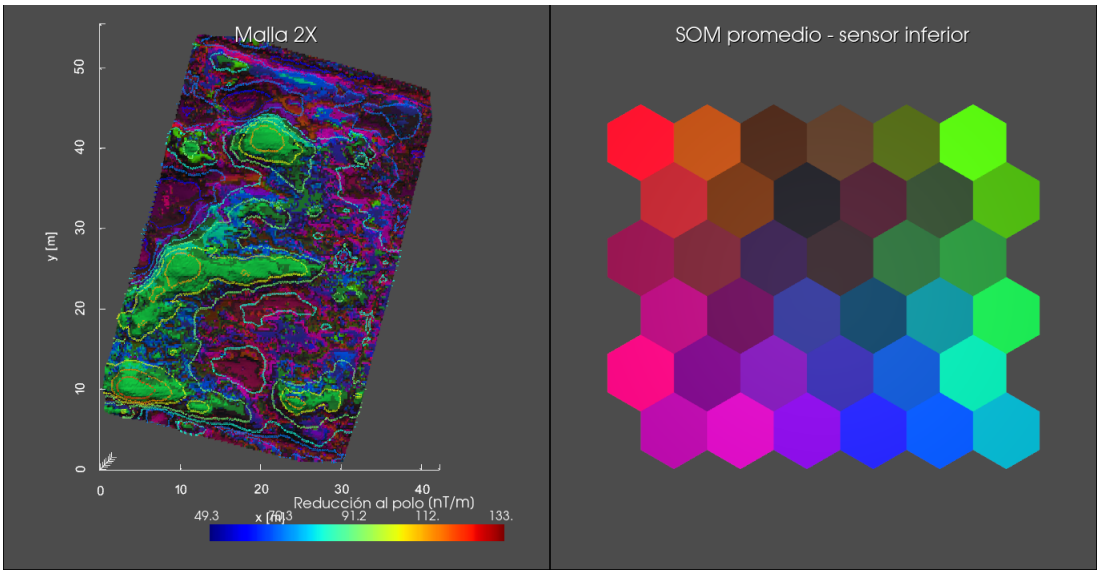


Figura 5.2: Agrupamiento aplicado a la malla 2X usando *SOM* hexagonal.

- Paleta de colores: los colores principales son verde y morado, la distribución de los tonos produce una imagen con suficiente contraste para ser interpretada.
- estructura de las neuronas: los colores agrupados por el *SOM* se encuentran bien distribuidos, pudiendo definir de tres a cuatro grupos principales.
- correlación con exavaciones: de acuerdo con Juárez et al. (2017) se realizaron excavaciones en la zona norte y en la zona central de la malla 2X, encontrando estructuras antropogénicas que corresponden respectivamente a un muro rodeado de suelo, aunado a piezas de interés arqueológico, y para la anomalía central la sección de una escalinata.

Dada la forma que tiene el mapa de reducción al polo, se puede atribuir que los tonos verdes corresponden a la anomalía principal, que de acuerdo al mapa autoorganizado se trata de elementos muy diferentes a las estructuras descritas por los tonos morados-rosas. Existe correlación entre las descripciones realizadas por Juárez et al. (2017) y la geometría de los elementos en tonos verdes. La estructura al norte corresponde con el muro y el suelo hallados durante las excavaciones, sin embargo en la posición donde detallan se encuentran las piezas sepultadas no se encuentran estructuras en el agrupamiento realizado. Por otro lado, la sección de escalinata que

pormenorizan al centro-este se distingue claramente conformada por elementos del conjunto verde.

A partir de la imagen obtenida se puede inferir que las estructuras delimitadas por los elementos en tonos verdes sugieren continuidad que corresponde a lo esperado arqueológicamente, evocando al suelo, escalinatas y suelo antropogénicos. Así mismo se observan esrtructuras más aisladas al suroeste y sureste.

Finalmente, una interpretación del *SOM* es:

- tonos verdes: la anomalía magnetométrica principal que corresponde a las estructuras arqueológicas, presenta características muy diferentes a los elementos en tonos morados-rosas;
- tonos azules: estructuras secundarias asociadas a una transición de las anomalías principales, pueden dar la sensación de continuidad en estructuras que pueden presentar cierto grado de erosión;
- tonos morados-rosas: elementos que caracterizan las estructuras más profundas dentro del levantamiento, dada la diferencia con los elementos en tonos verdes y al análisis de la imagen de reducción al polo; y
- tonos rojos-cafés: elementos con menos presencia en la imagen agrupada, los tonos oscuros se deben a la *discrepancia* en los colores de las cinco redes entrenadas, a diferencia de los tonos rojos cuyo agrupamiento siempre fue similar, es posible que correspondan a vectores anómalos.

Malla 2F

La figura 5.3 muestra un agrupamiento aplicado a la malla 2F, los parámetros usados fueron:

- entrenamiento y agrupación usando los datos del sensor inferior sin considerar vecinos y sin ecualizar el histograma;

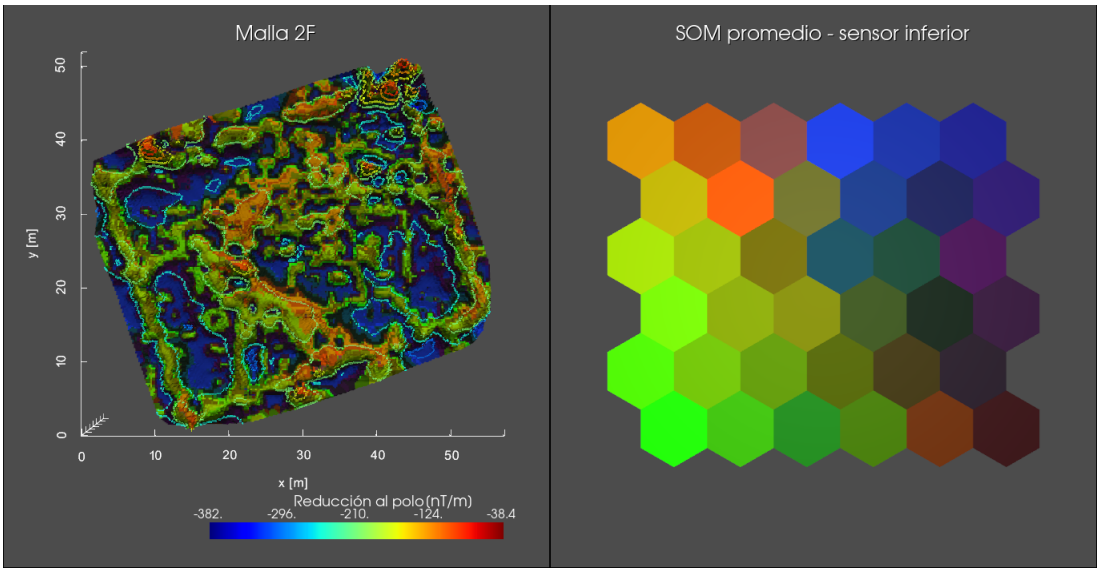


Figura 5.3: Agrupamiento aplicado a la malla 2F usando *SOM* hexagonal.

- *SOM* hexagonal de 6×6, obtenida a partir de cinco redes de 6×6 con agrupación de neuronas usando k-medoides en 3 grupos;
- el error topológico se situó entre el 30 % al 40 %, mientras que el error de cuantización se mantuvo cerca de 0,15; y
- el porcentaje de uso de la red fue superior al 95 %.

Respecto a la validación cualitativa:

- Nivel de detalle: se aprecian zonas de transición que delimitan las estructuras principales, las variaciones entre los colores verde y naranja distinguen estructuras que pueden ser diferentes.
- Definición de anomalías: el agrupamiento resultante marca estructuras con bordes muy bien definidos. Se identifica como efecto regional a los elementos de tonos azules, esto a partir de la reducción al polo.
- Paleta de colores: los colores principales son verde, naranja y azul, la distribución de los tonos produce una imagen con suficiente contraste para ser interpretada.

- estructura de las neuronas: los colores agrupados por el *SOM* se encuentran bien distribuidos, pudiendo definir de tres a cuatro grupos principales.
- correlación con exavaciones: aun no hay evidencia de excavaciones en trabajos publicados.

Dada la forma que tiene el mapa de reducción al polo, se puede atribuir que los tonos naranja corresponden a la anomalía principal. La estructura formada por los elementos de estos tonos parece describir una geometría angular poco discontinua, pudiendo tratarse de algun elemento antropogénico superficial.

A partir de la imagen obtenida se puede inferir que las estructuras delimitadas por los elementos en tonos verdes sugieren una estructura continua con geometría similar a la anomalía principal, presentando poca transición a los elementos con características regionales (tonos azules).

Finalmente, una interpretación del mapa auto-organizado es:

- tonos naranja: la anomalía magnetométrica principal teniendo relación directa con los valores de reducción al polo;
- tonos verdes: estructuras secundarias asociadas a una transición de las anomalías principales, sin embargo poseen características diferentes a los elementos naranja; y
- tonos azules: elementos que caracterizan las estructuras más profundas dentro del levantamiento, dada la diferencia con los elementos en tonos verdes y al análisis de la imagen de reducción al polo.

Malla 4F

La figura 5.4 muestra un agrupamiento aplicado a la malla 4F, los parámetros usados fueron:

- entrenamiento y agrupación usando los datos del sensor inferior considerando vecinos y sin ecualizar el histograma;

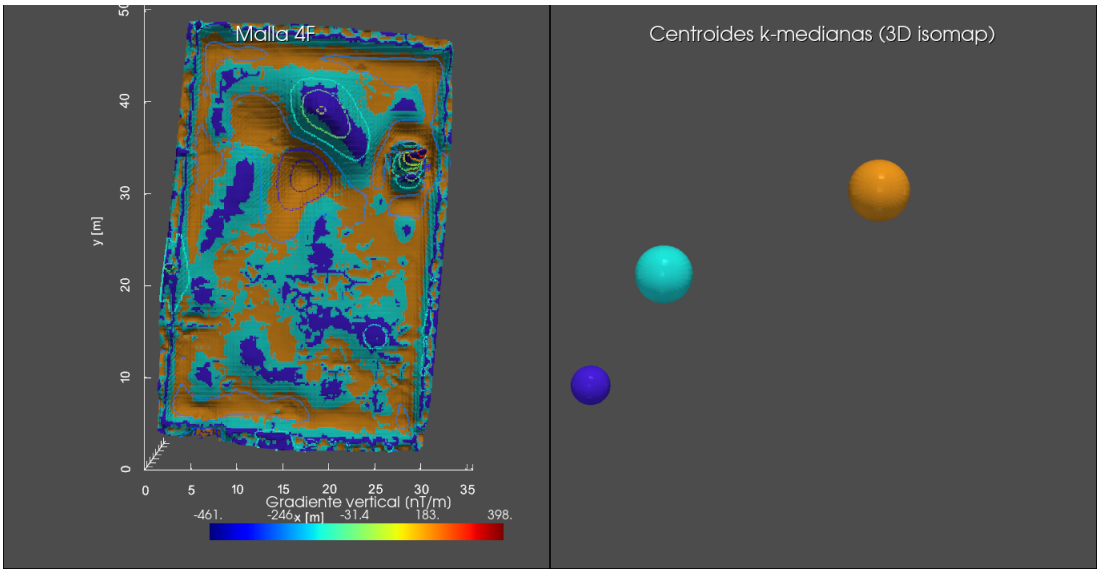


Figura 5.4: Agrupamiento aplicado a la malla 4F usando k-medianas.

- k-medianas con tres centroides; y
- las configuraciones anteriores no resultan en los mejores índices calculados (tabla 5.1), obteniendo 1,5 y 0,18 para el *índice de Davies-Bouldin* y para el *índice de Silhouette* respectivamente, sin embargo esta configuración presenta un resultado que resulta lógico para el fenómeno analizado.

Respecto a la validación cualitativa:

- Nivel de detalle: a pesar de que k-medianas es una metodología menos robusta comparada con *SOM*, la imagen delinea las estructuras principales sin agregar elementos de distintas clases con un patrón aparentemente aleatorio.
- Definición de anomalías: el agrupamiento resultante define de forma clara y precisa tres regiones diferentes.
- Paleta de colores: la imagen resultante tiene tres colores (azul, cyan y naranja) cuyo contraste es suficiente para poder interpretarla.
- estructura de los centroides: mediante la proyección de *isomap* se aprecia que los centroides estimados no se encuentran encimados, siendo entonces elementos

que por si mismos pueden describir porciones específicas del fenómeno analizado.

- correlación con exavaciones: aun no hay evidencia de excavaciones en trabajos publicados.

Dada la forma que tiene el mapa de gradiente vertical, se puede atribuir que el color azul corresponde a la anomalía principal. La estructura formada por estos elementos presentan alineaciones con orientación N-S hacia el oeste y con orientación W-E hacia el sur, además de una estructura con diemnsiones importantes al norte. Se puede concluir que las tres estructuras diferenciadas en este agrupamiento son más profundas que las analizadas en las mallas anteriores, por ello un algoritmo cuya resolución tiende a ser menor tiene un mejor desempeño.

Finalmente, una interpretación de los centroides es:

- color azul: la anomalía magnetométrica principal teniendo relación directa con los valores del gradiente vertical;
- color cyan: estructuras secundarias asociadas a una transición de las anomalías principales; y
- color naranja: elementos que caracterizan las estructuras más profundas dentro del levantamiento, dada la diferencia con los elementos en tonos verdes y al análisis de la imagen del gradiente vertical.

5.3. *TRE2D*

Retomando lo mencionado en la sección 2.7.2, los datos de *TRE2D* se interpretan analizando la imagen obtenida a partir de la *inversión* de datos, una imagen por cada arreglo electródico aplicado. Generalmente a partir de análisis visuales se concluye sobre las características de las posibles estructuras presentes en el subsuelo. Se proponen entonces dos puntos de partida diferentes para efectuar el entrenamiento y agrupamiento de estos datos:

- aplicando una transformación logarítmica, y
- no aplicando una transformación logarítmica;

lo anterior para cada variable de la base de datos, ya que este *pre-proceso* puede ayudar a mejorar el contraste en este tipo de datos (sección 2.5.1). Por otro lado, se desconoce de antemano los niveles de importancia que tiene cada variable para definir los patrones que pueden caracterizar una anomalía geofísica, por lo que se proponen los siguientes casos de conjuntos de variables a evaluar:

- considerando únicamente las variables de resistividad para cada arreglo electródico usado; y
- mismas variables del punto anterior junto a las variables de los respectivos 3 vecinos espacialmente más cercanos, mismos que serán reducidos mediante *PCA*.

Uniendo las dos listas anteriores resulta en 4 combinaciones a evaluar. Los parámetros fijos en cada metodología de entrenamiento son:

- El conjunto de entrenamiento usado es el correspondiente para cada *TRE2D* a agrupar, obtenido a partir de los modelos sintéticos semi-automáticos.
- Si aplica, el número de componentes principales resultante del empleo de *PCA* se elige de forma automática condicionando que dichas componentes representen el 80 % de varianza, esto con el fin de reducir de forma sustancial la cantidad de variables involucradas sin perder demasiada información.
- El escalamiento aplicado a cada variable fue la estandarización (sección 3.3.1).

5.3.1. Métodos de agrupamiento

Con la finalidad de elegir el mejor número de centroides que representa a las bases de datos se realizó una evaluación basada en los índices de *Davies-Bouldin* y de *Silhouette*.

La tabla 5.3 muestra los centroides con los que se obtuvieron los mejores índices para todas las configuraciones de entrada.

Tabla 5.3: Mejores índices para cada configuración de entrada de los datos de *TRE2D*.

Configuración	**Número de centroides**	**Índice Davies-Bouldin**	**Índice Silhouette**
Transformada y sin vecinos	2 - 3	0,93 - 0,99	0,42 - 0,32
Sin transformar y sin vecinos	2 - 3	0,88	0,33 - 0,27
Transformada y con vecinos	2 - 3	1,12 - 1,38	0,12 - 0,1
Sin transformar y con vecinos	2 - 3	0,96 - 1,6	0,28 - 0,14

5.3.2. *SOM*

Para esta metodología se realizó un análisis cualitativo que involucró los siguientes puntos:

- los errores de cuantización y topológico,
- la estructura de la *u-matrix*, y
- la distribución del *hit-map*.

con las siguientes variables sujetas a la validación:

- dimensiones del mapa auto-organizado,
- número de épocas,
- coeficiente de vecindad, y
- tasa de aprendizaje.

La tabla 5.4 muestra los mejores resultados a partir de ensayos realizados. A continuación se realizará una descripción de cada uno:

Tabla 5.4: Mejores resultados de los parámetros de los *SOM* para los datos de las *TRE2D*, obtenidos a partir de ensayos experimentales.

Variable	**Errores**	**u-matrix**	**hit-map**
Dimensiones	Mayores a 3×3	Mayores a 4×4	Menores a 6×6
Épocas		100 iteraciones	
Vecindad	50 %	30 %	30 %
Aprendizaje	0,1	0,1	0,1

1. Dimensiones:

   - respecto a los errores, mapas con dimensiones mayores a 3×3 tienden a disminuir de forma considerable sus errores, esto debido a que hay una mayor cantidad de neuronas para caracterizar la base de datos;
   - respecto a la *u-matrix* y al *hit-map*, una red de 4×4 muestra una distribución más homogénea de los grupos de neuronas, así como de los aciertos durante la clasificación.

2. Épocas: usar más de 100 iteraciones tiende a aumentar el error topológico en la red de manera considerable, pudiendo alcanzar valores mayores al 50 %. Así se mantienen los errores de cuantización y topológico por debajo del 10 % y del 20 % respectivamente, de esta forma los mapas resultantes tienden a estar mejor estructurados.

3. Vecindad:

   - respecto a los errores, comenzar a actualizar a partir del 50 % de la red los disminuye de forma considerable;
   - respecto a la *u-matrix* y al *hit-map*, con la finalidad de distribuir de forma más homogenea los grupos de neuronas se puede ajustar el radio de vecindad al 30 %.

4. Aprendizaje: usar una taza de aprendizaje de 0,1 disminuye los errores de cuantización y topológico, comparado con el uso de una taza más alta. Además mejora la distribución de las neuronas de acuerdo con la *u-matrix* y el *hit-map*.

5.3.3. Interpretación

Para elegir la mejor imagen se realizó una evaluación efectuada por intérpretes con experiencia en el método. Los aspectos evaluados fueron:

- nivel de detalle: visualización de elementos que puedan definir zonas de transición entre estructuras;
- definición de anomalías: las estructuras conformadas por los elementos del modelo generado presentan geometrías simples y generales;
- paleta de colores: los colores del modelo generado presentan un contraste que facilita la visualización; y
- estructura de los grupos o neuronas: los centroides o el *SOM*, según sea el caso, presenta una estrutura homogénea.

A continuación se muestran las imágenes que exhiben los resultados más representativos.

Tamazulápam

La figura 5.5 muestra un agrupamiento aplicado a la *TRE2D* realizada en la localidad de Tamazulápam, Oaxaca. Los parámetros usados fueron:

- entrenamiento y agrupación usando los datos de resistividad de los dos arreglos electródicos (Wenner-Schlumberger y dipolo-dipolo) sin considerar vecinos y aplicando transformación logarítmica;
- k-medianas con tres centroides; y
- las configuraciones anteriores no resultan en los mejores índices calculados (tabla 5.3), obteniendo 0,99 y 0,32 para el *índice de Davies-Bouldin* y para el *índice de Silhouette* respectivamente, sin embargo esta configuración presenta un resultado que resulta lógico para el fenómeno analizado.

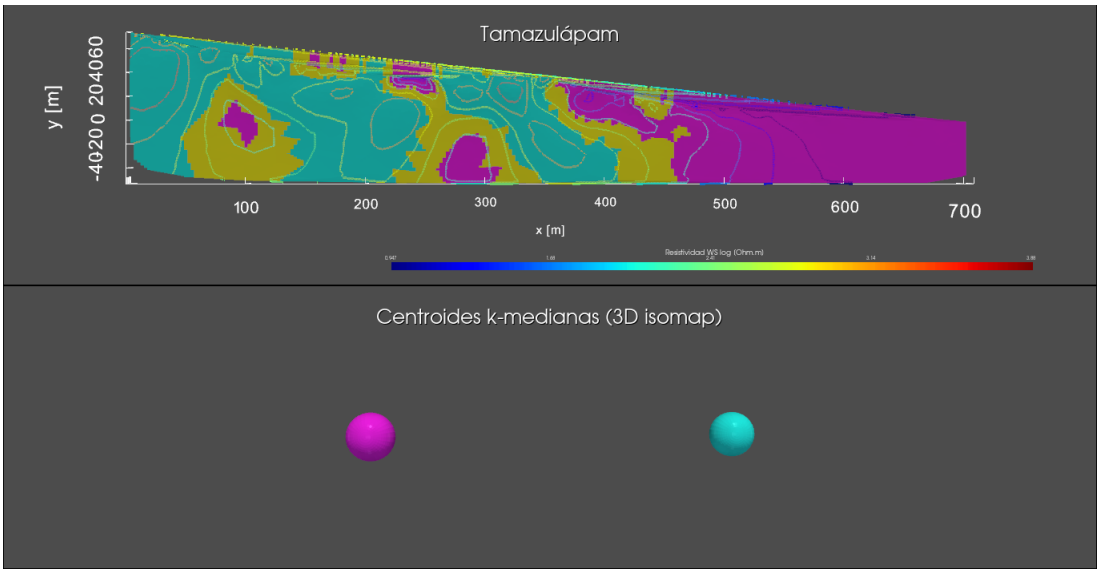


Figura 5.5: Agrupamiento aplicado a la *TRE2D* de Tamazulápam usando *k-medianas*.

Respecto a la validación cualitativa:

- Nivel de detalle: en este método geofísico se busca delimitar las estructuras con la geometría más simple, sin buscar demasiado detalle en la transición de una estructura a otra, el algoritmo de k-medianas es capaz de generalizar el comportamiento del conjunto de entrenamiento de forma que la imagen agrupada exhibe, en este caso, las estructuras principales que se pueden interpretar a partir de los datos *invertidos*.

- Definición de anomalías: el agrupamiento resultante preserva las geometrías generales con presencia de continuidad.

- Paleta de colores: los colores presentes son magenta, cyan y amarillo, la distribución de los tonos produce una imagen con suficiente contraste para ser interpretada.

- estructura de los centroides: los centroides ocupan un lugar bien definido en el espacio de atributos, cumpliendo su función al describir el comportamiento de la base de datos agrupada.

La aplicación del algoritmo de k-medianas da como resultado una imagen simple y fácil de interpretar, pudiendo definir de forma clara las posibles estructuras anómalas. Este resultado puede ser más útil que el homólogo obtenido usando *SOM*, ya que en este caso se busca un agrupamiento más general, además de que el número .

La interpretación de los centroides y grupos es:

- color magenta: representa las estructuras que se comportan como cuerpos de resistividad baja, llama la atención la geometría del contacto en la zona derecha, estructura que es dificil de interpretar observando únicamente las imágenes *invertidas*;
- color amarillo: estructura de transición, caracterizada por las resistividades medias, posiblemente intemperizada y con un importante grado de saturación;
- color cyan: estructuras definidas por resistividades altas, define a una litología marcadamente diferente que la explicada por los elementos en color magenta.

De acuerdo con los antecedentes geológicos y estudios previos de perfiles magnetométricos (Velasco Lindero, 2021), los elementos color cyan corresponden a Andesita - Brecha andesítica y los elementos color magenta corresponden a aluvión. En conclusión, se pudo estimar la ubicación del contacto litológico usando esta herramienta de agrupamiento.

Calpulalpan

La figura 5.6 muestra un agrupamiento aplicado a la *TRE2D* realizada en la localidad de Calpulalpan, Tlaxcala. Los parámetros usados fueron:

- entrenamiento y agrupación usando los datos de resistividad de los dos arreglos electródicos (Wenner-Schlumberger y dipolo-dipolo) sin considerar vecinos y aplicando transformación logarítmica;
- *SOM* hexagonal de 4×4, obtenida a partir de 5 redes de 4×4 con agrupación de neuronas usando *k-medoides* en tres grupos;

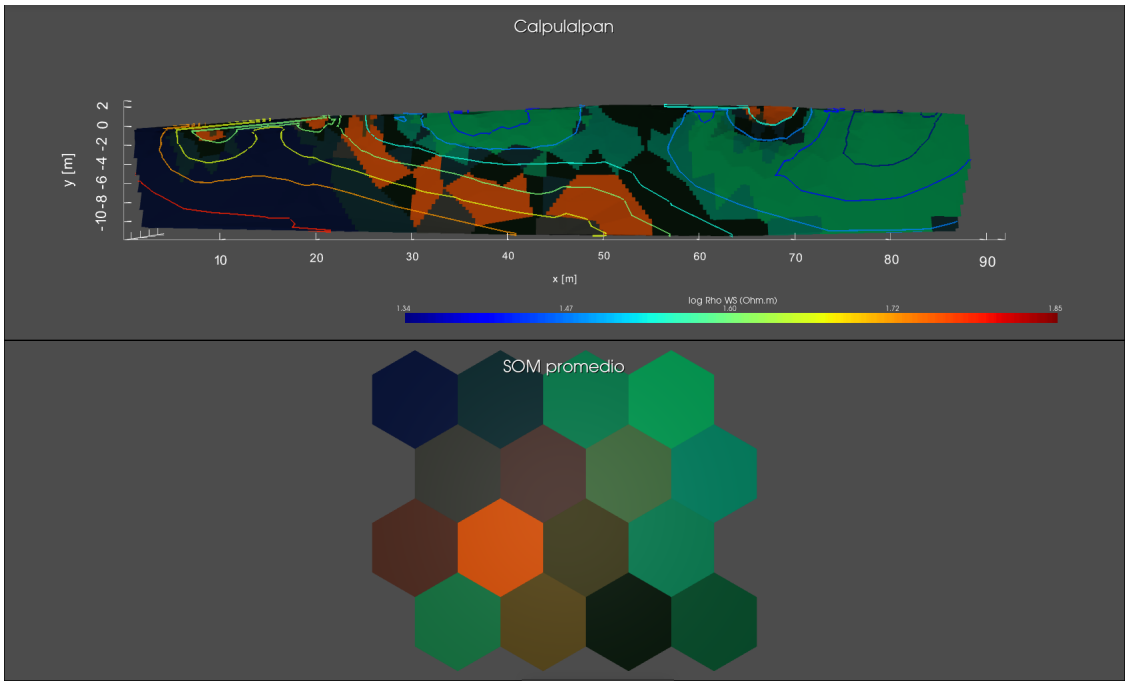


Figura 5.6: Agrupamiento aplicado a la *TRE2D* de Calpulalpan usando *SOM* hexagonal.

- el error topológico se situó entre el 30 % y 40 %, mientras que el error de cuantización se mantuvo cerca de 0,1; y
- el porcentaje de uso de la red fue superior al 95 %.

Respecto a la validación cualitativa:

- Nivel de detalle: en esta imagen se pueden interpretar estructuras de transición, marcando discontinuidades que pueden ser de importancia para interpretar.
- Definición de anomalías: el agrupamiento resultante muestra cambios tanto grandes como sutiles en las estructuras principales, pudiendo facilitar la interpretación.
- Paleta de colores: existen tonos verdes, naranjas y azules, la distribución de los tonos produce una imagen con suficiente contraste para ser interpretada.

- Estructura de las neuronas: los centroides ocupan un lugar bien definido en el espacio de atributos, cumpliendo su función al describir el comportamiento de la base de datos agrupada.

La aplicación del algoritmo de *SOM* da un resultado más complejo que el obtenido por un método de agrupamiento (*k-medias* o *k-medianas*), sin embargo en este caso el sistema delimita de mejor forma la continuidad de las estructuras.

La interpretación del *SOM* es:

- tonos verdes: son los más abundantes, atribuyéndolos a las estructuras de baja resistividad, los tonos más oscuros corresponden a zonas de transición, pudiendo entonces delimitar estructuras indpendientes, separadas por una interfaz ligeramente diferente;
- tonos naranja: elementos que forman estructuras de transición de bajas a altas resistividades;
- tonos azules: siendo un solo tono en este caso, conforma las estructuras de alta resistividad.

Capítulo 6

Conclusiones

La metodología propuesta es capaz de generalizar las estructuras y patrones en las bases de datos usadas. Las imágenes resultantes presentan estructuras que resultan lógicas a la experiencia de los intérpretes, realizando entonces una interpretación automática pudiendo ser una alternativa a la *inversión conjunta*.

Los resultados del agrupamiento aplicado a la base de datos de magnetometría muestra correlación con las excavaciones realizadas en la zona de *Xalasco, Tlaxcala*, pudiendo ayudar a elegir de forma más eficiente los lugares de futuras excavaciones. Para la zona de *La Ferrería, Durango*, no se cuenta con información de excavaciones en los sectores estudiados, sin embargo se pueden observar alineaciones que resultan lógicas de acuerdo a las estructuras antropogénicas esperadas.

Los resultados del agrupamiento efectuado en la base de datos de *TRE2D* muestra estructuras que son lógicas de acuerdo con lo esperado para cada zona, pudiendo reducir el grado de incertidumbre al momento de presentar la interpretación final.

En este trabajo se presentaron las mejores imágenes, usando los parámetros descritos oportunamiente. Durante las pruebas se observó que la inclusión de valores de los n vecinos más cercanos no agrega información útil para el agrupamiento de los vectores, así mismo ecualizar el histograma de valores no ayuda a mejorar la distribución de las neuronas en el mapa auto-organizado ni en la distribución de los

centroides para estas bases de datos y con la metodología propuesta.

Los algoritmos de agrupamiento: k-medias y k-medianas, producen resultados con patrones muy generales, que pueden ser lo deseado de acuerdo al objetivo de cada levantamiento; estos algoritmos tienden a enmascarar estructuras pequeñas. Por otro lado, el algoritmo de *SOM* es capaz de detectar cambios sutiles en los patrones, pudiendo ser de interés cuando el objetivo requiere de la interpretación de estructuras someras; este algoritmo es más robusto y requiere de más poder computacional, sin embargo el agrupamiento es más versátil debido a que se distinguen *subgrupos* dentro del agrupamiento principal. Ambas metodologías son aplicables para la interpretación automática, aplicados de forma conjunta pueden reducir incertidumbre durante la toma de decisiones.

Con este sistema es posible distinguir elementos que presentan características diferentes dentro de las bases de datos estudiadas, sin embargo a estos elementos no se les ha otorgado alguna etiqueta que refleje la característica buscada, por ejemplo la resistividad eléctrica. Así mismo no se ha trabajado con datos de múltiples prospecciones realizadas en la misma zona, mezclando datos de distintas fuentes físicas. La propuesta para trabajo a futuro es:

- Realizar un procesamiento extra a los resultados de magnetometría, con la finalidad de que el sistema sea capaz de resaltar las alineaciones en las estructuras de interés. Esto puede realizarse usando algoritmos de procesamiento digital de imágenes y herramientas de aprendizaje computacional.

- Realizar un procesamiento extra a los resultados de *TRE2D*, con la finalidad de calibrar el mapa auto-organizado de forma que cada neurona represente un valor de resistividad, para entonces etiquetar cada vector con base en un promedio pesado de las resistividades de las *BMU* para cada red entrenada.

- Crear un sistema autónomo que pueda manipular modelos de varios métodos. Dicha manipulación dependerá de la experiencia aprendida por el sistema para encontrar el mejor modelo, con fundamento físico, que explique los resultados de cada prospección, siendo entonces la versión automática de un *intérprete experto*. Para esto es necesario realizar múltiples pruebas con modelos sintéticos, de forma que puedan definirse las mejores reglas para el desempeño del sistema.

Anexo A

Filtrado *adaptativo* de las mallas de magnetometría

El ojetivo de la metodología de filtrado propuesta en este trabajo es mejorar la calidad de los datos que son entrada a los algoritmos de aprendizaje computacional, retirando efectos asociados a altas y medias frecuencias, generalmente ruido, y de artefactos no deseados presentes durante la adquisición de los datos.

La metodología empleada es *adaptativa* en el sentido de que a partir de la transformada de ondícula se obtiene una representación en distintos niveles, cuyas características dependen en la forma en que la ondícula usada se *comprime* y desplaza a lo largo de la señal de interés. Para mejorar esta metodología se propone un sistema que sea capaz de elegir automáticamente los pesos para atenuar cada nivel de la señal. Para el caso de la base de datos de magnetometría se busca eliminar efectos asociados a la adquisición de los datos, y a los artefactos relacionados con elementos metálicos independientes del interés arqueológico, efectos normalmente asociados a frecuencias medias y altas.

Se espera que mediante la combinación de las dos metodologías *básicas*: filtro de ventana y filtro de malla completa, se obtengan datos cuya representación en un mapa de contornos muestre una continuidad más clara en los valores de campo, hablando de dipolos, y la atenuación de elementos no deseados.

Retomando los parámetros mencionados en la sección 4.2:

1. Columnas: campo magnético en sensor inferior, campo magnético en sensor superior y el gradiente vertical.
2. Ondícula: Daubechies de cuarto orden.
3. Componentes: 3.

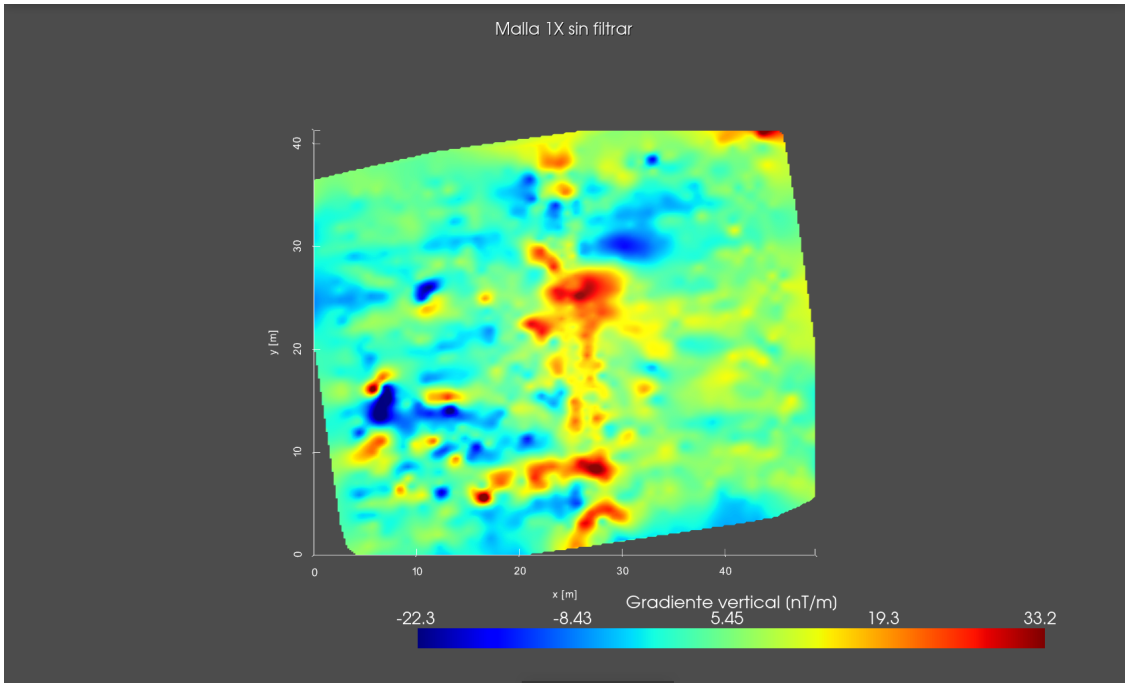
Para el punto 1, se filtran todas las variables antes de someterse al realce de anomalías, el gradiente vertical no recibe más procesamiento.

Respecto al punto 2, se realizaron pruebas usando distintos órdenes de la ondícula de Daubechies, se eligió el cuarto orden debido a que su geometría es similar a los dipolos magnéticos.

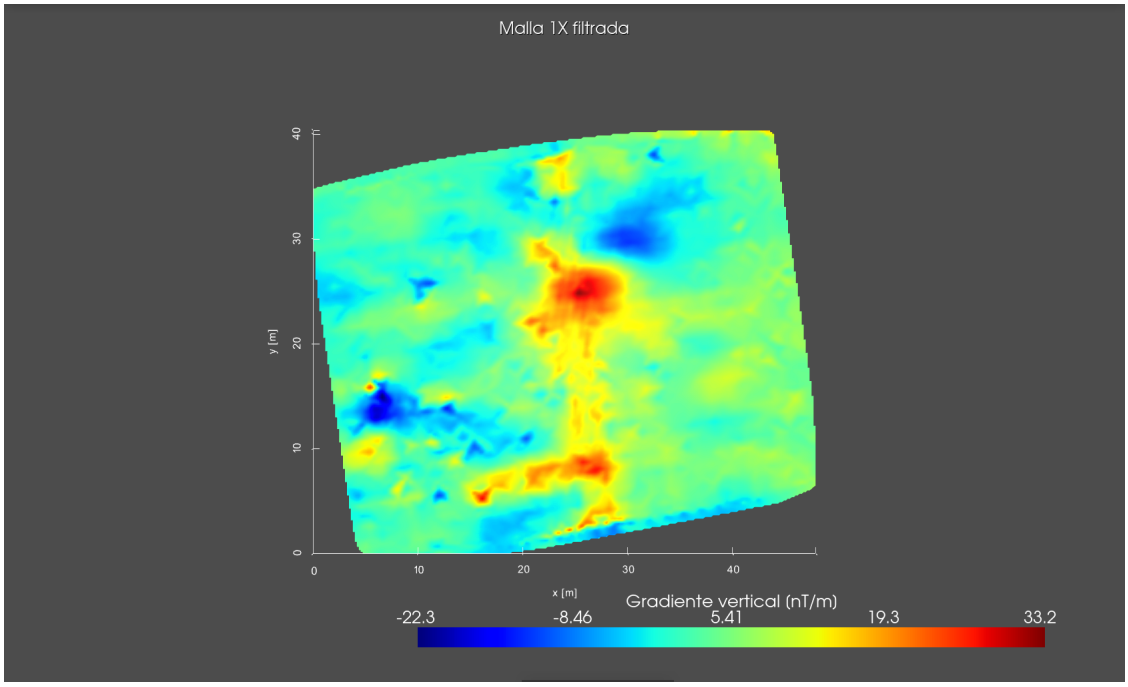
Finalmente para el punto 3, se seleccionaron tres componentes porque niveles mayores a este número no parecen contener información útil, obteniendo una imagen prácticamente constante.

Las figuras A.1 y A.2 muestran la comparación de las imágenes filtrada y sin filtrar de las mallas 1X y 2X respectivamente. Ambas fueron filtradas de la misma forma, tanto el tipo de filtro como los pesos usados. Para estos datos se usó únicamente el filtrado de malla completa, ya que no se aprecia un artefacto no deseado dentro del área levantada, es decir una region con valores de campo atípicamente altos o bajos. Se buscó reducir en su mayoría la información del tercer nivel, asociado a componentes no deseados, equivalentes a las altas frecuencias; se atenuó el nivel dos, esperando una imagen con discontinuidades menos marcadas; finalmente se mantuvo intacto el primer nivel y la componente de aproximación.

La figura A.3 muestra la comparación de las imágenes filtrada y sin filtrar de la malla 2F. Se usó el filtro de malla completa, ya que el artefacto situado en el extremo noreste puede ser una anomalía de interés. Se buscó reducir en su mayoría la

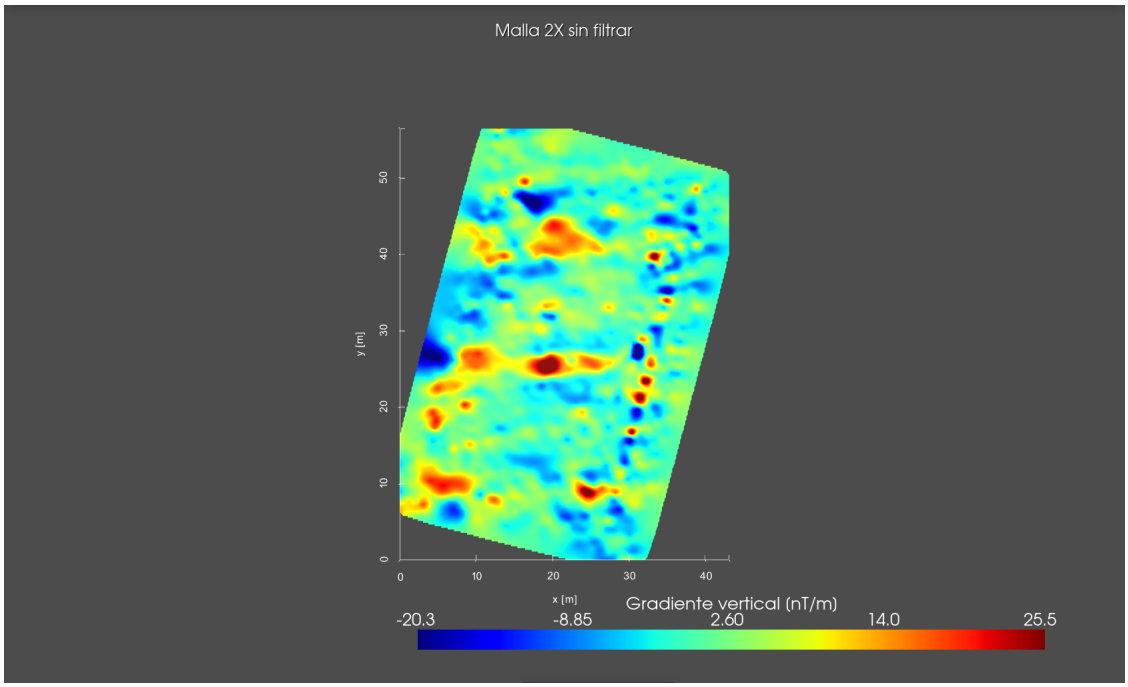


(a) Gradiente vertical sin filtrar.

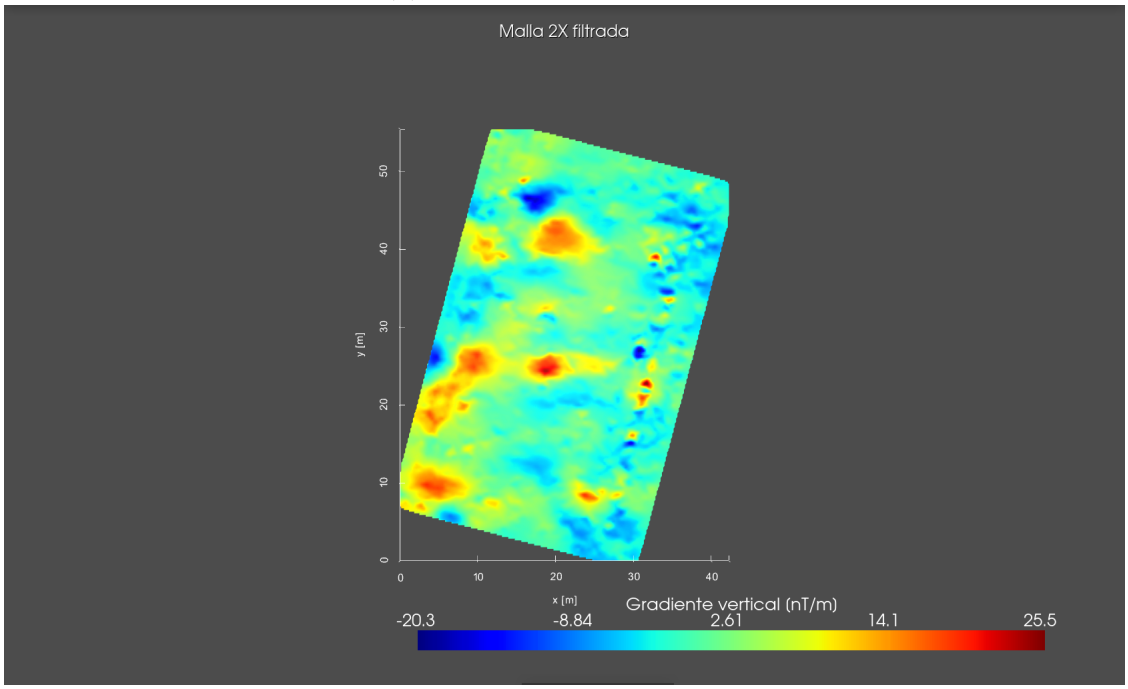


(b) Gradiente vertical filtrado.

Figura A.1: Comparación entre la versión filtrada y sin filtrar de la malla 1X.



(a) Gradiente vertical sin filtrar.

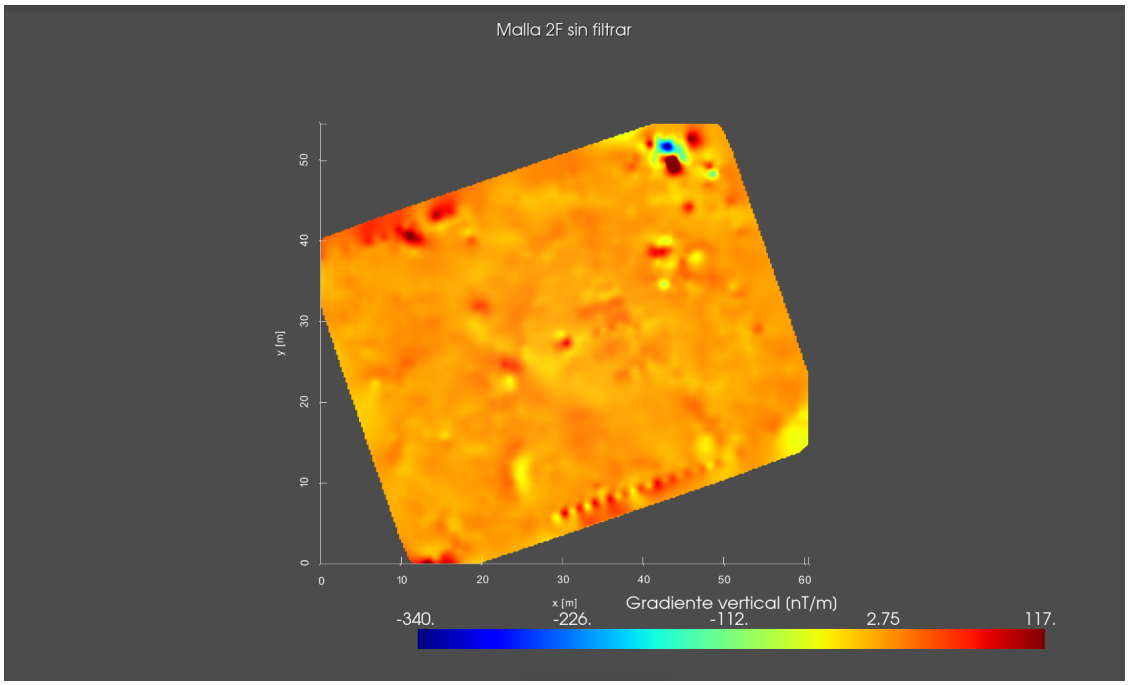


(b) Gradiente vertical filtrado.

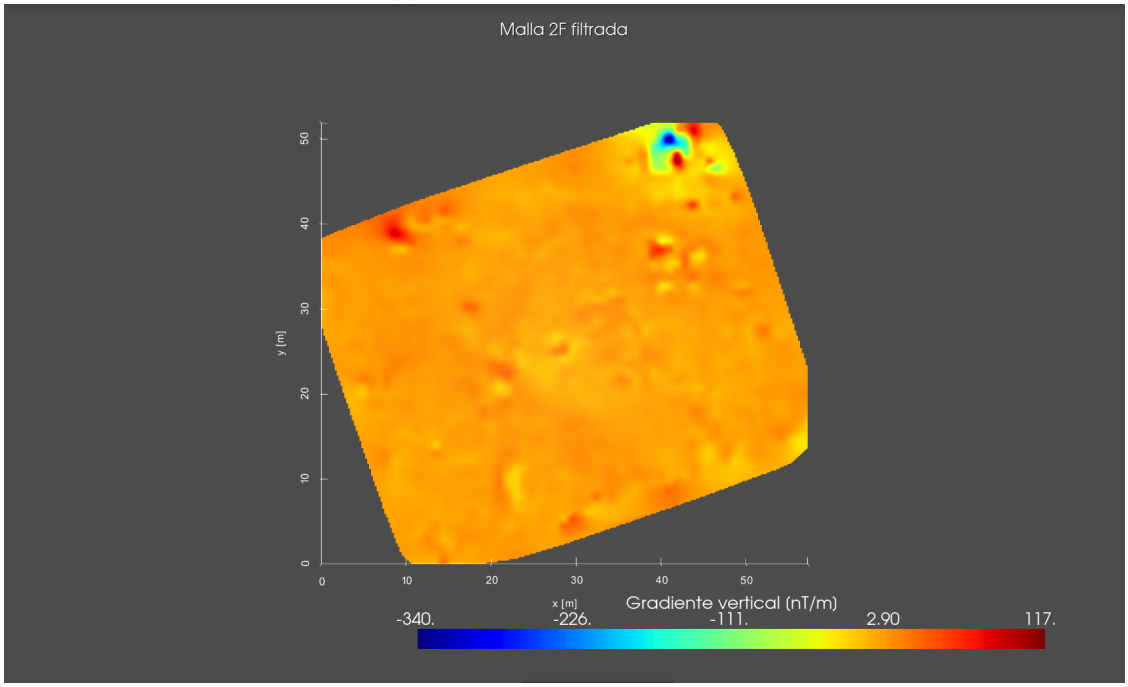
Figura A.2: Comparación entre la versión filtrada y sin filtrar de la malla 2X.

información del tercer nivel, asociado a componentes no deseados, equivalentes a las altas frecuencias; se atenuó el nivel dos, esperando una imagen con discontinuidades menos marcadas; se aplicó una atenuación al primer nivel, procurando suavizar un poco más la imagen resultante; se mantiene completamente la componente de aproximación.

La figura A.4 muestra la comparación de las imágenes filtrada y sin filtrar de la malla 4F. Se usó el filtro de ventana seguido del filtro de malla completa, ya que el artefacto situado en el centro-este no es deseado. En la primer etapa del filtro se buscó eliminar la mayor parte de los niveles bajos y eliminar la componente de aproximación; a continuación se realiza el filtro de malla completa, atenuando los niveles altos de la misma forma que para la imágen A.3.

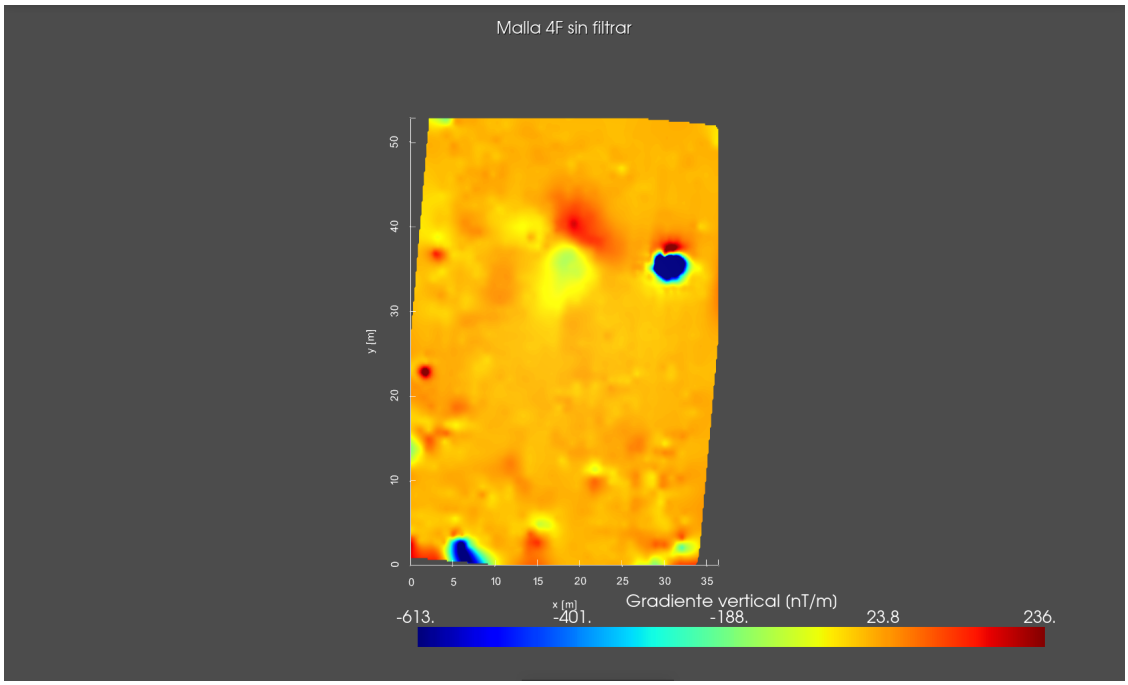


(a) Gradiente vertical sin filtrar.

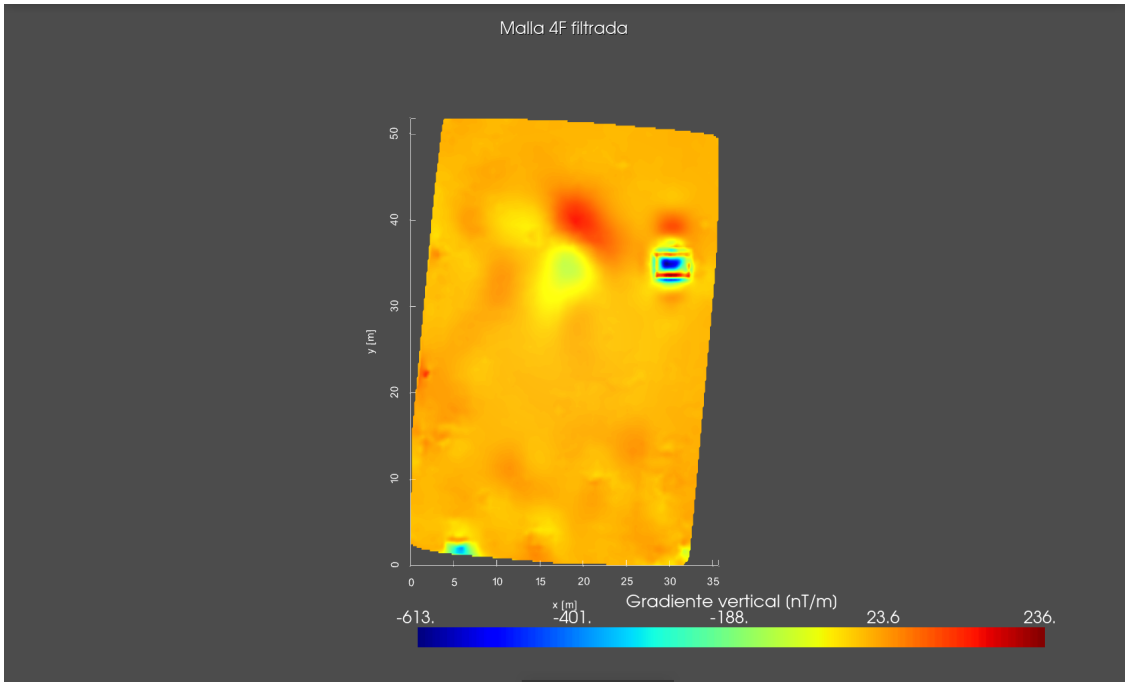


(b) Gradiente vertical filtrado.

Figura A.3: Comparación entre la versión filtrada y sin filtrar de la malla 2F.



(a) Gradiente vertical sin filtrar.



(b) Gradiente vertical filtrado.

Figura A.4: Comparación entre la versión filtrada y sin filtrar de la malla 4F.

Anexo B

Filtrado *adaptativo* de las *TRE2D* y generación de los modelos sintéticos

El ojetivo de la metodología de filtrado propuesta en este trabajo es mejorar la calidad de los datos que son entrada a los algoritmos de aprendizaje computacional, retirando efectos asociados a altas y medias frecuencias. Para el caso de la base de datos de TRE2D, se busca eliminar efectos asociados a la adquisición de los datos.

Se espera que mediante la combinación de las dos metodologías básicas: transformada de ondícula y detección de anomalías, se obtengan series de datos (representadas por nivel de la TRE2D) que simultáneamente preserven una tendencia general manteniendo los contrastes de resistividad que puedan definir una anomalía geoeléctrica.

Retomando los parámetros mencionados en la sección 4.3:

1. Variable a filtrar usando la DWT: Diferencia de potencial V_{MN}.
2. Ondícula: Daubechies de segundo orden.
3. Descomposiciones para la DWT: 5.

4. Recursiones para la DWT: 1.
5. Variable para detección de anomalías: Resistividad aparente ρ_a.
6. Algoritmo de detección de anomalías: IF.
7. Umbral de nivel de anomalía: $> 0{,}9$.

Para el punto 1, se filtra la variable registrada por el resistivímetro.

Respecto al punto 2, se realizaron pruebas usando distintos órdenes de la ondícula de Daubechies, se eligió el segundo orden porque con este se obtuvieron los mejores resultados, esperando una señal que mantenga la tendencia de la original.

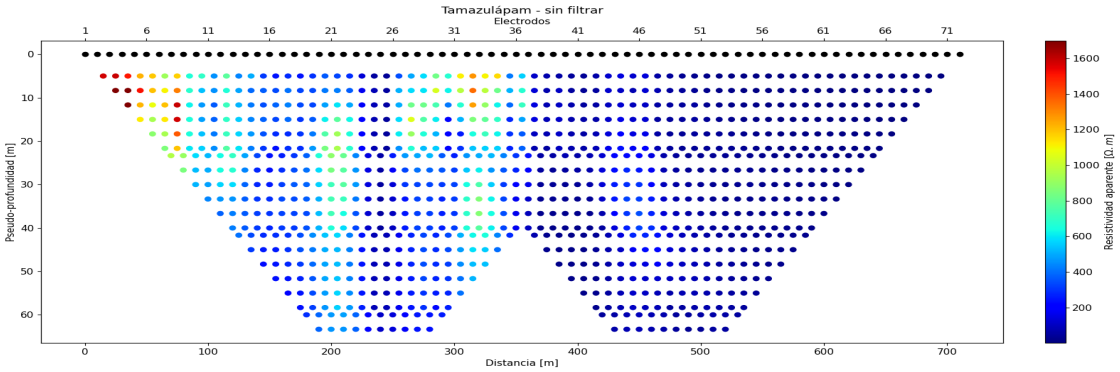
Se eligieron cinco descomposiciones, de acuerdo con el punto 3, ya que niveles mayores exhiben señales sin variaciones significativas.

Para el punto 4, una recursión es suficiente para filtrar este tipo de datos usando la metodología implementada.

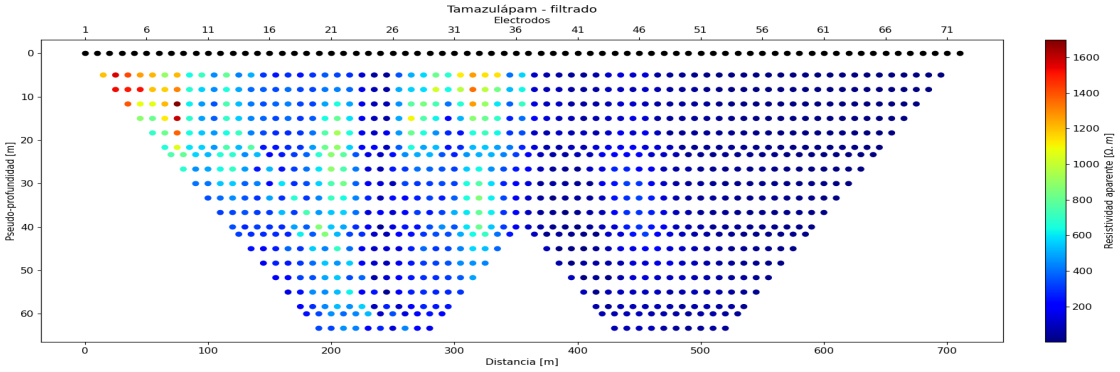
Respecto al punto 5, la variable analizada por los algoritmos de detección de anomalías es la resistividad aparente, ya que es el parámetro interpretable.

El algoritmo aplicado fue bosques de aislamiento, de acuerdo con el punto 6. No hay mucha diferencia entre el desempeño de los algoritmos usados en este punto, la elección del umbral de nivel de anomalía (punto 7) se eligió recordando que los valores cercanos a 1 se consideran como atípicos de acuerdo al algoritmo de bosques de aislamiento 2.3.2.

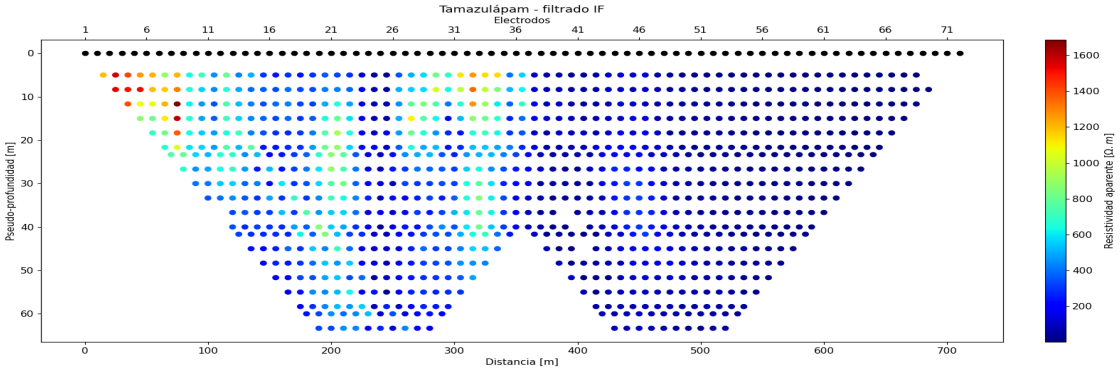
La figura B.1 muestra la comparación entre las imágenes sin filtrar, filtrada y aplicando el algoritmo de detección de anomalías posterior a la metodología de filtrado. Se buscó mantener la tendencia general de cada nivel de la *TRE2D* en cuestión con los siguientes pesos: $[0{,}1, 0{,}3, 0{,}5, 0{,}7, 0{,}9]$.



(a) Pseudosección sin filtrar.

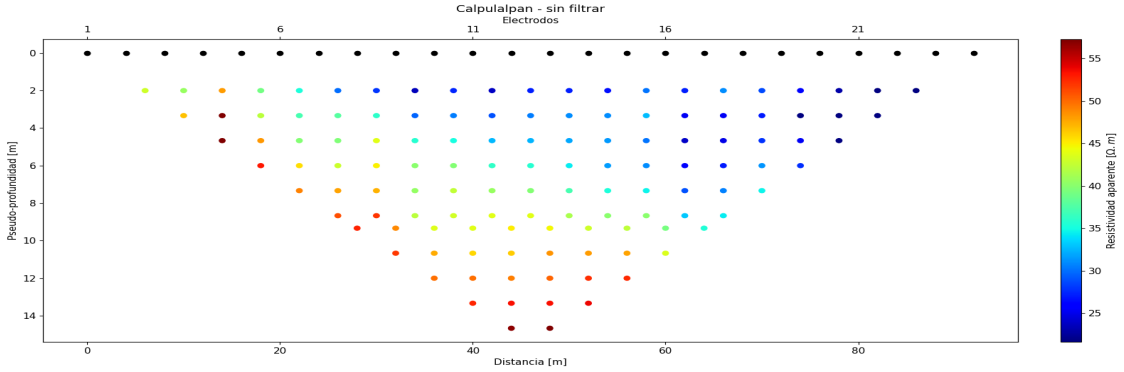


(b) Pseudosección filtrada usando la transformada de ondícula.

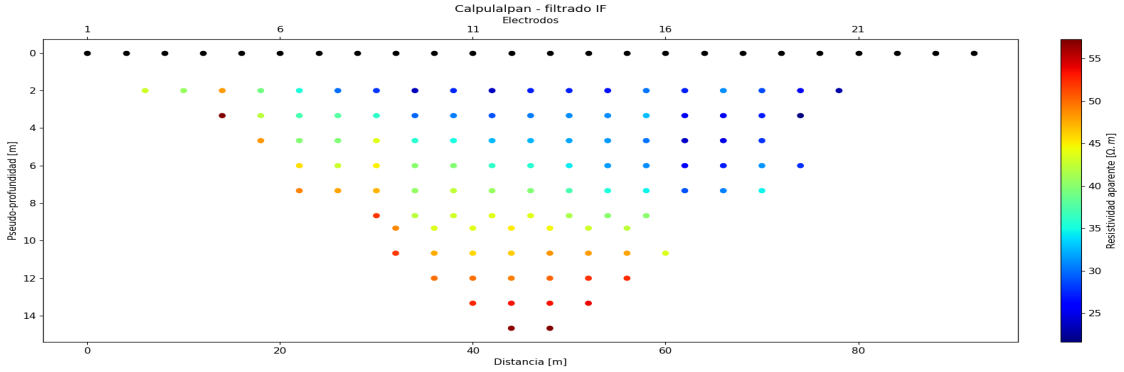


(c) Pseudosección filtrada mediante la transformada de ondícula y posterior remoción de vectores atípicos.

Figura B.1: Comparación entre pseudosecciones de resistividad aparente de la TRE2D de Tamazulápam, Oaxaca.



(a) Pseudosección sin filtrar.



(b) Pseudosección retirando los vectores atípicos.

Figura B.2: Comparación entre pseudosecciones de resistividad aparente de la TRE2D de Calpulalpan, Tlaxcala.

La figura B.2 muestra la comparación entre las imágenes sin filtrar y la resultante de retirar los vectores atípicos. En este caso no se aplicó la metodología de filtrado usando la transformada de ondícula, ya que no se aprecia ruido significativo en los datos.

Respecto a los modelos sintéticos el desarrollo es el siguiente:

1. Generación de la malla y regiones para la generación de n modelos sintéticos usando la biblioteca ResIPy (Blanchy et al., 2020). Las figuras B.3 y B.4 muestran los modelos base para las bases de datos de Tamazulápam y Calpulalpan, respectivamente.

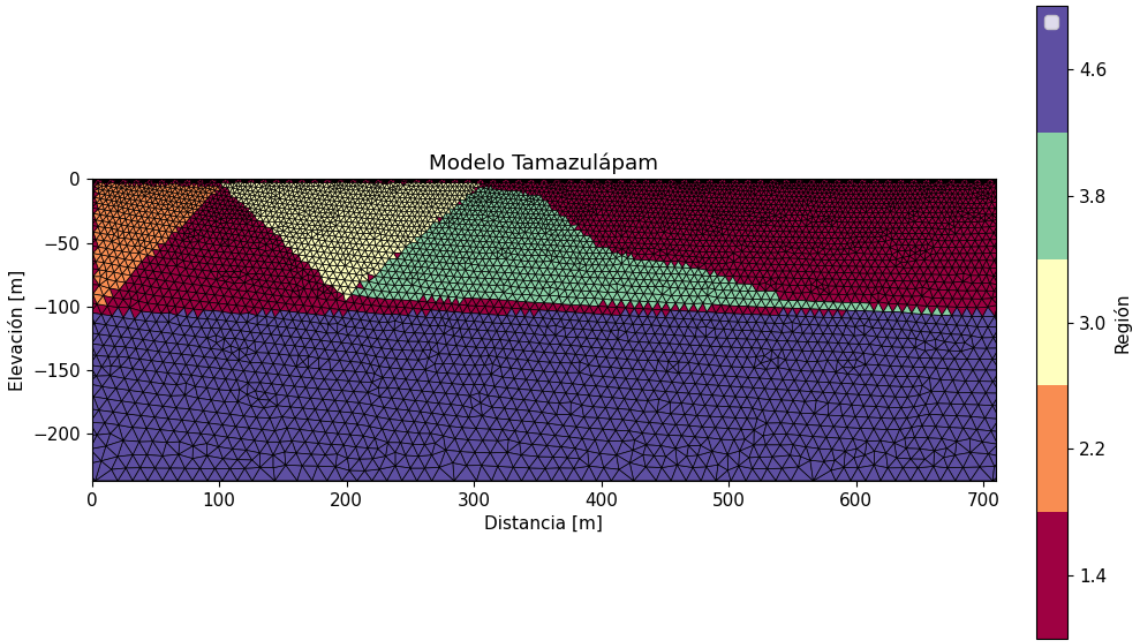


Figura B.3: Modelo base para los modelos sintéticos de Tamazulápam, Oaxaca.

2. Se programan las secuencias de lectura para los arreglos electródicos deseados, mismos con los que se genenrarán los modelos sintéticos. Para ambos modelos se programaron secuencias Wenner-Schlumberger y dipolo-dipolo con todos sus posibles cuadripolos, para el modelo de la figura B.3 se definieron 72 electrodos, mientras que el mostrado en la figura B.4 fueron 24; coincidentes en ambos casos a las secuencias realizadas durante cada levantamiento.

3. Se define el número de modelos y el intervalo dentro del cual se elegirán de forma aleatoria los valores de resistividad de cada región. Para ambos modelos base se crearon 10 modelos, para el modelo mostrado en la figura B.3 los intervalos fueron: $[100, 320]$, $[1000, 3200]$, $[300, 1000]$, $[300, 1000]$, $[30, 100]$, $[10, 30]]$, mientras que el por presentado por la figura B.4 fueron: $[[30, 40]$, $[45, 60]$, $[20, 30]]$, con unidades ωm.

4. Finalmente se seleccionan los vecinos espacialemente más cercanos, en caso de ser requerido para el entrenamiento.

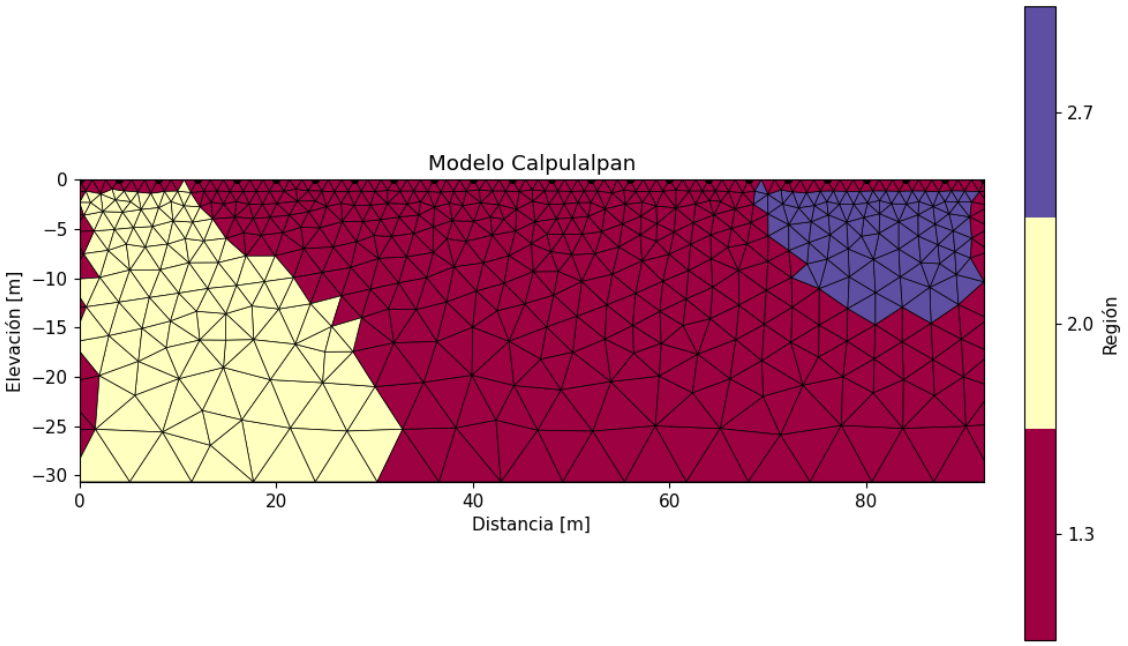


Figura B.4: Modelo base para los modelos sintéticos de Calpulalpan, Tlaxcala.

Referencias

Arango-Galván, C., Torre-González, B. D. l., Chávez-Segura, R. E., Tejero-Andrade, A., Cifuentes-Nava, G., and Hernández-Quintero, E. (2011). Structural pattern of subsidence in an urban area of the southeastern mexico basin inferred from electrical resistivity tomography. *Geofísica internacional*, 50(4):401–409.

Bachri, I., Hakdaoui, M., Raji, M., and Benbouziane, A. (2020). Geological mapping using random forests applied to remote sensing data: a demonstration study from msaidira-souk al had, sidi ifni inlier (western anti-atlas, morocco). In *2020 IEEE International conference of Moroccan Geomatics (Morgeo)*, pages 1–5. IEEE.

Balasubramanian, M. and Schwartz, E. L. (2002). The isomap algorithm and topological stability. *Science*, 295(5552):7–7.

Bedini, E. (2009). Mapping lithology of the sarfartoq carbonatite complex, southern west greenland, using hymap imaging spectrometer data. *Remote Sensing of Environment*, 113(6):1208–1219.

Bishop, C. M. and Nasrabadi, N. M. (2006). *Pattern recognition and machine learning*, volume 4. Springer.

Black, G. A. and Johnston, R. B. (1962). A test of magnetometry as an aid to archaeology. *American Antiquity*, 28(2):199–205.

Blakely, R. J. (1996). *Potential theory in gravity and magnetic applications*. Cambridge university press.

Blanchy, G., Saneiyan, S., Boyd, J., McLachlan, P., and Binley, A. (2020). ResIPy,

an intuitive open source software for complex geoelectrical inversion/modeling. page 104423.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., and Sander, J. (2000). Lof: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 93–104.

Butler, D. K., Bennett, H. H., and Ballard, J. H. (2006). Overview of multimethod geophysical system development for enhanced near-surface target detection, discrimination, and characterization. *The Leading Edge*, 25(3):352–356.

Carneiro, C. d. C., Fraser, S. J., Crósta, A. P., Silva, A. M., and Barros, C. E. d. M. (2012). Semiautomated geologic mapping using self-organizing maps and airborne geophysics in the brazilian amazon. *Geophysics*, 77(4):K17–K24.

Carrasco, A. A., Andrade, A. T., and González, A. L. (2021). Concepto de anomalía en la exploración geofísica.

Chen, L., Wang, L., Miao, J., Gao, H., Zhang, Y., Yao, Y., Bai, M., Mei, L., and He, J. (2020). Review of the application of big data and artificial intelligence in geology. In *Journal of Physics: Conference Series*, volume 1684, page 012007. IOP Publishing.

Costa, I. S. L., Tavares, F. M., and de Oliveira, J. K. M. (2019). Predictive lithological mapping through machine learning methods: a case study in the cinzento lineament, carajás province, brazil. *Journal of the Geological Survey of Brazil*, 2(1):26–36.

Cracknell, M., Reading, A., and McNeill, A. (2014). Mapping geology and volcanic-hosted massive sulfide alteration in the hellyer–mt charter region, tasmania, using random forests™ and self-organising maps. *Australian Journal of Earth Sciences*, 61(2):287–304.

Cracknell, M. J. and Reading, A. M. (2013). The upside of uncertainty: Identification of lithology contact zones from airborne geophysics and satellite data using random forests and support vector machines. *Geophysics*, 78(3):WB113–WB126.

de Oviedo, U. (2009). Correlación de pearson. https://web.archive.org/web/20091215105427/http://www.psico.uniovi.es/Dpto_Psicologia/metodos/tutor.6/fcope.html. Accedido: 2022-08-20.

Edwards, T. (1991). Discrete wavelet transforms: Theory and implementation. *Universidad de*, pages 28–35.

Fassbinder, J. W. (2017). Magnetometry for archaeology. *Encyclopedia of geoarchaeology*, pages 499–514.

Flach, P. (2012). *Machine learning: the art and science of algorithms that make sense of data*. Cambridge university press.

Gaffney, C. (2008). Detecting trends in the prediction of the buried past: a review of geophysical techniques in archaeology. *Archaeometry*, 50(2):313–336.

Gallardo, L. A. and Meju, M. A. (2004). Joint two-dimensional dc resistivity and seismic travel time inversion with cross-gradients constraints. *Journal of Geophysical Research: Solid Earth*, 109(B3).

Ghimire, B., Rogan, J., and Miller, J. (2010). Contextual land-cover classification: incorporating spatial dependence in land-cover classification models using random forests and the getis statistic. *Remote Sensing Letters*, 1(1):45–54.

GISI (2018). Transformada wavelet.

Goncalves, M., Netto, M., Costa, J., and Zullo Junior, J. (2008). An unsupervised method of classifying remotely sensed images using kohonen self-organizing maps and agglomerative hierarchical clustering methods. *International Journal of Remote Sensing*, 29(11):3171–3207.

Gonzalez, R. C. (2009). *Digital image processing*. Pearson education india.

Google (2020). Embeddings. https://developers.google.com/machine-learning/crash-course/embeddings/video-lecture. Accedido: 2022-06-24.

Grant, F. S. (1972). Review of data processing and interpretation methods in gravity and magnetics, 1964–71. *GEOPHYSICS*, 37.

Halkidi, M., Batistakis, Y., and Vazirgiannis, M. (2001). On clustering validation techniques. *Journal of intelligent information systems*, 17(2):107–145.

Harris, J. and Grunsky, E. C. (2015). Predictive lithological mapping of canada's north using random forest classification applied to geophysical and geochemical data. *Computers & geosciences*, 80:9–25.

Huang, C. and Shibuya, A. (2020). High accuracy geochemical map generation method by a spatial autocorrelation-based mixture interpolation using remote sensing data. *Remote Sensing*, 12(12):1991.

IAGA (2019). International geomagnetic reference field. https://www.ngdc.noaa.gov/IAGA/vmod/igrf.html. Accedido: 2022-06-07.

Jimenez, L. O. and Landgrebe, D. A. (1998). Supervised classification in high-dimensional space: geometrical, statistical, and asymptotical properties of multivariate data. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 28(1):39–54.

Jolliffe, I. (2005). Principal component analysis. *Encyclopedia of statistics in behavioral science*.

Juárez, K., López-García, P., Argote-Espino, D. L., Tejero-Andrade, A., Chávez, R. E., and García-Serrano, A. (2017). Magnetic and electrical prospections in the archaeological site of xalasco northeast of tlaxcala, mexico. *Global J. Archaeol. Anthropol*, 2(2):555–581.

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480.

Kuhn, S., Cracknell, M. J., and Reading, A. M. (2018). Lithologic mapping using random forests applied to geophysical and remote-sensing data: A demonstration study from the eastern goldfields of australia. *Geophysics*, 83(4):B183–B193.

Kuhn, S., Cracknell, M. J., and Reading, A. M. (2019). Lithological mapping in the central african copper belt using random forests and clustering: Strategies for optimised results. *Ore Geology Reviews*, 112:103015.

Lamsal, I., Ghimire, S., and Acharya, K. K. (2020). Geological and geophysical study in udheri khola area, nalgad hydroelectric project, jajarkot district, lesser himalaya, western nepal. *Bulletin of the Department of Geology*, 22:11–16.

Lee, G. R., Gommers, R., Waselewski, F., and Wohlfahrt, K. (2019). Pywavelets: A python package for wavelet analysis. *Journal of Open Source Software*, 4(36):1237.

Liu, F. T., Ting, K. M., and Zhou, Z.-H. (2008). Isolation forest. In *2008 eighth ieee international conference on data mining*, pages 413–422. IEEE.

Loke, M. H. (2004). Tutorial: 2-d and 3-d electrical imaging surveys.

Mallat, S. (1999). *A wavelet tour of signal processing*. Elsevier.

Markou, M. and Singh, S. (2003). Novelty detection: a review—part 1: statistical approaches. *Signal processing*, 83(12):2481–2497.

Márquez, E. F., Chávez, R., Serrano, R. M., Barrientos, J. H., Andrade, A. T., and Belmonte, S. (2001). Geophysical characterization of the etla valley aquifer, oaxaca, mexico. *Geofísica internacional*, 40(4):245–257.

Mitchell, T. M. (1997). *Machine learning*, volume 1. McGraw-hill New York.

Molino-Minero-Re, E., Rubio-Acosta, E., Benítez-Pérez, H., Brandi-Purata, J. M., Pérez-Quezadas, N. I., and García-Nocetti, D. F. (2018). A method for classifying pre-stack seismic data based on amplitude–frequency attributes and self-organizing maps. *Geophysical Prospecting*, 66(4):673–687.

Montiel, N. H., Mota, S., and Neme, A. (2021). Las anomalías:¿ qué son?,¿ dónde surgen?,¿ cómo detectarlas? *Tecnología e Innovación en Educación Superior*.

Murray, T. W., P., G. L., and E., S. R. (1990). *Applied Geophysics*. Cambridge University Press.

Nathan, D., Aitken, A., Holden, E.-J., and Wong, J. (2020). Imaging sedimentary basins from high-resolution aeromagnetics and texture analysis. *Computers & Geosciences*, 136:104396.

Novikov, A. (2019). PyClustering: Data mining library. *Journal of Open Source Software*, 4(36):1230.

Okpoli, C. C. (2013). Sensitivity and resolution capacity of electrode configurations. *International Journal of Geophysics*, 2013.

Oppenheim, A. V., Willsky, A. S., Mata Hernández, G., et al. (1998). *Señales y sistemas*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Redhaounia, B., Ilondo, B. O., Gabtni, H., Sami, K., and Bédir, M. (2016). Electrical resistivity tomography (ert) applied to karst carbonate aquifers: case study from amdoun, northwestern tunisia. *Pure and Applied Geophysics*, 173(4):1289–1303.

Reynolds, J. M. (2011). *An introduction to applied and environmental geophysics*. John Wiley & Sons.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

SGM (2017). Geofísica. https://www.sgm.gob.mx/Web/MuseoVirtual/Geofisica/Introduccion-geofisica.html. Accedido: 2021-11-23.

Sheriff, R. E. (2002). *Encyclopedic dictionary of applied geophysics*. Society of exploration geophysicists.

Sullivan, C. B. and Kaszynski, A. (2019). PyVista: 3D plotting and mesh analysis through a streamlined interface for the Visualization Toolkit (VTK). *Journal of Open Source Software*, 4(37):1450.

Temprano, M. A. F. (2020). Mapas autoorganizados (self-organizing maps).

Tenenbaum, J. B., Silva, V. d., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323.

Uieda, L. (2018). Verde: Processing and gridding spatial data using Green's functions. *Journal of Open Source Software*, 3(29):957.

Ultsch, A. (1990). Kohonen's self organizing feature maps for exploratory data analysis. *Proc. INNC90*, pages 305–308.

Valens, C. (1999). A really friendly guide to wavelets. *ed. Clemens Valens*.

Van Der Maaten, L., Postma, E., Van den Herik, J., et al. (2009). Dimensionality reduction: a comparative. *J Mach Learn Res*, 10(66-71):13.

Velasco Lindero, M. A. (2021). Caracterización y modelo geológico a partir de la exploración magnetométrica en san andrés lagunas, oaxaca.

Verleysen, M. and François, D. (2005). The curse of dimensionality in data mining and time series prediction. In *International work-conference on artificial neural networks*, pages 758–770. Springer.

Vettigli, G. (2018). Minisom: minimalistic and numpy-based implementation of the self organizing map.

Waske, B. and Braun, M. (2009). Classifier ensembles for land cover mapping using multitemporal sar imagery. *ISPRS journal of photogrammetry and remote sensing*, 64(5):450–457.

Yu, L., Porwal, A., Holden, E.-J., and Dentith, M. C. (2012). Towards automatic lithological classification from remote sensing data using support vector machines. *Computers & Geosciences*, 45:229–239.