
**UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO**
Posgrado en Ciencias de la Tierra
Instituto de Geofísica
Sismología

# Caracterización de la actividad del volcán Popocatépetl por medio de sus señales sísmicas

**Tesis**
QUE PARA OPTAR POR EL GRADO DE:
MAESTRÍA EN CIENCIAS DE LA TIERRA

PRESENTA:
**ALEJANDRO REYES ROMERO**

TUTOR PRINCIPAL
Dr. Denis Legrand. Instituto de Geofísica

MIEMBROS DEL COMITÉ TUTOR
Dr. Luis Antonio Dominguez. Instituto de Geofísica, UNAM
Dra. Clara Yoon. USGS, EUA

Ciudad de México, Agosto, 2024.

**PROTESTA UNIVERSITARIA DE INTEGRIDAD Y HONESTIDAD ACADÉMICA Y PROFESIONAL**
**(Graduación con trabajo escrito)**

De conformidad con lo dispuesto en los artículos 87, fracción V, del Estatuto General, 68, primer párrafo, del Reglamento General de Estudios Universitarios y 26, fracción I, y 35 del Reglamento General de Exámenes, me comprometo en todo tiempo a honrar a la Institución y a cumplir con los principios establecidos en el Código de Ética de la Universidad Nacional Autónoma de México, especialmente con los de integridad y honestidad académica.

De acuerdo con lo anterior, manifiesto que el trabajo escrito titulado:

Caracterización de la actividad del volcán Popocatépetl por medio de sus señales sísmicas

que presenté para obtener el grado de -----Maestria----- es original, de mi autoría y lo realicé con el rigor metodológico exigido por mi programa de posgrado, citando las fuentes de ideas, textos, imágenes, gráficos u otro tipo de obras empleadas para su desarrollo.

En consecuencia, acepto que la falta de cumplimiento de las disposiciones reglamentarias y normativas de la Universidad, en particular las ya referidas en el Código de Ética, llevará a la nulidad de los actos de carácter académico administrativo del proceso de graduación.

**Atentamente**

Alejandro Reyes Romero, 417083191

**(Nombre, firma y Número de cuenta de la persona alumna)**

**Con amor,**
**Para mis padres**
**Para mi hermana**
**Para mi abuelita**
**Para mi solecito**

**Witness me!**

## Agradecimientos:

A la UNAM por volver a permitirme estudiar en sus aulas y a utilizar sus instalaciones.

Al Posgrado en Ciencias de la Tierra-UNAM por la formación académica.

Al Consejo Nacional de Humanidades Ciencias y Tecnologías (CONAHCyT) y al pueblo de México por la beca otorgada para la realización de estudios de Maestría.

A la Fundación Telmex-Telcel por la beca otorgada para la realización de estudios de Maestría.

Al CENAPRED y al SSN por los datos e información proporcionados.

Le agradezco al Dr. Luis Antonio Domínguez Ramírez por el préstamo de sus GPUs así como el acceso a su computadora y por el apoyo y la paciencia.

Le agradezco al PAEP el apoyo para realizar una corta estancia de investigación en octubre de 2023 en la ciudad de Toulouse, Francia.

Le agradezco al Dr. Sebastien Chevrot y al personal del Geosciences Environnement Toulouse por la oportunidad de realizar una breve estancia de investigación con ellos.

Le agradezco al Dr. Bastien Plazolles por su apoyo para entender la metodología *de Template Matching* utilizada y por su paciencia al solucionar mis dudas.

Agradezco a los administradores de las computadoras de alto rendimiento de los Institutos de Geofísica y de Ingeniería, por su apoyo y paciencia.

Le agradezco a los profesores que tuve durante mis estudios de maestría, sus conocimientos y enseñanza fueron valiosos para mi desarrollo profesional.

Le agradezco a los miembros de mi comité tutor, la Dra. Clara Yoon y el Dr. Luis Antonio Domínguez Ramírez, por sus comentarios respecto a mi trabajo y al apoyo brindado.

Le agradezco a mi tutor, el Dr. Denis Legrand por su apoyo y guía durante el desarrollo del proyecto de tesis, así como su paciencia al momento de redactarla.

Le agradezco a mis padres, la oportunidad de hacer un posgrado y por apoyarme durante este camino.

Le agradezco a mi hermana por escucharme, por apoyarme y por su amor. Gracias Will Smith.

A Eugenia le agradezco el estar a mi lado durante el proceso de esta tesis, por darme ánimos cuando lo necesitaba y por quererme mucho. Te amote.

Le agradezco a mi familia, por apoyarme durante el desarrollo de mi tesis.

Le quiero agradecer igual a mis amigos que hice durante la maestría, Carlos, Fernando, Dayna, Gaby, Gustavo, Karina, Ali y a Karina. Por el apoyo que nos dimos para que la maestría fuese más llevadera y por las carnitas asadas.

Le quiero agradecer a mis amigos: Christian, Joel, Kuswara, Ilya, Estefania, Jasso, Gaby, Emma, Ludwig, Yess, Fany y Alicia. Por su amistad y su apoyo durante la elaboración de esta tesis.

Le quiero agradecer a los miembros de mi jurado de sinodales, conformado por el Dr. Mathieu Perton, el Dr. Carlos Miguel Valdes Gonzáles, el Dr. Javier Francisco Lermo Samaniego y el Dr. Luis Antonio Domínguez Ramírez por aceptar formar parte de la evaluación de este trabajo y por sus comentarios.

Y a ti lector, te agradezco por tu tiempo para leer este trabajo.

# Contenidos

Resumen ..... 10
Abstract ..... 12
1. Introducción ..... 14
2. Objetivo ..... 17
3. Contexto geológico ..... 18
3.1 Marco tectónico ..... 18
3.2 Historia eruptiva reciente ..... 19
4. Clasificación de señales sismo-volcánicas ..... 22
5. Datos ..... 24
5.1 Descripción de las estaciones y de los datos ..... 24
5.2 Pre-procesado de los datos ..... 27
*5.2.1 Filtrado de los datos* ..... 27
*5.2.2 Decimado de los datos* ..... 28
*5.2.3 Llenado de lagunas en los datos* ..... 28
*5.2.4 Cálculo de la energía* ..... 28
6. Metodología ..... 31
*6.1 Template Matching* ..... 32
*6.1.1 Primera etapa de TM (modo Sistemático)* ..... 34
*6.1.2 Segunda etapa de TM (Modo Estándar)* ..... 35
6.2 FAST ..... 37
*6.2.1 Generación de huellas binarias* ..... 38
*6.2.2 Búsqueda de huellas similares* ..... 39
6.3 Postprocesamiento de las detecciones ..... 41
*6.3.1 Eliminar detecciones duplicadas* ..... 41
*6.3.2 Eliminar detecciones falsas* ..... 42
6.4 Clasificación de las detecciones con el método de agrupamiento jerarquizado ..... 43
7. Resultados. ..... 47
7.1 Primera etapa de TM (Modo Sistemático) ..... 47
7.2 Segunda etapa de TM (Modo Estándar) utilizando las tres componentes ..... 53
7.3 Segunda etapa de TM utilizando la energía ..... 55
7.4 FAST ..... 58
7.5 Comparación de detecciones con las tres componentes y la energía para TM ..... 60

7.6 Comparación entre detecciones hechas por TM en energía y por FAST .....62
7.7 Clasificación de señales de TM con energía.....64
8. Discusión. .....66
8.1 Selección del mejor filtro.....66
8.2 Estadísticas sobre los *templates* obtenidos en la primera etapa.....66
8.3 Ventajas del uso de la energía .....67
8.4 Comparación con detecciones manuales .....68
8.5 Comparación entre FAST y TM .....69
8.6 Clasificación de las señales.....70
9. Conclusión .....71
Bibliografía.....73

# Índice de Figuras

Figura 1:. Ubicación de la FVTM (parte gris) y del volcán Popocatépetl (punto rojo).....19
Figura 2: Ubicación de la zona de estudio y de las estaciones utilizadas. .....24
Figura 3: Ejemplo de un registro con lagunas en los datos registrado en la estación PCP, el 20 de Julio de 2016.....25
Figura 4: Registro del enjambre sísmico del 8 de julio de 2016 en la estación PPIG .....26
Figura 5: Comparación de como un VT se observa en las diferentes bandas de frecuencia. .....27
Figura 6: Ejemplo de la señal en energía a partir de las tres señales en velocidad de las componentes NS, EW y vertical.....29
Figura 7: Diagrama de flujo del funcionamiento básico de la primera etapa de TM. .....34
Figura 8: Diagrama de flujo del funcionamiento básico de la segunda etapa. .....36
Figura 9: Etapas de FAST.....37
Figura 10: Comparación entre una detección verdadera y una detección falsa, utilizando las tres componentes (izquierda) y su energía(derecha). .....43
Figura 11: Ejemplo de dendrograma. .....45
Figura 12:  Templates encontrados por la primera etapa el 8 de julio (en rojo) .....52
Figura 13: Templates encontrados por la primera etapa el 10 de julio (en rojo). .....52

Figura 14: Ejemplos de detecciones hechas durante la segunda etapa usando las tres componentes. El template utilizado (en azul), el apilado de todas las detecciones realizadas (en rojo) y las mejores detecciones realizadas por el template (en verde). .....53

Figura 15: Detecciones realizadas por la segunda etapa con las tres componentes el 8 de julio (en azul). .....54

Figura 16: Detecciones realizadas por la segunda etapa con las tres componentes el 10 de julio (en azul). .....55

Figura 17: Ejemplos de detecciones realizadas usando energía. El Template utilizado (en azul), el apilado de todas las detecciones realizadas (en rojo) y las mejores detecciones realizadas por el template (en verde). .....56

Figura 18: Detecciones realizadas por la segunda etapa usando energía para el 8 de julio (en verde). .....57

Figura 19: Detecciones realizadas por la segunda etapa usando energía para el 10 de julio (en verde). .....57

Figura 20: Detecciones realizadas por FAST el 8 de julio (en naranja). .....58

Figura 21: Detecciones realizadas por FAST el 10 de julio (en naranja). .....59

Figura 22: Detecciones hechas con TM en velocidad y con energía el 8 de julio. Detecciones con las tres componentes (en azul). Detecciones hechas con la energía (en verde). Detecciones hechas por ambos métodos (magenta).....60

Figura 23: Detecciones hechas con TM en velocidad y con energía el 10 de julio. Detecciones con las tres componentes (en azul). Detecciones hechas con la energía (en verde). Detecciones hechas por ambos métodos (magenta).....61

Figura 24: Detecciones de TM en energía y de FAST el 8 de julio. Detecciones de Template Matching (verde), FAST (naranja). Detecciones hechas por ambos métodos (magenta). ....62

Figura 25: Detecciones de TM en energía y de FAST el 10 de julio. Detecciones de Template Matching (verde), FAST (naranja). Las detecciones hechas por ambos métodos (magenta). .....63

Figura 26: Dendrograma con las principales familias .....64

Figura 27: Señales VT clasificadas en la familia naranja (#ff7005). Apilado de las señales(rojo) Elementos de la familia (verde).....65

Figura 28: Señales LP clasificadas en la familia naranja (#ff7005). Apilado de las señales(rojo) Elementos de la familia (verde).....65

# Índice de Tablas

Tabla 1: Lista de las estaciones utilizadas con sus características.....24

Tabla 2: Número de templates por estación, donde se observa cuantos coinciden entre estaciones.....48

Tabla 3: Número de templates por estación, después de revisar la energía de los templates para cada una de sus ventanas, donde se observa cuantas templates coinciden entre estaciones.....50

Tabla 4: Comparación entre las templates crudas y después de revisar la energía de los templates. .....51

Tabla 5: Familias del dendrograma. .....64

Tabla 6: Comparación de las detecciones hechas durante el enjambre del 8 de julio entre las detecciones a mano reportadas por el primer catálogo y las automáticas por TM y por FAST. .....68

Tabla 7: Comparación de las detecciones hechas durante el mes de julio, incluyendo el enjambre del 8 de julio, entre las detecciones a mano reportadas por el segundo catálogo y las automáticas por TM y por FAST. .....69

# Resumen

Una de las maneras más eficientes de entender cómo funciona un volcán es a través de señales sísmicas. Para una misma clase de señales sísmicas se supone la asociación con una actividad volcánica especifica. Por eso es importante clasificar las señales en familias ya que permite predecir o conocer la actividad del volcán, como por ejemplo ciclos de construcción y destrucción de domos, desplazamiento de magma, desgasificación, o una reactivación del volcán. Se propone estudiar el volcán Popocatépetl en el mes de Julio de 2016 debido a la existencia de un enjambre sísmico junto a una actividad magmática importante en ese periodo.

La observación continua de un volcán genera una cantidad importante de datos, tardados de procesar manualmente y por ende es importante desarrollar métodos de detección y clasificación automáticos de señales sísmicas. En particular, los métodos de aprendizaje de máquina permiten realizar un mayor número de detecciones que las hechas manualmente. La capacidad computacional actual favorece el continuo desarrollo de técnicas de aprendizaje de máquina. La aplicación de estas técnicas de detección y clasificación automáticas a señales sísmicas registradas sobre volcanes facilita la toma de decisiones, sobre todo durante una erupción volcánica cuando la cantidad de datos es importante y tardado en procesar y analizar.

La primera etapa de este proyecto consiste en detectar automáticamente las señales sísmicas relacionadas a la actividad del volcán Popocatépetl. Para eso, se utilizaron dos metodologías de inteligencia artificial donde el aprendizaje no depende de la intervención humana. Una consiste en una adaptación del “*Template Matching*” (TM), y otra consiste en la metodología “*Fingerprint and Similarirty Thresholding*” o FAST. El método TM usa las tres componentes EW, NS y vertical de la señal sísmica, mientras que FAST comprime las formas de onda de estas tres componentes en tres

huellas, lo que permite procesar los datos de manera rápida. Adicionalmente se ha mejorado el método TM para utilizar la energía de la señal a partir de las tres componentes, lo que reduce de tres a uno a los datos de entrada y el tiempo de proceso de los datos por un factor de 2.

La segunda etapa consiste en realizar una clasificación automática de las señales detectadas en la primera etapa. La clasificación se hace con agrupación de señales de misma forma de onda que recibirán el nombre de familias de señales, usando una correlación cruzada y un agrupamiento jerarquizado.

Con el método TM, se pudieron detectar 2589 señales utilizando las tres componentes NS, EW y vertical en velocidad, con 1320 detecciones verdaderas (sin falsas detecciones) y 1269 falsas. Usando la energía, se hicieron 2269 detecciones de señales, con 885 falsas alarmas, obteniendo 1384 señales reales. Entonces, el uso de la energía disminuye el número de falsas detecciones. Con FAST se detectaron 1709 señales, incluyendo 1075 detecciones verdaderas y 634 falsas, es decir menos detecciones exactas que el método TM con la energía. Al final se tomaron las 1384 detecciones (sin falsas detecciones) obtenidas usando el método TM y la energía para clasificarlas, ya que las detecciones con TM son las más relacionadas con la sismicidad volcánica al poder detectar algunos LPs. FAST por su parte es la mejor para detectar eventos tectónicos. Al clasificar estas señales, se obtuvieron 9 familias las cuales incluyen eventos como los sismos volcano-tectonicos del enjambre del 8 de julio 2016 y eventos de largo periodo a lo largo de julio 2016.

# Abstract

One of the most efficient ways to estimate the current activity of a volcano is through its seismic signals. Each class of seismic signals can be associated with a specific type of volcanic activity. Thus, it is important to classify these signals into families that allow us to know or predict the activity of the volcano, such as cycles of dome construction and destruction, magma movement or reactivation of the volcano. This work proposes a study of the Popocatépetl volcano in the month of July of 2016 due to the existence of a seismic swarm, as well as important magmatic activity during this period.

Uninterrupted monitoring of a volcano can generate a significant amount of data, that is challenging to process manually. Therefore, it is important to develop techniques that can automatically detect and classify seismic signals. Notably, machine learning techniques can make more detections than those usually made manually. Recent breakthroughs in computational science benefit the ongoing development of these machine learning techniques. Applying them to seismic signals recorded at volcanoes facilitates decision taking, especially during periods of volcanic unrest when an eruption is more likely to happen, and the amount of data is large and time-consuming to properly process and analyze.

This project has two stages, the first of which consists of the detection of seismic signals related to the activity of the Popocatépetl volcano. Two artificial intelligence methods that do not require human interaction during the learning process were used to achieve this goal. One was an implementation of *Template Matching* (TM), and the other was the Fingerprint and Similarity Thresholding (FAST) method. TM uses the three components of the signal: EW, NS and vertical. While FAST compresses the waveforms of the three components in three sparse binary fingerprints resulting in low execution times. An improvement to the TM implementation has been made that uses the energy of the three components, reducing the input data from three to one and speeding up the process by a factor of 2.

The second stage consists in the automatic classification of the signals detected in the first stage. The classification is done by clustering the signals that have similar waveforms and they are given the name of signal families, this was achieved by using cross correlation and hierarchical clustering.

Using the TM method and the three components, 2589 detections were made, of which 1320 were true detections and 1269 were false alarms. Meanwhile, using energy resulted in a total of 2269 detections, of which 1384 were true and 885 were false. This means that the use of energy in TM reduces the number of false detections. FAST had 1709 detections, of which 1075 were true and 634 were false. FAST had fewer true detections than the energy-based TM method. Because energy-based TM detections were more focused to volcanic seismicity than those made by FAST, they were classified using hierarchical clustering. A total of 9 families were obtained, these families include events such as the VT swarm of July 8, 2016, and long period events throughout the month of July.

# 1. Introducción

México es un país con un contexto geológico complejo, con gran actividad sísmica y volcánica, por lo que un porcentaje alto de la población mexicana se encuentra de una manera u otra en una zona de riesgo. Por eso es de suma importancia que se monitoree a los fenómenos naturales de manera adecuada y continua. Actualmente, el volcán más riesgoso para la población mexicana es el volcán Popocatépetl, principalmente porque se encuentra ubicado cerca de zonas altamente pobladas y presenta una constante actividad desde 1994 tras 67 años de reposo desde su última erupción de 1927 (De la Cruz y Siebe, 1997).

La actividad típica del volcán Popocatépetl, incluye manifestaciones como la formación y la destrucción de domos, el movimiento de magma adentro o de lavas o basalto afuera del edificio volcánico, la reactivación de la actividad sísmica y/o explosiva del volcán. Estas manifestaciones generan señales sísmicas típicas y diferentes entre sí (Chouet, 1996) que se pueden clasificar mediante su forma de onda temporal o su contenido espectral. El monitoreo sísmico de un volcán consiste en detectar y clasificar estas señales que cambian antes de una erupción y permite así elaborar un pronóstico de la actividad eruptiva del volcán. Por ejemplo, antes de una erupción se puede observar un incremento de la actividad sísmica alrededor del volcán o la emergencia de tremor volcánico. Por este motivo es importante contar con una red de monitoreo sísmico alrededor de un volcán.

Desde hace varias décadas, los datos sísmicos se registran en continuo y con mayores tasas de muestreo, con un número creciente de estaciones y con una cobertura espacial cada vez más grande, lo que genera una gran cantidad de datos, tardados y complejos de analizar e interpretar manualmente. Esto genera un volumen grande de datos con variedad importante de señales. Lo que durante un periodo largo de tiempo puede desarrollar un problema de “Big Data” (Nielsen, 2016). Estos problemas necesitan desarrollar métodos nuevos de detecciones y clasificaciones

automáticas de señales sísmicas que complementen a los métodos manuales. En el Popocatépetl se han utilizado estos métodos automáticos para detectar y clasificar, el CENAPRED usó durante años un programa llamado *Classification_Seismic_Signal*, desarrollado por Lesage (2009) (C. Valdes-Gonzalez, comunicación personal, junio 2024).

El uso de la inteligencia artificial es una manera de enfrentar la problemática del Big Data, como por ejemplo el aprendizaje de máquina que es capaz de realizar tareas automáticas con información obtenida de los datos. Lo atractivo de este método es que permite a los algoritmos tener un buen funcionamiento incluso con datos adquiridos en ambientes donde la relación señal ruido es baja. A pesar de que estos algoritmos existen desde años, han proliferado recientemente en las geociencias, en particular en sismología para la detección de eventos, sus picados de fases (Mousavi y Beroza, 2023). Otra aplicación de las metodologías que utilizan el aprendizaje de máquina es la clasificación de señales volcánicas detectadas (Bueno et al., 2020, Malfante et al., 2018). Los algoritmos de aprendizaje de máquina son usados ampliamente para facilitar tareas de clasificación, regresión, agrupamiento, y detección de anomalías dentro de los datos (Varshney, 2023). Se pueden clasificar el tipo de aprendizaje realizado en dos categorías: uno supervisado (con intervención humana) y otro no supervisado (sin intervención humana).

El aprendizaje supervisado consiste en generar un set de señales que se quieren buscar en los datos continuos y que son previamente etiquetados en clases determinadas por el usuario. Los programas de este tipo usan las nuevas detecciones inyectadas en el set de datos etiquetados y mejora poco a poco las nuevas detecciones. Con esta iteración se minimizan las diferencias entre lo etiquetado a priori y lo clasificado por los algoritmos (Jo, 2021). El aprendizaje supervisado va a encontrar solamente las señales entradas en el set de señales y no va a encontrar señales diferentes a este set de señales (Delua, 2021).

El método de clasificación de señales utiliza software desarrollado con aprendizaje de máquina supervisado (Malfante et al., 2008), cuyo funcionamiento necesita de información o de un entrenamiento a priori, es decir un conjunto de datos previamente seleccionados manualmente.

Por otra parte, el aprendizaje no supervisado es aquel donde ninguna información previa acerca de la base de datos o de las etiquetas de los datos es proporcionada previamente (Delua, 2021). Tiene como propósito realizar un reconocimiento de patrones con objetividad, para permitir a los usuarios tomar decisiones a partir de resultados obtenidos sin aporte humano.

# 2. Objetivo

El objetivo de esa tesis es entender la dinámica del volcán Popocatépetl mediante la asociación de cada proceso físico con una señal sísmica que se supone lo caracteriza. En particular, se busca establecer patrones de relación entre las señales sísmicas y los periodos de actividad del volcán. Por ello, es importante detectar y clasificar estas señales sísmicas. Pues cada fase de actividad de un volcán puede producir señales características, las cuales se pueden agrupar en diferentes familias de señales.

Para realizar este trabajo, se necesita primero detectar de manera automática las señales sísmicas volcánicas. Se usarán dos técnicas diferentes de detección automática de señales sísmicas que utilizan aprendizaje que no necesita generar previamente una base de señales a buscar: una implementación de *Template Matching* y "*Fingerprint And Similarity Thresholding*" o FAST (Yoon et al, 2015). Se observa que FAST detecta principalmente a un tipo de señales (VTs), así que este trabajo se enfocará en sus detecciones. Se podrá aplicar el método que se propone en esta tesis a otros tipos de señales volcánicas. Posteriormente se hará una clasificación automática de las señales detectadas con una técnica de agrupamiento jerarquizado. Se definirán así familias de señales con formas parecidas.

# 3. Contexto geológico

El volcán Popocatépetl, es un estratovolcán activo de composición andesítica a dacítica con una elevación promedio de 5,452 m.s.n.m. siendo la segunda montaña más alta de México, después del Pico de Orizaba. Junto con el volcán de Colima es uno de los volcanes más activos del país y se encuentra localizado cerca de una de las zonas más densamente pobladas del mundo. Alrededor de 25 millones de personas viven en un rango de 80 kilómetros del volcán, por lo que una posible erupción representa un gran riesgo para la sociedad mexicana.

## 3.1 Marco tectónico

El volcán Popocatépetl se encuentra dentro de la Faja Volcánica Trans Mexicana (FVTM), la cuál es un arco continental de más de 1000 km de extensión que se extiende de este a oeste. La FVTM se encuentra encima de la fosa de la compleja subducción de la placa oceánica de Cocos bajo la placa continental Norte Americana (Ferrari et al., 2012). La FVTM cuenta con una gran variedad de manifestaciones volcánicas que se formaron en el último millón de años (Ferrari et al., 2012). Estas manifestaciones volcánicas incluyen: campos de volcanes monogenéticos, cadenas de estratovolcanes, calderas y campos geotérmicos (Macías, 2005), con una orientación oblicua respecto a la dirección de la fosa. La migración de la actividad volcánica desde el Norte hacia el Sur (parte frontal del arco) genera una concentración de la mayoría de los volcanes activos en la parte sur de la FVTM (Macías, 2005).

El volcán Popocatépetl se encuentra en la parte central de la FVTM (Fig. 1). El volcán se encuentra en el extremo sur de la Sierra Nevada, que tiene una extensión de más de 80 kilómetros, con una orientación de Norte a Sur. La Sierra Nevada cuenta con los volcanes Tlaloc (en la parte Norte), Telapón e Iztaccíhuatl-Popocatépetl (en la

parte Sur). Macías (2005) propuso que la actividad volcánica empezó en el Norte con el volcán Tlaloc y se ha desplazado hasta el Sur con los volcanes Iztaccíhuatl-Popocatépetl durante los últimos 2 millones de años.

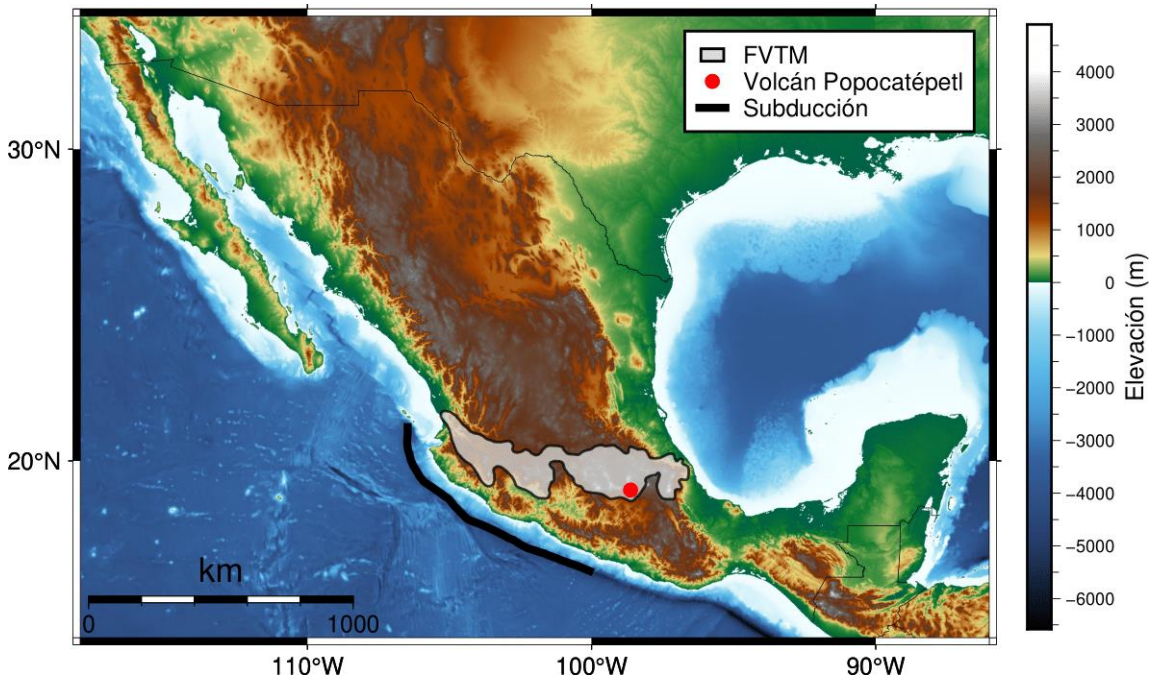


*Figura 1:. Ubicación de la FVTM (parte gris) y del volcán Popocatépetl (punto rojo).*

### 3.2 Historia eruptiva reciente

El inicio de la historia moderna del volcán Popocatépetl empezó hace aproximadamente 23,000 años con una erupción cataclísmica de tipo Bezymianni (Rusia). El cono moderno se encuentra emplazado sobre un gran edificio volcánico antiguo (Robin y Boudal, 1987).

El volcán Popocatépetl, ha tenido múltiples periodos de actividad en los últimos 23,000 años. Algunos periodos de actividad han tenido una gran magnitud al tener erupciones de tipo pliniano, mientras que otros han tenido episodios vulcanianos de menor magnitud (Ferres y Fonseca, 2017). Tres distintos episodios volcánicos que ocurrieron hace 10,000 años, entre 10,000 y 8,000 años y aproximadamente 3,800 años destruyeron parcialmente el cono más antiguo del Popocatépetl, llamado “El

Fraile" ubicado al norte de la cima actual (Robin y Boudal, 1987). El cono más joven se localiza en la cima del volcán, el cual ha presentado actividad desde antes de tiempos prehispánicos y durante este periodo, la actividad consistió en flujos de lava, así como episodios eruptivos y algunas explosiones cataclísmicas (Robin y Boudal, 1987).

A partir de la época prehispánica, la actividad fue efusiva en la cima, con presencia de erupciones plinianas. Aunque pareciera permanecer en calma, el volcán seguía activo (Robin y Boudal, 1987). En la época colonial, la actividad del volcán se caracterizó por periodos efusivos, distinguidos por ciclos de construcción y destrucción de domos dentro del cráter. En el siglo XVII, hubo una erupción de gran magnitud, entre pliniana y vulcaniana. (Delgado, 2017).

Durante el siglo XX, el volcán presentó una notable reactivación, teniendo un episodio de actividad explosiva con intensidad moderada de 1919 a 1927, para posteriormente tener una relativa quietud por décadas. No fue sino hasta finales de 1994, que el volcán Popocatépetl presentó otro episodio de actividad explosiva, después de casi 70 años de quietud (De la Cruz y Siebe, 1997). La actividad fue precedida por un incremento en la sismicidad, emisión de gases, temperatura y pH del lago del cráter y emisiones de ceniza que alcanzaron hasta los 3 km de altura (Martin-del-Pozzo, 2017; Macías, 2005). La actividad siguió esporádicamente hasta marzo de 1995, cuando se reportó que la emisión de ceniza había decrecido.

Desde marzo de 1996, la actividad del volcán Popocatépetl ha sido representada por ciclos de construcción y destrucción de domos de lava en la cima del cráter. La construcción de domos tiene como rasgos distintivos episodios individuales que pueden durar desde horas hasta días, los que se pueden asociar a trenes de exhalación. Su destrucción es por periodos de actividad explosiva que puede durar varios días.

En el mes de diciembre del 2000, el volcán Popocatépetl presentó un estado de gran actividad sísmica, fumarólica y con lanzamiento de fragmentos incandescentes relacionados a destrucción de domos. Durante el 18 de diciembre del 2000 ocurrió una erupción vulcaniana con lanzamiento de fragmentos incandescentes (Macías, 2005). La actividad que concluyó esta etapa fue la destrucción de domo formado el 18 de diciembre, esto ocurrió con una gran erupción vulcaniana el 21 de enero de 2001, explosión que alcanzó más de 6 km sobre el cráter, siendo la erupción con mayor liberación de energía desde 1994 (Macías, 2005).

Desde 1994, la actividad ha continuado con ciclos de construcción y destrucción de domos, usualmente por explosiones vulcanianas, la actividad también incluye emisión de columnas de cenizas, emisión de gases (Siebe et al., 1996), y periodos con abundante sismicidad (Matoza et al., 2019). Otros episodios de actividad notable han ocurrido en 2004, 2012, 2014 y más recientemente en 2023. La actividad presente desde la reactivación en 1994 ha superado a la magnitud del periodo eruptivo de 1919 a 1927 (Macías, 2005) y no hay señales que se vaya a detener pronto (Ferres y Fonseca, 2017). Como resultado de la reactivación del volcán Popocatépetl en 1994, se decidió establecer una red de monitoreo sísmico en el volcán Popocatépetl. Esta red de monitoreo es administrada por el CENAPRED y por el Instituto de Geofísica de la UNAM.

# 4. Clasificación de señales sismo-volcánicas

Varias disciplinas se dedican a la observación de cambios en los comportamientos medibles de un volcán, tales como deformaciones, emisión de gases, incandescencia o modificaciones en los patrones de sismicidad. En un volcán activo hay muchos tipos de señales sísmicas diferentes, con contenidos frecuenciales y formas de ondas muy diferentes a las de un sismo tectónico.

Desde el inicio de la sismología volcánica, se han propuesto muchos tipos de clasificaciones de señales volcánicas (i.e, Sassa, 1935; Minakami, 1974; McNutt 1986). Desde algunas décadas, la clasificación más usada es la propuesta por Chouet (1996) que clasifica a las señales en cuatro tipos en relación con un proceso físico del volcán. Los cuatro tipos de señales son los siguientes:

- Volcanotectónico (VT): Este tipo de señales tiene ondas P y S con una banda amplia de frecuencias, generalmente en las altas frecuencias. Su principal característica es tener formas de onda similares a los sismos de origen tectónico. Está asociado a fallas o fractura de rocas dentro del edificio volcánico (McNutt y Roman, 2015).
- Largo Periodo (LP): Es un tipo de señales con una onda P, pero sin onda S. Se caracterizan por tener un contenido espectral en las bajas frecuencias (<6Hz), por lo general más bajas que los VTs. Se relaciona con el movimiento y presurización de fluidos hidro-magmáticos dentro del volcán (McNutt y Roman, 2015).
- Explosiones: Son generadas por expulsiones súbitas de magma, gas y/o cenizas, acompañando por lo general a las erupciones explosivas. Sus contenidos espectrales varían dependiendo del tipo de explosión. Es posible que señales de este tipo se puedan componer de varias explosiones pequeñas consecutivas. Debido a que se trata de un evento súbito y con causas complejas, es muy raro que existan dos explosiones que se parezcan entre sí (McNutt y Roman, 2015).

- Tremores: tienen una amplitud sostenida, que puede durar desde minutos hasta días. Su rango de frecuencia varía mucho en función del tipo de tremor: algunos son armónicos (una sola frecuencia predominante $f_0$ o varias frecuencias con una relación armónica entre ellas ($f_0$, $2f_0$, $3f_0$…), o sin relación entre ellas, mientras que otros tienen un rango ancho de frecuencias. En algunos volcanes su origen se debe a la superposición de eventos LP, por lo que se pueden relacionar con el movimiento y presurización de fluidos hidro-magmáticos dentro del volcán (McNutt, 2005).

# 5. Datos

## 5.1 Descripción de las estaciones y de los datos

Se utilizaron datos sísmicos registrados en julio de 2016 en cuatro sismómetros de banda ancha, 3 componentes (Tabla 1, Fig. 2). Los sismómetros registran señales en continuo con una tasa de muestreo de 100 Hz.

| Código | Nombre | Latitud (°) | Longitud (°) | Altitud (m) | Equipo | Tasa muestreo (Hz) |
|---|---|---|---|---|---|---|
| PPIG | Tlamacas | 19.067 | -98.628 | 3980 | Nanometrics Trillium Post Hole | 100 |
| PBP | Canario | 19.041 | -98.627 | 4170 | Güralp 40T | 100 |
| PCP | Colibrí | 18.986 | -98.557 | 2650 | Güralp 6TD | 100 |
| PXP | Chipiquixtle | 19.009 | -98.656 | 3980 | Güralp 40T | 100 |

*Tabla 1: Lista de las estaciones utilizadas con sus características.*

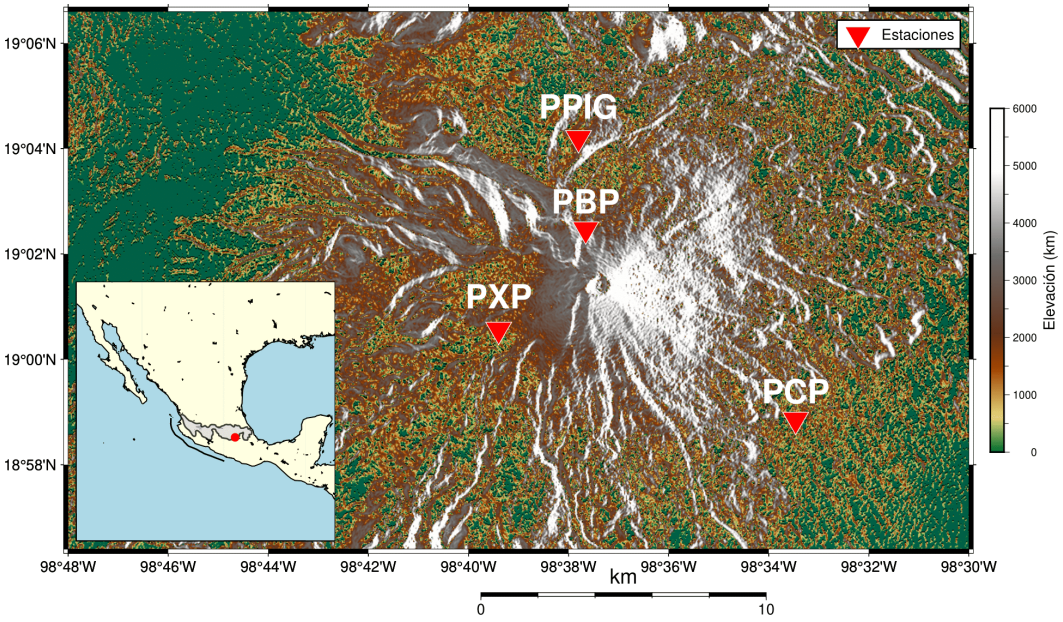


*Figura 2: Ubicación de la zona de estudio y de las estaciones utilizadas.*

En los registros hay periodos de tiempo sin mediciones, estos reciben el nombre de lagunas en los datos. Pueden ser el resultado de problemas con la carga de la batería de los instrumentos o de transmisión de los datos y pueden representar un problema por la posible pérdida de algunas señales. Un ejemplo de un registro de una estación con lagunas en los datos se muestra en la Figura 3.

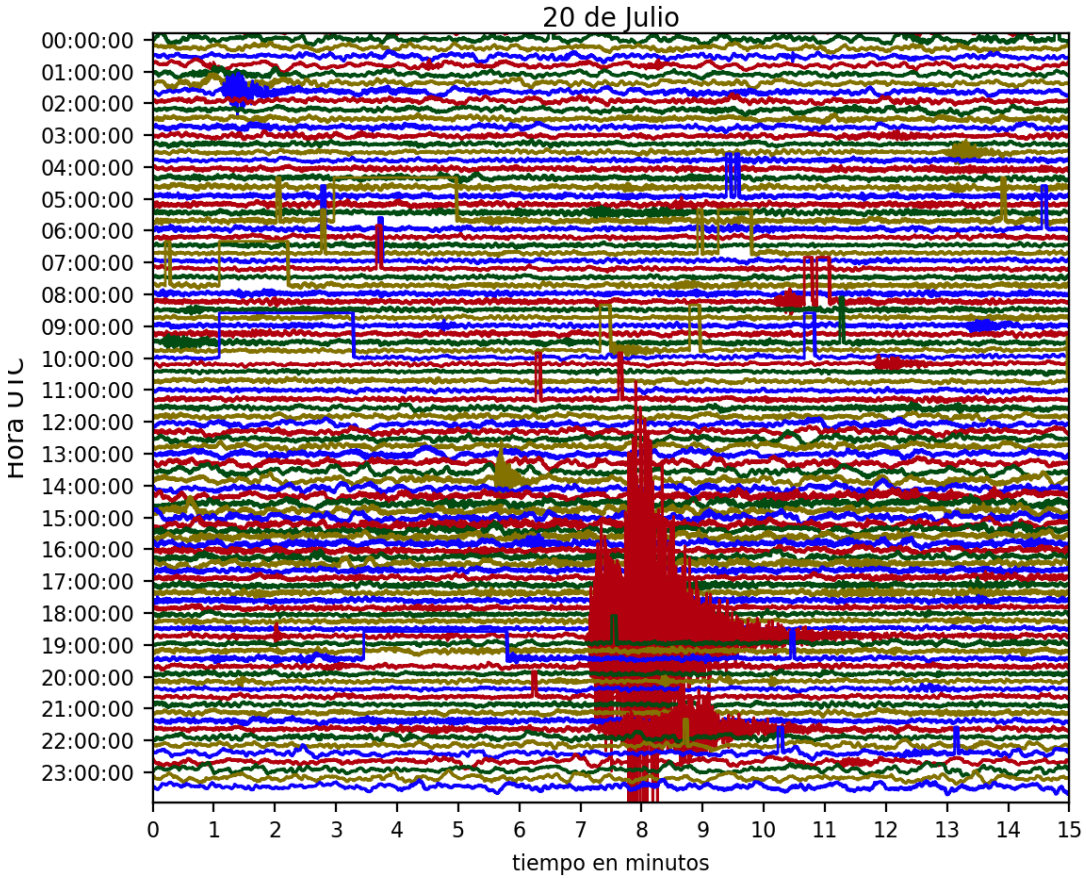


*Figura 3: Ejemplo de un registro con lagunas en los datos registrado en la estación PCP, el 20 de Julio de 2016.*

Se analizarán datos del mes de julio 2016 porque el 8 de julio de 2016 ocurrió un enjambre de VTs (Fig. 4) debajo del Popocatépetl. Este enjambre contiene 39 de los sismos ocurridos durante el mes, con magnitudes entre 1.3 y 3.8 y por ende la mayor energía acumulada por VTs, aportando $1.3147 \times 10^{11}$ J (CENAPRED, 2017). El enjambre comenzó a las 16:51 UTC y terminó a las 18:54 UTC. También se reportaron

dos episodios de actividad moderada estromboliana los días 10 y 31 de julio 2016, con columnas eruptivas y expulsión de fragmentos incandescentes.

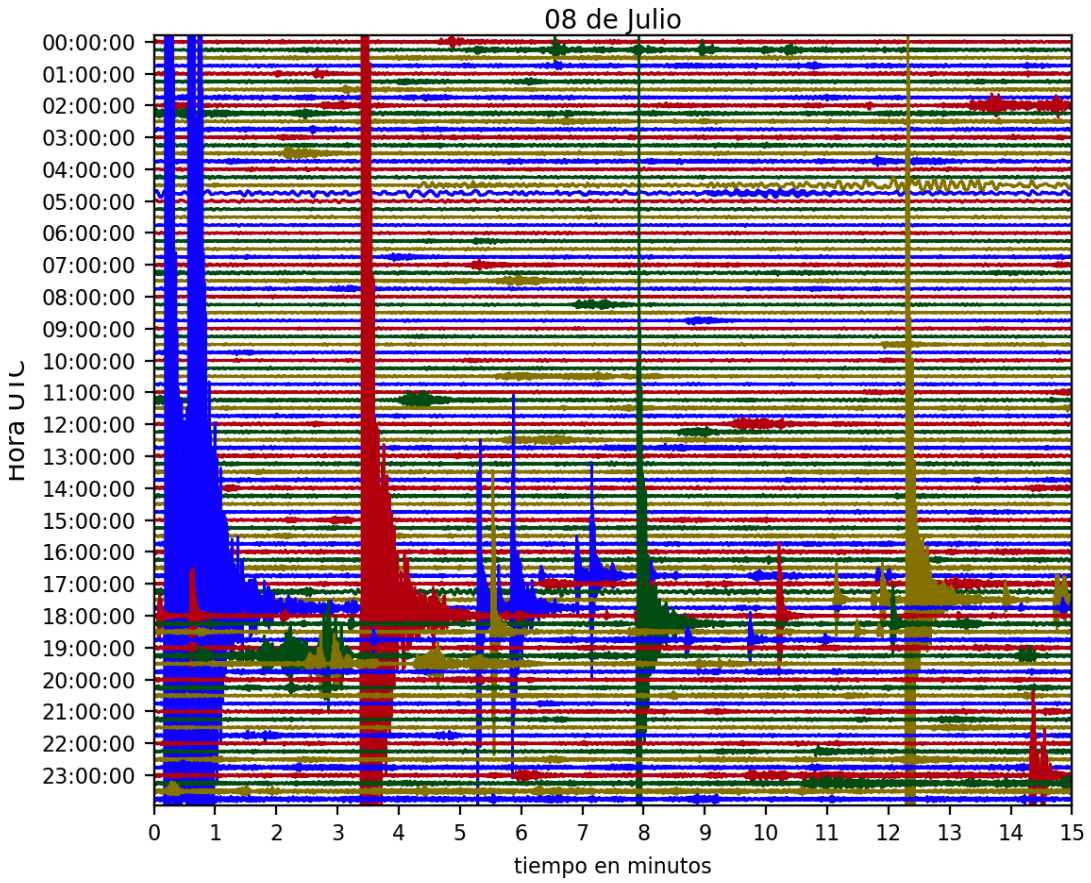


*Figura 4: Registro del enjambre sísmico del 8 de julio de 2016 en la estación PPIG*

## 5.2 Pre-procesado de los datos

Para aplicar las técnicas de detección automática sobre los datos, es mejor realizar una serie de tratamientos de la señal antes (que se llama un pre-proceso de los datos) para buscar un tipo de señal especifico (VT, tremor, explosión…), reducir el tiempo de cálculo del programa, llenar las lagunas de señales con ruido para tener una señal continua, minimizar el número de falsas detecciones etc... Se va a hacer un pre-procesado a las señales con las etapas siguientes: filtrado de frecuencias, decimado, llenado de lagunas en los datos y cálculo de la energía.

### *5.2.1 Filtrado de los datos*

Como se mencionó anteriormente, la sismicidad local de un volcán tiene diferentes tipos de señales registrados en diferentes rangos de frecuencias. ~~Entonces~~ el uso de filtros con diferentes bandas de frecuencias permite buscar los diferentes tipos de señales (VTs, tremor, explosiones y LPs). Los datos de Julio 2016 contienen básicamente VTs, por eso este trabajo se va a concentrar en buscar este tipo de eventos. Para buscar cual es la mejor banda de frecuencia para detectar los VTs, se probaron tres bandas de frecuencias: de 0.05 a 1 Hz, de 5 a 10 Hz y de 0.05 a 10 Hz. Se observó que la banda de frecuencias donde se detectan mejor los VTs es en la de 5 a 10 Hz (Fig. 5).

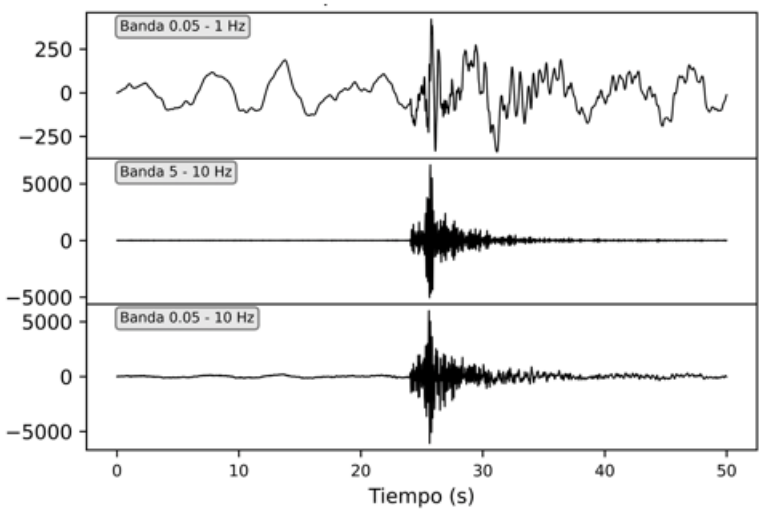


*Figura 5: Comparación de como un VT se observa en las diferentes bandas de frecuencia.*

No se utilizaron otras bandas de frecuencias porque durante el mes de estudio no hubo señales en estas bandas de frecuencias. Pero se pueden usar las metodologías de detección por ejemplo para buscar LPs, tremores o explosiones en otras bandas de frecuencias. Por ejemplo, es posible buscar LPs en la banda de frecuencia 1 a 5 Hz.

### *5.2.2 Decimado de los datos*

Para disminuir el tiempo de cálculo y mejorar la eficiencia de los programas de detección, se va a decimar las señales. Esa operación tiene un límite en función de la frecuencia de la señal que se busca. Para el caso de este trabajo, se desea detectar VTs entre 5 y 10 Hz, así que se decidió decimar la señal hasta 25 Hz (tasa de muestreo de 50 Hz), valor un poco más grande que 10 Hz para tener un margen de seguridad. Entonces se realizó un decimado de los datos para pasar de una tasa de muestreo de 100 Hz a una 50 Hz que corresponde a un factor de decimación de dos.

### *5.2.3 Llenado de lagunas en los datos*

Las lagunas en los datos son segmentos con valores iguales a 0 que van a generar falsas detecciones a largo de toda la laguna, generando una familia de señales de valores de 0. Esa familia puede tener muchas señales que puede disminuir la velocidad de ejecución de los programas (Yoon et al., 2015). Los programas de detección utilizados en este trabajo permiten llenar las lagunas de datos con ruido no correlacionable, de forma que no se hagan detecciones dentro de las lagunas.

### *5.2.4 Cálculo de la energía*

Los métodos de detección y clasificación se aplican tradicionalmente a las tres componentes NS, EW y vertical de la señal sísmica (Withers et al., 1998; Ramirez et

al., 2011). En este trabajo, se usarán las tres componentes para realizar detecciones en dos métodos de automáticos (FAST y TM que son descritos más abajo).

En este trabajo, se usó el cálculo de la suma de las energías de las tres componentes (Perton et al., 2024) para uno de los métodos de TM y para la clasificación de las señales. Este método permite pasar de 3 componentes a una, lo que reduce por tres la cantidad de datos a procesar y así minimizar el tiempo de cálculo.

. La energía se calcula de la siguiente manera:

$$Energía = \frac{E^2 + N^2 + Z^2}{3}$$

donde E, N y Z representan las componentes EW, NS y vertical respectivamente.

Este proceso generalmente aumenta la relación señal ruido. En el caso particular de tener un sismómetro de una sola componente, se podrá calcular la energía y aplicar el método.

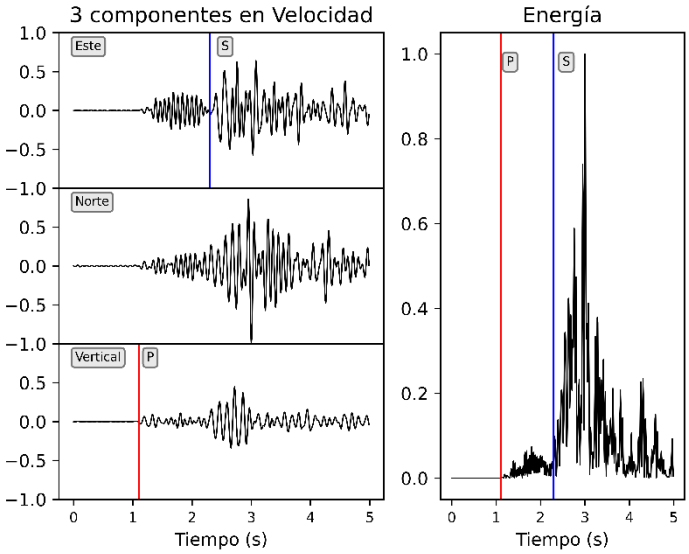


*Figura 6: Ejemplo de la señal en energía a partir de las tres señales en velocidad de las componentes NS, EW y vertical.*

Se puede observar el resultado de esta operación (Fig. 6) mediante un ejemplo de una forma de onda de una señal VT. Del lado izquierdo se puede ver un segmento de la serie de tiempo. Las amplitudes de las tres componentes son divididas por la amplitud máxima de las 3 componentes, de tal manera que se conserva la amplitud relativa entre sí.

En las tres componentes de velocidad, el arribo de la onda P es bastante claro, alrededor del segundo 1, mientras que el arribo de la onda S no lo es. En cambio, el arribo de la onda S es muy claro cuando se usa la energía, se puede visualizar alrededor del segundo 2.

Otra ventaja del uso de la energía es la siguiente. Si una o dos de las tres componentes esta dañada, la energía corresponde principalmente a la componente sana y se podrá aplicar el método. En cambio, al usar las tres componentes separadas (sin el uso de la energía total) en el caso que haya un problema en alguna de las componentes este método no podrá ser utilizado.

Se utilizó también el cálculo de la energía para verificar manualmente y a posteriori que las detecciones no son falsas detecciones (caracterizadas por una energía pequeña), este proceso se describirá a detalle en el capítulo siguiente.

# 6. Metodología

Se usaron dos metodologías de detección automática de señales. La primera metodología es una implementación de *Template Matching*, que funciona en dos etapas. La primera etapa (modo Sistemático) genera automáticamente una base de señales preliminares detectadas por el programa sin introducir ninguna información a priori sobre las señales que se buscan (se habla de aprendizaje no supervisado). Las señales obtenidas como salida de esta etapa se llaman "*templates*", que serán usados en la segunda etapa para buscar señales parecidas dentro de la base de datos continuos. La primera etapa (modo Sistemático) tiene como ventaja que permite obtener los *templates* de manera automática, y no manualmente como se hacía en los primeros códigos de detecciones automáticas (Withers et al., 1998). La segunda etapa (modo Estándar) busca señales parecidas a los *templates* dentro de los datos continuos, mediante un aprendizaje "supervisado", es decir que usa la información previa que son los *templates* determinados en la primera etapa.

La segunda metodología es FAST la cual busca señales similares (que se repiten) adentro de una señal continua sin ninguna información a priori. Eso significa en particular que, si dentro de los datos continuos, hay una señal grande que no se repite, esta no será detectada. Cada señal será transformada en una huella que tiene la misma información que la señal, pero de manera comprimida. FAST corta la señal continua en N ventanas de duración predeterminada. Luego calcula una huella para cada ventana y almacena todas las huellas. FAST hace familias de huellas similares. Después escoge una huella y la compara con todas las otras huellas por familia y determina si esta huella es una detección (que se determina con un número grande de huellas parecidas). FAST funciona en dos etapas: 1) generación de huellas y 2) búsqueda de huellas similares. Como salida, se tiene una lista de tiempos de llegada de las señales similares detectadas.

Después de realizar las detecciones con TM, se clasificarán en diferentes familias usando un método de agrupamiento jerarquizado.

### *6.1 Template Matching*

La correlación cruzada de formas de onda es uno de los métodos más comunes para realizar detecciones automáticas de señales sísmicas. Por lo general, se trata de un algoritmo de aprendizaje supervisado, que utiliza como información a priori de la base de datos en forma de *templates*. Los *templates* son segmentos de interés de una serie de tiempo, los que pueden ser obtenidos manualmente o de manera automática (Withers et al., 1998; Gibbons et al., 2004; Shelly et al., 2007). Fue diseñado para realizar detecciones a partir del cálculo del coeficiente de correlación normalizado (CCN), entre el segmento de consulta, y una serie de tiempo más larga. Se obtiene como resultado segmentos dentro de la serie de tiempo, que, por el resultado del cálculo del CCN, son similares al segmento de consulta.

Se usa la correlación cruzada porque se buscan formas de ondas parecidas, independientemente de su amplitud. El cálculo del CCN tiene como resultado valores que van desde el -1 hasta el 1, donde +1 significa la similitud exacta entre las señales A y B y -1 significa las señales A y -B son similares. Se puede representar mediante la siguiente fórmula:

$$CCN_{A,B} = \frac{corr(A,B)}{\|A\| \cdot \|B\|}$$

donde $corr(A,B)$ es la correlación cruzada de ambas dos series de tiempo A y B, y $\|A\| \cdot \|B\|$ es la multiplicación de las dos normas de A y B.

Sin embargo, es importante resaltar que cualquier programa de detección automática puede generar falsas detecciones. Gibbons et al. (2004), las define como señales que tienen un valor de CCN lo suficientemente alto para activar el umbral de detección. En realidad, se trata de marcas de tiempo correspondientes a señales que son ruido con una forma de onda parecida y que no son de interés. Por lo que ajustar los umbrales de detección para que las falsas detecciones sean mínimas es una labor importante. No obstante, se debe tener en cuenta que siempre existirán falsas

detecciones o habrá señales útiles no detectadas por el programa en función del nivel del umbral. Por ejemplo, si el umbral de detección es muy grande, se van a detectar muy pocas buenas detecciones, que serán muy parecidas, y casi no se harán falsas detecciones. Si el umbral disminuye, se tendrá una mezcla de detecciones verdaderas muy buenas y detecciones falsas. La dificultad es encontrar un buen valor del umbral para no tener muchas falsas detecciones. Por esta razón, es importante recordar que ninguna metodología de detección automática existe para reemplazar al 100% al analista, su existencia es para complementar y facilitar su trabajo.

El *Template Matching* (TM) utilizado en este trabajo es una implementación realizada por el laboratorio *"Géosciences Environnement Toulouse"* (GET) del "*Observatoire Midi-Pyrénées"* en Toulouse, Francia. La implementación aprovecha las unidades de procesamiento gráfico (GPUs) para realizar cálculos del coeficiente de correlación. El TM aprovecha que los GPUs tienen un alto poder computacional, memorias grandes de banda ancha, bajos costos computacionales y un relativamente bajo consumo de energía (Mu et al., 2017). Esta técnica tiene dos etapas de funcionamiento. La primera etapa (modo Sistemático) funciona como un algoritmo no supervisado para obtener *templates* que después se utilizan en la segunda etapa (modo Estándar) que busca señales similares a los *templates* y funciona como un algoritmo de TM tradicional.

Ambos modos utilizan la librería *RapidAlignerSeismic* (RAS) desarrollada por el GET y construida usando una herramienta para acelerar procesos al ejecutarlos usando varios GPUs llamada CUDA. La librería RAS, a la fecha de la elaboración de esta tesis, sigue en desarrollo. La librería está diseñada para mejorar el rendimiento de análisis de TM al usar GPUs. La librería toma como punto de partida otra librería de minería de datos llamada *RapidAligner*, la cual fue desarrollada por NVIDIA (Hundt, 2021). La librería *RapidAligner* permite alinear un segmento corto de una serie de tiempo con series de tiempo completas y hacer cálculos eficientes de similitud entre segmentos.

### *6.1.1 Primera etapa de TM (modo Sistemático)*

La primera etapa del código TM (modo Sistemático) realiza una búsqueda no supervisada dentro de toda la base de datos. Los resultados son una lista de tiempos de inicio de los *templates*.

La búsqueda se realiza sistemáticamente, parecida a una autocorrelación en una base de datos. Sigue un orden cronológico de las ventanas, al calcular el CCN entre una ventana y las demás ventanas dentro de la base de datos. Al momento en el que se calcula un coeficiente de correlación que supera un umbral de detección, el programa continúa con la siguiente ventana. El funcionamiento de la primera etapa de TM (modo Sistemático) se ilustra en la Figura 7.

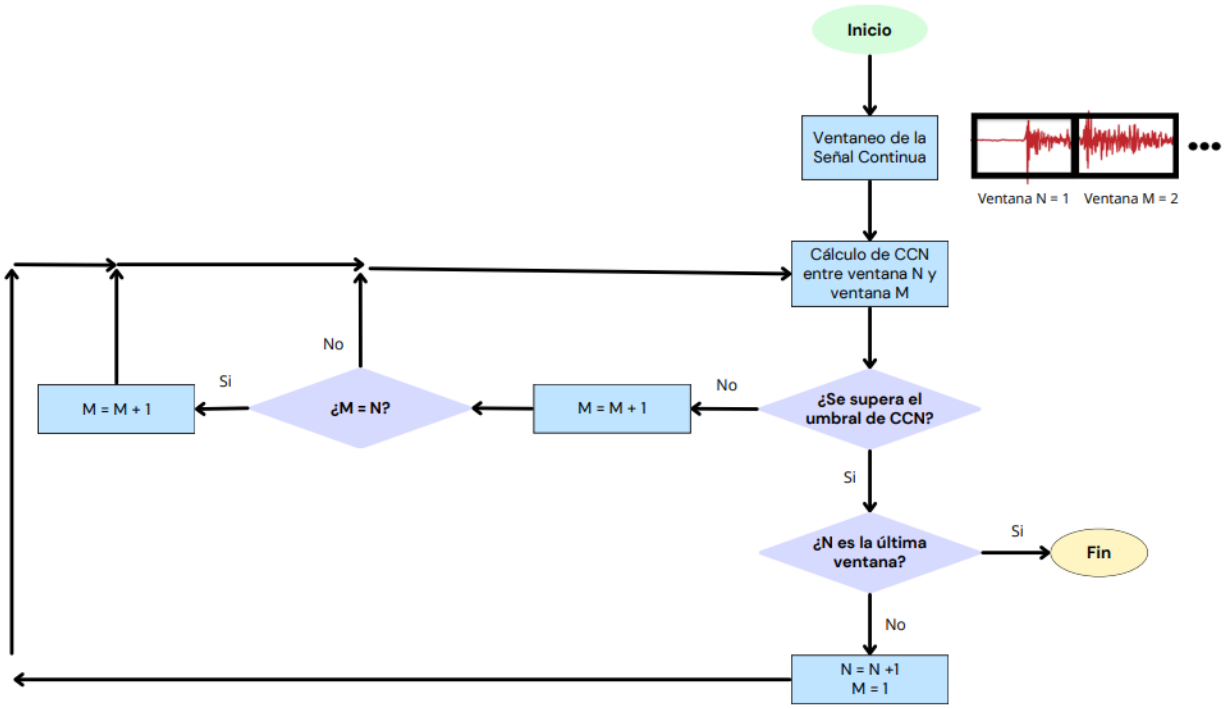


*Figura 7: Diagrama de flujo del funcionamiento básico de la primera etapa de TM.*

La primera etapa (modo Sistemático) necesita contar con nodos con al menos dos GPUs de NVIDIA, ya que este algoritmo se ejecuta en varios GPUs al mismo tiempo. Para la ejecución de esta etapa se utilizó un nodo con dos GPUs de NVIDIA, de modelo Quadro RTX 4000. Se recomienda que el usuario utilice tantos nodos y GPUs de NVIDIA modernos como pueda disponer. El código ejecuta tantas instancias de la etapa (Fig. 7) como se tengan GPUs, siempre y cuando sean más de dos. Cada GPU iniciará el ventaneo con un ligero desplazamiento, para no perder señales por el ventaneo.

La ejecución de la primera etapa (modo Sistemático) (Fig. 7) se repite para cada una de las componentes y guarda el tiempo de cada detección en una lista por componente. Por ejemplo, si una misma detección aparece en las listas de las tres componentes esta se considera como una detección verdadera y cuenta como una detección. Pero si una detección se ve solamente en una o dos componentes, aunque esta detección se trate de una detección verdadera, el programa no la toma como detección. Por esta razón, si una componente no funciona, ninguna señal será detectada. Al final se obtiene una lista con las detecciones que aparecen en las tres componentes.

### *6.1.2 Segunda etapa de TM (Modo Estándar)*

La primera etapa (modo **Sistemático**) genera una lista de *templates*. La segunda etapa (modo **Estándar**) busca señales parecidas a estos *templates* dentro de los datos continuos. Se calcula la correlación cruzada entre cada *template* y cada ventana (de duración definida por el operador) móvil dentro de los datos continuos. Esta segunda etapa tiene un tiempo de ejecución más rápido que la primera. El funcionamiento de la segunda etapa (modo Estándar) se ilustra en la Figura 8.

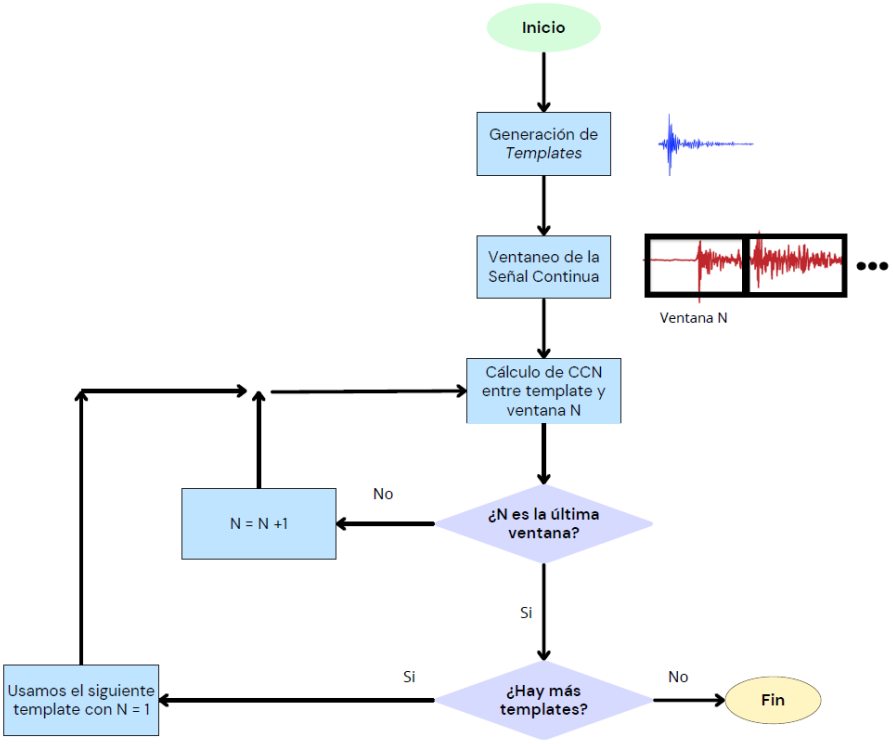


*Figura 8: Diagrama de flujo del funcionamiento básico de la segunda etapa.*

La segunda etapa del programa TM utiliza las tres componentes NS, EW y vertical para realizar las detecciones. Se quiere usar a la energía para hacer las detecciones, es decir con una sola componente. Por esta razón, el código TM es modificado en la segunda etapa para poder funcionar con una sola componente (la energía).

La segunda etapa necesita menos recursos computacionales para ser ejecutado (al menos un GPU de NVIDIA), y se puede ejecutar en la mayoría de las computadoras portátiles modernas. Esta facilidad permite su uso tanto con el nodo con los dos GPUs Quadro RTX 4000, así como con un GPU GeForce RTX 3050 de una computadora portátil personal. De todas maneras, para esta segunda etapa, es recomendable utilizar tantos GPUs como le sea posible al usuario, para generar tantas instancias del programa (Fig. 8) como se tengan GPUs, lo que hace que el programa sea más rápido en su ejecución, a comparación de usar sólo un GPU.

### 6.2 FAST

FAST (Yoon et al., 2015) es un algoritmo de detección automática de señales sísmicas no supervisado. Este programa adapta el funcionamiento del programa *"Waveprint"* (Baluja et al., 2008), el cual a partir de recibir segundos de una canción puede identificarla, siempre y cuando este dentro de una base de datos. Este programa, primero segmenta la forma de onda en ventanas, para generar huellas binarias. Las huellas binarias son una representación comprimida que contiene las características más importantes de la señal. Luego, hace una búsqueda de similitudes construyendo una base de datos donde estas huellas son almacenadas y poder hacer las detecciones de señales similares. El flujo de estas etapas se muestra en la Figura 9. Algunos estudios que entran en detalle en usos y aplicaciones de FAST estos estudios incluyen a los hechos por: Baluja et al., 2008; Yoon et al., 2015; Yoon et al., 2017; Reyes-Romero, 2022; Garza-Girón et al., 2023.

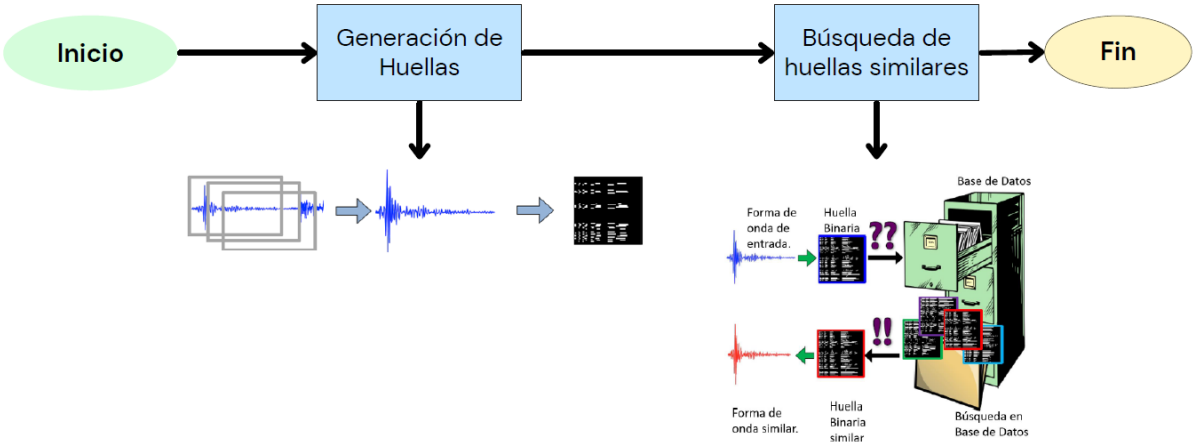


*Figura 9: Etapas de FAST*

Entre las ventajas principales de FAST se incluyen: 1) su portabilidad, es decir que es posible implementarlo en sistemas compatibles, ya que dentro de la paquetería se encuentra lo necesario para usarlo; 2) su escalabilidad, es decir que se puede implementar con cualquier base de datos de cualquier red sismológica, sin importar la duración; 3) su no dependencia de catálogos de eventos ya conocidos; 4) y su rápida velocidad de ejecución, ya que permite analizar rápidamente bases de datos de larga

duración en solamente algunas horas de procesamiento para encontrar resultados aceptables.

La fortaleza de FAST es detectar sismos similares, pero solamente si estos existen. Si existe en la base de datos un evento que ocurre una vez, este no será detectado por FAST. Otra de sus desventajas es que no realiza la detección de fases, como lo son la onda P y la onda S, por lo que se tiene que utilizar otro software para realizar localización de eventos. Para que FAST funcione de manera adecuada, necesita de computadoras con una gran cantidad de memoria disponible, por ejemplo, en este trabajo FAST necesito más de 171 Gb de almacenamiento para las cuatro estaciones del mes de datos.

### *6.2.1 Generación de huellas binarias*

Este paso se puede asociar a como cada persona tiene una huella dactilar única, con la que a partir de ciertos datos biométricos un sistema puede determinar la identidad de una persona. El programa hace una compresión de la forma de onda para generar huellas binarias, lo suficientemente compactas, para ser almacenadas en una base de datos sin ocupar mucho espacio en memoria y dispersas para poder ser únicas dentro de la base de datos.

A partir de los datos continuos, FAST genera ventanas de una longitud fija que pueden traslaparse entre ellas con un espacio de segundos entre cada una. Cada ventana corresponde a una huella, por lo que el número de ventanas es directamente proporcional al tiempo de ejecución del programa. Por lo tanto, el tamaño de duración de las huellas es importante, ya que tiene que ser la adecuada para albergar la información de las señales de interés para el usuario. El tamaño de las ventanas representa una constante que todas las componentes y estaciones que se utilizarán deben mantener. Para este trabajo, se ajustaron los parámetros de entrada para que las huellas binarias fueran de 20 segundos, teniendo un segundo de traslape entre sí. Esta duración se escogió debido a que se puede tener la mayor información de las

señales propias del volcán y para consistencia entre el tiempo de ventanas de ambos métodos.

Una vez calculadas las ventanas, el siguiente paso es comprimir la forma de onda, primero se obtiene una imagen espectral, la cual es el resultado de aplicar al segmento una ventana suavizadora de Hamming, y posteriormente realizar una transformada de Fourier en tiempo corto, la que calcula el contenido frecuencial para ese segmento de tiempo y no el de toda la señal (Kehtarnavaz, 2008), obteniendo como resultado la imagen espectral. El siguiente paso consiste en hacer un remuestreo con un valor en las potencias de dos, para obtener la potencia de la imagen espectral, y así distinguir las señales por su potencia, ya que las débiles usualmente representan ruido mientras que las fuertes podrían ser de un terremoto. Se utiliza la transformada de ondícula de Haar, la cual es una operación que se utiliza para obtener imágenes de menor resolución al comprimir la imagen original. De esta transformación se obtienen coeficientes que permiten la reconstrucción de la imagen original (Stollniz, 1995). Para poder comprimir aún más la imagen espectral, se conserva un número de estos coeficientes, preferiblemente los que tengan mayor energía, y se descarta el resto. De los coeficientes restantes, se transforman a 1 o -1 dependiendo si su valor es positivo o negativo, mientras que los coeficientes descartados, toman un valor de 0, obteniendo como resultado una huella binaria y lo suficientemente dispersa para ser única.

### *6.2.2 Búsqueda de huellas similares*

En esta etapa, se buscan huellas similares a cada una de las huellas generadas en la etapa anterior. Cuando el número de huellas similares es superior a un límite definido por el operador, se les considera como detecciones. Se guardan los tiempos de inicio de cada huella en un fichero que serán los tiempos de inicio de las detecciones. Si el número de huellas similares es inferior al límite mencionado anteriormente, no se considera esa huella. Así un evento muy grande pero único no será detectado por FAST.

Para buscar huellas similares, se podrían calcular correlaciones cruzadas entre las huellas, pero este proceso es muy costoso en tiempo de cálculo y podría tomar días en procesarse. Para mejorar el tiempo de cálculo, se usa el algoritmo de "Min-Hash" que reduce cada huella a una firma (definido por un arreglo de enteros) y estima la similitud entre las firmas de dos huellas A y B, acelerando así el proceso de búsqueda de huellas similares. Yoon (2015) mostró que esa similitud es igual a la similitud de Jaccard. La similitud de Jaccard $J(A,B)$ es la tasa del tamaño de la intersección de dos conjuntos A y B entre el tamaño de la unión entre ambos conjuntos (Leskovec et al., 2014).

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

La similitud de Jaccard en FAST equivale a comparar los bits que existen en las dos huellas binarias A y B y así obtener una medida de su similitud (Yoon et al., 2015). Se ha establecido que la probabilidad que dos firmas "Min-Hash" sean iguales es igual a la similitud de Jaccard de estas dos huellas (Yoon et al., 2015).

$$\Pr(h(A) = h(B)) = J(A,B)$$

donde $h(A)$ y $h(B)$ corresponden a las firmas de las huellas A y B.

A partir de la base de datos de las huellas obtenidas en 6.2.1, se busca sistemáticamente a las huellas que tengan una alta probabilidad de parecerse a otra huella, evitando realizar la búsqueda con huellas que no se parezcan entre sí. Lo que permite encontrar eventos similares es que la búsqueda utiliza a todas las huellas binarias dentro de la base de datos, lo que maximiza la probabilidad de encontrar huellas similares.

Una analogía para explicar este proceso es como una búsqueda en un archivero, donde se va a buscar un archivo similar a un archivo en específico. Pero la búsqueda será eficiente en la forma en que sólo se hará sobre las carpetas donde podría encontrarse un archivo similar, en lugar de buscar en todas las carpetas en el archivero.

En resumen, tener una alta probabilidad de que las firmas $h(A)$ y $h(B)$ se parezcan, significa que las huellas serán similares, al tener un valor de similitud Jaccard alto. Por lo que el algoritmo “Min-Hash” coloca ambas huellas en el mismo grupo de huellas similares.

Al final, FAST determina los tiempos de arribo de las detecciones más repetidas, es decir caracterizadas por un mayor número de apariciones en el mismo grupo.

## 6.3 Postprocesamiento de las detecciones

Una vez obtenidos las detecciones, se tiene que verificar si están duplicadas, o si son verdaderas o falsas. Esta etapa de postprocesado de las detecciones obtenidas se hace a mano.

### *6.3.1 Eliminar detecciones duplicadas*

Dentro de la lista de detecciones, al utilizar más de una estación para hacer detecciones puede haber detecciones duplicadas. Al ejecutar los métodos por estación se puede observar que un mismo evento puede llegar a tiempos diferentes en diferentes estaciones debido al tiempo de propagación de las ondas hasta cada estación. Eso genera diferentes detecciones que en realidad corresponden al mismo evento. Por lo que se tienen que buscar a las detecciones únicas y descartar de la lista de detecciones a las que son duplicadas para un mismo evento. Para eliminar estas detecciones duplicadas se eliminan las detecciones que se encuentren dentro de una misma ventana de tiempo para la lista. En este caso se conservaron las detecciones que tienen como diferencia 20 segundos entre sí o más, debido a que este fue el tamaño de ventana utilizada por los programas. Sin embargo, se pueden utilizar ventanas con un tamaño en función de la distancia que separa a cada una de las estaciones del volcán.

Este proceso se utilizó tanto para las detecciones de la primera etapa (modo Sistemático) de TM como para las de la segunda etapa (modo Estándar) de TM, debido a que estos fueron ejecutados en cada estación por separado. En el caso de la segunda etapa, es posible que dos *templates* detecten a la misma señal, por lo que además de eliminar las detecciones duplicadas por estación, se eliminaron las detecciones duplicadas por *template*. FAST, se ejecutó con las cuatro estaciones al mismo tiempo, por lo que los tiempos obtenidos fueron ya considerando que se puede haber detecciones duplicadas por más de una estación.

### *6.3.2 Eliminar detecciones falsas*

Con los tiempos de detección de las detecciones no repetidas obtenidos en 6.3.1, se calculó la energía (Fig. 6) para las ventanas de tiempo correspondientes a las detecciones para verificar si estas detecciones son verdaderas o falsas. Esta revisión manual observando la energía de las señales se realizó con los resultados de todas las metodologías.

Se verifica manualmente que la energía de cada detección corresponde a una señal verdadera, es decir que no es una energía de forma aleatoria (Fig. 10). En el caso de las detecciones verdaderas, se observará un aumento de la señal respecto al ruido, mientras que para las falsas detecciones no habrá mucha diferencia entre el ruido y la señal (Fig. 10).

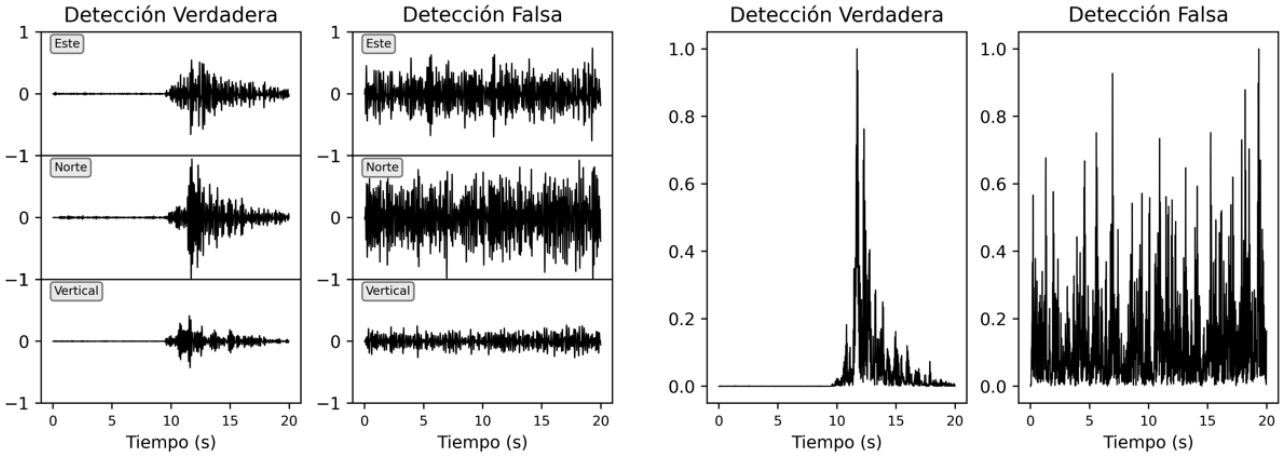


*Figura 10: Comparación entre una detección verdadera y una detección falsa, utilizando las tres componentes (izquierda) y su energía(derecha).*

Se ve que es más fácil identificar falsas detecciones en la energía que en las tres componentes NS, EW y vertical porque en alguna de las tres componentes los niveles de ruido pueden ser más grandes que los niveles de la señal, mientras que, en las otras componentes, no. Al realizar el apilado de las energías, se evita este problema, ya que, al calcular la energía de las señales detectadas, se observa que la diferencia entre ruido y señal se intensifica. Así que para esta revisión manual se decidió usar la energía para discriminar a las señales verdaderas de las falsas.

### 6.4 Clasificación de las detecciones con el método de agrupamiento jerarquizado

Algunos programas necesitan una base previa de familias para después buscarlas en las señales detectadas. Se llama un método supervisado. Estas familias previas se hacen manualmente, donde el usuario da etiqueta a una familia dada, por ejemplo, la llama familia VT, la familia explosión… El programa busca después automáticamente señales parecidas a estas familias. La manera de clasificar las señales previamente al uso de un programa de clasificación de aprendizaje supervisado es subjetiva porque depende altamente de como el usuario etiqueta a cada familia. Por ejemplo, algunos LPs están etiquetados VTs por un usuario y LPs

por otro usuario (McNutt y Roman, 2015). Además, generar tales familias previamente es tardado.

Otros programas como los no-supervisados, es decir aquellos que no necesitan crear previamente bases de familias. Estas bases de familias se generan automáticamente por el programa, así las familias se generan de manera más objetiva y rápida. Es una técnica útil para el descubrimiento de clases nuevas sin necesidad de aporte humano previo (Nielsen, 2016, Saxena et al., 2017, Ran et al., 2023). Esta clasificación se hace de manera posterior a la detección y Nielsen (2016) la define como un agrupamiento, es decir un conjunto de técnicas que buscan lotes compactos de datos que definen familias de datos.

Una de estas técnicas de agrupamiento es el agrupamiento jerarquizado la que consiste en realizar una clasificación binaria, al establecer relaciones jerárquicas entre los datos. Dichas relaciones se obtienen a través de cálculos de distancia o de similitud entre muestras (Ran et al., 2023). Como resultado se obtiene una clasificación en forma de árbol, un dendrograma, donde cada hoja representa a una muestra y cada rama representa una familia. Los dendrogramas se arman utilizando una función de distancia de enlace, esta función se puede calcular mediante el método de pares de grupos pesados con promedio (WPGMA). La WPGMA permite calcular la distancia entre un grupo A y un nuevo grupo D. El grupo D es el resultado de agregar los grupos B y C. Se promedian las distancias de los grupos B y C desde el grupo A y este promedio será la distancia entre el grupo A y el grupo D (Carr, 2007). Es una operación computacionalmente simple y se representa de la siguiente forma:

$$distancia_{A,D} = \frac{1}{2}(distancia_{A,B} + distancia_{A,C})$$

Tradicionalmente se representa el resultado de un agrupamiento jerarquizado por medio de un dendrograma (Fig. 11). Las muestras se separan en diferentes ramas del dendrograma, utilizando un umbral de distancia, con cada una de las ramas representando a una familia de muestras. Viendo al dendrograma de arriba hacia abajo, un valor alto significa que las ramas se separaran más rápido, lo que significa que más señales serán clasificadas, aunque, podría haber familias que posiblemente abarquen a más de una familia.

No se necesita conocer el número de grupos en los que clasificará las muestras, debido a que es un algoritmo de aprendizaje no supervisado. Sin embargo, tiene la desventaja de ser computacionalmente costoso, sobre todo al aplicarse en bases de datos con un gran número de muestras (Kotsiantis y Pintelas, 2004).

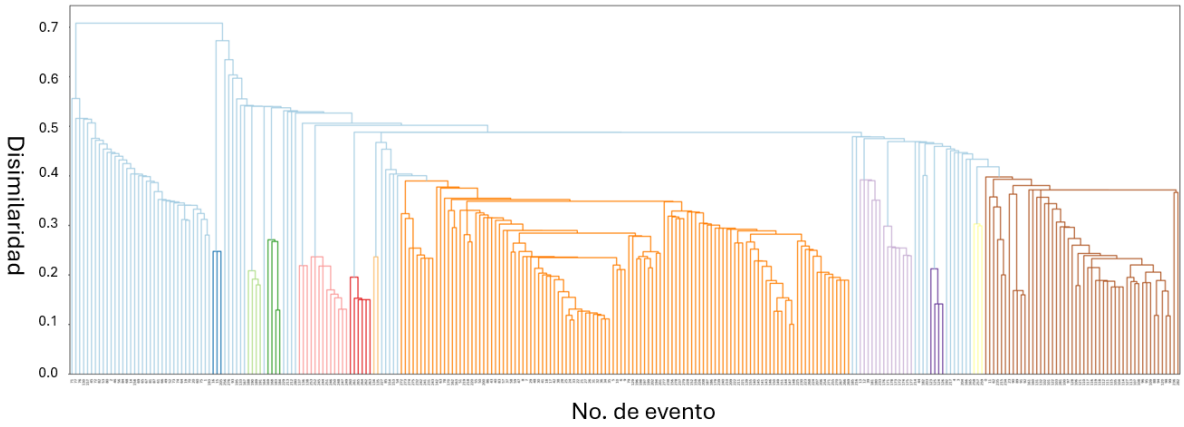


*Figura 11: Ejemplo de dendrograma.*

En esta etapa, se tienen solamente a las detecciones verdaderas, ya sin falsas detecciones. Estas son clasificadas con un método de agrupamiento jerarquizado que el código *Template Matching* contempla. FAST no tiene esta rutina de clasificación.

La rutina de agrupamiento jerarquizado incluida en el paquete de *RapidAlignerSeismic* de TM hace un cálculo del CCN, con ayuda del GPU, entre todas las muestras para obtener la matriz de covarianza. La rutina construye una matriz de covarianza por componente, para después juntarlas en una sola. Para este trabajo de tesis se utiliza el cálculo de la energía (Fig. 6) en lugar de las 3 componentes para la clasificación, lo que reduce el número de operaciones que se tienen que realizar en un factor de tres, ya que sólo se construye una matriz de covarianza.

Después de construir la matriz de covarianza la rutina de agrupamiento jerarquizado calculó la función de distancias de enlace utilizando una función WPGMA. Con esta función de distancias de enlace y un umbral de distancia, el cual para este trabajo tuvo un valor de 0.88, se construyó el dendrograma. Finalmente se extrajeron las distintas familias del dendrograma usando la función de enlace.

# 7. Resultados.

Primero se muestran las detecciones obtenidas en la primera etapa (modo Sistemático) de *Template Matching* (TM), que genera los *templates*. Se menciona el tiempo de ejecución para las tres bandas de frecuencias utilizadas. Se muestran después los resultados correspondiendo al filtro más conveniente (más rápido y con menos falsas detecciones). Se continua con la estadística de detecciones crudas obtenidas por el programa por estación y como cambia esta después de realizar el control de calidad manual con la energía. Posteriormente se muestran las detecciones realizadas por la segunda etapa (modo Estándar) de TM, comenzando con las obtenidas al usar las tres componentes (EW, NS y vertical) en velocidad, continuando con las detecciones usando la energía. Luego, se muestran las detecciones realizadas con la metodología FAST. Se muestra luego una comparación entre las detecciones obtenidas con TM usando velocidad y usando energía, así como las obtenidas por TM en energía y por FAST. Finalmente se muestran los resultados obtenidos por la clasificación de las detecciones obtenidas por TM en energía.

## 7.1 Primera etapa de TM (Modo Sistemático)

Los tiempos de ejecución para los 3 filtros usados para la primera etapa son:

- Para el filtro entre 0.05 a 1 Hz:  Tardó en ejecutarse un promedio de 12 minutos por estación.
- Para el filtro entre 5 a 10 Hz: Tardó en ejecutarse un promedio de 14 horas por estación.
- Para el filtro entre 0.05 a 10 Hz: Tardó en ejecutarse entre 10 y 14 horas dependiendo la estación.

El segundo filtro es el más adecuado para detectar señales sin muchas falsas detecciones y con un tiempo de ejecución relativamente corto. Se mostrarán a continuación los resultados obtenidos utilizando ese segundo filtro.

Se ha aplicado la primera etapa (modo Sistemático) en las tres componentes de cada una de las cuatro estaciones, es decir que al final se obtienen 12 listas de *templates*. La suma de los *templates* de estas 12 listas es 36,077. Para saber a qué número de señales efectivas corresponden, se selecciona las detecciones que aparecen en las tres componentes en cada estación lo que corresponde a 2,530 señales efectivas.

Se observaron los *templates* encontrados para cada estación y sus tiempos para ver si un mismo *template* se puede encontrar en más de una estación. Estos resultados se recopilan en la Tabla 2:

| Estación | PPIG | PBP | PCP | PXP |
|---|---|---|---|---|
| *Templates* Totales | 87 | 1,651 | 184 | 608 |
| *Templates* exclusivos de la estación | 28 | 1531 | 131 | 486 |
| *Templates* compartidos con: PPIG | | 18 | 4 | 8 |
| *Templates* compartidos con: PBP | 18 | | 7 | 68 |
| *Templates* compartidos con: PCP | 4 | 7 | | 9 |
| *Templates* compartidos con: PXP | 8 | 68 | 9 | |
| *Templates* compartidos con: PPIG y PBP | | | 3 | 6 |
| *Templates* compartidos con: PPIG y PCP | | 3 | | 12 |
| *Templates* compartidos con: PPIG y PXP | | 6 | 12 | |
| *Templates* compartidos con: PBP y PCP | 3 | | | 11 |
| *Templates* compartidos con: PBP y PXP | 6 | | 11 | |
| *Templates* compartidos con: PCP y PXP | 12 | 11 | | |
| *Templates* compartidos con todas | 8 | 8 | 8 | 8 |
| *Templates* Totales (%) | 3.43 | 65.25 | 7.27 | 24.03 |

*Tabla 2: Número de templates por estación, donde se observa cuantos coinciden entre estaciones*

Un mismo *template* puede ser encontrado por más de una estación, por lo que hay que eliminar a los *templates* repetidos, para obtener los tiempos de los *templates* no repetidos. De estos *templates*, 2,327 de los 2,530 no fueron detectados por las demás estaciones, mientras que las 153 *templates* restantes fueron detectados en

más de una estación, es decir el 6% de los *templates* tiene la característica de ser detectado por más de una estación.

Los *templates* hechos por la estación Canario (PBP), representan al 65% de los *templates* de esta primera etapa, mientras que las otras estaciones: Tlamacas (PPIG), Colibrí (PCP) y Chipiquixtle (PXP) hicieron 3%, 7% y 24% respectivamente. Las estaciones que más *templates* compartieron entre ellas fueron PBP y PXP respectivamente, al tener un mayor número de detecciones.

Se realizó una revisión manual de los *templates* mediante el cálculo de la energía (Fig. 6) para cada uno de los *templates* para observar cuales son detecciones verdaderas y cuales son falsas (Fig. 10). En la Tabla 3 se puede observar un ejemplo del análisis de los *templates* encontrados en la primera etapa. Después del proceso, los *templates* disminuyeron a un valor de 740 en las 4 estaciones, es decir que 70.75% de los *templates* encontrados en esta primera etapa corresponden a una falsa detección (*template*).

| Estación | PPIG | PBP | PCP | PXP |
|---|---|---|---|---|
| *Templates* Totales | 77 | 457 | 52 | 154 |
| *Templates* exclusivos de la estación | 28 | 367 | 13 | 71 |
| *Templates* compartidos con: PPIG | | 18 | 5 | 7 |
| *Templates* compartidos con: PPPB | 18 | | 6 | 45 |
| *Templates* compartidos con: PPPC | 5 | 6 | | 8 |
| *Templates* compartidos con: PPPX | 7 | 45 | 8 | |
| *Templates* compartidos con: PPIG y PPPB | | | 3 | 6 |
| Templates compartidos con: PPIG y PPPC | | 3 | | 5 |
| *Templates* compartidos con: PPIG y PPPX | | 6 | 5 | |
| *Templates* compartidos con: PPPB y PPPC | 3 | | | 7 |
| *Templates* compartidos con: PPPB y PPPX | 6 | | 7 | |
| *Templates* compartidos con: PPPC y PPPX | 5 | 7 | | |
| *Templates* compartidos con todas | 5 | 5 | 5 | 5 |
| *Templates* Totales (%) | 10.40 | 49.59 | 7.02% | 20.81 |

*Tabla 3: Número de templates por estación, después de revisar la energía de los templates para cada una de sus ventanas, donde se observa cuantas templates coinciden entre estaciones*

Después de buscar los *templates* que se repiten para las demás estaciones, se encontraron 115 de los 594 *templates* con esta característica, es decir un 20%. En la Tabla 4 se presenta un resumen de la revisión manual de la energía de los *templates*.

| Estación | PPIG | PBP | PCP | PXP | *Templates* totales | *Templates* no repetidos | *Templates* en común |
|---|---|---|---|---|---|---|---|
| *Templates* crudos | 87 | 1,651 | 184 | 608 | 2,530 | 2,327 | 153 |
| *Templates* después de calcular su energía | 77 | 457 | 52 | 154 | 740 | 594 | 115 |
| *Templates* eliminados (%) | 11.49 | 72.31 | 71.73 | 74.60 | 70.75 | 74.47 | 24.83 |

*Tabla 4: Comparación entre las templates crudas y después de revisar la energía de los templates.*

En las siguientes figuras se observan algunos ejemplos de los *templates* detectados en esta primera etapa. Se mostrará el registro, con las ventanas de tiempo en las que abarca el *template*. Los días 8 y 10 de julio son mostrados, ya que son algunos de los días con mayor actividad sísmica y volcánica durante el mes de julio (CENAPRED, 2017).

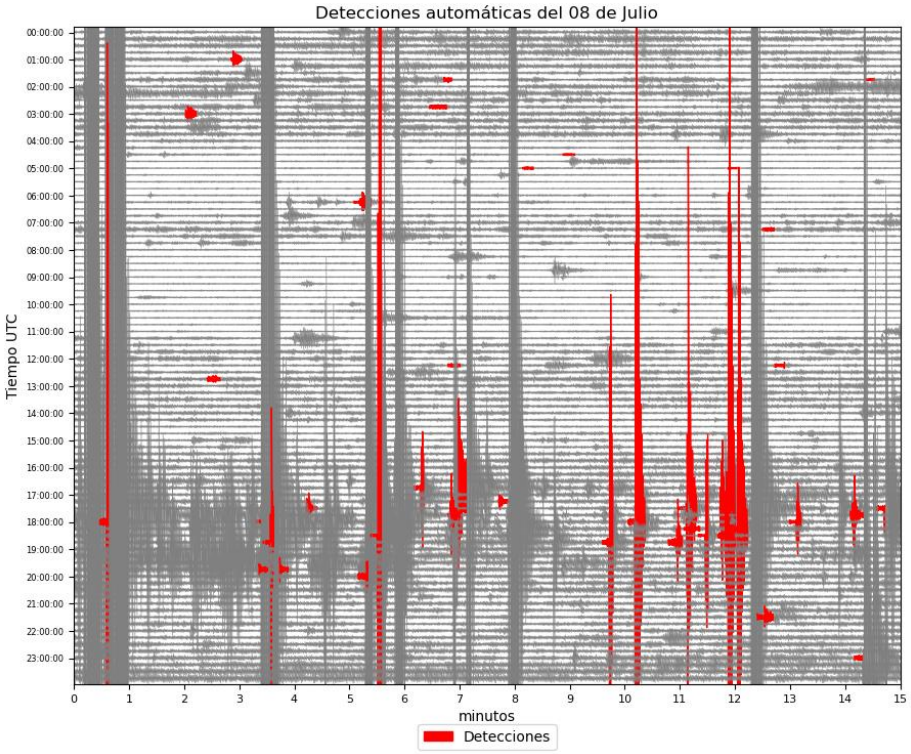


*Figura 12: Templates encontrados por la primera etapa el 8 de julio (en rojo)*

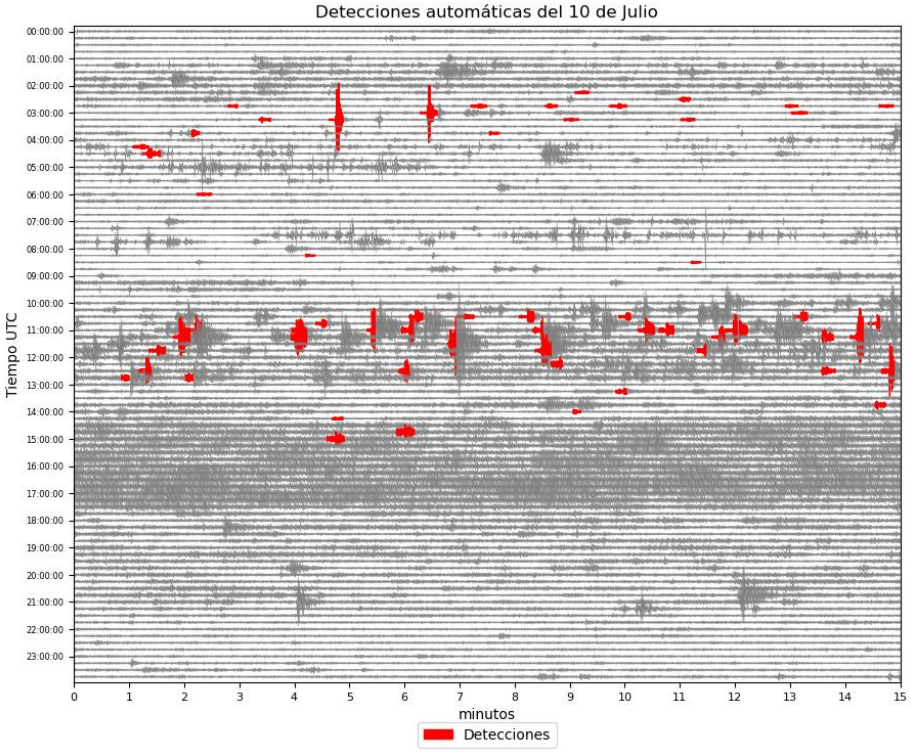


*Figura 13: Templates encontrados por la primera etapa el 10 de julio (en rojo).*

### 7.2 Segunda etapa de TM (Modo Estándar) utilizando las tres componentes

Con los *templates* generados en la primera etapa, se ejecutó la segunda etapa (modo **Estándar**) para detectar señales similares a los *templates* dentro de la señal continua. Utilizando las tres componentes en velocidad el programa tarda 6 horas y media para obtener las detecciones para las 4 estaciones durante el mes de julio. En esta segunda etapa de detecciones (modo **Estándar**), se obtuvieron 3,588 detecciones para las cuatro estaciones. A continuación, en la Figura 14 se muestran ejemplos de las detecciones realizadas en la segunda etapa.

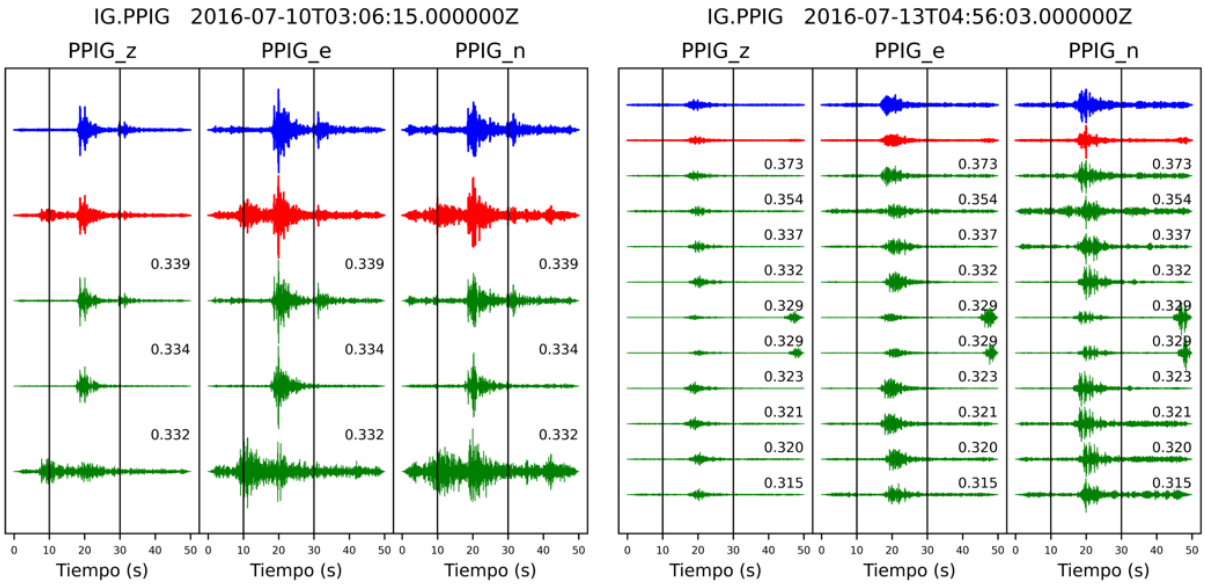


*Figura 14: Ejemplos de detecciones hechas durante la segunda etapa usando las tres componentes. El template utilizado (en azul), el apilado de todas las detecciones realizadas (en rojo) y las mejores detecciones realizadas por el template (en verde).*

En esta segunda etapa, se tiene que hacer un postprocesado, al igual que en la primera etapa. El postprocesado recopila todas las detecciones para eliminar las detecciones duplicadas, esto hace que se reduzca a 2,589 detecciones de los 3,588 originales, es decir un 72% de las detecciones realizadas correspondieron a detecciones no repetidas. De estas detecciones se realizó una rutina de control de calidad manual al calcular la energía de las detecciones y de esta manera observar cuales son verdaderas y cuales falsas. Con este proceso se obtuvieron 1,320 detecciones verdaderas y 1,269 falsas, es decir el 51% de las detecciones fueron

verdaderas. En las siguientes figuras se muestran las detecciones no repetidas y verdaderas para las detecciones encontradas en la segunda etapa con las tres componentes.

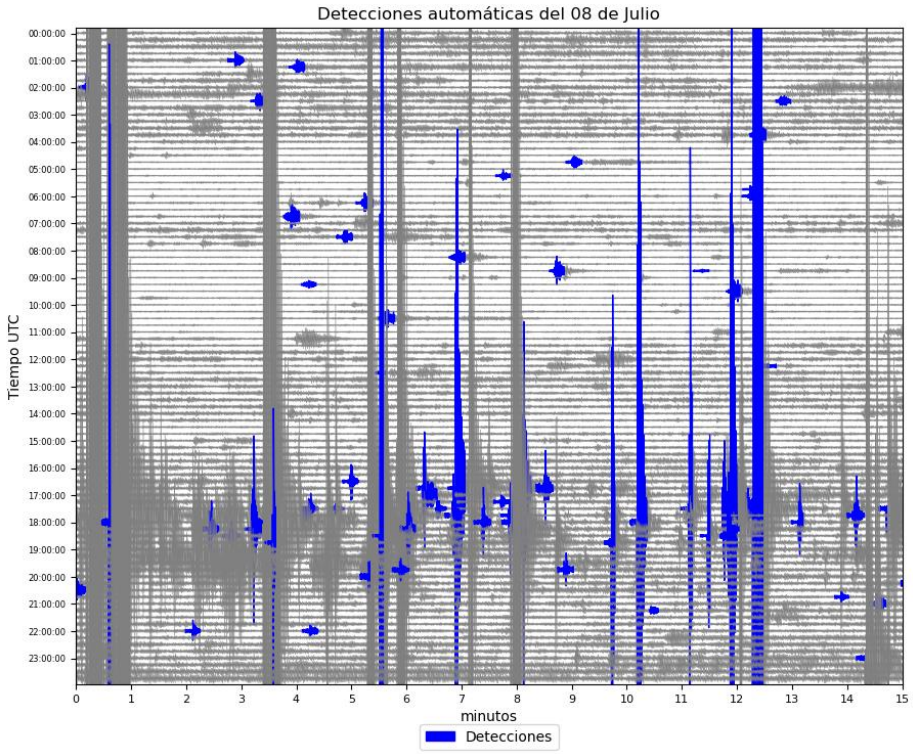


*Figura 15: Detecciones realizadas por la segunda etapa con las tres componentes el 8 de julio (en azul).*

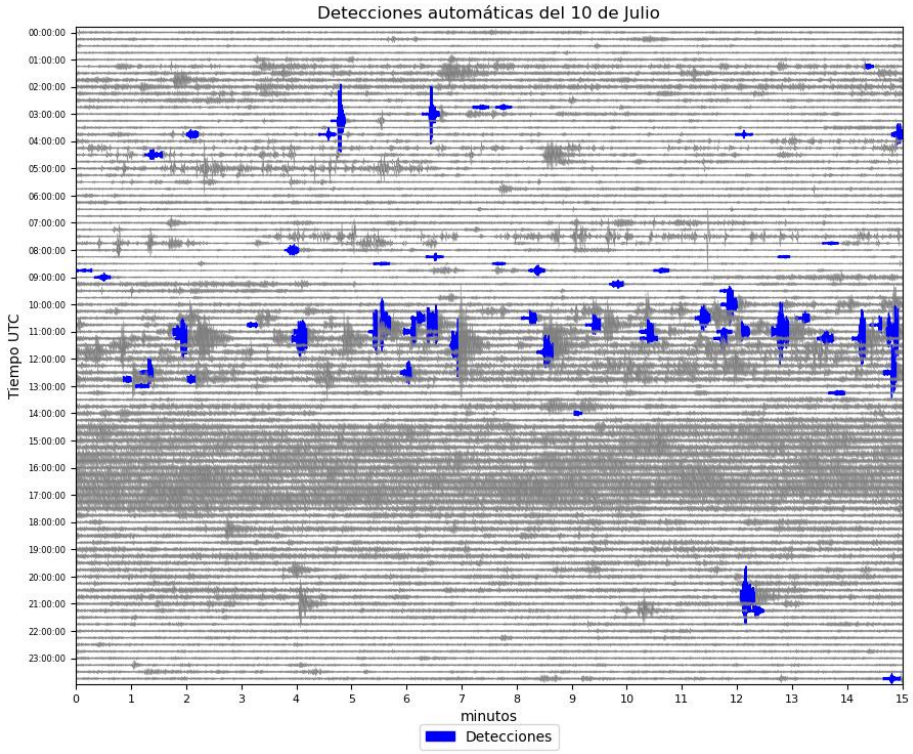


*Figura 16: Detecciones realizadas por la segunda etapa con las tres componentes el 10 de julio (en azul).*

### 7.3 Segunda etapa de TM utilizando la energía

La segunda etapa es repetida, pero ahora usando la energía en lugar de las tres componentes. Se tardó poco menos de 3 horas para realizar las detecciones con las 4 estaciones durante el mes de julio. Se obtuvieron en total de 5080 detecciones, utilizando la energía. Algunos ejemplos de las detecciones se muestran en la Figura 17 (a pesar de que fueron detectadas con la energía, se muestran las detecciones en velocidad para poder comparar con los resultados de 7.2).

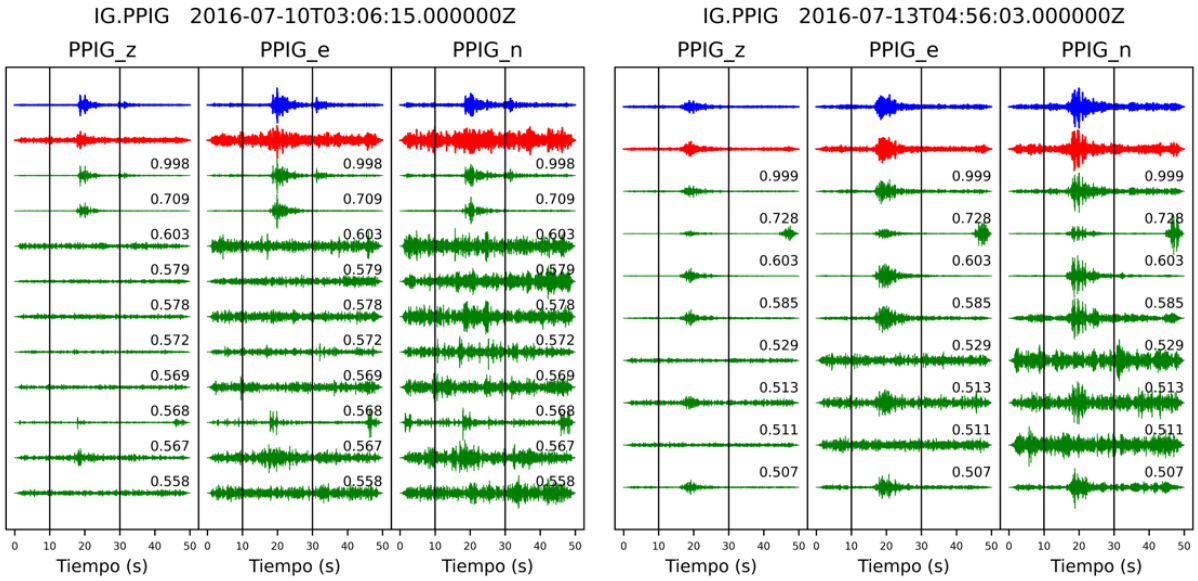


*Figura 17: Ejemplos de detecciones realizadas usando energía. El Template utilizado (en azul), el apilado de todas las detecciones realizadas (en rojo) y las mejores detecciones realizadas por el template (en verde).*

Después de hacer detecciones usando la energía, se tiene que hacer otra vez el postprocesado para eliminar las detecciones duplicadas, las detecciones se redujeron a 2,269. Un 45% de las detecciones de la segunda etapa usando la energía corresponde a detecciones no repetidas. Luego de realizar la revisión manual con la energía de las detecciones se obtuvo un total de 1,384 detecciones. Lo que significa que el 61% de detecciones no repetidas con el modo Estándar utilizando energía corresponde a detecciones verdaderas. Estas detecciones se pueden observar, para los días previamente mencionados, en las siguientes figuras.

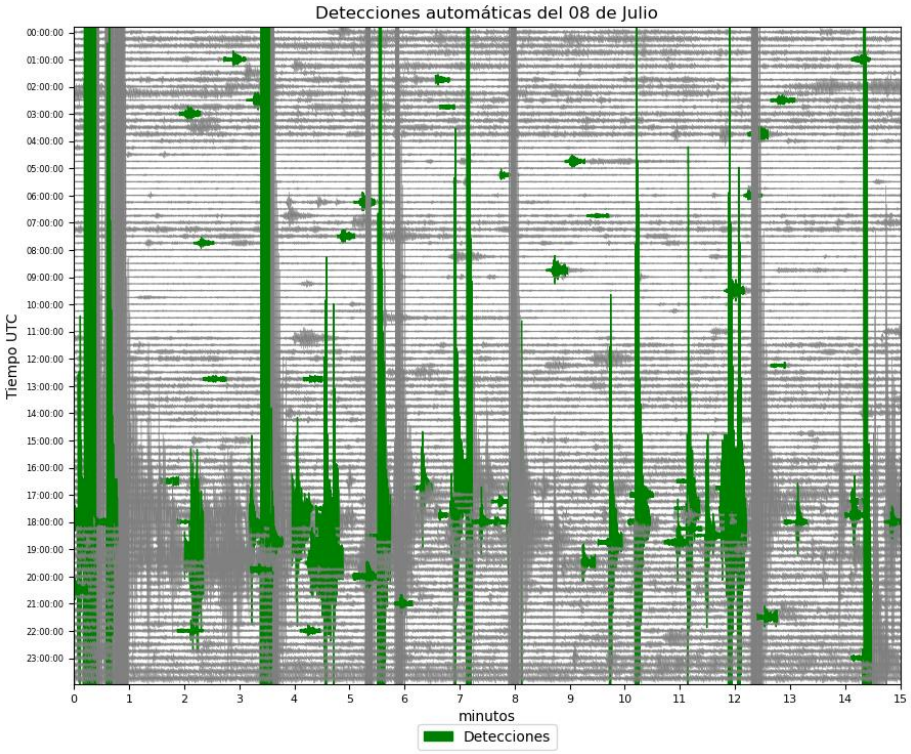


*Figura 18: Detecciones realizadas por la segunda etapa usando energía para el 8 de julio (en verde).*

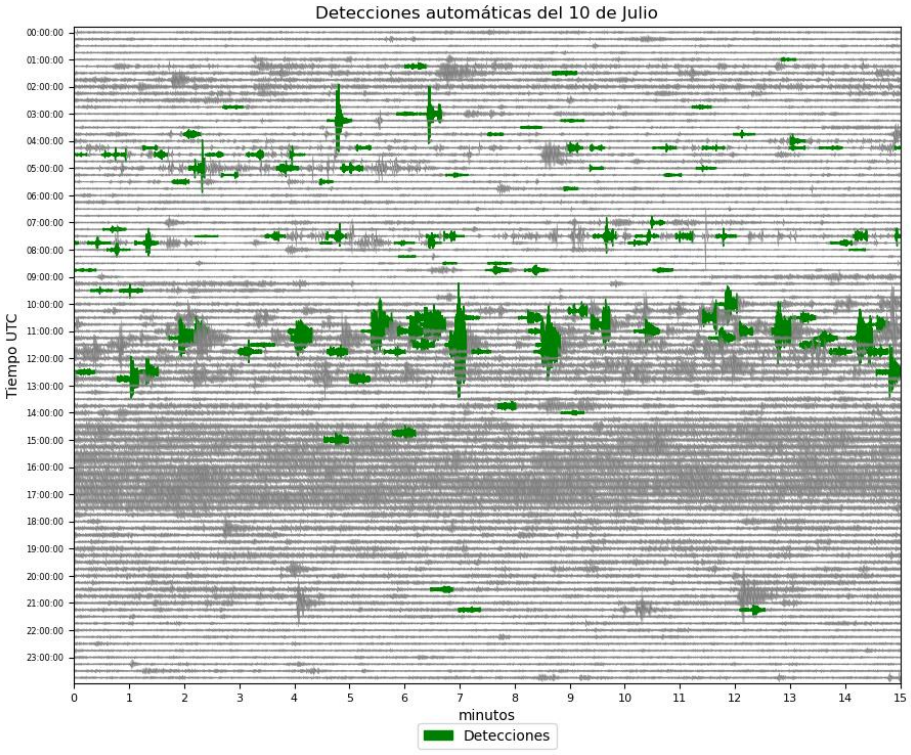


*Figura 19: Detecciones realizadas por la segunda etapa usando energía para el 10 de julio (en verde).*

## 7.4 FAST

FAST se ejecutó para las cuatro estaciones de la base de datos, y se tardó aproximadamente 7 horas y media para realizar las 1,709 detecciones. Para postprocesar a estas detecciones, no fue necesario remover detecciones duplicadas, ya que el programa las elimina. Sin embargo, si se realizó una revisión manual de las ventanas marcadas por las detecciones al observar su energía (de la misma manera como se describió anteriormente) y se vio que 1,075 de las 1,709 fueron detecciones verdaderas, es decir alrededor de un 62%. En las siguientes figuras se muestran algunos ejemplos de las detecciones que FAST puede realizar.

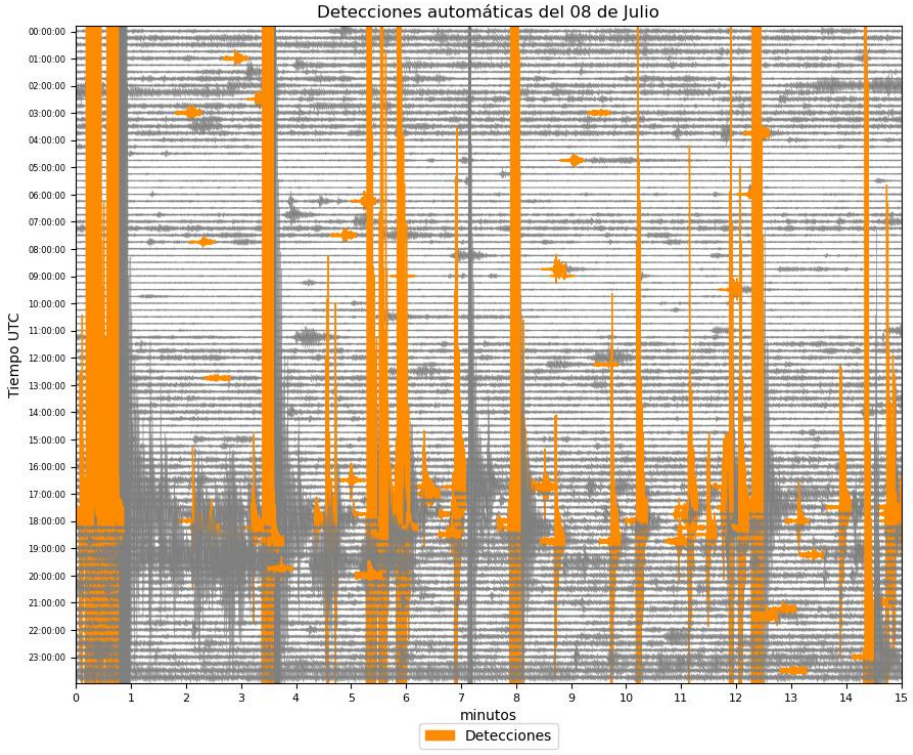


*Figura 20: Detecciones realizadas por FAST el 8 de julio (en naranja).*

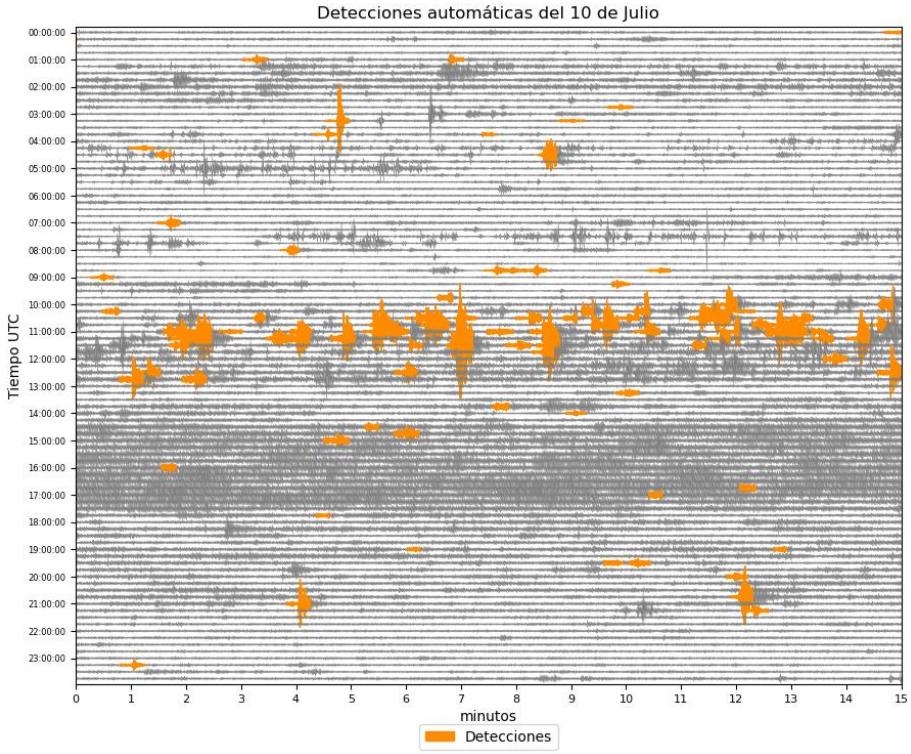


*Figura 21: Detecciones realizadas por FAST el 10 de julio (en naranja).*

### 7.5 Comparación de detecciones con las tres componentes y la energía para TM

Se muestra una comparación de las capacidades de detección de las dos maneras de TM, es decir a partir de la forma de onda de las tres componentes y de la energía.

De las 1,320 detecciones realizadas en las tres componentes (en azul), y de las 1,384 detecciones hechas con la energía (en verde), hay una coincidencia de 450 detecciones (magenta). Se muestran las detecciones en las siguientes figuras para los días ya mencionados:

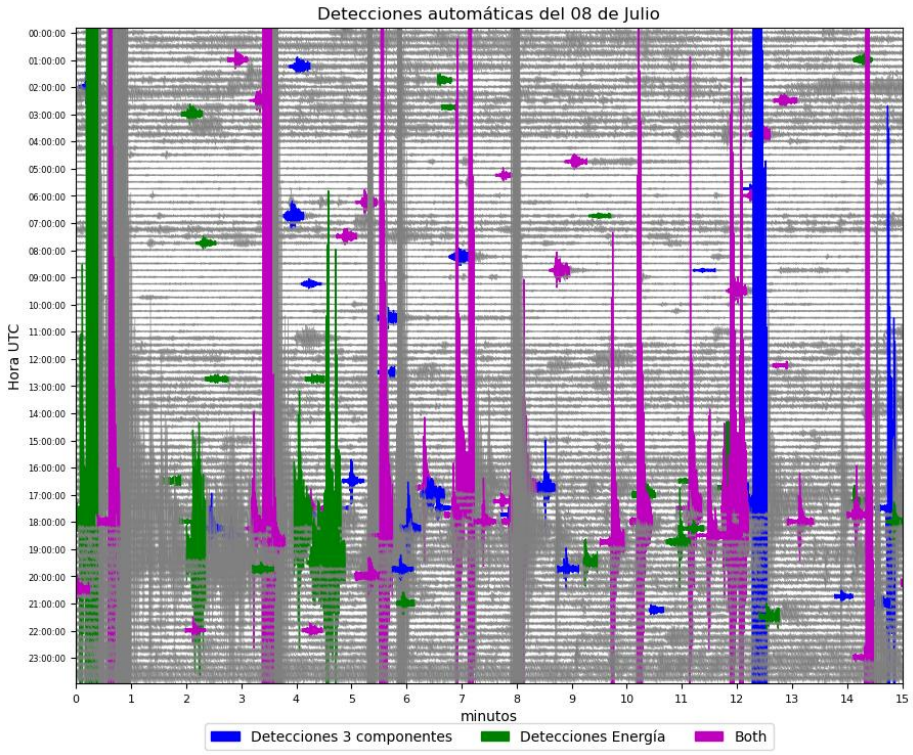


*Figura 22: Detecciones hechas con TM en velocidad y con energía el 8 de julio. Detecciones con las tres componentes (en azul). Detecciones hechas con la energía (en verde). Detecciones hechas por ambos métodos (magenta).*

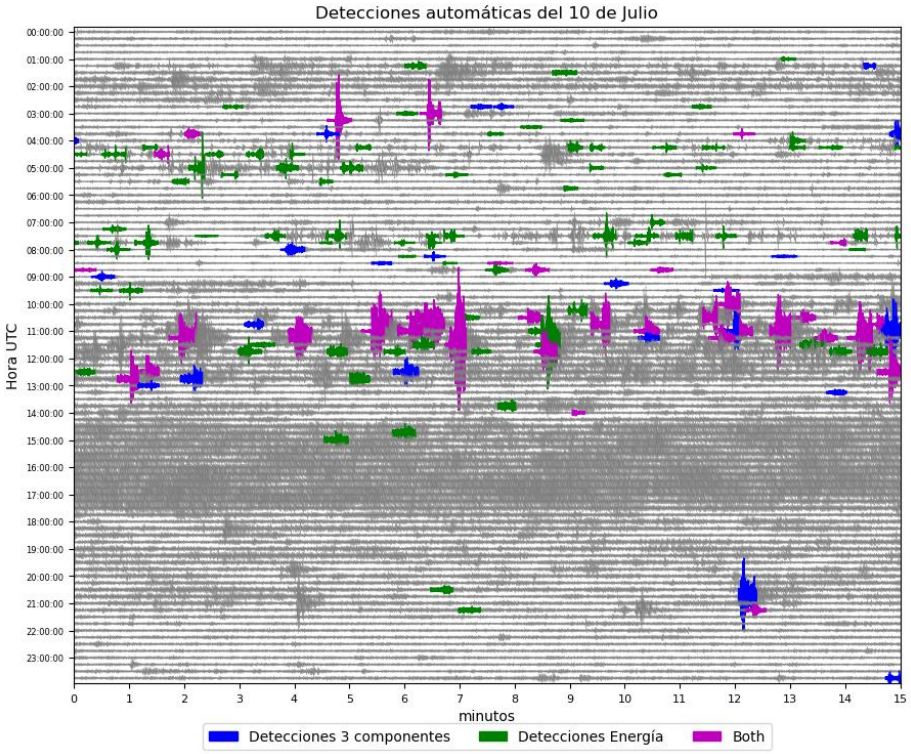


*Figura 23: Detecciones hechas con TM en velocidad y con energía el 10 de julio. Detecciones con las tres componentes (en azul). Detecciones hechas con la energía (en verde). Detecciones hechas por ambos métodos (magenta).*

### 7.6 Comparación entre detecciones hechas por TM en energía y por FAST

Se compararon las detecciones obtenidas a partir de los dos métodos, las 1,384 de *Template Matching (verde)* y las 1,075 de FAST (naranja). Los dos métodos hicieron 394 detecciones en común (magenta).

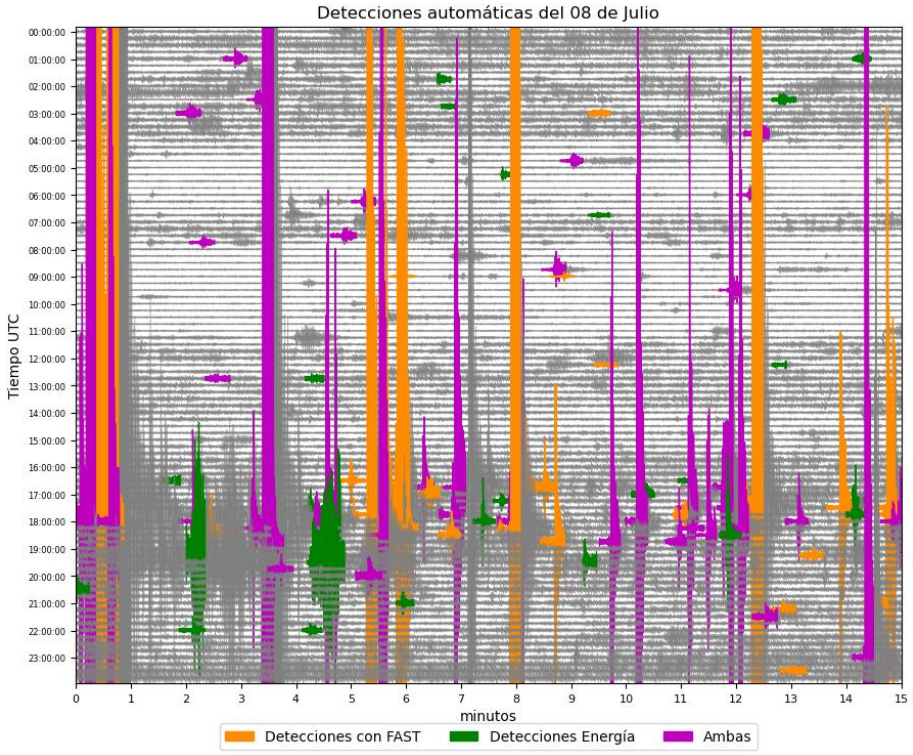


*Figura 24: Detecciones de TM en energía y de FAST el 8 de julio. Detecciones de Template Matching* (verde)*, FAST (naranja). Detecciones hechas por ambos métodos (magenta).*

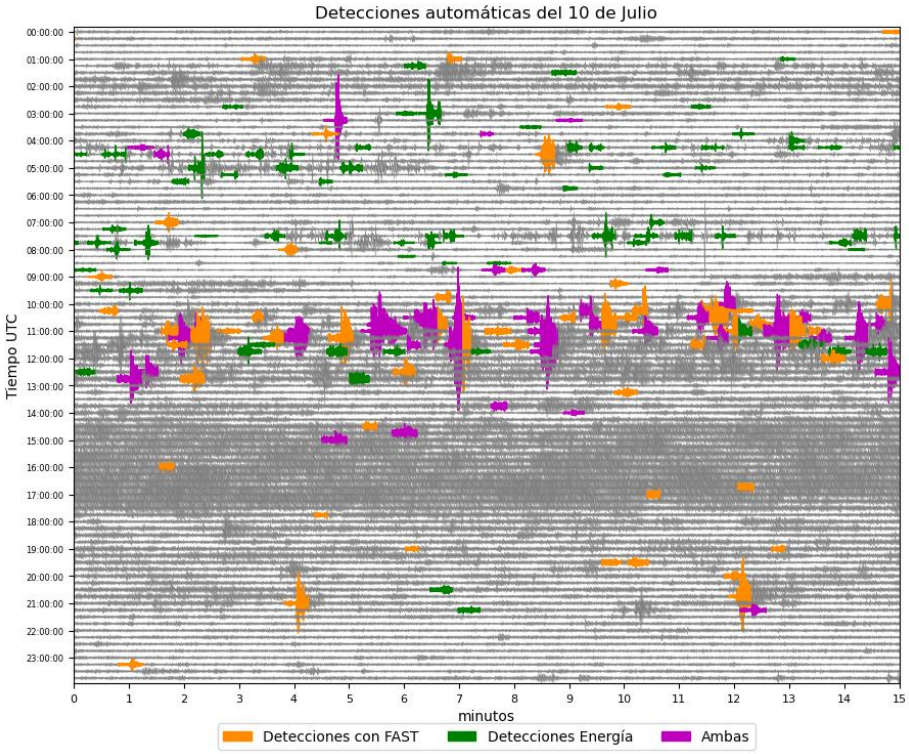


*Figura 25: Detecciones de TM en energía y de FAST el 10 de julio. Detecciones de Template Matching* (verde)*, FAST (naranja). Las detecciones hechas por ambos métodos (magenta).*

### 7.7 Clasificación de señales de TM con energía.

Con las 1,384 detecciones obtenidas con el *Template Matching* usando la energía, se obtuvieron 9 familias. Dado a que este es un proceso de agrupamiento automatizado, sin etiquetado de familias previo, las familias reciben un nombre genérico, es decir su color y su código de color.

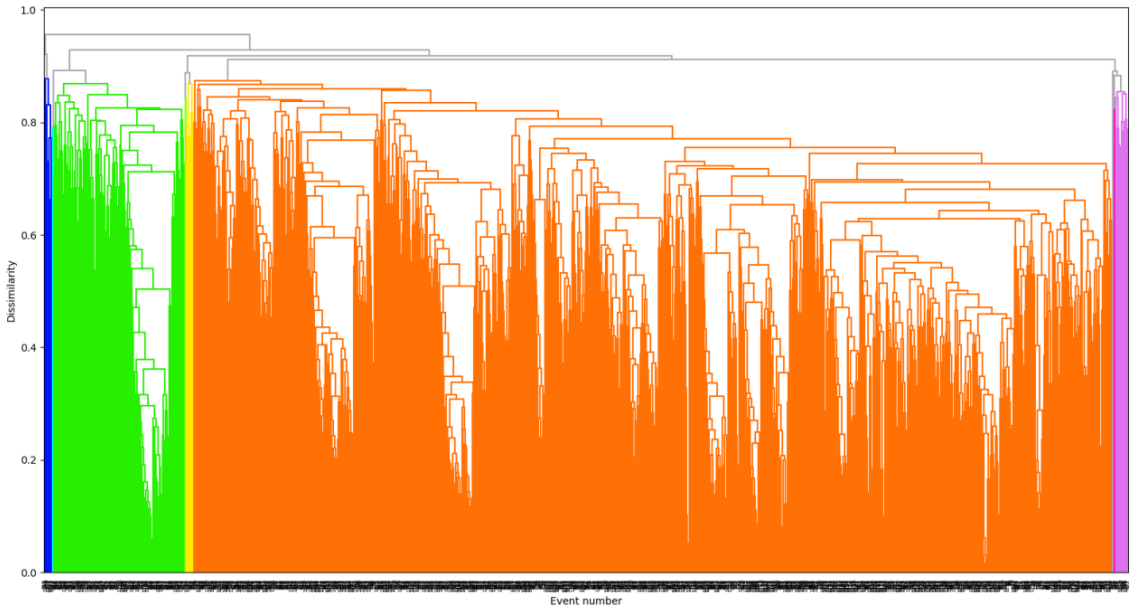


*Figura 26: Dendrograma con las principales familias*

Las familias son las siguientes:

| Color | Código de color | Señales en la familia |
|---|---|---|
| Azul | #0012ff | 11 |
| Azul claro | #00fee0 | 2 |
| Verde | #26ef00 | 169 |
| Verde Claro | #97ff1d | 2 |
| Amarilla | #ffe900 | 9 |
| Naranja | #ff7005 | 1170 |
| Rosa | #ff00b7 | 3 |
| Lila | #dd71f0 | 16 |
| Sin clasificar | #aaaaaa | 2 |

*Tabla 5: Familias del dendrograma.*

A continuación, se muestran ejemplos de las señales dentro de algunas de las familias, así como su apilado para establecer un promedio de las señales dentro de las familias. Se muestran ejemplos de la familia naranja (#ff7005).

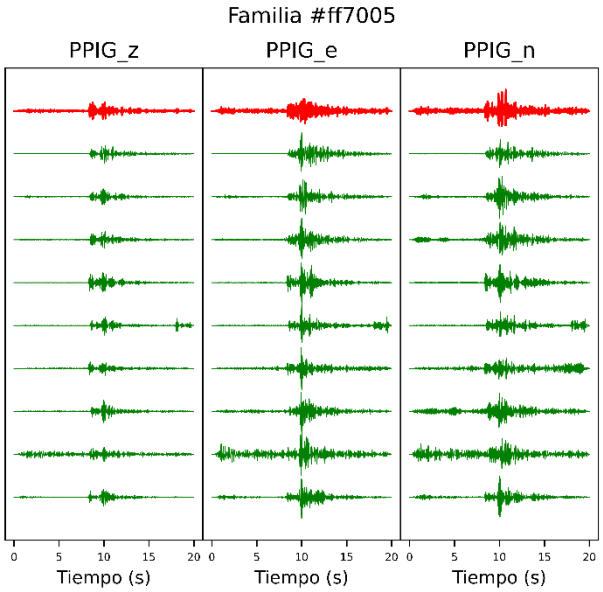


*Figura 27: Señales VT clasificadas en la familia naranja (#ff7005). Apilado de las señales(rojo) Elementos de la familia (verde).*

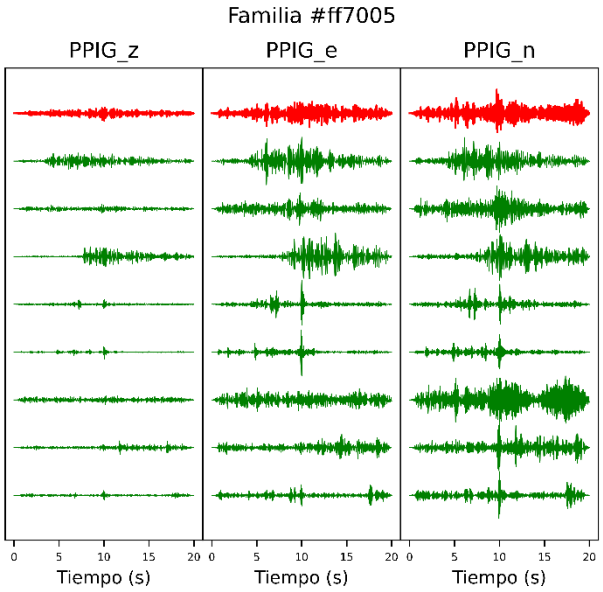


*Figura 28: Señales LP clasificadas en la familia naranja (#ff7005). Apilado de las señales(rojo) Elementos de la familia (verde).*

# 8. Discusión.

## 8.1 Selección del mejor filtro

El uso de las diferentes técnicas se puede hacer sobre los datos crudos o sobre los datos filtrados en diferentes bandas de frecuencias en función de las señales buscadas. El objetivo del trabajo era comparar los dos métodos TM y FAST. El uso del primer filtro (entre 0.05 Hz y 1 Hz) tiene como objetivo detectar eventos de largo periodo, incluyendo el tremor. Pero ninguna señal fue detectada en este rango de frecuencia con TM y FAST en el mes de julio de 2016. El segundo filtro (entre 5 Hz y 10 Hz) fue escogido para detectar principalmente los VTs, y estos fueron efectivamente detectados por FAST y TM (Fig. 5). El tercer filtro (entre 0.05 Hz y 10 Hz) era otro intento para detectar los VTs y señales con frecuencias más bajas como los LPs. Para el *Template Matching* (TM), el primer filtro fue el más rápido que los dos otros filtros. Los cálculos se hicieron en menos de 12 minutos para cada estación y por un mes de datos continuos con este primer filtro, pero no detecto ningún evento volcánico. El tiempo de ejecución con el segundo y tercer filtro fue de 14 horas por estación. A pesar de ser más tardado, se decidió utilizar el segundo filtro porque este es el que detecta más señales (VTs) sin tantas falsas alarmas. Se usará también este segundo filtro para FAST. El segundo filtro permite detectar de mejor manera señales VTs, por lo que las señales de este tipo son las principales señales para detectar en este trabajo.

## 8.2 Estadísticas sobre los *templates* obtenidos en la primera etapa

En la primera etapa, 70% de los *templates* realizadas en PBP, PCP y PXP son falsos (Tabla 4). En cambio, solamente 11% de los *templates* antes del postprocessdo detectados en PPIG son falsos. Pero cuidado, eso no significa que PPIG haya detectado la mayor cantidad de *templates* verdaderos después del postprocesado. De hecho, la estación que detecto la mayor cantidad de *templates* después del postprocesado es la estación PBP (Canario). En este sentido, Canario es la mejor estación para realizar *templates* con la primera etapa

En la primera etapa, la mayor parte (93.5%) de los *templates* verdaderos y no repetidos obtenidos fueron detectados en una sola estación. Sólo 4.9% están detectados en 2 estaciones, 1.3% están detectados con 3 estaciones y 0.3% con 4 estaciones (Tabla 2) Una de las explicaciones es que los *templates* detectados por una estación son sismos pequeños grabados solamente en una estación. Al tener una detección realizada en más de una estación, se puede tener más certeza de que sea una señal verdadera.

El otro punto es que la primera etapa no detecta todas las señales posibles. Por ejemplo, si la señal no se ve en dos o tres componentes, no la va a detectar. De hecho, esa es la razón por la cual el uso de la energía es mejor ya que evita ese problema. En conclusión, para poder detectar muchos eventos pequeños cerca de una estación, es mejor usar los *templates* detectados en esta estación y así detectar los eventos de esta misma estación. En cambio, si se quiere detectar eventos más grandes detectados en varias estaciones, es mejor usar los *templates* grabados en varias estaciones, a pesar de que existan en menor cantidad.

### 8.3 Ventajas del uso de la energía

La primera ventaja de usar la energía es aumentar la razón señal entre ruido, lo que permite detectar más eventos verdaderos. La segunda ventaja es poder detectar eventos cuando uno o dos componentes de un sismómetro no funciona. La tercera es usar la energía para detectar falsos *templates* o falsas detecciones en la etapa del postprocesado. La cuarta ventaja es que se usa una componente (la energía) en vez de las tres componentes (NS, EW, vertical), es decir que reduce la cantidad de datos a procesar por 3, lo que reduce por un factor de aproximadamente dos el tiempo de ejecución. La quinta ventaja es que los coeficientes de correlación normalizado (CCN) son más altos usando la energía respecto al uso de las tres componentes (Figs. 14 y 17). Las señales detectadas que usan la energía son mejores cuanto a la razón señal sobre ruido y es más probable que estén relacionadas a eventos verdaderos. Por ejemplo, al usar la energía se detectan más señales en el

enjambre de VT's (Figs. 18 y 22) y en la secuencia de LP's (Figs. 19 y 23) que cuando se usan las tres componentes (Figs. 15 y 16).

### 8.4 Comparación con detecciones manuales

La eficacia de los métodos de detección automática se puede comprobar al compararlo con detecciones realizadas con un catálogo de detecciones manuales. En este caso se compararon las detecciones automáticas con un catálogo de detecciones manuales de VTs (J. Lermo-Samaniego, comunicación personal, febrero 2024) correspondiendo al 8 de julio 2016. Este catálogo detecta 30 VTs mientras que el método TM detecta solamente 22 VTs registrados en 4 estaciones, de los reportados en el catálogo de Lermo. En cambio, TM detecta más eventos cuando se detecta los VTs en una sola estación. FAST detecta a los 30 VTs con 4 estaciones detectados manualmente (Tabla 6).

| Método | Manual | TM | FAST |
|---|---|---|---|
| VTs detectados durante el enjambre | 30 | 22 | 30 |

*Tabla 6: Comparación de las detecciones hechas durante el enjambre del 8 de julio entre las detecciones a mano reportadas por el primer catálogo y las automáticas por TM y por FAST.*

Adicionalmente se compararon las detecciones realizadas automáticamente con las realizadas manualmente de otro catálogo de VTs (CENAPRED, comunicación personal, junio 2024), este catálogo corresponde al mes de julio de 2016. Este catálogo registra 75 VTs en todo el mes, el método TM detecta 56 VTs usando las 4 estaciones de los reportados por el catálogo de Valdés, es decir un 74%. FAST usando las cuatro estaciones puede detectar 65 VTs de los reportados manualmente, es decir un 84%. El catálogo del CENAPRED reporta 31 eventos durante el enjambre del 8 de julio de 2016, de los que TM puede detectar a 20, mientras que FAST puede detectar a 30 de los VTs del enjambre (Tabla 7).

| Método | Manual | TM | FAST |
|---|---|---|---|
| VTs detectados durante el mes | 75 | 56 | 65 |
| VTs detectados durante el enjambre | 31 | 20 | 30 |

*Tabla 7: Comparación de las detecciones hechas durante el mes de julio, incluyendo el enjambre del 8 de julio, entre las detecciones a mano reportadas por el segundo catálogo y las automáticas por TM y por FAST.*

Al comparar los catálogos manuales de VTs del mes de julio de 2016 con las detecciones automáticas, se observa que el mejor método para detectar VTs es FAST. FAST puede detectar la mayoría de los VTs reportados por los catálogos manuales, detectando casi todos los VTs del enjambre del 8 de julio en ambos catálogos. Los resultados de TM dependen mucho de los templates detectados durante la primera etapa, por lo que es necesario tener una base de datos más grande para hacer más detecciones. El código TM está todavía en desarrollo y aún puede mejorar.

Es posible que los eventos reportados manualmente que no son detectados automáticamente se deban a que no existen señales similares a los *templates* utilizados en TM o a las huellas generadas por FAST. Para mejorar las detecciones automáticas y obtener resultados similares a la detección manual se puede utilizar una base de datos con más estaciones o de mayor duración.

### 8.5 Comparación entre FAST y TM

El enjambre de VT's del 8 de julio fue detectado por los dos códigos FAST y TM (Figs. 12, 15, 18 y 20) pero no con la misma eficiencia. FAST detectó más eventos (30 VTs) y es entonces es el mejor método para detectar VTs (Figs. 20 y 24). El código TM detectó solamente 22 VTs (Fig. 18).

Durante la secuencia de LPs del 10 de julio (Figs. 13, 16, 19, 21) reportada por CENAPRED (2017) FAST detectó una menor cantidad de LPs (34) que el código TM (43) (Fig. 25). El código TM, puede detectar más LPs durante esta secuencia usando la energía (43) que usando las tres componentes (32) (Fig. 23). Eso se puede atribuir

al aumento de la razón señal sobre ruido en los LPs, ya que estos tienen generalmente una pequeña razón S/R en cada una de las tres componentes.

El volcán Popocatepétl tiene diferentes tipos de eventos. Para los VTs se vio que FAST es el código más adaptado para detectar estos eventos. El volcán tiene también LPs, algunos se pudieron detectar con FAST. En este caso FAST no es el mejor programa para detectarlos. El mejor método para detectar los LPs es el TM con la energía. Cabe mencionar que este trabajo se concentra en detectarlos VTs, pero se puede mejorar la detección de LPs usando TM, por ejemplo, usando otros filtros o ventanas de tiempo de diferente duración.

En los registros en un volcán, también se puede encontrar sismicidad regional. Es importante aislarlos y en ese caso FAST los detecta muy bien. Así que se puede usar FAST a posteriori para rechazar los sismos regionales y usar el TM para los sismos volcánicos.

### 8.6 Clasificación de las señales

La clasificación se hace con el agrupamiento jerarquizado que toma un valor de correlación a partir del cual separará a las muestras en diferentes familias. Al haber utilizado un valor bastante bajo (88%), la mayoría de las señales serán colocadas en una misma familia, pero pueden existir familias diferentes agrupadas dentro de una misma. Por ejemplo, la familia naranja (#ff7005 de la Fig. 26) contiene señales tanto de tipo VT como algunas LP (Figs. 27 y 28. Tabla 5). Al tratarse de una clasificación con un algoritmo no supervisado se hizo un compromiso, con el que se estableció un umbral que clasificara el mayor número de señales, sin generar demasiadas familias. De esta manera, se obtuvieron 9 familias a partir del agrupamiento jerarquizado. Se debe mejorar este método de clasificación, por ejemplo, usando parámetros adicionales a la correlación, como la duración de la señal, su contenido frecuencial, entre otros.

# 9. Conclusión

Se utilizaron dos programas que utilizan Inteligencia Artificial para hacer detecciones automáticas de señales volcánicas del Popocatepetl. Uno es *Template Matching* (TM) que es semi-supervisado (requiere generar una base de datos previa con las familias que se quiere buscar, la que se hace en este trabajo de manera automática) y el otro es el FAST que es no-supervisado (no requiere hacer una base de datos previa). El programa TM usando la energía de las señales detectó en total 2,259 eventos, compuestos de 1,384 eventos verdaderos y 885 falsas detecciones. FAST detectó 1,709 señales en total, 1,075 siendo verdaderas y 634 falsas detecciones. TM tomó un total de 59 horas de tiempo de ejecución, ya incluyendo ambos modos de ejecución, mientras que FAST tardó 7 horas. FAST es más rápido, pero detectó menos eventos.

El método TM detectó señales principalmente de las familias VT y LP mientras que FAST detectó principalmente señales de la familia VT y de tipo sismicidad regional. El método FAST está diseñado para detectar réplicas de sismos grandes y sobre volcanes puede detectar VTs, pero no está diseñado para detectar más tipos de señales volcánicas, principalmente con baja energía. Debido a que ambos métodos detectaron mejor las señales de tipo VT, este trabajo se concentró en buscar estas señales, sin embargo, estos métodos pueden realizar detecciones de otro tipo de señales. Hay que mencionar que el método TM se puede adaptar más fácilmente que FAST para detectar diferentes familias de señales en el volcán Popocatépetl, y a pesar de que es más tardado que el método FAST, es más automático porque no requiere hacer una base de datos de las familias que uno quiere buscar dentro de los datos.

El enfoque principal de la tesis fue la comparación de ambas metodologías de detección. Se realizó una modificación a la metodología de TM para mejorar su tiempo de ejecución, se introdujo el uso de la energía para realizar detecciones en vez de usar las tres componentes NS, EW y vertical, lo que reduce el tiempo de ejecución

por un factor de 2, además de hacer detecciones en señales con energía alta relacionadas con la actividad del volcán.

Los diferentes programas de detección automática como FAST o TM se pueden usar de manera conjunta y no exclusiva. Por ejemplo, se puede empezar a usar FAST para detectar los sismos regionales y después el TM para detectar los eventos volcánicos. Si el TM detecta sismos regionales, estos se pueden eliminar fácilmente a posteriori gracias a las detecciones realizadas con FAST.

# Bibliografía


Ammirati, J.B. *RapidAligner Seismic (RAS): A fast GPU-based approach for template matching*

Baluja, S., & Covell, M. (2008). *Waveprint: Efficient wavelet-based audio fingerprinting.* Pattern Recognition, 41(11), 3467–3480.

Berry, M. W., Mohamed, A., & Yap, B. W. (Eds.). (2019). *Supervised and unsupervised learning for data science* (1st ed.). Springer Nature.

Bueno, A., Zuccarello, L., Díaz-Moreno, A., Woollam, J., Titos, M., Benítez, C., Álvarez, I., Prudencio, J., & De Angelis, S. (2020). *PICOSS: Python Interface for the Classification of Seismic Signals*. In Computers & Geosciences (Vol. 142, p. 104531). Elsevier BV.

Carr, S. M. (2007). *UPGMA vs WPGMA.* Mun.Ca. Retrieved May 02, 2024, from https://www.mun.ca/biology/scarr/UPGMA_vs_WPGMA.htm.

Centro Nacional de Prevención de Desastres. (2017). *ACTIVIDAD DEL VOLCÁN POPOCATÉPETL EN 2016*. CENAPRED, México, 2017

Chouet, B. (1996*). Long-Period volcano seismicity: its source and use in eruption forecasting.* Nature. (1996) Vol 380, pp 309-315

De la Cruz-Reyna, S. y Siebe, C. (1997). *The giant Popocatépetl stirs*. Nature 388, 227 (1997).

Delgado-Granados, H. (2017). *"La erupción del siglo XVII"*. Memoria técnica del mapa de peligros del volcán Popocatépetl. Monografías Instituto de Geofísica, vol. 22, México.

Delua, J. (2021). *Supervised vs. Unsupervised learning: What's the difference?* IBM Blog. https://www.ibm.com/blog/supervised-vs-unsupervised-learning/

Ferrari, L., Orozco-Esquivel, T., Manea, V., & Manea, M. (2012). *The dynamic history of the Trans-Mexican Volcanic Belt and the Mexico subduction zone*. Tectonophysics (Vols. 522–523, pp. 122–149). Elsevier BV.

Ferres, D. y Fonseca, D. (2017). *Estudios geológicos y actualización del mapa de peligros del volcán Popocatépetl.* Memoria técnica del mapa de peligros del volcán Popocatépetl. Monografías Instituto de Geofísica, vol. 22, México.

Garza-Girón, R., Brodsky, E. E., Spica, Z. J., Haney, M. M., & Webley, P. W. (2023). *A specific earthquake processing workflow for studying long‐lived, explosive volcanic eruptions with application to the 2008 Okmok Volcano, Alaska, eruption.* Journal of Geophysical Research. Solid Earth, 128(5).

Gibbons, S. J., & Ringdal, F. (2006). *The detection of low magnitude seismic events using array-based waveform correlation.* Geophysical Journal International, 165(1), 149–166.
Hundt, C. (2021). *Aligning time series at the speed of light*. NVIDIA Technical Blog. https://developer.nvidia.com/blog/aligning-time-series-at-the-speed-of-light/
Jo, T. (2021). *Machine learning foundations: Supervised, unsupervised, and advanced learning*. Springer International Publishing.
Kehtarnavaz N. (2008). *Digital Signal Processing System Design* (Second Edition). Elsevier.
Kotsiantis, S. B. and Pintelas, P. E. (2004). *Recent advances in clustering: A brief survey.* WSEAS Transactions on Information Science and Applications, 1, 73-81.
Macías, J.. (2005). *Geología e historia eruptiva de algunos de los grandes volcanes activos de México.* Boletín de la Sociedad Geológica Mexicana, 57(3), 379-424.
Lesage, P. (2009). Interactive Matlab software for the analysis of seismic volcanic signals. Computers & Geosciences, 35(10), 2137–2144.
Leskovec J., Rajaraman A., Ullman J.D. (2014). Finding Similar Items. In: Mining of Massive Datasets. Cambridge University Press. pp. 68-122.
Malfante, M., Dalla Mura, M., Mars, J. I., Métaxian, J.-P., Macedo, O., & Inza, A. (2018). *Automatic classification of volcano seismic signatures*. Journal of Geophysical Research: Solid Earth, 123, 10,645–10,658.
Martin-Del-Pozzo A. L. y Torres A. N. (2017). *Actividad reciente en el Popocatépetl 1993-2016.* Memoria técnica del mapa de peligros del volcán Popocatépetl. Monografías Instituto de Geofísica, vol. 22, México.
Matoza, R. S., Arciniega-Ceballos, A., Sanderson, R. W., Mendo-Pérez, G., Rosado-Fuentes, A., & Chouet, B. A. (2019). *High-broadband seismoacoustic signature of Vulcanian explosions at Popocatépetl volcano*. Mexico. Geophysical Research Letters, 46, 148–157.
McNutt, S. R.(1986). Observations and analysis of B-type earthquakes, explosions, and volcanic tremor at Pavlof volcano, Alaska, Bull. Seismol. Soc. Am. 76, 153–175
McNutt, S. R. (2005). *Volcanic Seismology*. Annual Review of Earth & Planetary Sciences, 33(1), 461-C-3.
McNutt, S. R., & Roman, D. C. (2015). Volcanic seismicity. En The Encyclopedia of Volcanoes (pp. 1011–1034). Elsevier.
Minakami, T.: 1974, Seismology of volcanoes in Japan, in L. Civetta, P. Gasparini, G. Luongo, and A. Rapolla (eds), Physical Volcanology — Developments in Solid Earth Geophysics, Elsevier, Amsterdam, pp. 1–27.

Mousavi, S. M., & Beroza, G. C. (2023). *Machine learning in earthquake seismology.* Annual Review of Earth and Planetary Sciences, 51(1), 105–129.
Mu, D., Lee, E.-J., & Chen, P. (2017). *Rapid earthquake detection through GPU-Based template matching.* Computers & Geosciences, *109*, 305–314.
Nielsen, F. (2016). *Introduction to HPC with MPI for Data Science* (1st ed.). Springer International Publishing.
Perton, M., Legrand, D., Macías, J. L., Cisneros, G., & Yañez-Sandoval, R. (2024). *Magma migration below Tancítaro and Paricutin volcanoes revealed by seismology.* Geophysical Journal International, 236(3), 1699–1715.
Ramirez, J., Jr, & Meyer, F. G. (2011). *Machine learning for seismic signal processing: Phase classification on a manifold*. 2011 10th International Conference on Machine Learning and Applications and Workshops.
Ran, X., Xi, Y., Lu, Y., Wang, X., & Lu, Z. (2023). *Comprehensive survey on hierarchical clustering algorithms and the recent developments.* Artificial Intelligence Review, 56(8), 8219–8264.
Reyes-Romero, A. (2022). *Detección y Clasificación de Automática de Señales Volcánicas en el volcán Popocatépetl.* [Tesis de Licenciatura]. UNAM.
Robin, C., Boudal, C. (1987) *A gigantic Bezymianny-type event at the beginning of modern volcan Popocatepetl*. J Volcanol Geotherm Res 31:145-130
Sassa, K. (1935). Volcanic micro-tremors and eruption-earthquakes (Part 1 of the geophysical studies on the volcano Aso). Mem. Coll. Sci. Kyoto Univ. Series A 18, 255±293.
Saxena A., Prasad, M., Gupta A., Bharill, N., Patel, O. P., Tiwari, A., Er, M. J., Ding, W., & Lin, C.-T. (2017). *A review of clustering techniques and developments*. Neurocomputing, 267, 664–681.
Shelly, D. R., Beroza, G. C., & Ide, S. (2007*). Non-volcanic tremor and low-frequency earthquake swarms.* Nature, 446(7133), 305–307.
Siebe, C., Abrams, M., Luis Macías, J., & Obenholzner, J. (1996). *Repeated volcanic disasters in Prehispanic time at Popocatépetl, central Mexico: Past key to the future?* Geology, 24(5), 399.
Stollnitz, E. J., DeRose, T. D. & Salesin, D. H. (1995) *Wavelets for computer graphics: A primer, part 1*. IEEE Computer Graphics and Applications, 15(3):76-84.
Tibco.com. (2014). *What is the Hierarchical Clustering Tool?* Retrieved May 02, 2024, from https://docs.tibco.com/pub/spotfire/6.5.3/doc/html/hc/hc_what_is_the_hierarchical_clustering_tool.htm.

Varshney, S., Shreya, S., Kumar, R. R., Kumar, V., & Chauhan, A. S. (2023). *A review paper on big data analytics: Tools problems and future reseach difficulties.* 2023 International Conference on Sustainable Emerging Innovations in Engineering and Technology (ICSEIET).

Withers, M., Aster, R., Young, C., Beiriger, J., Harris, M., Moore, S., & Trujillo, J. (1998). *A comparison of select trigger algorithms for automated global seismic phase and event detection.* Bulletin of the Seismological Society of America, 88(1), 95–106.

Yoon, C. E., O'Reilly, O., Bergen, K. J., & Beroza, G. C. (2015). *Earthquake detection through computationally efficient similarity search.* Science Advances, 1(11), e1501057.

Yoon, C. E., Huang, Y., Ellsworth, W. L., & Beroza, G. C. (2017). *Seismicity during the initial stages of the Guy-Greenbrier, Arkansas, earthquake sequence.* Journal of Geophysical Research. Solid Earth, 122(11), 9253–9274.