**UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ**

**Colegio de Posgrados**

# Clasificación no supervisada de eventos sísmicos en el volcán Tungurahua - Ecuador

**Juan Camilo Anzieta Reyes**

**Carlos Jiménez Mosquera, Ph.D.**
**Director de Trabajo de Titulación**

Trabajo de titulación de posgrado presentado como requisito
para la obtención del título de Magister en Matemáticas Aplicadas

Quito, 22 de diciembre de 2016

**UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ**
**COLEGIO DE POSGRADOS**

# HOJA DE APROBACIÓN DE TRABAJO DE TITULACIÓN

**Clasificación no supervisada de eventos sísmicos en el volcán Tungurahua - Ecuador**

**Juan Camilo Anzieta Reyes**

| | firmas |
|---|---|
| Carlos Jiménez Mosquera, Ph.D.<br>Director del Trabajo de Titulación | ..... |
| Julio Ibarra Fiallo M.Sc.<br>Miembro del Comité de Trabajo de Titulación | ..... |
| Carlos Jiménez Mosquera, Ph.D.<br>Miembro del Comité de Trabajo de Titulación<br>Director de la Maestría en Matemáticas Aplicadas | ..... |
| César Zambrano, Ph.D.<br>Decano del Colegio de Ciencias e Ingeniería | ..... |
| Hugo Burgos, Ph.D.<br>Decano del Colegio de Posgrados | ..... |

Quito, 22 de diciembre de 2016

Firma: .....

Nombre: Juan Camilo Anzieta Reyes

Código de estudiante: 00125348

C.I.: 0103024782

Fecha: Quito, 22 de diciembre de 2016

# DEDICATORIA

*A la mujer más importante de mi vida, mi mamá*

*A mi familia y amigos*

# AGRADECIMIENTOS

*Primero quiero agradecer a la Universidad San Francisco de Quito pues la consecución de este trabajo se logró gracias al programa de becas del Colegio de Posgrado para Maestría en Matemáticas Aplicadas. En lo intelectual no puedo dejar de agradecer a los profesores de esta maestría pues aunque este proyecto estuvo en mi mente tiempo antes de empezar el viaje, ellos supieron orientarme y darme herramientas de maneras a veces insospechadas o indirectas, pero siempre válidas; particularmente Luca Guzzardi quien sentó bases indispensables para este trabajo. También debo agradecer al Instituto Geofísico no sólo por permitirme usar los datos con los que se elaboró esta tesis, sino porque fue trabajando ahí que surgió en mí la necesidad de mejorar mis habilidades para servir mejor al mundo.*

*A mis amigos, todos quienes han aportado en mayor o menor manera a este proyecto. Mencionaré algunos sin ningún orden en particular y seguro de que me faltan muchos más, pero explícita o implícitamente, gracias a todos. A mis compañeros de maestría con los que pasamos incontables horas de estudio tratando de mejorar. A Antonina Calahorrano pues sin su ayuda este proyecto hubiera sido mucho más complicado de elaborar. A Daniel Pacheco, Hugo Ortiz y Jake Anderson pues sus comentarios respecto a los resultados finales me ayudaron a afinar ideas. Y finalmente a los maestros y científicos que dedicaron y dedican sus vidas a tratar de hacer de nuestra especie una especie mejor.*

# Resumen

En este trabajo se planteó examinar el estado de actividad del volcán Tungurahua (Ecuador) durante el año 2014, mediante el análisis de diferentes eventos sísmicos registrados en una estación geofísica permanente del Instituto Geofísico de la Escuela Politécnica Nacional ubicada en dicho volcán.

En los procedimientos estándar de monitoreo volcánico existe una clasificación de eventos sísmicos que se realiza de manera supervisada (un ser humano asigna una clase para cada evento en función de su percepción y ciertos criterios fijados). Sin embargo a pesar de que esta clasificación entrega cierta información de los procesos volcánicos que pueden estar ocurriendo dentro de un volcán, no es determinante al momento de usarla como método para predecir una erupción como tal. Por eso en este trabajo se propone agrupar las señales sísmicas de forma no supervisada en búsqueda de posibles nuevas clases de eventos que podrían indicar el inicio de fases eruptivas o que permitan describir mejor el estado del volcán. El análisis como tal se realizó utilizando algunas técnicas de clasificación no supervisada de objetos (k-medias, análisis de arquetipos y mapas auto-organizados) sobre señales sísmicas "discretas", es decir que no superaron más de 2 minutos de duración. Para realizar la clasificación, a cada una de las señales se le aplicó la transformada de Fourier para poder distinguirlas en función de su contenido espectral. Finalmente se obtuvieron algunos grupos o clases de señales sísmicas que aparecieron únicamente durante períodos de actividad volcánica superficial/visible.

Palabras clave: clasificación no supervisada, espacio de características, señales sísmicas volcánicas, actividad volcánica, volcán Tungurahua, k-medias, análisis de arquetipos, mapas auto-organizados

# Abstract

In this work the state of activity/unrest of Tungurahua volcano (Ecuador) during 2014 was examined through analysis of different seismic events recorded on a permanent geophysical station from Instituto Geofísico EPN located at the volcano.

In standard volcanic monitoring procedures there exists a classification for seismic events performed in a supervised manner (a human being assigns a class to each event based on perception and some fixed criteria). However even if this classification yields some information on the possible ongoing volcanic processes inside a volcano, it is not determinant when used as a method to predict an actual volcanic eruption. Therefore in this work an unsupervised seismic signal classification is proposed so that possible new classes of events are found which could indicate the beginning of eruptive phases or that allow a better description of the volcano's state. The analysis was performed using some unsupervised classification techniques (k-means, archetypal analysis and self-organizing maps) on "discrete" seismic signals, signals that were not longer than 2 minutes in time. To achieve the classification, the Fourier transform was applied to each of the signals so that they were distinguished in regard of their spectral content. Finally some groups of signals were formed that appeared only during superficial/visual volcanic activity.

Keywords: unsupervised classification, feature space, volcanic seismic signals, volcanic activity, Tungurahua volcano, k-means, archetypal analysis, self-organizing maps

# TABLA DE CONTENIDO

# 1. Introducción

Los volcanes y su actividad han sido estudiados durante siglos: desde la descripción por Plinio el joven de la erupción del volcán Vesubio en el año 79 d.c. hasta la época actual. Por lo mismo las diferentes formas de describir o caracterizar la actividad volcánica han ido mejorando con el paso de los años en búsqueda de entender mejor los procesos volcánicos y así tratar de mitigar posibles desastres.

Además de las conclusiones que se pueden sacar acerca de la actividad volcánica con base en las percepciones sensoriales (observación de emisiones, audición de explosiones, olfato de gases, etc.), hoy en día existen un sinnúmero de técnicas y aparatos que permiten monitorear el estado de actividad de un volcán. Dichas técnicas se basan en distintos principios y parámetros naturales. Hay técnicas que se realizan estudiando fenómenos ya ocurridos (como el análisis físico-químico de la ceniza de las primeras emisiones de un proceso eruptivo, o el estudio de los procesos históricos del volcán basado en sus depósitos) y otras que se realizan observando parámetros en tiempo real (como el monitoreo sísmico, de deformación o de emisión de gases). Mientras más información se tiene de dichos parámetros, más confiabilidad se tiene al tratar de predecir la actividad de un volcán. Por esta razón los volcanes más activos tienden a tener un número cada vez más alto de equipos y personal encargado de su monitoreo.

La sismicidad en los volcanes es un parámetro altamente estudiado y controlado pues refleja en tiempo real procesos al interior de los mismos. Los eventos al interior de un volcán pueden tener distintos orígenes y por esta razón se han definido distintos tipos de eventos volcánicos que son clasificados conforme ocurren (sea de forma automatizada o manual) para tratar de identificar diferentes estados de actividad. Hasta el día de hoy a pesar de esta clasificación parcialmente exitosa de eventos volcánicos la predicción certera de erupciones no ha sido posible del todo. En este estudio se menciona el estado de monitoreo del volcán Tungurahua, se propone un sistema de clasificación no supervisado (generado por computadora y no basado en criterios/percepciones humanas anteriores) y se presenta el resultado de aplicar algunos algoritmos en la clasificación de eventos discretos (cortos en el tiempo) del volcán Tungurahua ocurridos en el 2014.

# 2. Preliminares

## 2.1. Volcanes activos y sismos volcánicos

### 2.1.1. El volcán Tungurahua

El Ecuador es un país con características geológicas y geódinamicas muy particulares, entre las cuales destaca un volcanismo muy activo que ha tenido consecuencias en poblaciones aledañas al volcán desde tiempos históricos hasta la actualidad[1]. Uno de los volcanes más activos y potencialmente destructivos del país es el volcán Tungurahua ubicado en la provincia del mismo nombre[2]. Este volcán ha presentado una actividad eruptiva fuerte desde 1999 hasta la fecha y es uno de los mejor monitoreados del país. La presencia de algunas estaciones geofísicas en este volcán implica la disponibilidad de una base de datos extensa y creciente que ha logrado captar diferentes episodios de actividad del mismo.

### 2.1.2. Eventos sísmicos en volcanes

Uno de los parámetros más controlados en los volcanes es su sismicidad. Debido a que la reactivación de un volcán o el inicio de un proceso eruptivo implica el movimiento o emisión de material volcánico (magma o gases), este fenómeno causa vibraciones en el edificio volcánico que se pueden detectar con sismómetros. Un aumento en el número de dichos eventos indica un incremento en la actividad volcánica y por eso este monitoreo se realiza continuamente. La creciente cantidad de datos volcánicos de diversa naturaleza ha llevado a la creación de muchas técnicas que tienen como objetivo tratar de predecir erupciones volcánicas[3]. Cambios en las características asociadas a la sismicidad pueden llegar a ser precursores (es decir indicadores previos) de un nuevo proceso eruptivo; el estudio y control de dichos parámetros ha permitido en casos concretos predecir erupciones volcánicas con cierto éxito[4].

La sismicidad en volcanes es diferente de la sismicidad en contextos no volcánicos pues su origen puede estar asociado a diversos fenómenos físicos o mecanismos de fuente[5] (tales como movimiento de fluídos y ruptura de rocas en conductos). Con base en el análisis de la forma y las frecuencias de las señales sísmicas, además de su posible origen, los eventos se han separado principalmente en las siguientes categorías: Eventos tipo volcanotectónicos (VT), eventos tipo largo periodo (LP) y los eventos de tipo tremor

(TR).

Los eventos tipo volcano-tectonicos están asociados a ruptura de material rígido debido a algún incremento de presión en el medio por el ingreso o salida de fluidos y por lo tanto son eventos discretos (cortos en el tiempo) que empiezan generalmente de manera impulsiva y poseen altas frecuencias. Los eventos tipo LP están asociados a movimientos de fluidos sin ruptura de material y se originan por resonancia o expansión y relajación elástica de grietas, cámaras o conductos dentro del volcán y por esto poseen bajas frecuencias; aunque son también eventos discretos -cortos- empiezan de manera emergente. Los eventos tipo tremor son sostenidos en el tiempo y pueden deber su origen a muchas fuentes (explosiones, resonancia de otros eventos, secuencia acelerada de eventos discretos cada vez más cercanos en el tiempo, etc.) además de poder presentar múltiples frecuencias[6].

Además de éstos, existen otros tipos de eventos que pueden aparecer en registros sísmicos de volcanes y que pueden estar o no estar relacionados a la actividad volcánica. Estos son los eventos Híbridos (HB) que comparten características de los eventos tipo VT y LP, rupturas de glaciar (ICEQ) que se presentan en volcanes con glaciares, eventos de muy largo período (VLP) que son similares a LPs pero de frecuencias características aún menores a los LPs usuales, eventos regionales (sismos o terremotos no volcánicos alejados del volcán pero suficientemente grandes para detectarse), explosiones con sus debidas señales asociadas (sísmicas y acústicas), entre otros [7]. Parte de la diversidad de la sismicidad en volcanes se puede visualizar en la figura 1.

Aunque la asignación de eventos a estas clases ha producido una cantidad importante de técnicas y estudios que han tenido éxito en la predicción de erupciones volcánicas en algunos casos [3, 4, 6, 8], cada volcán posee características particulares que hacen que dentro de las clases de eventos haya subclases de los mismos, o que ciertos rangos fijos de parámetros para la clasificación de un evento (basados en la frecuencia o la forma por ejemplo) provoquen ambigüedad e incluso arbitrariedad al momento de "etiquetar" un evento dentro de una u otra categoría [9]. Además al momento de asignar una clase a un evento la percepción de cada individuo puede influir en la clasificación dentro de uno u otro grupo. En términos de monitoreo en general esto es aún más destacable cuando la cantidad de datos es muy grande y diferentes operarios pueden ser necesarios para un mismo volcán o un mismo individuo podría estar afectado por la cantidad de

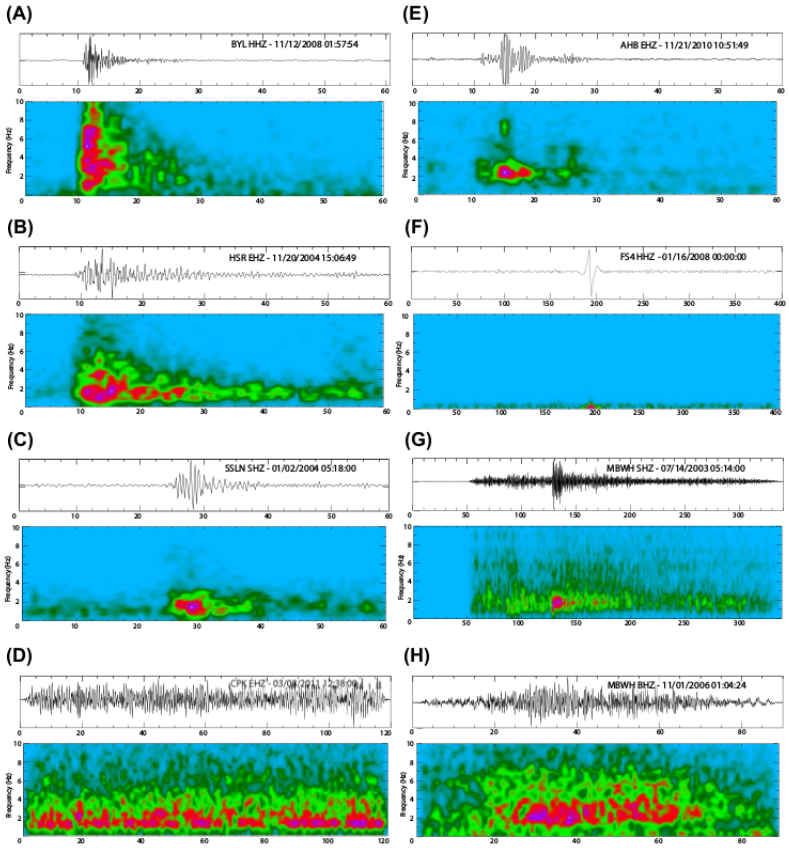



Figura 1: Ejemplo de formas de onda (arriba) y espectrogramas (abajo) de sismicidad en volcanes: A) Evento volcano-tectónico VT B) Evento híbrido C) Evento de largo período LP D) Tremor volcánico E) Evento de argo período profundo F) Evento de muy largo período G) Terremoto de explosión H) Derrumbe. Tomado de "The Encyclopedia of Volcanoes"[7].

trabajo necesaria para llevar a cabo una clasificación apropiada de demasiados eventos. En general mantener una forma objetiva o consistente de clasificar eventos puede ser una tarea exigente para una o varias personas y una ejecución pobre de esta tarea puede disminuir la eficacia del monitoreo o las predicciones relacionadas a la ocurrencia de erupciones volcánicas.

## 2.2. Clasificación automática de eventos volcánicos

Debido a que la problemática expresada es común en muchos observatorios vulcanológicos alrededor del mundo, se han propuesto algunos métodos para realizar el monitoreo volcánico y la clasificación de eventos de forma automatizada. La clasificación de objetos

en general se puede hacer de dos formas: de manera supervisada o no-supervisada.

La clasificación supervisada de objetos se basa en crear un sistema matemático-algorítmico que "aprende" cómo asignar categorías a los objetos con base en ejemplos ya clasificados. Más formalmente, al sistema se le provee de un conjunto de datos u objetos con categorías previamente asignadas por un ser humano (conjunto de entrenamiento) y el sistema organiza las características determinantes de los objetos de cada categoría. Después, al recibir nuevos datos no asignados pero de los que se conoce su asignación (conjunto de prueba) los enmarca dentro de alguna de las categorías predefinidas. Con este conjunto se verifica que el algoritmo realiza bien su trabajo antes de implementarlo con datos enteramente desconocidos. En el contexto de la clasificación supervisada automática de eventos volcánicos existe un número cada vez mayor de trabajos que se basa en las clases mencionadas en la sección anterior [10]. Algunos ejemplos de volcanes estudiados con diferentes métodos de clasificación supervisada en diversos artículos de publicación y tesis de posgrado son: Cotopaxi[11] (Ecuador), Galeras[12] y Nevado de Ruiz[13] (Colombia), San Cristobal y Telica[14] (Nicaragua), Colima[15] (México), Merapi[9, 15] (Indonesia), Etna y Estómboli[14] (Italia), Piton de la Fournaise[15] (Francia) e Isla Decepción[16] (Antártica).

La clasificación no supervisada de objetos difiere de la clasificación supervisada en que los sistemas "encuentran por sí solos" un agrupamiento para los datos sin necesidad de que se les proporcione un conjunto de datos con una clasificación creada de antemano. Esto se realiza analizando la similitud entre los objetos y agrupando los objetos más parecidos entre sí. En el caso de clasificación no supervisada de eventos volcánicos existe un número también creciente aunque menor de estudios que utilizan este enfoque. Su aplicación se ha realizado en menos volcanes alrededor del mundo pues históricamente, la clasificación no supervisada de eventos no es una práctica estándar en observatorios vulcanológicos. Algunos ejemplos de estos estudios se han realizado en volcanes como: Stromboli[17, 18, 19], Etna[20, 21] y Vesubio[22] (Italia), Merapi[23] (Indonesia) e Isla Raoul[24] (Nueva Zelanda).

En este trabajo se plantea hacer un estudio mediante clasificación no supervisada utilizando algunas técnicas de agrupamiento (clustering): K-means, análisis de arquetipos y mapas auto-organizados. Estas técnicas se aplican sobre señales registradas en una estación sísmica en el volcán Tungurahua que no duraron más de 2 minutos y representan

eventos “discretos” o esporádicos en el volcán. El objetivo principal de este estudio es tratar de encontrar clases de eventos diferentes o contenidas como sub-clases de las clases tradicionales, que quizás puedan reflejar comportamientos más complejos que los visibles basados en las clases típicas.

# 3. Datos y metodología

La clasificación de objetos sea de forma supervisada o no supervisada requiere que éstos sean caracterizados de alguna manera, sea asignándoles atributos, categorías o más usualmente valores numéricos para poder compararlos o distinguirlos entre sí[25]. Este conjunto de descriptores de los objetos se denominan variables o características. En la clasificación de objetos es crucial tratar los datos y definir sus atributos de manera consistente antes de aplicar cualquier algoritmo para lograr buenos resultados; este procedimiento se denomina preprocesamiento de datos y se detallará después de describir los datos utilizados como tales.

## 3.1. Datos

Los datos utilizados en este estudio provienen de una estación geofísica permanente en el volcán Tungurahua perteneciente al Instituto Geofísico de la Escuela Politécnica Nacional (Ecuador) llamada Mazón (BMAS). Esta estación se encuentra a 5.51 km de la cumbre en el flanco sur-occidental del volcán a una altura de 2965 metros sobre el nivel del mar. Dicha estación posee un sensor sísmico de banda ancha Guralp CMG-40T con una respuesta plana de 50Hz a 60 s y un digitalizador Geotech SMART-24D con 24 bits de resolución muestreando a 50 Hz. Este sensor se encuentra registrando vibraciones de la región continuamente y sus datos se envían con una red de telecomunicaciones continuamente al Instituto Geofísico para su análisis y monitoreo. Los sensores sísmicos como el mencionado registran las vibraciones en 3 componentes (equivalente a registrar el movimiento en un sistema de coordenadas tridimensional): registran movimientos en sentido Norte-Sur, Este-Oeste y vertical respecto a la base del sensor. Estos datos constan de señales digitales (series de tiempo) con un número de cuentas cada 0.02 segundos que representan la velocidad de desplazamiento del suelo bajo el sensor medido cada instante.

En este estudio se utilizaron los datos sísmicos de la estación mencionada correspon-

dientes al año 2014, analizando únicamente su componente vertical para evitar posible redundancia en los datos, usar la componente con mejor relación señal-ruido y aligerar el tiempo de procesamiento[17, 20, 22]. Se escogió dicho año por haber presentado episodios de actividad volcánica importante bien marcados. En estudios anteriores de clasificación no-supervisada de eventos volcánicos las bases de datos utilizadas han correspondido a una sola clase de eventos seleccionada (como tremor[17, 20, 21, 24] o eventos de muy largo período[19]), a clases de eventos de diversos tipos preseleccionadas[18, 22] o incluso a toda la señal sísmica vista como un continuo[23]. El objetivo en los estudios en los que se escogió una sola clase de evento conocida y se aplicó clasificación no supervisada fue determinar si dentro de esa clase de eventos se identificaban sub-clases que aparecieran en distintos episodios de actividad volcánica. En los trabajos en los que se aplicó clasificación no supervisada a distintas clases también preseleccionadas de eventos fue corroborar la consistencia de la clasificación supervisada comparándola con los grupos encontrados por la clasificación no supervisada. Finalmente el trabajo en el que se usó toda la señal sísmica continua tenía como objetivo identificar diferentes regímenes del fondo sísmico y la detección de sismicidad anómala; la clasificación de eventos aparece de todas maneras contrastado con eventos clasificados de forma supervisada en un estudio anterior[23, 9].

Aquí se decidió escoger los datos de manera diferente, seleccionando todas las señales que duraran menos de 2 minutos o cuyo comportamiento más importante fuera menor que 2 minutos. Comportamiento más importante se refiere a todo tramo de señal que contenga un inicio impulsivo/claro y un decaimiento exponencial (típico en las "codas" de los eventos sísmicos) a niveles estables -aún por encima del ruido-. En este sentido algunas señales largas son "cortadas" cuando alcanzan una cierta estabilidad pues aquí se ha considerado que es el inicio de la señal la que contiene información del posible mecanismo que la origina, más que su remanente en el tiempo. Hacer este tipo de clasificación permite recolectar tanto eventos pequeños usuales (que coincidirán con VT, LP, Híbridos, sismos asociados a explosiones, etc.) que no duran más de 1 minuto, como eventos medianos/grandes (usualmente llamados tremores). Esta recolección permite realizar una exploración lo menos sesgada posible pues el objetivo intrínseco de la clasificación no supervisada de objetos es que la intervención del investigador sea la mínima posible[23]. El razonamiento para haber hecho esta elección es la idea de que los eventos impulsivos contienen información concreta de cambios repentinos en un volcán

independientemente de si estos desencadenan una señal más larga (generalmente estable en forma de tremor) después de ocurrir. Por la misma razón (evitar posible sesgo por parte del investigador) las señales se escogieron a partir del registro sísmico continuo en BMAS de 2014 utilizando algoritmos de picado automático de eventos.

Los algoritmos de picado fueron implementados con el paquete ObsPy[26] en Python que permite leer señales sísmicas en muchos formatos y analizarlas de distintas maneras. Después de filtrar las señales diarias (formato en el que se encuentran los datos originalmente) se usaron varios algoritmos basados en el algoritmo STA/LTA[27] (short term average-long term average ratio) con diferentes parámetros para identificar eventos suficientemente grandes respecto a la señal de fondo. Los algoritmos de detección automática utilizados son STA/LTA clásico y recursivo, carlSTAtrig y z-detect[26, 28] debido a que no existe un algoritmo universal infalible para detección de eventos[28]. Después todos los eventos detectados mayores a dos minutos fueron rechazados. La combinación de estos algoritmos permitió detectar una gama de señales sísmicas con inicios claros respecto a la señal de fondo y con decaimientos de señal relativamente rápida. De esta primera selección de eventos resultaron 13223 eventos picados automáticamente.

Esta forma de proceder en la elección de los datos tiene dos inconvenientes: Los datos picados de forma automática contienen eventos y señales de todo tipo que cumplen con las especificaciones mencionadas, es decir, señales de ruido de cualquier tipo que cumplan las restricciones de picado también entran en los eventos seleccionados para la clasificación. Esto puede incluir posibles eventos regionales (ajenos al volcán), truenos, animales cercanos a la estación, ruido instrumental o humano, ruido sin señal debido a inestabilidades en el picado automático, etc. Este inconveniente se resolvió parcialmente haciendo una segunda elección de los eventos de forma visual/manual. Aunque este hecho va en contra del argumento de que en este trabajo se busca que la interferencia humana sea la menor posible, esto se realiza por el hecho de que de todas formas la base de datos con la que se va a trabajar debe estar "limpia" de señales demasiado pequeñas o evidentemente anómalas si se quiere que los algoritmos de clasificación tengan resultados consistentes. Esto no quiere decir que se quitaron todos los eventos ajenos a la actividad del volcán, solamente aquellos eventos que no poseían señal clara o que eran ruidosos/contaminados. El número final de eventos después de esta depuración fue de 4006.

El segundo problema está relacionado a la variabilidad de los datos en cuanto a su tamaño. Al realizar clasificación de objetos, es necesario que los mismos sean comparables. Es decir que el conjunto de sus atributos sea igual en cada objeto para distinguirlos o compararlos consistentemente. Esto hizo impráctico el uso de las propias señales sísmicas para su comparación como en otros trabajos[19]. A continuación se explica cómo se trató con este problema.

### 3.1.1. Preprocesamiento de datos

El preprocesamiento de datos es un paso central en la clasificación de objetos. Consiste en transformar el conjunto de datos de un objeto (los atributos o valores que lo describen) de un espacio $S$ a un espacio $S'$ llamado espacio de características, en donde cada objeto tendrá un respectivo vector de características que lo representa:

$$x_i = (x_{i1}, x_{i2}, \ldots, x_{im}) \tag{1}$$

donde $i$ es el índice del i-ésimo objeto de los $N$ que se analizan y $m$ es la dimensión del vector (y por ende del espacio $S'$) de características. Todos los objetos que se van a analizar deben tener vectores de características de la misma dimensión y sus componentes deben coincidir con las mismas características. Este procedimiento de asignación/selección de características se puede hacer incluso si todos los objetos son en inicio comparables por medio de sus atributos, con el fin de remover redundancia, extraer información robusta y además expresar la información de cada objeto en forma comprimida[18].

Se mencionó que en este trabajo se seleccionaron señales sísmicas de distintos tamaños (a diferencia de trabajos anteriores[17, 20, 21, 24, 19]) por lo que usar las señales sísmicas en sí para su comparación no es posible. En términos de la determinación de las características que pueden usarse para describir los objetos, hay algunos estudios que hacen una investigación exhaustiva sobre qué parámetros pueden servir para describir señales sísmicas de forma comprimida[23, 29]. Sin embargo dependiendo del alcance y objetivos de cada investigación seleccionar características adecuadas para describir los objetos es una elección y búsqueda de cada investigador. Así por ejemplo en algunos estudios a pesar de que las señales son construídas de tal forma de que sus tamaños son iguales y por lo tanto podrían ser comparadas directamente, se recurre a usar el espectro

de potencias de la señal promediado en el tiempo como vector de características [20, 17]. Aquí se calculó directamente el espectro de frecuencias de cada señal con la transformada rápida de Fourier (siglas FFT en inglés) y se usó como vector de características de cada evento. Este procedimiento ya ha sido utilizado anteriormente aunque únicamente para señales de tremor[24]. Se optó por tomar esta estrategia para construir los vectores de características de los eventos, puesto que en particular en la clasificación manual de eventos volcánicos los criterios que los expertos usan para definir las clases de eventos son sus formas y su contenido espectral (visualizado en un espectrograma o en un espectro de frecuencias).

Al aplicar la transformada de Fourier a cada señal independientemente de su tamaño se obtuvieron amplitudes espectrales para frecuencias entre 0,05 y $25Hz$ con resoluciones diferentes según el tamaño de la señal original. Como las señales escogidas comprenden un rango amplio de formas y tamaños, hay señales cuyas amplitudes son varios órdenes de magnitud más grandes que otras. Sin embargo esto no necesariamente significa que pertenezcan a clases diferentes. Por esto, para poder agrupar señales independientemente de su magnitud se normalizó cada espectro de frecuencias respecto a la máxima amplitud espectral. Las amplitudes de frecuencias se muestrearon cada $0{,}5Hz$. Además de los valores de amplitud espectral correspondientes a cada frecuencia, se utilizaron los máximos valores en intervalos de $0{,}5Hz$ para tomar en cuenta también amplitudes que no ocurrieran justo cada $0{,}5Hz$ como se indica en la figura 2. Aunque utilizar este par de vectores de amplitud de frecuencias basadas en el mismo espectro original agrega cierta redundancia al vector de características de la señal, este procedimiento se realiza por la siguiente razón: aunque en principio parecería razonable utilizar únicamente el espectro muestreado cada $0{,}5Hz$ comparándolo con el espectro verdadero (tercer gráfico comparado con segundo gráfico en la figura 2), si se observa cuidadosamente el pico de máxima amplitud de frecuencia en ese gráfico corresponde a $10{,}5Hz$ mientras que el pico del espectro original está alrededor de $9{,}65Hz$. Esto implica que al comparar este espectro con el de otra señal la amplitud más importante no es la verdadera, cosa que sí sucede escogiendo máximos por intervalos (cuarto gráfico de la figura 2). Por otro lado si sólo se usara el vector de máximo por intervalos, pasaría algo similar respecto a frecuencias bajas (por ejemplo analizando las amplitudes cercanas a $5Hz$ del segundo y cuarto gráfico). También se exploró la posibilidad de usar promedios por intervalos pero los resultados fueron similares.

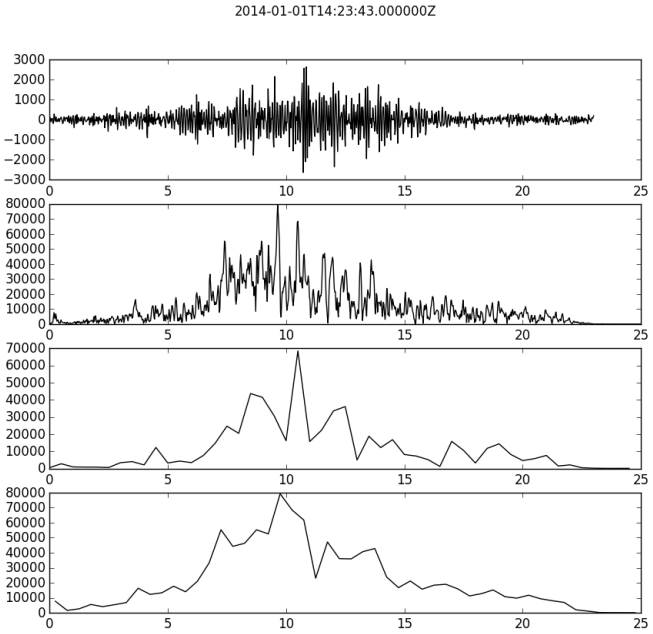



Figura 2: Señal sísmica, espectro, espectro muestreado cada $0{,}5Hz$ y espectro máximo en intervalos de $0{,}5Hz$.

Al unir estos dos vectores de amplitud de frecuencia para cada señal se genera un vector de características de dimensión 100 que está en el mismo orden de otros estudios[17, 18, 19, 20, 22]. Como último paso antes del procesamiento como tal la varianza de los vectores resultantes se normalizó a uno ya que esto mejora los procesos de agrupamiento posteriores al dar igual peso a todas las variables[20, 22, 25].

Una vez definidos los vectores de características de cada señal se procedió a aplicar los diferentes algoritmos de clasificación no supervisada descritos a continuación.

## 3.2. Métodos

La clasificación no supervisada de objetos se puede entender como un proceso en el que se pretende encontrar una partición en el espacio de características que produce grupos lo más homogéneos posible. Esto se realiza tratando de estimar la densidad de probabilidad de los datos en el espacio de características[25]. Existen muchos métodos para realizar clasificación no supervisada; una clase de estos métodos se refiere al análisis de conglomerados (cluster analysis en inglés) o sementación de datos. Este análsis con-

siste en crear subconjuntos o grupos tal que aquellos objetos dentro de un grupo están más relacionados entre sí que con objetos en otros grupos. También se usa para crear estadísticas descriptivas para definir o no si los datos tienen o no subgrupos distintos donde cada grupo representa objetos con propiedades substancialmente diferentes[25]. Este análisis se basa en medidas de similitud (disimilitud) entre los objetos. Esto sigifica que los datos son presentados cada uno con sus caractertísticas (como en este trabajo) o se puede presentar directamente la similitud (o disimilitud) directamente en una matriz de proximidad. En el caso en que las características del objeto sean valores numéricos, la disimilitud entre dos objetos $x_i$ y $x'_i$ se puede calcular definiéndola como la suma de las disimilitudes/distancias entre los atributos:

$$D(x_i, x'_i) = \sum_{j=1}^{m} d(x_{ij}, x'_{ij})$$

donde $j$ es el índice de la j-ésima característica y $d(x_{ij} - x'_{ij})$ es la distancia entre los valores de dicha característica. Esta distancia se puede escoger dependiendo del estudio pero la distancia más utilizada usualmente es la cuadrática:

$$d(x_{ij}, x'_{ij}) = (x_{ij} - x'_{ij})^2$$

Los métodos de los algoritmos de agrupamiento se pueden basar en 3 enfoques diferentes que requieren especificar de antemano un número $K$ de grupos: Algoritmos combinatorios, modelamiento de mixturas (mixture modelling en inglés) y buscadores de moda (mode seekers en inglés)[25]. En los algoritmos combinatorios se busca encontrar el mejor agrupamiento en $K$ grupos probando todas las combinaciones posibles de asignaciones de los elementos a cada grupo y designando como agrupamiento final (o ganador) aquel que minimiza la dispersión intragrupal de los datos. Aunque este enfoque encuentra la asignación que minimiza la diferencia de los objetos que pertenecen a un mismo grupo (agrupamiento óptimo) se vuelve inmanejable para un número alto de objetos. En el modelamiento de mixturas se asume que los datos son muestras independientes idénticamente distribuídas de poblaciones descritas por una función de densidad de probabilidad. Esta función se describe como un modelo parametrizado que se asume como una mixtura de componentes de funciones de densidad donde cada componente de densidad describe uno de los grupos. Después el modelo se ajusta a los datos por máxima verosimilitud u

otros enfoques Bayesianos[25]. Los buscadores de moda tratan de encontrar regiones de alta densidad de probabilidad de los datos de forma no paramétrica. Existen formas de "relajar" la búsqueda en los algoritmos combinatorios con estrategias manejables para números de datos grandes basadas en descendidas codiciosas iterativas. Estas estrategias se basan en iniciar con particiones arbitrarias de los datos que son perturbadas en cada iteración de tal forma que la asignación a los grupos disminuya la dispersión intragrupal respecto al estado anterior. Estos algoritmos paran cuando ya no hay mejora y encuentran una asignación mínima local (pues no exploran todas las combinaciones posibles de asignación sino aquellas que empiezan desde una partición inicial arbitraria)[25]. A continuación se describen los algoritmos escogidos para su implementación en este trabajo. La implementación se realizó con base en distintos paquetes del software estadístico libre R. Además se presenta para cada algoritmo un ejemplo y discusiones acerca de su funcionamiento basado en un mismo conjunto de datos.

### 3.2.1. Algoritmo k-means

El algoritmo k-medias (k-means en inglés) es uno de los algoritmos más populares de descendida iterativa. Se basa en encontrar $K$ vectores (también llamados prototipos) en el espacio de características a partir de los cuales se particiona el mismo en función de las regiones más cercanas a cada vector. Por esto, en este algoritmo se requiere definir una cantidad de grupos (o clusters en inglés) deseados $K$ antes de ejecutarlo. Este algoritmo se puede emplear cuando todos los datos son cuantitativos-numéricos y usa la distancia cuadrática Euclideana:

$$D(x_i, x_i') = \sum_{j=1}^{m} (x_{ij} - x_{ij}')^2 = ||x_i - x_i'||^2 \tag{2}$$

para medir la disimilitud entre objetos. En general para efectuar el agrupamiento se debe definir una cantidad a minimizar: la dispersión intragrupal de los datos para los $K$ grupos que se conformarán. Se define $C(i)$ como el mapeo que asigna la i-ésima observación al k-ésimo grupo. Entonces, dado un mapeo $C(i)$ la dispersión total de los datos es:

$$T = \frac{1}{2} \sum_{i=1}^{N} \sum_{i'=1}^{N} d_{ii'} = \frac{1}{2} \sum_{k=1}^{K} \sum_{C(i)=k} \left( \sum_{C(i')=k} d_{ii'} + \sum_{C(i')\neq k} d_{ii'} \right)$$

donde $d_{ii'} = d(x_i, x'_i)$. Esta separación de la dispersión entre objetos pertenecientes a una categoría y a otra ayuda a definir la dispersión intragrupal como:

$$W(C) = \frac{1}{2} \sum_{k=1}^{K} \sum_{C(i)=k} \sum_{C(i')=k} d(x_i, x'_i) \tag{3}$$

En el algoritmo k-means la dispersión intragrupal se puede expresar como:

$$W(C) = \frac{1}{2} \sum_{k=1}^{K} \sum_{C(i)=k} \sum_{C(i')=k} ||x_i - x'_i||^2$$

y arreglando un poco los términos:

$$W(C) = \sum_{k=1}^{K} N_k \sum_{C(i)=k} ||x_i - \overline{x}_k||^2 \tag{4}$$

donde $\overline{x}_k = (\overline{x}_{k1}, \ldots, \overline{x}_{km})$ es el vector promedio asociado al k-ésimo grupo y $N_k = \sum_{i=1}^{N} I(C(i) = k)$ con $I$ la función indicadora. La optimización se logra cuando la disimilitud promedio de los datos con la media del grupo es minimizada buscando un $C$ adecuado, es decir:

$$C^* = \min_{C} \sum_{k=1}^{K} N_k \sum_{C(i)=k} ||x_i - \overline{x}_k||^2$$

esta expresión se puede extender notando que para cualquier conjunto de observaciones (en este caso las del grupo $k$):

$$\overline{x}_S = \operatorname*{argmin}_{m} \sum_{i \in S} ||x_i - m||^2 \tag{5}$$

entonces se puede encontrar $C^*$ resolviendo el problema de optimización aumentado:

$$C^* = \min_{C, m_k} \sum_{k=1}^{K} N_k \sum_{C(i)=k} ||x_i - m_k||^2 \tag{6}$$

Esto se realiza con el procedimiento de optimización alternante descrito a continuación:

- Para una asignación a grupos $C$ cualquiera la varianza total intragrupos (6) se minimiza con respecto a $\{m_1, \ldots, m_k\}$usando (5).
- Dado un conjunto de medias $\{m_1, \ldots, m_k\}$, se minimiza (6) asignando cada obser-

vación a la nueva media de grupo más cercana, es decir se actualiza $C$ haciendo:

$$C(i) = \underset{1 \leq k \leq K}{argmin} \; ||x_i - m_k||^2$$

- Se repiten los pasos anteriores secuencialmente hasta que no haya cambio en las asignaciones.

Este procedimiento converge pues el valor del criterio (6) disminuye cada vez que se realizan los primeros pasos. Como se mencionó el resultado puede recaer en un mínimo local subóptimo por lo mismo este algoritmo debería ser efectuado con muchas elecciones aleatorias para las medias iniciales y se debería escoger aquella que entrega el valor más pequeño de varianza intragrupal[25].

Finalmente cuando el algoritmo ha convergido y quedan definidas las k-medias finales (óptimos locales o globales), el espacio de características queda particionado en regiones en las que cada punto del espacio pertenece al k-ésimo grupo si está más cercano al la k-ésima media (esta partición se llama teselado de Voronoi). Como se mencionó al inicio de la descripción del algoritmo el número $K$ de grupos es un parámetro que debe ser fijado de antemano por el investigador. En caso de buscar el número de grupos más conveniente existen algunos criterios (como el del "codo" o el índice de Davies-Bouldin[30] que se explicarán en la sección de resultados) para su elección pero al final esta depende de cada investigación/aplicación.

Ahora se muestra el ejemplo del funcionamiento de este algoritmo con un conjunto de datos generados aleatoriamente.

Los datos de la figura 3 se generaron a partir de 5 distribuciones Gaussianas multivariadas. Cuatro de ellas localizadas y separadas en los 4 cuadrantes con 50 muestras cada una y la quinta en el medio con 20 muestras para darle cierta continuidad a los datos. Estos puntos ejemplifican las coordenadas que representan a cada objeto en el espacio de características. En este ejemplo se usa un espacio 2 dimensional simple con fines de visualización, pero en general la distribución de los datos se manifiesta en espacios de gran dimensionalidad y puede ser compleja.

En la figura 4 se muestra el agrupamiento realizado por el algoritmo k-means. En este caso el algoritmo encontró de manera automática de forma aproximada los centros de las distribuciones Gaussianas que se utilizaron para generar los 4 grupos principales (en

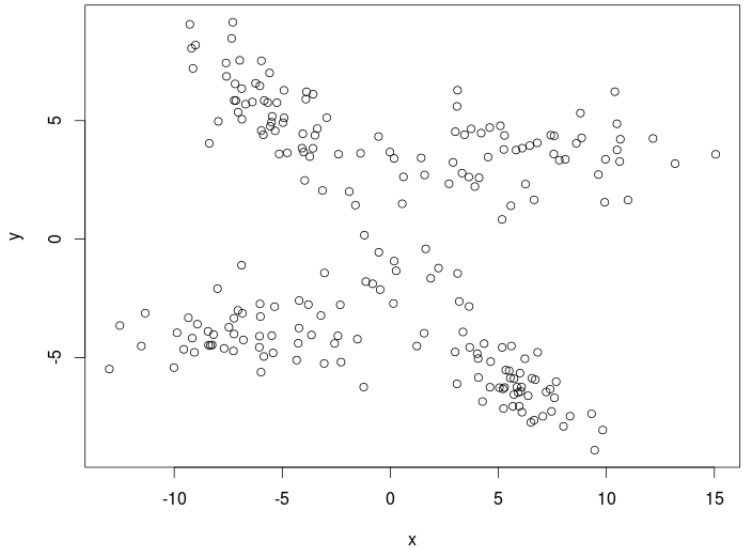



Figura 3: Datos fabricados para los ejemplos

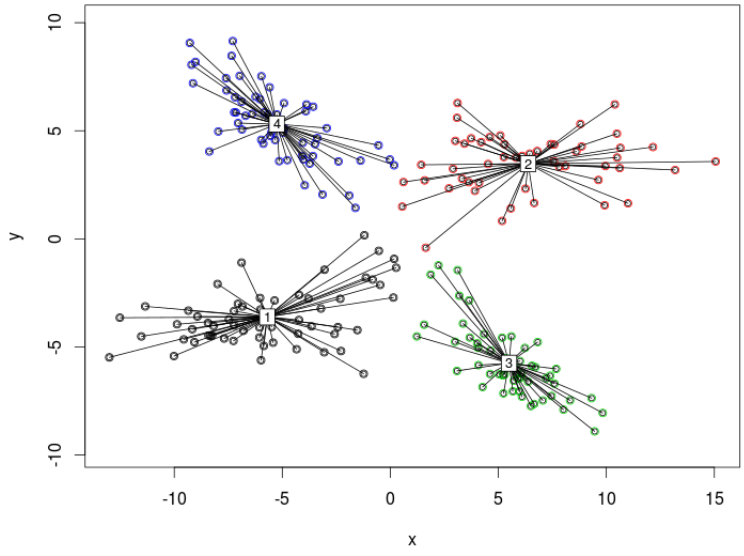



Figura 4: Agrupamiento mediante k-means con $k = 4$ para los datos generados

los 4 cuadrantes). Cada punto del espacio es agrupando según la distancia al centroide más cercano. Esto genera una partición del espacio de características que permite la clasificación de cada objeto posible.

### 3.2.2. Análisis de arquetipos

El análisis de arquetipos[31] se basa en encontrar $K$ elementos en el espacio de características que representan a los datos pero que además son una combinación convexa de los mismos. A su vez, cada dato se puede expresar como una combinación convexa de dichos elementos. Como los vectores que sirven para particionar el espacio son combinaciones convexas de datos, yacen en la envolvente convexa de la nube de datos en el espacio de características, es decir representan tipos "extremales" de datos, y de ahí su denominación como "arquetipos". Este método no representa en sí un algoritmo de agrupamiento combinatorio[25] "relajado" como k-means, aunque el proceso de búsqueda resulta parecido.

La busqueda se describe formalmente de la siguiente manera. Dada una matriz $X$ de tamaño $N \times m$ que representa $N$ datos con $m$ características, para un número de vectores $K$ dado, el análisis de arquetipos encuentra la matriz $Z$ de $K$ arquetipos en el espacio de características m-dimensional de acuerdo a los siguientes fundamentos[32]:

1. Se buscan vectores que aproximan los datos de la mejor manera con combinaciones convexas, es decir que minimizan:

$$SCR = \left|\left|X - \alpha Z^T\right|\right|_2 \tag{7}$$

   con $\alpha$ la matriz $N \times K$ de coeficientes de dichos vectores, donde $\alpha_{ij} \geq 0$ y $\sum_j^K \alpha_{ij} = 1$ para cada $i$ y donde $||.||_2$ es una norma matricial apropiada. Aquí SCR se usa para expresar suma del cuadrado de los residuos.

2. Los vectores son combinaciones convexas (arquetipos) de los datos:

$$Z = X^T \beta \tag{8}$$

   con $\beta$ la matriz $N \times K$ de coeficientes de los datos, donde $\beta_{ji} \geq 0$ y $\sum_i^N \beta_{ji} = 1$ para cada $j$.

Para lograr que estas condiciones se cumplan se realiza un procedimiento alternante para minimizar (7): dados arquetipos $Z$ se encuentran los $\alpha$ que minimizan la expresión y dados los $\alpha$ se encuentran los arquetipos $Z$ que minimicen (7) en problemas de mínimos cuadrados convexos. Así, se minimiza SCR cada vez que se realizan estos pasos de forma alternada llegando a un mínimo local (dependiendo de la inicialización). La aplicación del algoritmo tal como fue implementado en el paquete "archetypes" del software estadístico R [32] se resume a continuación:

1. Se preparan los datos $X$ normalizándolos, se añade una fila extra falsa y se inicializan los $\beta$ que cumplan la restricción tal que se obtienen arquetipos iniciales $Z$.

2. Se ejecutan los siguientes pasos en un bucle hasta que SCR sea suficientemente pequeña o hasta que un máximo de iteraciones se alcance:

   *a*) Encontrar los mejores $\alpha$ para los $Z$ dados: Se resuelven $n$ problemas de mínimos cuadrados convexos:

   $$\min_{\alpha_i} \frac{1}{2} ||X_i - Z\alpha_i||_2 \quad \alpha_i \geq 0 \quad \sum_{j=1}^{K} \alpha_{ij} = 1 \quad i = 1, \ldots, n$$

   *b*) Recalcular nuevos arquetipos $Z'$ resolviendo el sistema de ecuaciones lineales: $X = \alpha Z^T$

   *c*) Encontrar los mejores $\beta$ para $Z'$: se resuelven $K$ problemas de mínimos cuadrados convexos:

   $$\min_{\beta_j} \frac{1}{2} \left|\left|Z'_j - X\beta_j\right|\right|_2 \quad \beta_j \geq 0 \quad \sum_{i=1}^{N} \beta_{ji} = 1 \quad j = 1, \ldots, k$$

   *d*) Recalcular arquetipos $Z$ resolviendo el sistema de ecuaciones lineales: $Z = X\beta$

   *e*) Calcular la suma del cuadrado de los residuos SCR

3. Después de verificar la condición de parada del lazo remover la fila extra falsa y reescalar los arquetipos

Como se mencionó anteriormente, después de ejecutar este algoritmo se obtiene un conjunto de arquetipos $Z$ que minimiza SCR pero que puede o no ser un mínimo global.

Por lo mismo se deben realizar varias inicializaciones aleatorias para los vectores $Z$ y se debe escoger el mejor modelo (el que encuentre el valor más pequeño de SCR al final). Este algoritmo no define regiones en el espacio como en el algoritmo k-means, sino que los datos se clasifican en función de su mayor componente respecto a los arquetipos. Es decir, recordando que cada dato se expresa como una combinación convexa de los arquetipos, si la componente más grande en dicha combinación pertenece al k-ésimo arquetipo entonces se puede clasificar bajo el k-ésimo grupo. Otro hecho notable de los arquetipos es que aunque yacen en la envolvente convexa de la nube de datos, pueden coincidir o no con datos que la definen. Al igual que con el algoritmo k-means el número de arquetipos $K$ no viene dado por el algoritmo y debe ser un parámetro explorado por el investigador. Cabe mencionar que debido a que se ejecutan múltiples procesos de minimización en este algoritmo para realizar una sola iteración, el mismo es moderadamente costoso en términos computacionales[32] y la exploración de resultados adecuados puede ser extensa.

Ahora se realiza una muestra del resultado de aplicación de este algoritmo a los datos fabricados de la figura 3.

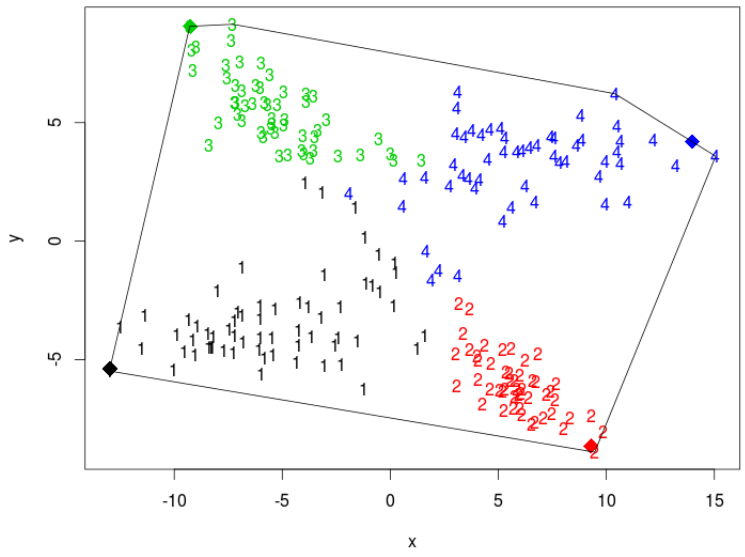



Figura 5: Agrupamiento mediante $k = 4$ arquetipos para los datos generados

En la figura 5 se muestra la asignación de cada punto de la distribución de ejemplo al arquetipo con mayor componente en su respectiva combinación convexa. Además se

grafica la envolvente convexa de la nube de datos. Los arquetipos se muestran como rombos y yacen en la envolvente convexa. Como se mencionó anteriormente los arquetipos no necesariamente son datos que definen la envolvente convexa, por ejemplo el arquetipo del grupo 4 (azul) no es un dato original. Se debe observar que aunque el agrupamiento resultante de este algoritmo es bastante parecido al del algoritmo k-means en la figura 4, no se basa en distancias a los arquetipos, y esto se evidencia claramente con algunos de los puntos asignados al arquetipo 1 (negro). Esto se interpreta como que existen datos arquetipales que representan de manera "ideal" al resto de datos y aquellos no son nada más que una mezcla de estos elementos prototípicos, luego cada asignación se realiza respecto al arquetipo que representa de mayor manera al dato.

### 3.2.3. Mapas auto-organizados

Los mapas auto-organizados (self-organizing maps o SOM en inglés) también llamados mapas de Kohonen en honor a su descubridor[33], constituyen una técnica de clasificación no supervisada de objetos altamente visual, pues su producto final entrega un mapa que aproxima en dos dimensiones la distribución de los datos en el espacio de características.

Los mapas-autoorganizados se pueden entender como un tipo de regresión no paramétrica en el que se ajustan un número fijo de vectores discretos ordenados a la distribución de los datos. Estos vectores de referencia definen los nodos de una "red elástica" en donde se preserva un orden topológico entre ellos además de un grado de regularidad entre nodos vecinos gracias a interacciones locales[33] que definen esta "elasticidad". Los nodos como tal definen un grupo, pues datos cercanos a cada vector de referencia pertenecen o están asociados al nodo correspondiente. El mapa se construye con un algoritmo que ajusta los vectores referenciales o prototipos a la densidad de los datos de forma competitiva y cooperativa a la vez y que entrega una estructura que se auto-organiza. Este algoritmo se puede ver como una versión restringida del algoritmo k-means en el que los prototipos son "obligados" a yacer en una variedad dos dimensional del espacio de características[25] preservando un orden entre ellos. Este proceso se puede interpretar también como un plano bidimensional que se deforma para ajustarse a la distribución de los datos.

En sí un mapa auto-organizado consiste de una red neuronal artificial de una sola

capa en donde la j-ésima neurona (o nodo) está numerada de manera ordenada en una grilla bidimiensional y que tiene asociado un vector m-dimensional $w_j$ (vector código) con $m$ la dimensión del espacio de características de los objetos. Cada neurona está conectada con sus vecinas definiendo una vecindad rectangular o hexagonal. Cuando un dato se le presenta a la red genera un proceso de auto-organización en una región de la misma y después de muchos datos e iteraciones permite la aproximación de la red (los nodos) a la distribución de los datos en el espacio de características. Los pasos que describen el procedimiento para realizar el ajuste son los siguientes [33, 24]:

1. Se define el tamaño (número de nodos y dimensiones) y la topología (plana, cilíndrica, toroidal) de la red bidimensional. Para descripciones posteriores supongamos que el número total de nodos es $l$. Cada nodo de la red tiene un índice que lo identifica y localiza según la geometría definida. Se inicializan los vectores en los nodos (también llamados "vectores código"). A cada vector se le asignan valores aleatorios o se escogen vectores arbitrarios de los datos (sin repetición).

2. Se empieza a "alimentar" a la red para que aprenda con los datos. Para cada dato ingresado, se calcula la distancia (Euclideana por lo general) del nodo de la red a dicho dato en función de su vector código. La neurona con la menor distancia al dato se declara "ganadora" (se denomina unidad de mejor ajuste). Sea $m$ la dimensión del espacio de características, $x_i = (x_{i1}, \ldots, x_{im})$ el dato $i$ alimentado a la red, $w_j = (w_{j1}, \ldots, w_{jm})$ el vector código de la j-ésima neurona de las $l$ definidas. Se busca la unidad de mejor ajuste al dato $x_i$ minimizando:

$$\min_{w_j} ||x_i - w_j|| = \sqrt{\sum_{p=1}^{m} (x_{ip} - w_{jp})^2}$$

   Se puede escribir la unidad de mejor ajuste respecto al vector $x$ como $w_{c(x)}$ y esta unidad determina el centro de una vecindad topológica $h_{c(x)\,j}$. Esta parte del proceso es competitiva pues sólo una neurona (o nodo) se declara "ganadora".

3. La unidad de mejor ajuste determina la localización en la red de una vecindad de nodos que se alterarán con la entrada del dato $x_i$. La función $h_{c(x)\,j}$ determina cuán fuertemente están conectadas neuronas cercanas a $w_{c(x)}$. Esta función es unimodal con una medida de la distancia que corre sobre las posiciones de los nodos en la red (no sobre los vectores código asociados) en función de una distancia

$d_{c(x)\,j} = ||r_c - r_j||$ con $r_k$ la posición de la k-ésimo nodo en el mapa. La elección de esta función no es tan importante para redes pequeñas o moderadas y suele ser Gaussiana:

$$h_{cj}(t) = \alpha(t) \times \exp\left(-\frac{||r_c - r_j||^2}{2\sigma^2(t)}\right)$$

esta elección supone una influencia suave alrededor de $w_{c(x)}$ con $\alpha(t)$ un factor de tasa de aprendizaje y $\sigma(t)$ el ancho de la vecindad. Tanto $\alpha(t)$ como $\sigma(t)$ son funciones monótonamente decrecientes de $t$ es decir, la iteración. En cada iteración todo el conjunto de datos es procesado. El hecho de que estas funciones decaigan con el paso de las iteraciones implica que tanto la vecindad como la magnitud de variación de la unidad de mejor ajuste son grandes al inicio del proceso (para aproximar la distribución de los datos de forma general) pero van disminuyendo conforme avanzan las iteraciones para que el algoritmo converja y el ajuste a los datos sea fino. Esta parte del proceso es cooperativo puesto que la neurona ganadora define un vecindario de influencia sobre otras neuronas.

4. La combinación de estos pasos permite la modificación de los vectores código de los nodos involucrados al momento de la entrada del dato $x_i$. En este paso la unidad de mejor ajuste se modifica acercándose (es decir replicando de mejor manera) al dato de entrada y con varias iteraciones implica el ajuste de los nodos a los datos y una eventual partición del mapa en grupos (cada grupo asociado a un nodo). Con la entrada de cada dato la actualización de los nodos ocurre siguiendo el siguiente modelo adaptativo:

$$w_j(t+1) = w_j(t) + h_{c(x)\,j} \times [x_i(t) - w_j(t)]$$

donde el tiempo $t$ indica la iteración. Esta expresión implica que el valor del vector asociado al nodo $j$ cambia según su localización respecto la unidad de mejor ajuste y además la magnitud del cambio está controlada según sea el inicio del proceso o no (en la función $h_{c(x)\,j}$). Esto depende de cuántas veces todo el conjunto de datos será presentado al sistema y por ende el número de iteraciones total es un parámetro que se debe fijar de antemano. Si el número total de iteraciones es $I$ entonces se hace que cuando $t = I/2$, $\sigma(t)$ sea cero, por lo que la primera mitad de las iteraciones constituyen un proceso de ajuste general de la distribución de

puntos que preserva el orden topológico en toda la escala de los datos y la segunda mitad representa ajustes finos (solo la unidad de mejor ajuste será modificada).

Este conjunto de pasos asegura que los vectores código de cada nodo se aproximen a la distribución de datos entregada al algoritmo y que cada nodo tenga como vecinos nodos con vectores código similares/parecidos preservando un orden topológico. Esto quiere decir que los nodos se ordenan en la grilla tal que prototipos similares están asociados a nodos cercanos entre sí. Este resultado logra que los datos se agrupen según los vectores código (agrupamiento) y que además estén ordenados en el espacio (dos datos similares caerán en nodos similares -sino en el mismo-). El hecho de que se preserve un orden para los datos en una grilla bidimensional implica que se pueden hacer visualizaciones e interpretaciones rápidas que se detallarán en la sección de resultados. Hay que mencionar que la ejecución de este algoritmo es bastante liviana y puede usarse para la exploración de grandes bases de datos[25].

Por las razones expuestas los mapas auto-organizados representan una técnica de agrupamiento (cada dato está asociado a la unidad de mejor ajuste -y nodo- que le corresponda), pero también se pueden usar como una técnica de proyección [22]. Esto se logra pues cada dato tiene un lugar en el mapa y si se realizan agrupamientos de los vectores código, se pueden interpretar las propiedades de cada dato en función de la localización en estos grupos. Esto se visualiza de mejor manera a continuación con las diversas informaciones que nos dan los mapas auto-organizados. Primero se muestran los productos de ejecutar un mapa auto-organizado en los datos generados de ejemplo anteriormente.

Como se mencionó, en un mapa auto-organizado hacen falta definir algunos parámetros (a diferencia de solamente el número de grupos como en k-means o análisis de arquetipos): número de nodos total con su debida geometría, la topología del mapa, el número de veces que se van a presentar todos los datos al mapa, el factor de tasa de aprendizaje, el tipo y tamaño inicial de vecindad. De entre estos parámetros el factor de tasa de aprendizaje y el tamaño inicial de vecindad se dejan en valores por defecto que se han determinado como convenientes en otros estudios[33, 22] (factor de tasa de aprendizaje entre 0.05 y 0.01 y tamaño de vecindad $2/3$ la distancia total de las unidades). Se construyó una grilla de dimensiones $13 \times 10$ con vecindades hexagonales, con topología plana (no toroidal o cilíndrica). Además se presentaron 100 veces todos los

datos al mapa.

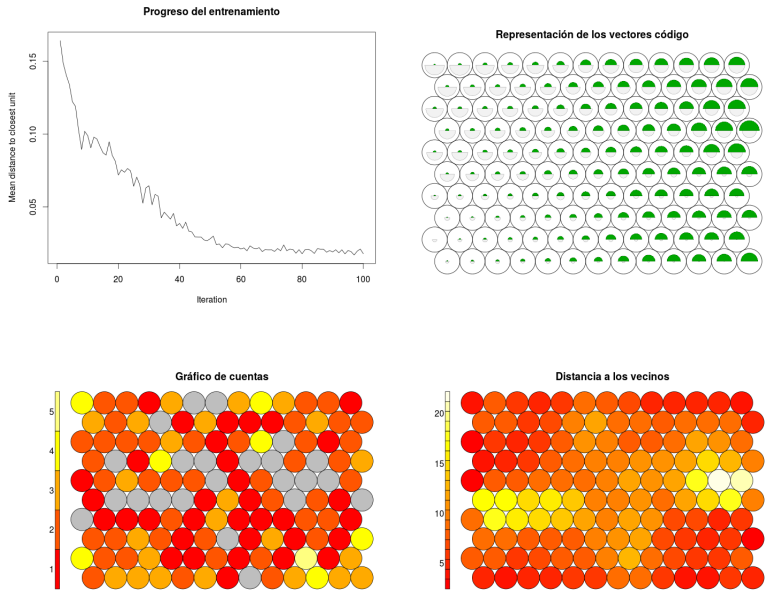



Figura 6: Productos de un mapa auto-organizado

Los productos de un mapa auto-organizado son varios. En la figura 6 se presenta el progreso del entrenamiento de la red para ajustarse con los datos en función de las iteraciones en la parte superior izquierda. En la parte superior derecha se muestra una representación de los vectores código y su ubicación en el mapa. Por la naturaleza de los datos como fueron construidos se puede interpretar el tamaño de la semiesfera superior verde como asociado al valor en $x$ de los datos y el de la semiesfera inferior blanca a $y$. Se puede notar que nodos cercanos tienen vectores código asociados parecidos. La figura inferior izquierda muestra cuántos datos "caen" en cada nodo. Los nodos grises indican que ningún dato es más cercano a su vector prototipo. Normalmente los mapas no deberían tener muchos de estos nodos "vacíos" pero en este caso son producto de una red construida demasiado grande para el ejemplo para efectos visuales. El gráfico inferior derecho muestra la distancia promedio a los vecinos cercanos de cada nodo. Este mapa también es llamado "U-matrix" y en problemas de mayor dimensión sirve para visualizar e identificar regiones de alta densidad de datos del espacio de características separadas por distancias grandes. Es con base en esta información que se pueden realizar técnicas de agrupamiento -como k-means- en espacios altamente reducidos (los mapas

auto-organizados) y facilitar una exploración de agrupamientos convenientes.

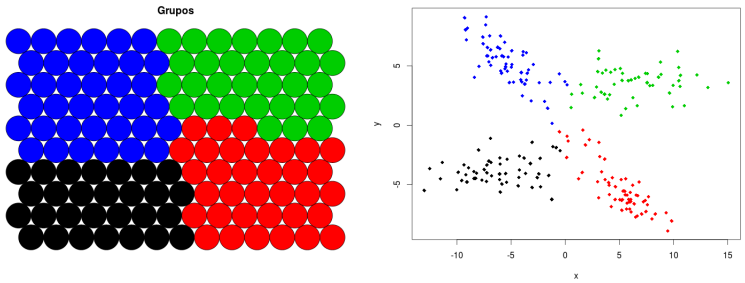



Figura 7: Agrupamiento por medio del algoritmo k-means de los vectores prototipo en un mapa auto-organizado y asignación posterior a los datos asociados a cada uno de ellos

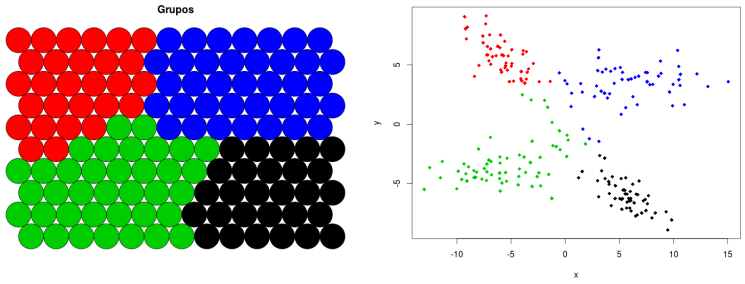



Figura 8: Agrupamiento con análsis de arquetipos de los vectores prototipo en un mapa auto-organizado y asignación posterior a los datos asociados a cada uno de ellos

La figura 7 y la figura 8 muestran el resultado de aplicar el algoritmo k-means y el análisis de arquetipos respectivamente a los vectores prototipo del mapa auto-organizado y después asignar el grupo a cada dato en función de su vector prototipo asociado. Se observa que a grandes rasgos el agrupamiento logra resultados similares a los anteriores pues el mapa auto-organizado preservó las propiedades topológicas de la distribución de los datos.

# 4. Aplicación de algoritmos y resultados

Para realizar la clasificación de eventos volcánicos se realizaron tres procesos secuenciales. El primero fue la determinación mediante exploración de agrupamientos de las variables que se usaron para la aplicación posterior de los algoritmos de clasificación no

supervisada. Después se efectuó un análisis de los posibles agrupamientos que se pudieran realizar en espacio de características con todos los datos mediante criterios propios de los algoritmos k-means y análisis de arquetipos. El último paso consistió en hacer exploraciones más rápidas y visualizaciones en varios mapas auto-organizados. De esta forma primero se presentan los pasos para la determinación de los agrupamientos para el algoritmo k-means junto con los resultados de aplicarlo a todos los datos. Luego se explican los métodos para definir los agrupamientos usando criterios propios del análisis de arquetipos y los resultados respectivos. Finalmente de manera separada se muestran los resultados de combinar los mapas auto-organizados con los otros dos métodos después de las exploraciones y análisis individuales.

## 4.1. Exploración de características y determinación de parámetros

El primer paso exploratiorio consistió en seleccionar variables adecuadas para la descripción de los eventos volcánicos en el espacio de características mediante los vectores de amplitud de frecuencia mencionados en la sección de preprocesamiento de datos. La exploración consistió en calcular el índice de Davies-Bouldin[30] (DB) para señales seleccionando sólo el vector del espectro muestreado o sólo el máximo por intervalo y luego la unión de ambos. El índice de Davies-Bouldin es una medida que compara la dispersión intragrupal y la distancia entre las medias de los grupos, es decir se usa solo con algoritmos de agrupamiento como k-means. El índice es menor mientras mayor sea la distancia entre grupos y más baja dispersión intragrupal. Así índices DB bajos indican agrupaciones mejores. Para valores de grupos ($k = 3, \ldots, 7$) y 4 inicializaciones aleatorias distintas se calculó el índice DB usando cada vez como vector de características de los datos: 1) la amplitud de frecuencia muestreada cada $0{,}5Hz$ 2)la amplitud máxima en intervalos de $0{,}5Hz$ y 3) ambos vectores unidos. Los valores más bajos de índices DB se obtuvieron para el caso en el que se usó solamente el vector de amplitud máxima en intervalos de $0{,}5Hz$ sin embargo al añadir el vector de amplitudes por frecuencia muestreada los índices no aumentaron sustancialmente. Utilizar solamente el vector de amplitudes por frecuencia muestreada dio los índices más altos (peor agrupamiento). Por esto, se usaron ambos vectores combinados pues se obtiene buen agrupamiento según el índice mencionado pero mejor caracterización de los datos en términos de lo discutido

en la sección de preprocesamiento de datos. Los índices DB calculados para los distintos $k$ se adjuntan en los apéndices de este documento.

Para conservar coherencia con lo a lo largo de la investigación y con fines de comparación entre los métodos se decidió usar entonces como vector de característica la unión de ambos vectores en todos los métodos subsiguientes.

### 4.1.1. Algoritmo k-means

Efectuado este análisis se procedió a aplicar el algoritmo k-means directamente sobre los datos. Ya que para este conjunto de datos el índice de DB aumentó para valores crecientes de $k$ con $k = 3, \ldots, 7$, se buscó definir el número de grupos adecuado mediante un gráfico de codo como en la figura 9.

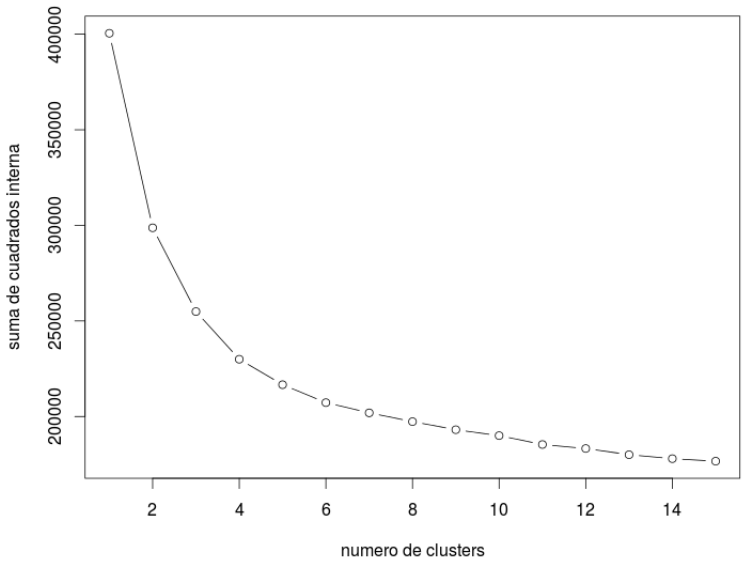



Figura 9: Suma de los cuadrados de la varianza por número de grupos

El criterio del codo consiste en seleccionar el menor $k$ que genere cambios notables en la varianza intragrupal, pues aumentar $k$ siempre generará una varianza intragrupal cada vez menor[25]. Es conveniente también escoger un $k$ pequeño por motivos de interpretación pues la existencia de varios grupos hace más difícil una descripción coherente/consistente de sus propiedades como grupos. Además de este criterio existen muchos más para la elección del número de grupos. Sin embargo, la determinación definitiva de

este número depende al final del tipo de estudio y de la experiencia del investigador. Utilizando lo expuesto entonces pareció razonable escoger $k = 5$ que permite la existencia de un número interesante de grupos que al mismo tiempo se puedan interpretar/explorar más fácilmente.

Seleccionado el número de grupos y aplicado el algoritmo se definió la pertenencia de cada dato a algún grupo. Las $k$ medias encontradas definen regiones de pertenencia en el espacio que pueden intentar visualizarse con base en las direcciones perpendiculares de máxima variabilidad de los datos (con análisis de componentes principales[25]) como se muestra en la figura 10.

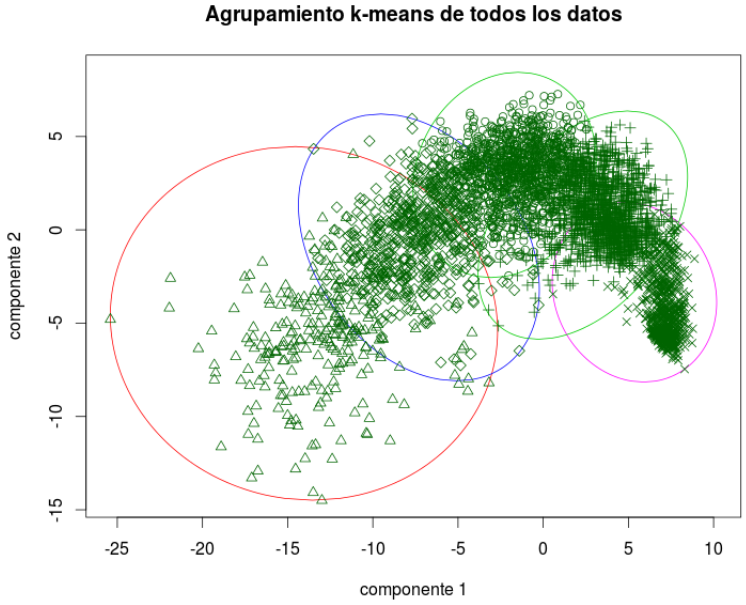



Figura 10: Distribución y asignación de los datos proyectados en las 2 direcciones perpendiculares de máxima variabilidad de los datos con un agrupamiento k-means con $k = 5$

En este gráfico se intentan visualizar los grupos que se formaron con diferentes símbolos para cada grupo y además con las mínimas elipses que contienen a todos los datos correspondientes a cada grups. Las componentes principales explican el 48,29 % de la variabilidad de los datos. Sin embargo esta visualización es insuficiente para exponer la distribución de los datos o su naturaleza.

### 4.1.2. Análisis de arquetipos

En el análisis de arquetipos la búsqueda del mejor número de agrupamientos se realiza ejecutando varias veces (en la medida de lo posible recordando lo intensivo del algoritmo) el análisis arquetipal para distintos valores de $k$ y se usa como medida de error de ajuste la suma de los cuadrados de los residuos de cada una de estas ejecuciones. El resultado de este análisis se adjunta también en los apéndices de este documento.

Después de calcular las SCR de cada iteración para cada grupo $k$ se realiza un gráfico de codo similar al del algoritmo k-means presentado en la figura 11.

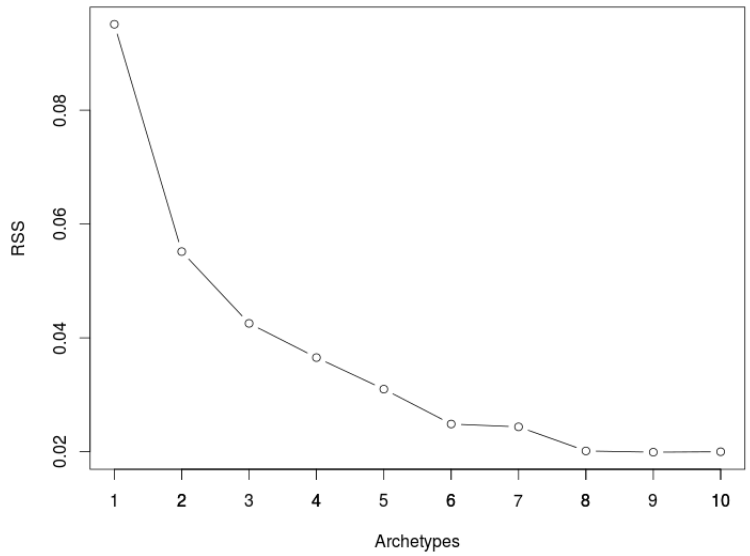



Figura 11: Suma del cuadrado de los residuos por número de arquetipos

En la figura 11 se muestra que la suma del cuadrado de los residuos disminuye conforme aumentan el número de arquetipos, pero la disminución deja de ser notable cuando $k = 6$. Por otro lado para mantener consistencia y poder comparar resultados se elije $k = 5$ con la confianza de que no se disminuirá mucho la calidad del análisis (esto se puede corroborar verificando la tabla de suma del cuadrado de los residuos en el apéndice correspondiente).

Al realizar el análisis de arquetipos con $k = 5$ se obtiene una asignación de cada dato al grupo definido por el arquetipo que lo representa de mejor manera. Esto se

visualiza nuevamente en la proyección de los datos en el espacio de las dos componentes perpendiculares de mayor variabilidad de los datos en la figura 12.

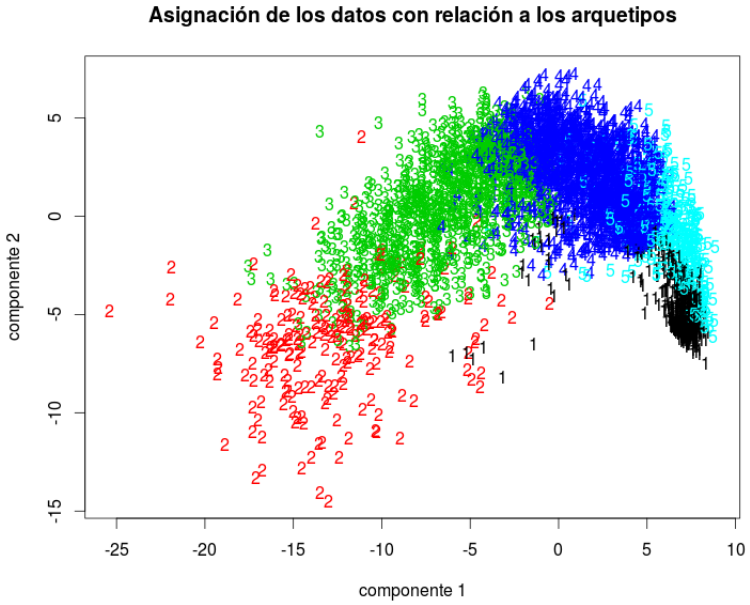



Figura 12: Distribución y asignación de los datos proyectados en las 2 direcciones perpendiculares de máxima variabilidad de los datos realizando un análisis de arquetipos con $k = 5$

En la figura 12 se observan las asignaciones tal como se realizaron a raíz del análisis de arquetipos. Es claro que la asignación es distinta que la presentada en el análisis k-means mas nuevamente la distribución mostrada con las componentes principales no otorga información acerca de la naturaleza de la asiganción o los grupos de datos pues el significado de las componentes principales en espacios de alta dimensión es de difícil interpretación.

Debido a la difícil visualización de la distribución y la asignación de datos tanto al aplicar el algoritmo k-means como el análisis de arquetipos (y además por la difícil capacidad de exploración de parámetros y asignaciones interesantes en el caso de análisis de arquetipos por el tiempo computacional) se decidió utilizar estos dos métodos en un mapa auto-organizado de los datos (similar a lo realizado como ejemplo en la sección de métodos) para obtener los resultados finales de este trabajo. Esto se describe a continuación.

## 4.2. Resultados

Los resultados finales de este proyecto se obtuvieron aplicando un mapa auto-organizado a todos los datos y obteniendo agrupamientos para los mismos indirectamente a través de dicho mapa. El mapa en sí consta de una red plana con dimensiones $13 \times 23$ y vecindad hexagonal. Los datos se utilizaron en el aprendizaje 200 veces. El resto de parámetros se conservaron con sus valores por defecto. En este caso se escogió una red con 299 nodos puesto que se ha demostrado en otros estudios[22] que un número adecuado de nodos es aproximadamente $5 \times \sqrt{n}$ donde $n$ es el número total de datos, en este caso 4006. El resultado de aplicar el algoritmo a los datos produjo el gráfico del progreso de entrenamiento y de número de datos por nodo que se muestran juntos en la figura 13, el gráfico con la representación de los vectores código en la figura 14 y el gráfico de las distancias a los vecinos inmediatos en la figura 15.

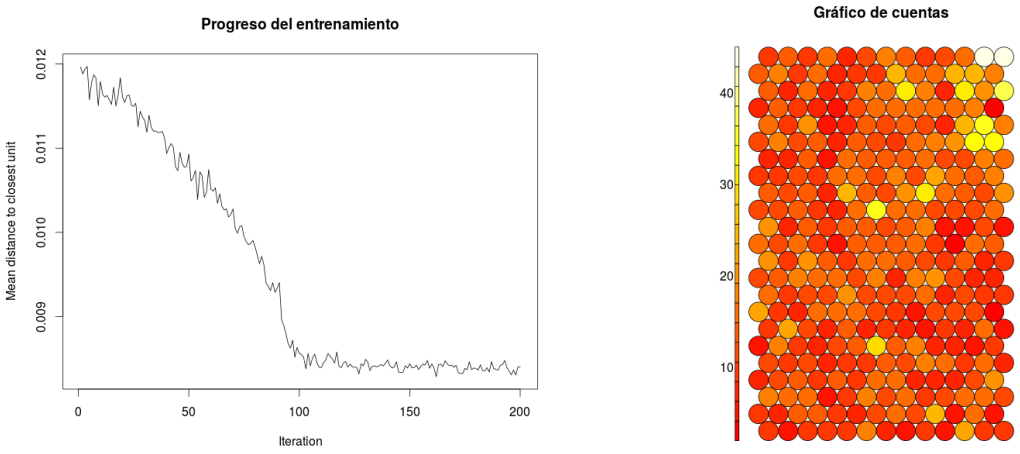



Figura 13: Izq: Progreso del entrenamiento de la red en 200 iteraciones. Der: Gráfico de número de datos asociados a cada nodo

En la figura 13 se muestra el progreso del entrenamiento a la izquierda y el gráfico con el número de datos asociados a cada nodo a la derecha. Se observa que la distancia promedio de los datos a las unidades de mejor ajuste es pequeña por lo que se puede decir en primera instancia que la red aproxima bien la distribución de los datos en el espacio de características. En el gráfico del número de datos asociados a cada nodo se verifica que no hay nodos vacíos por lo que la red es adecuada.

En la figura 14 se muestra la ubicación y la forma de cada vector código en los nodos del mapa auto-organizado. Esta visualización confirma que de forma aproximada los

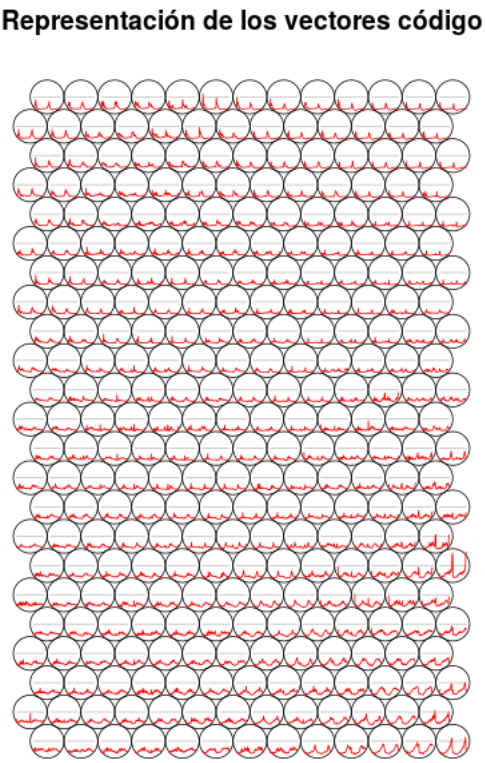



Figura 14: Vectores código asociados a cada nodo en el mapa auto-organizado

nodos poseen vecinos parecidos. Es importante observar que esto produce regiones en las que habrá datos con contenido espectral similar independientemente de su tamaño. Cabe recalcar que estos vectores código son idealizaciones de la unión de dos vectores de amplitud de frecuencia unidos (como fue explicado al inicio de la sección de exploración de características) y por eso se observa que los vectores código presentan una forma que usualmente se duplica (esto se puede observar claramente en el vector código más inferior a la derecha).

En la figura 15 se muestra la distancia promedio de cada vector código a sus vecinos inmediatos. A diferencia del ejemplo mostrado en la sección de los métodos, en este mapa es difícil identificar grandes zonas donde evidentemente estén "encerrados" un alto número de puntos (regiones de alta densidad -poca distancia entre vecinos- rodeadas de vectores alejados a vecinos). Esto se debe a la alta dimensionalidad del espacio de características ($m = 100$) que implica que hay lugar a complejas distribuciones de

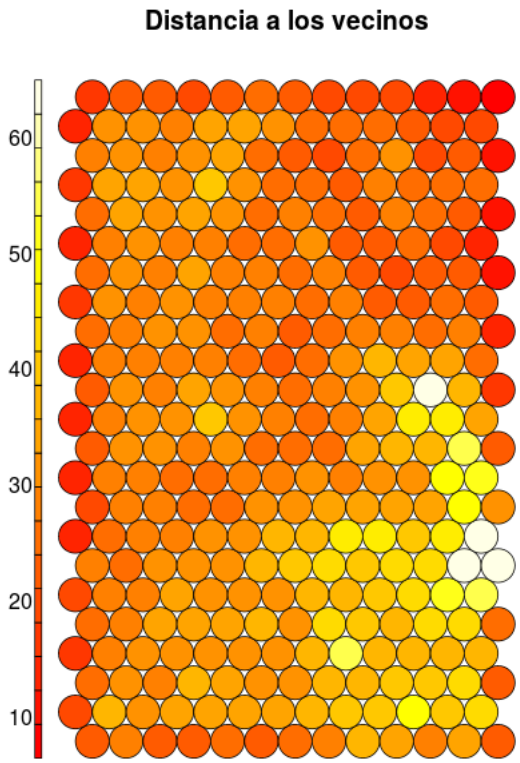



Figura 15: Distancia promedio a los vecinos de los vectores código asociados a cada nodo

los datos en el mismo. Sin embargo se puede reconocer al menos porciones de mapa distantes a otras (por ejemplo la esquina inferior derecha está claramente alejada de la parte superior del mapa en términos de la distancia promedio de los vectores. Además gráficamente cualquier agrupamiento que se efectúe en este espacio de la misma forma que en los ejemplos de la sección de los métodos se puede visualizar más fácilmente y es más fácilmente interpretable que la visualización de la proyección en las componentes principales gracias a la correspondencia de cada nodo con los vectores código mostrados en la figura 14.

Es así que se procedió a efectuar un agrupamiento k-means y luego un análisis de arquetipos sobre el mismo mapa auto-organizado con $k = 5$ con fines de comparación y se presentan los agrupamientos obtenidos a continuación. Las asignaciones para cada grupo en sus representaciones visuales se presentan con distintos colores: Grupo 1: negro; grupo 2: rojo; grupo 3: verde; grupo 4: azul; grupo 5: celeste.

### 4.2.1. Mapa auto-organizado y algoritmo k-means

Al aplicar el algoritmo k-means a los vectores código del mapa auto-organizado se obtuvo el agrupamiento mostrado en la figura 16.

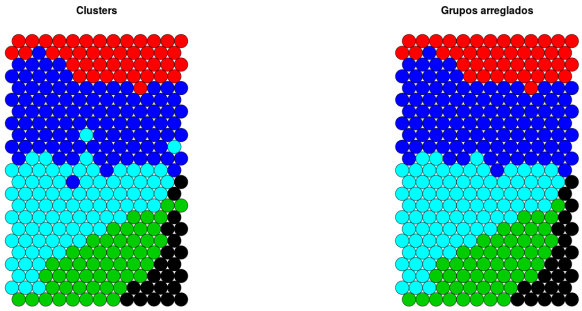



Figura 16: Izq: Agrupamiento de los vectores código del mapa auto-organizado con $k = 5$ crudo. Der: Mismo agrupamiento donde se "limpiaron" posibles puntos atípicos

En la figura 16 de la derecha se muestra el agrupamiento obtenido con el método k-means y aunque la asignación de los datos es visualmente directa y comparable con los vecores código de la figura 14, las fronteras de los grupos son en cierta forma ambiguos puesto que como se explicó antes, nodos cercanos poseen vectores código parecidos. Por esto para darle una interpretación más fuerte a los agrupamientos se exploró la naturaleza de los datos pertenecientes a cada grupo examinándo sus formas de onda directamente. Antes de realizar el análisis de cada grupo se muestra la sismicidad total utilizada en este trabajo correspondiente a los 4006 eventos utilizados en este trabajo en la figura 17.

En la figura 17 se muestra la frecuencia diaria de eventos utilizados en este trabajo en barras negras y además se indica con líneas veticales rojas el inicio de distintos episodios eruptivos que el volcán Tungurahua presentó en el año 2014. Los espacios en blanco fueron días en los que la estación sísmica tuvo problemas de adquisición o transmisión y no se tuvieron registros sísmicos.

Con base en la figura 17 se analizó la naturaleza de los grupos obtenidos en el agrupamiento mediante k-means sobre el mapa auto-organizado graficando sus miembros en el tiempo y una forma de onda típica observada en dicho grupo.

Las figuras 18, 19, 20, 21 y 22 muestran la distribución diaria (arriba) de los eventos asignados al grupo 1, 2, 3, 4 y 5 respectivamente por el algoritmo k-means en el mapa auto-organizado (arriba). Además el segundo gráfico dentro en estas figuras (abajo) consta de 3 partes: la primera (superior) muestra una señal sísmica típica o usual de ese

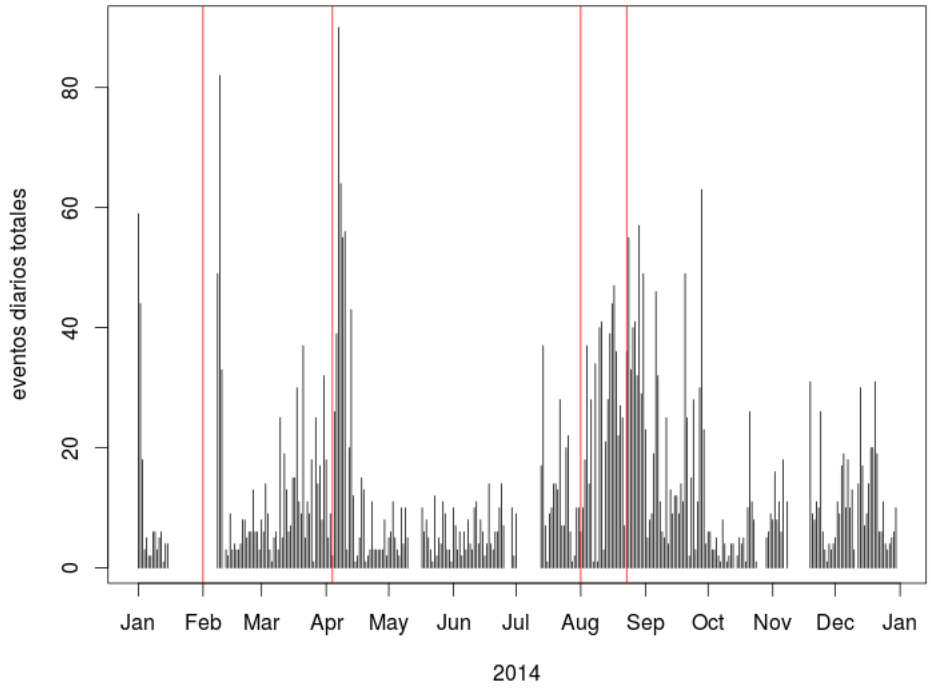



Figura 17: Frecuencia diaria de todos los eventos de 2014 utilizados en este estudio e indicadores de inicio de actividad eruptiva del volcán Tungurahua

grupo tal como es adquirida, la segunda (medio) muestra el espectro de la señal y la tercera (inferior) muestra una componente de infrasonido captada en la misma estación para tratar de revisar su naturaleza más allá de la sísmica. Este modo de presentar los datos fue compartido y discutido con expertos del Instituto Geofísico - EPN que ayudaron a interpretar los agrupamientos con base en dichas observaciones y su experiencia en monitoreo volcánico.

En el caso del grupo 1 (color negro en la asignación sobre el mapa auto-organizado) hubo consenso en que este grupo consistía de señales puramente ruidosas pues las frecuencias de los mismos no caían en regiones naturales para eventos volcánicos. Además las formas de onda de los eventos de este grupo tampoco poseían rasgos típicos de eventos volcánicos, por lo que podrían haberse originado como producto de desperfectos electrónicos, rayos, movimiento de animales o personas u otros fenómenos atípicos cercanos a la estación.

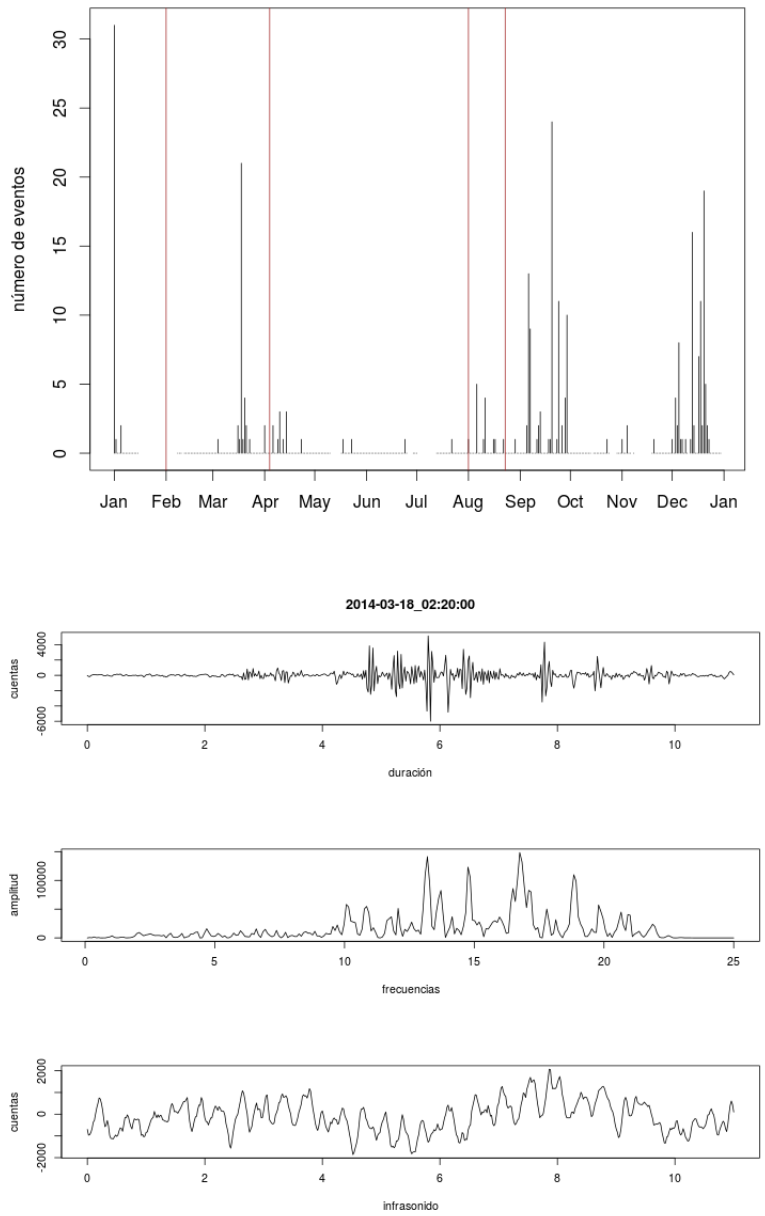



Figura 18: Distribución en el tiempo (arriba) y evento típico perteneciente al grupo 1 según el algoritmo k-means aplicado al mapa auto-organizado

En el caso del grupo 2 (color rojo en la asignación sobre el mapa auto-organizado) es destacable que las señales poseen un impulso claro en infrasonido y durante las discusiones con los especialistas se corroboró que este grupo identifica señales asociadas a

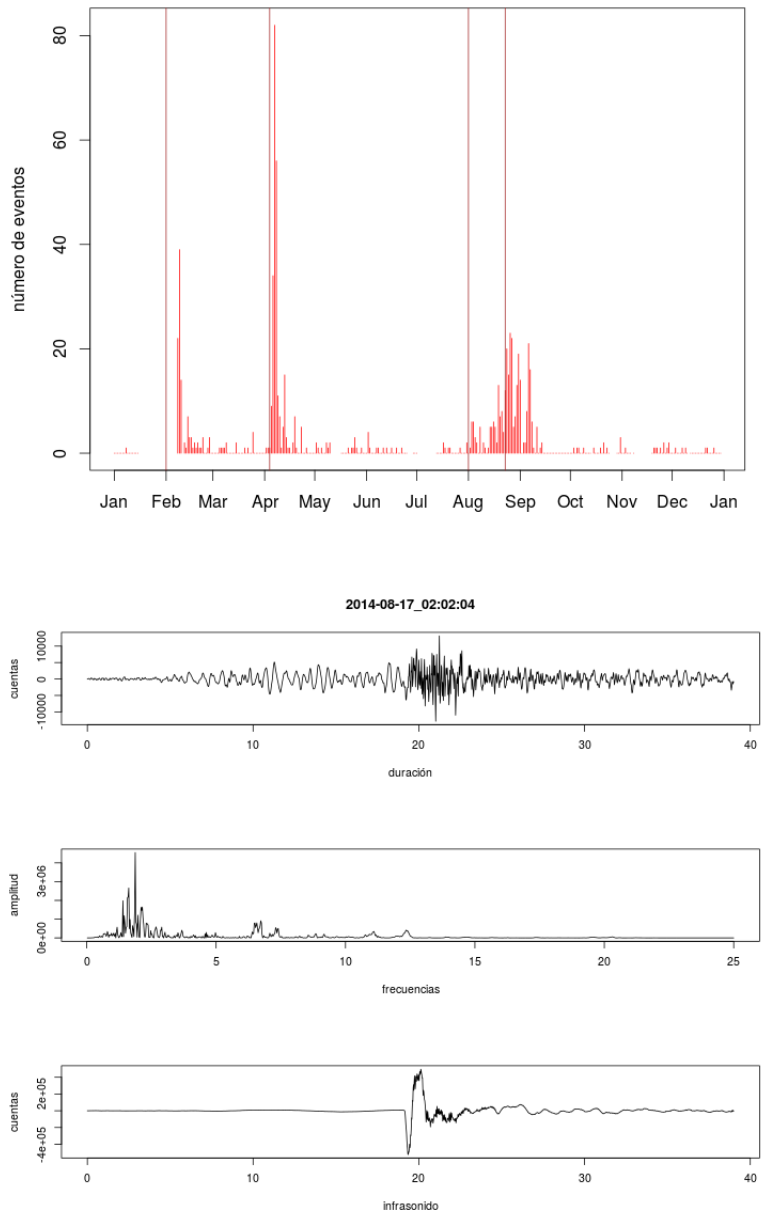



Figura 19: Distribución en el tiempo (arriba) y evento típico perteneciente al grupo 2 según el algoritmo k-means aplicado al mapa auto-organizado

explosiones o eventos de emisión (por eso aparecen de manera casi exclusiva durante períodos de alta actividad volcánica en el tiempo).

Para los eventos del grupo 3 (color verde según la asignación de grupos del mapa auto-

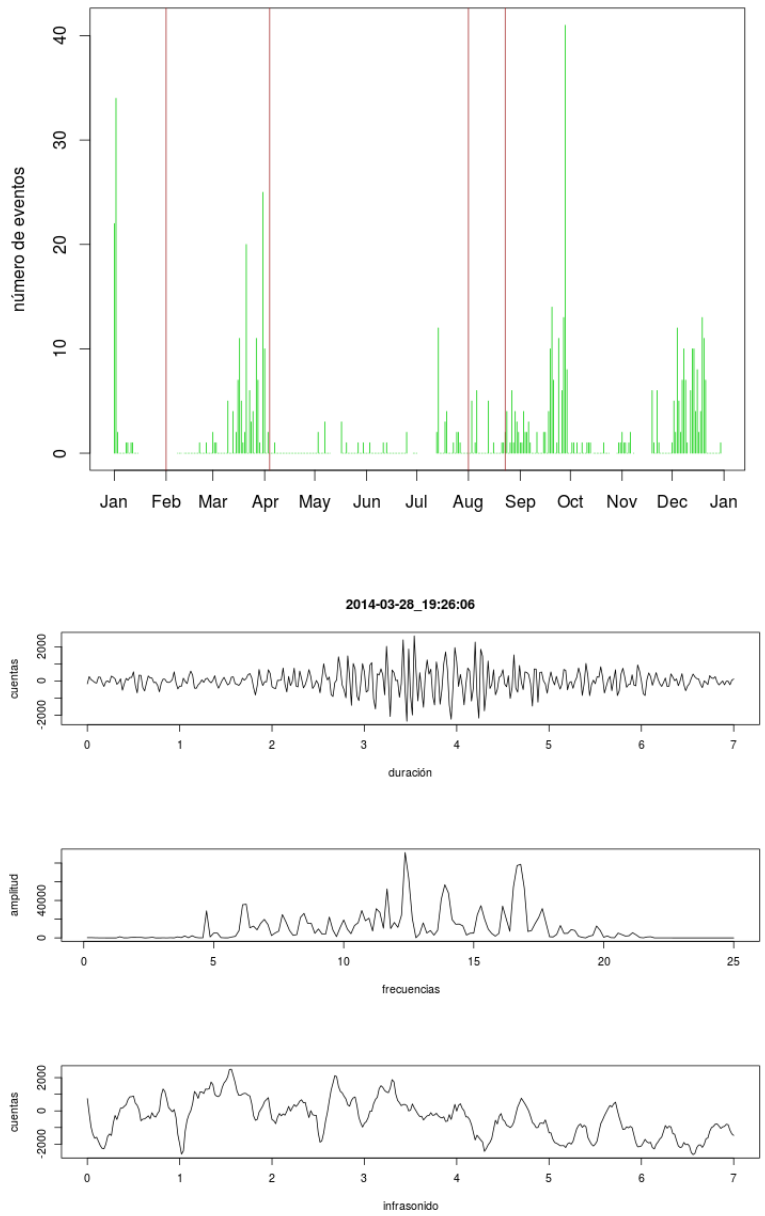



Figura 20: Distribución en el tiempo (arriba) y evento típico perteneciente al grupo 3 según el algoritmo k-means aplicado al mapa auto-organizado

organizado) no se identificaron rasgos particulares notables o conocidos como eventos de origen volcánico. Al ser analizados, discutidos y comparados con tramos de tiempo medidos en otras estaciones estos eventos aparecieron de forma local (es decir no se

observaron en otras estaciones) y podrían representar otra clase de ruido menos atípico (en el sentido de que tienen frecuencias más parecidas a las de eventos volcánicos) quizás asociado a pequeños deslaves.

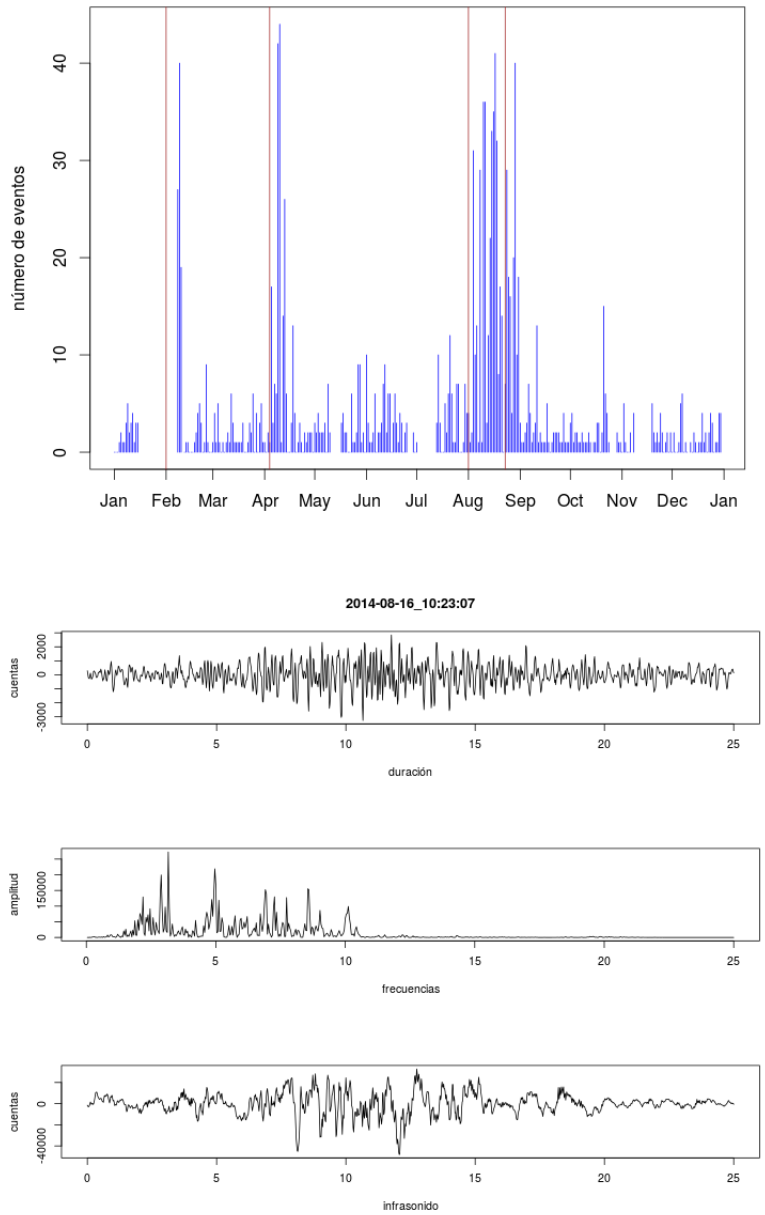



Figura 21: Distribución en el tiempo (arriba) y evento típico perteneciente al grupo 4 según el algoritmo k-means aplicado al mapa auto-organizado

Para los eventos de la categoría 4 (azul en la asignación mediante k-means en el mapa auto-organizado) es interesante notar que guardan cierta similaridad con los eventos del grupo 2 pues aparecieron mayormente en épocas de alta actividad eruptiva, aunque también en cantidades moderadas fuera de ellas. Los eventos de esta categoría en sí mostraron formas de onda que no son claras pues constituyen eventos relativamente pequeños como se indica en la figura 21. Estos eventos parecen tener altas frecuencias y también algún contenido en la componente de infrasonido, por lo que podrían representar emisiones de alta frecuencia, pequeños derrumbes o quizás fenómenos locales parecidos a los de la clase 3 -opción debatible por la distribución en el tiempo de estos eventos-. Después de las discusiones con los expertos no quedó clara la naturaleza de estos eventos (no se asocian inmediatamente con ruido con base en su forma y contenido espectral).

Finalmente los eventos del grupo 5 (aparecen celestes en la asignación sobre el mapa auto-organizado) mostraron características claras en su forma y espectro, y además por su manifestación a lo largo del año llevaron a la conclusión de que se trata de eventos tipo VT o terremotos regionales.

De los agrupamientos obtenidos se tiene que el 6,8 % de los datos pertenecen al grupo 1 (ruido puro), el 18,1 % al grupo 2 (explosiones/emisiones), el 14,4 % al grupo 3 (ruido local), el 34,8 % al grupo 4 (eventos no volcánicos locales) y el 25,9 % al grupo 5 (VTs u otros de origen volcánico).

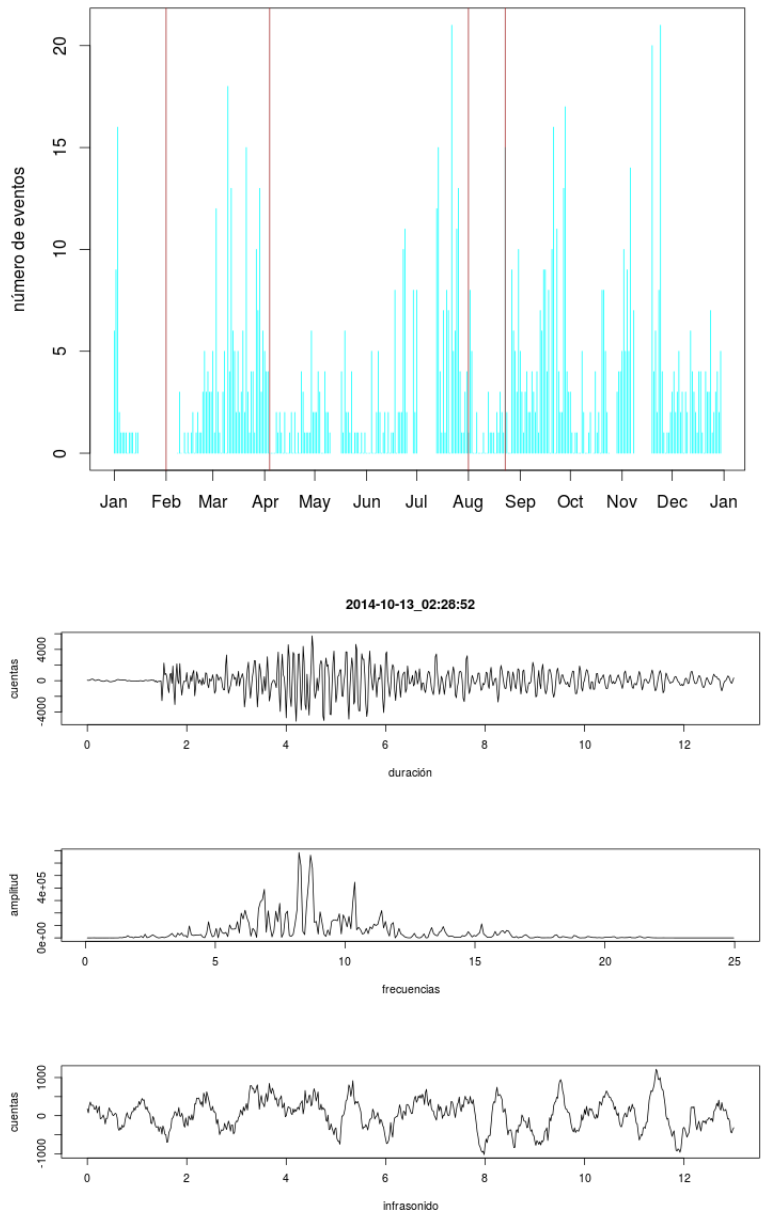



Figura 22: Distribución en el tiempo (arriba) y evento típico perteneciente al grupo 5 según el algoritmo k-means aplicado al mapa auto-organizado

### 4.2.2. Mapa auto-organizado y análisis de arquetipos

Para el análisis de arquetipos del mapa auto-organizado se procedió de forma totalmente similar al caso con el algoritmo k-means. Los razonamientos para la caracterización

de los grupos fueron parecidos cuando fue posible.

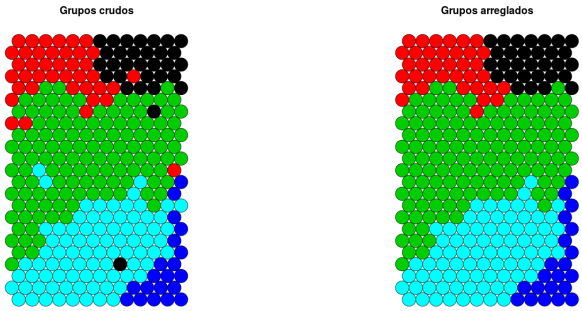



Figura 23: Izq: Análsis de arquetipos de los vectores código del mapa auto-organizado con $k = 5$ crudo. Der: Mismo agrupamiento donde se "limpiaron" posibles puntos atípicos

En la figura 23 se muestra la asignación de los vectores código del mapa auto-organizado según el análisis de arquetipos. Aunque el agrupamiento es similar al obtenido con k-means, hay diferencias notables como la división del grupo originalmente 2 (rojo en la figura 16) en el mapa con k-means. Un resultado también destacado es que el grupo de la esquina inferior derecha se mantiene casi idéntico al encontrado por el algoritmo k-means.

En las figuras 24, 24, 24, 24 y 24 se muestran la distribución de datos y un evento típico asociado a los nuevos grupos correspondientes 1, 2, 3, 4 y 5 respectivamente logrado con análisis de arquetipos aplicado sobre los vectores código del mapa auto-organizado.

Los eventos asociados al grupo 1 (en negro según la asignación del análisis de arquetipos sobre el mapa auto-organizado) vuelven a mostrar características y distribución temporal asociadas a explosiones o explosiones.

Los datos asociados al grupo 2 (en color rojo en el mapa auto-organizado) vuelven a mostrar la distribución de eventos y características asociadas a explosiones o emisiones de algún tipo aunque la separación en dos clases no es evidente según los análisis visuales. Analizando el espectro parece haber algún contenido de frecuencia intermedia en la clase 2 aunque el origen del evento sigue siendo algún tipo de emision (verificado en el infrasonido).

Los datos pertenecientes a la asignación 3 del análisis de arquetipos (en color verde según la asignación al arquetipo 3 en el mapa auto-organizado) poseen frecuencias medias/altas y también muestras de infrasonido. Contiene VTs, sismos regionales y posiblemente derrumbes u otros fenómenos locales aunque producidos por actividad volcá-

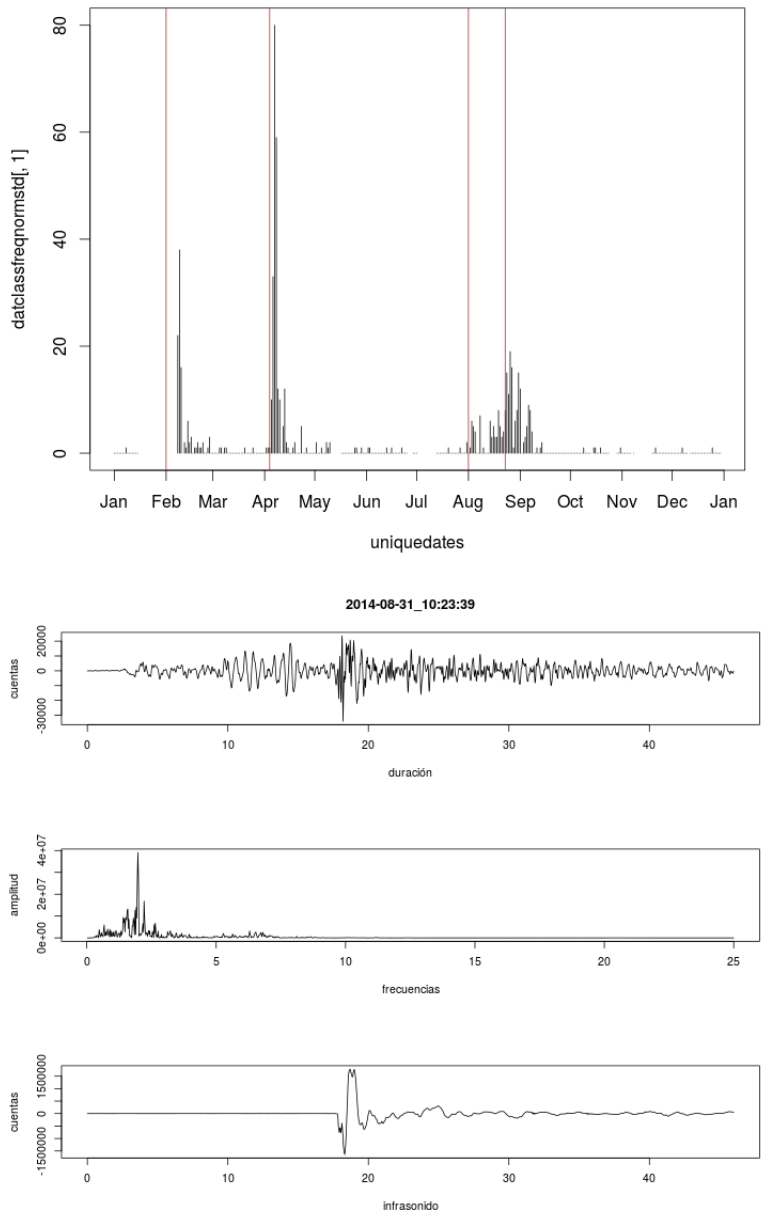



Figura 24: Distribución en el tiempo (arriba) y evento típico perteneciente al grupo 1 según el análisis de arquetipos aplicado al mapa auto-organizado

nica. Es una clase con cierto nivel de variabilidad entre los datos que pertenecen a esta categoría.

En el grupo 4 (de color azul en el mapa auto-organizado) los eventos asociados caen

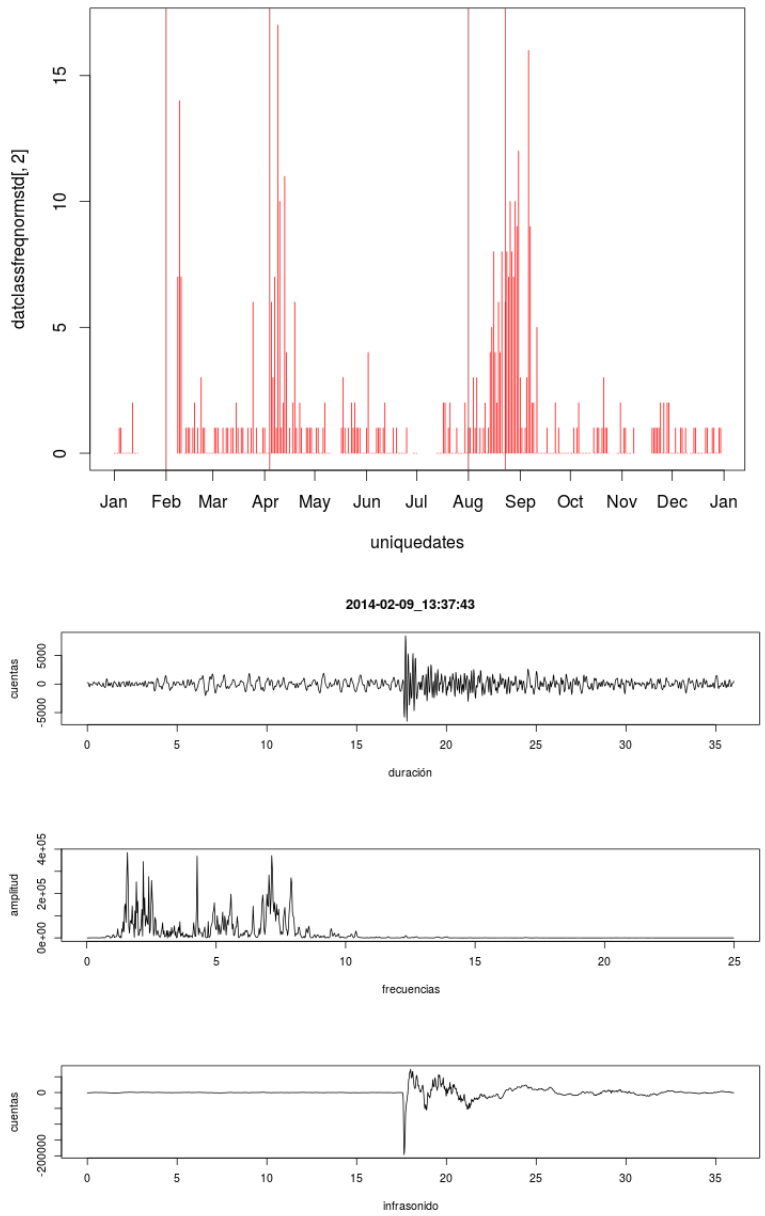



Figura 25: Distribución en el tiempo (arriba) y evento típico perteneciente al grupo 2 según el análisis de arquetipos aplicado al mapa auto-organizado

en la categoría de ruido como se esperaba comparándolo con resultados del algoritmo k-means (por forma y distribución temporal también).

Para terminar los eventos asignados al arquetipo 5 (asociados a los nodos celestes en

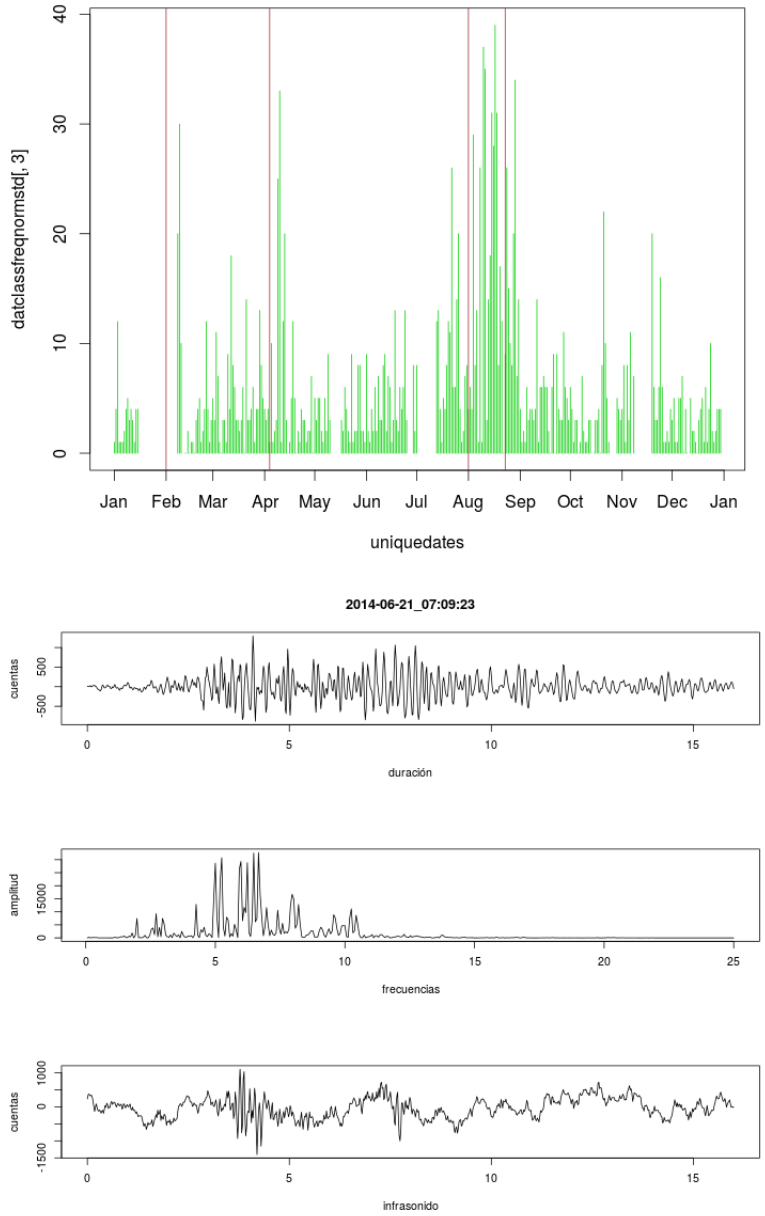



Figura 26: Distribución en el tiempo (arriba) y evento típico perteneciente al grupo 3 según el análisis de arquetipos aplicado al mapa auto-organizado

el mapa auto-organizado) parecen representar los eventos descritos anteriormente como de tipo "ruido local" ocasionado por fuentes generalmente ajenas al volcán pero naturales (pequeños deslaves, animales, seres humanos, carros, entre otros).

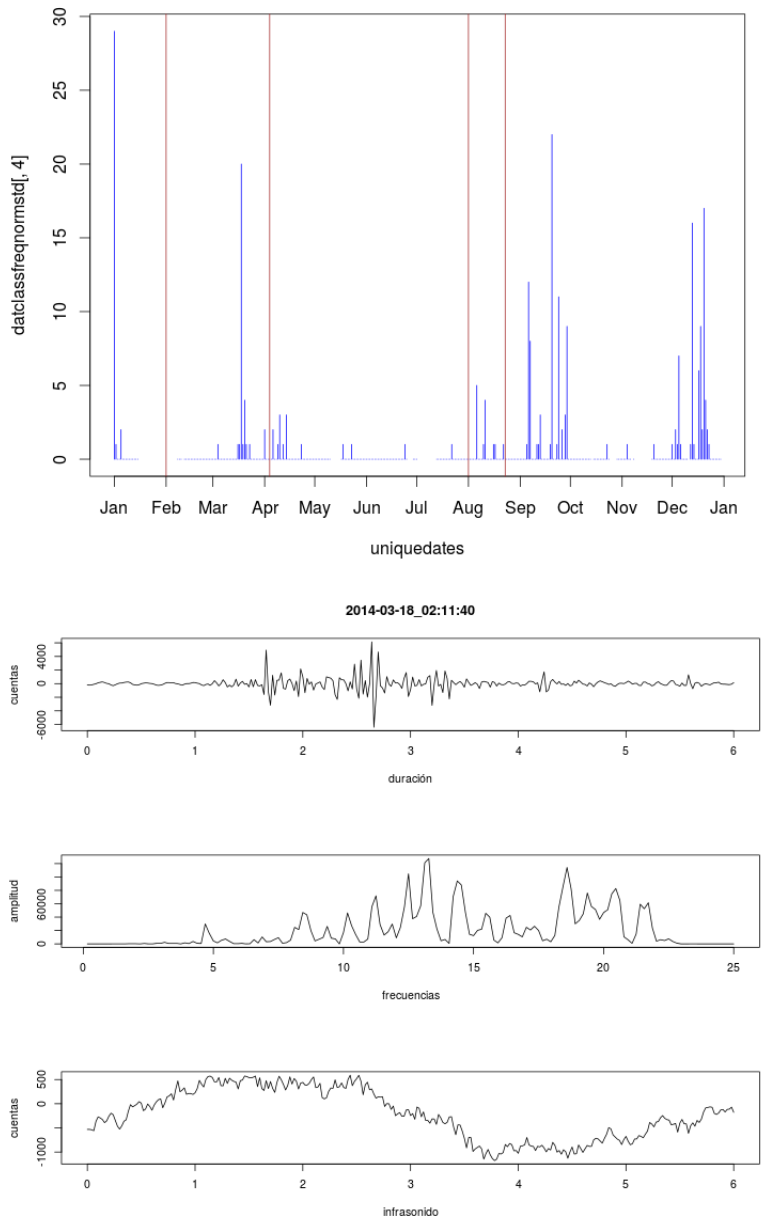



Figura 27: Distribución en el tiempo (arriba) y evento típico perteneciente al grupo 4 según el análisis de arquetipos aplicado al mapa auto-organizado

Para los agrupamientos obtenidos mediante el análisis de arquetipos se tiene que el 14,5 % de los datos pertenecen al grupo 1 y el 10,3 % al grupo 2 (ambos grupos asociados a explosiones o emisiones), el 45,9 % al grupo 3 (VTs, sismos regionales y derrumbes, no

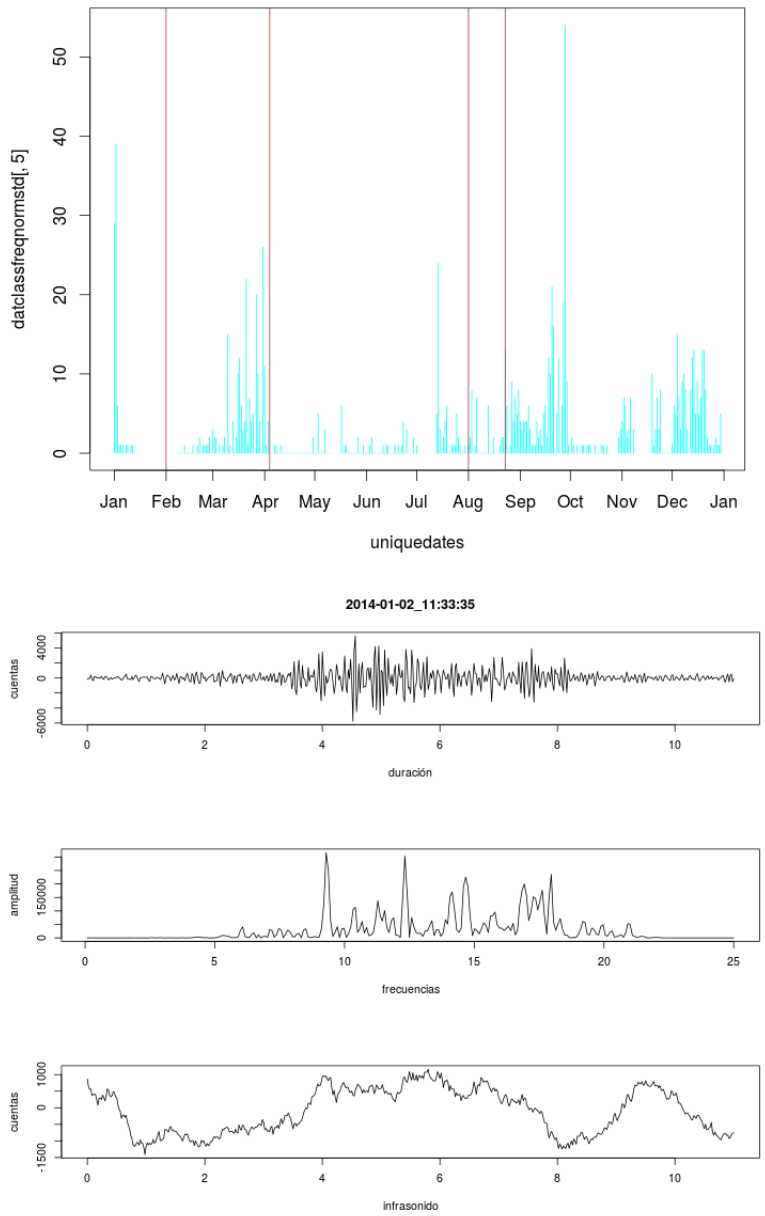



Figura 28: Distribución en el tiempo (arriba) y evento típico perteneciente al grupo 5 según el análisis de arquetipos aplicado al mapa auto-organizado

muy bien definido), el 6 % al grupo 4 (ruido puro) y el 23,3 % al grupo 5 (ruido local).

### 4.2.3. Comparación entre ambas aplicaciones

El resultado de aplicar estos análisis entregó ciertas intuiciones acerca de qué significan los grupos encontrados por los algoritmos. La caracterización y descripción de los agrupamientos se realizó con criterio de expertos y aunque entregó resultados mayormente positivos, existe cierto grado de ambigüedad por la naturaleza de los algoritmos así como de los propios datos. Un resultado notable es por ejemplo la consistencia de ambos algoritmos en encontrar ruido totalmente ajeno al volcán, así como que ambos identificaron eventos asociados netamente a períodos eruptivos. De esta forma se llegaron a identificar explosiones y distinguirlas de otros eventos sísmicos aún sin la utilización de información de infrasonido (esta información sólo se usó para explorar/encontrar la naturaleza de los grupos, no para encontrar los grupos en sí).

Las coincidencias y diferencias entre los grupos encontrados resumen en el cuadro 1.

| | arch1 | arch2 | arch3 | arch4 | arch5 |
|---|---|---|---|---|---|
| kmeans1 | 0 | 0 | 0 | 241 | 31 |
| kmeans2 | 529 | 197 | 0 | 0 | 0 |
| kmeans3 | 0 | 0 | 0 | 0 | 575 |
| kmeans4 | 54 | 214 | 1126 | 0 | 0 |
| kmeans5 | 0 | 0 | 711 | 0 | 328 |

Cuadro 1: Tabla de coincidencias entre los grupos formados por los algoritmos k-means y análisis de arquetipos

En el cuadro 1 se muestra el número de observaciones contrastados entre los grupos obtenidos por el algoritmo k-means por filas y por el análisis de arquetipos por columnas. En el caso de la familia 1 del agrupamiento según k-means la coincidecia es del 88,6 % con la familia 4 de arquetipos y el resto con la familia 5. A su vez los eventos de la familia 4 de arquetipos están contenidos en un 100 % en la familia 1 de k-means. Esto nos lleva a decir que esta es una familia bien definida por ambos algoritmos y corresponden a eventos puramente ruidosos ajenos al volcán u otras causas naturales. El 72,9 % de los eventos de la familia 2 de k-means correspondieron a eventos de la familia de arquetipos 1 y el resto a la famila 2 de arquetipos. Estos eventos eran los reconocidos como explosiones. De forma recíproca el 90,7 % de la familia 1 de arquetipos cae en la familia 2 de k-means y el resto de eventos en la familia 4 de k-means. La familia 2 de arquetipos consistió en un 47,9 % de eventos en la familia 2 y el resto en la familia 4 de k-means. Esto indica que la familia 1 de arquetipos pertenece casi en su totalidad a la familia 2 de k-means

y son eventos puramente explosivos o de emisión del volcán. Por otro lado la familia 2 de arquetipos se compone de eventos de la familia 2 y 4 de k-means que corresponden a emisiones/explosiones y posibles derrumbes o descensos de material volcánico también asociados con actividad (examinar figura 21). Los eventos de la familia 3 del algoritmo k-means se encuentran contenidos totalmente en la familia 5 obtenida por medio de análisis de arquetipos. Esta familia se encuentra dividida en un 61,6 % de eventos de la familia 3, 35,1 % de eventos en la familia 5 y 3,3 % en la familia 1 obtenidas en el algoritmo k-means. La familia 4 según el algoritmo k-means guarda un 80,77 % de coincidencias con la familia 3 de análisis de arquetipos, 15,4 % con la familia 2 de arquetipos y 3,9 % con la familia 1 de arquetipos. Finalmente la familia 5 de k-means tiene un 68,4 % de coincidencias con la familia 3 de arquetipos y el resto con la familia 5 de arquetipos.

Estos resultados implican que hubo grupos de señales bien definidos que fueron encontrados por ambos algoritmos: grupos asociados a eventos durante períodos de alta actividad volcánica (explosiones, emisiones, descenso de material) en el caso del grupo 2 de kmeans y los grupos 2 y 3 de arquetipos. Hubo otros grupos que también reflejaron eventos asociados a señales sísmicas volcánicas o con origen asociado a actividad volcánica aunque menos evidentes en cuanto a su naturaleza (grupo 4 en k-means y 3 en arquetipos). Notablemente el grupo individual de mayor coincidencia en ambos casos estuvo asociado a señales puramente ruidosas (no naturales) en el caso del grupo 1 de k-means y 4 de arquetipos. Estos resultados se pueden intuir basados en la asignación según los mapas auto-organizados de las figuras 16 y 23.

Los resultados mostrados se obtuvieron con la totalidad de los datos. Sin embargo, debido a que no existe actualmente un catálogo de eventos que comprenda los eventos seleccionados en este estudio (por el carácter automático con el que se los escogió en el que se incluyen todas las señales discretas del volcán), para verificar la robustez de lo obtenido y validar los agrupamientos se recurrió a utilizar un enfoque parecido a la validación cruzada. Aunque aquí no existen conjuntos o funciones objetivo conocidas, simplemente se analiza si los agrupamientos obtenidos son estables (y parecidos) para diferentes subconjuntos (muestreos) de los datos. Así se escogió repetir los pasos que entregaron los grupos mostrados 10 veces con muestras diferentes aleatorias del 90 % de los datos. Los resultados de este proceso se adjuntan en el apéndice 3. Dichas pruebas mostraron consistencia en la asignación de grupos por medio de los dos métodos (en los grupos conocidos hubo coincidencias mayores al 85 % más detalladas en el apéndice 3).

Además tanto en los resultados mostrados en esta sección como en la validación del apéndice 3 no se presentaron casos de familias o agrupamientos demasiado dispersos entre los algoritmos. Con esto se verificó la robustez de los algoritmos que produjeron resultados similares en varias pruebas diferentes y la existencia efectiva de clases bien definidas de eventos. Por el espacio de características relativamente pequeño que se utilizó, cada ejecución de los algoritmos tomó un par de minutos a lo sumo (condición necesaria para la extensa exploración demandada por el trabajo).

# 5. Conclusiones

En este trabajo se reconsideró aplicar algoritmos de clasificación no supervisada a señales sísmicas volcánicas prácticamente carentes de preprocesamiento o pre-selección en búsqueda de posibles eventos particulares precursores de episodios eruptivos. Aunque esta meta quizás fue demasiado ambiciosa considerando el enfoque minimalista que se le dio a la resolución del problema, otros objetivos alternativos se alcanzaron.

Se exploró el resultado de aplicar diversas técnicas de clasificación no supervisada bajo distintos parámetros y se comprobó la estabilidad de dichos procedimientos enfrentados a un problema aún abierto sobre datos los menos alterados posible. Además la utilización de un vector de características basado en el espectro de frecuencias como medida de caracterización de eventos volcánicos para su clasificación ofreció una vía simple, rápida y ligera computacionalmente.

Los agrupamientos encontrados tuvieron éxito en identificar claramente al menos 2 clases de señales: Ruidosas (ajenas al volcán) y de infrasonido (asociadas a períodos intensos de actividad del volcán). Este descubrimiento tiene implicaciones mayores a futuro en caso de emergencias volcánicas en las que el personal calificado no esté disponible para responder con velocidad adecuada ante la crisis y también para homologar los posibles intentos de clasificación de eventos. Incluso en los grupos en los que los eventos asociados tuvieron características no muy bien definidas, es posible hacer análisis más exhaustivos tanto en los parámetros de los algoritmos, como en esfuerzos por tratar de descubir la naturaleza propia de los eventos. Otro resultado interesante relacionado a las clases encontradas es que se pudo encontrar grupos de eventos sísmicos asociados a explosiones aún sin analizar componentes de infrasonido de las mismas. Esto puede

tener una aplicación directa en el caso de que los sensores de infrasonido sufrieran algún daño o desperfecto. Además podría confirmar la existencia de unas supuestas "emisiones silenciosas" que consisten en emisiones de ceniza confirmadas visualmente pero carentes de señal de infrasonido asociada (esto quizás porque la señal acústica es pequeña y se atenúa antes de llegar a la estación). En el caso de corroborarse esta hipótesis esto también significaría que estos resultados se podrían usar como una herramienta para detectar "emisiones silenciosas" en días nublados o durante la noche (en donde confirmación visual de las emisiones es improbable).

También se verificó que la velocidad de computo de los mapas auto-organizados (que permiten a su vez la aplicación de posibles algoritmos más complejos a grandes grupos de datos) y su carácter profundamente visual permite crear herramientas de monitoreo fáciles de implementar e interpretar para el uso futuro en observatorios vulcanológicos. Además permite hacer una exploración rápida de los catálogos de eventos pre-existentes en donde se puede tratar de buscar y corregir posibles errores humanos cometidos.

Este estudio también sugiere una investigación más minuciosa acerca de la naturaleza de los grupos conformados o incluso de nuevos agrupamientos más finos que podrían permitir encontrar precursores clave de erupciones volcánicas. También se puede plantear utilizar otros parámetros que caractericen de otras formas a las señales sísmicas para así obtener grupos más detallados y definidos.

Finalmente el alcance y potencia de este estudio se puede multiplicar aplicando las mismas técnicas a otras estaciones de un mismo volcán para tener ratificaciones y definiciones más consistentes de los grupos y los eventos asociados a cada uno. La misma idea a su vez se puede extender a otros volcanes y puede producir un sistema de monitoreo integral de los volcanes del país.

# Apéndice 1: Tablas con índices DB para análisis de agrupamiento de todos los datos

| | iDB1 | iDB2 | iDB3 | iDB4 |
|---|---|---|---|---|
| k3 | 1.939967 | 1.939967 | 1.939967 | 1.939967 |
| k4 | 2.056500 | 2.056500 | 2.061791 | 2.056500 |
| k5 | 2.170303 | 2.170303 | 2.170303 | 2.260204 |
| k6 | 2.227933 | 2.226699 | 2.223303 | 2.223303 |
| k7 | 2.438157 | 2.437357 | 2.276649 | 2.275349 |

Tabla con los índices de Davies-Bouldin para vectores de características basados en la amplitud de frecuencia muestreada cada $0{,}5Hz$ para 4 simulaciones con valores de $k = 3$ hasta $k = 7$.

| | iDB1 | iDB2 | iDB3 | iDB4 |
|---|---|---|---|---|
| k3 | 1.616536 | 1.616536 | 1.616536 | 1.499896 |
| k4 | 1.630393 | 1.627846 | 1.627846 | 1.630393 |
| k5 | 1.651817 | 1.797262 | 1.797262 | 1.650442 |
| k6 | 1.758914 | 1.758914 | 1.758914 | 1.758914 |
| k7 | 1.930239 | 1.933508 | 1.907239 | 1.933508 |

Tabla con los índices de Davies-Bouldin para vectores de características basados en la máxima amplitud de frecuencia en intervalos de $0{,}5Hz$ para 4 simulaciones con valores de $k = 3$ hasta $k = 7$.

| | iDB1 | iDB2 | iDB3 | iDB4 |
|---|---|---|---|---|
| k3 | 1.779197 | 1.779197 | 1.779197 | 1.779197 |
| k4 | 1.839129 | 1.849345 | 1.850948 | 1.849345 |
| k5 | 2.042589 | 2.042589 | 2.042589 | 1.952297 |
| k6 | 2.041762 | 2.447832 | 2.041762 | 2.041762 |
| k7 | 2.217290 | 2.322387 | 2.040871 | 2.391402 |

Tabla con los índices de Davies-Bouldin para vectores de características basados en la máxima amplitud de frecuencia en intervalos de $0{,}5Hz$ y en la máxima amplitud de frecuencia en intervalos de $0{,}5Hz$ para 4 simulaciones con valores de $k = 3$ hasta $k = 7$.

# Apéndice 2: Tabla con SCR para analisis de arquetipos de todos los datos

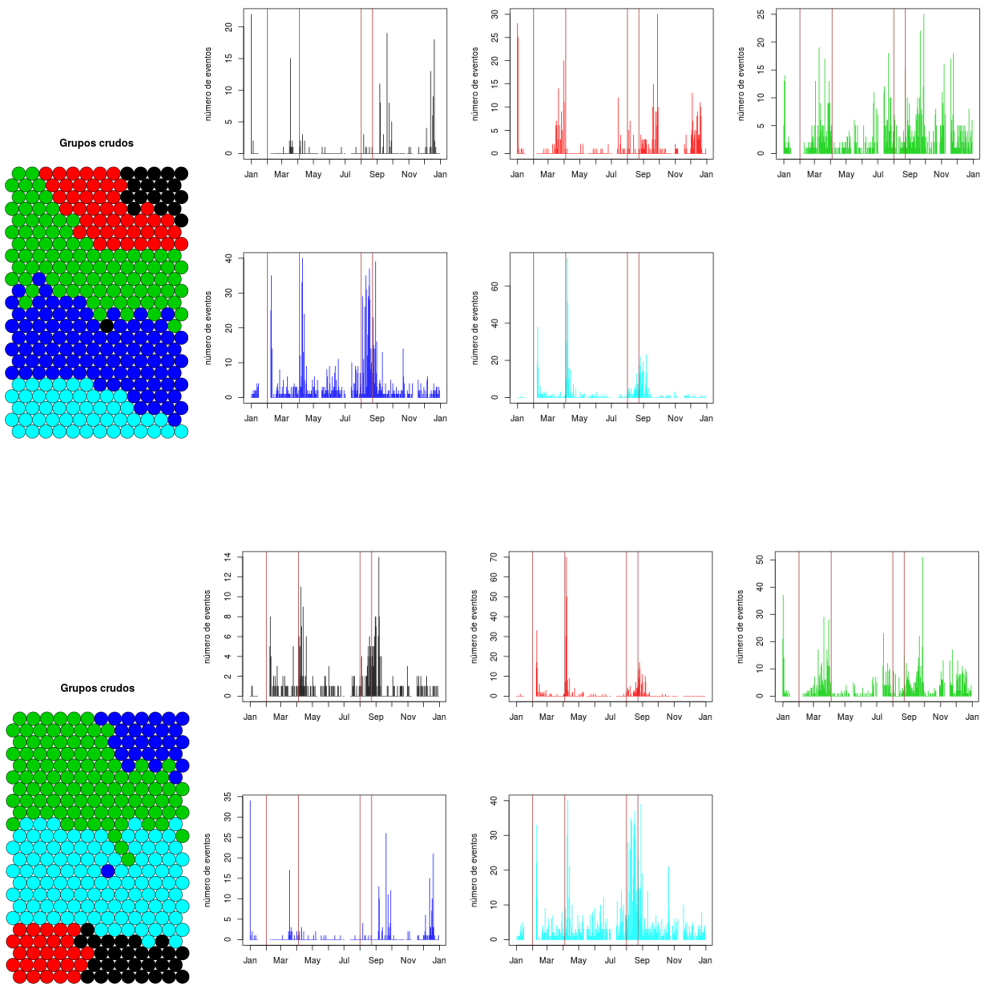
| | scr1 | scr2 | scr3 | scr4 |
|---|---|---|---|---|
| k1 | 0.09510244 | 0.09510244 | 0.09510244 | 0.09510244 |
| k2 | 0.05518031 | 0.05516121 | 0.05518209 | 0.05518280 |
| k3 | 0.04264067 | 0.04298428 | 0.04264080 | 0.04254997 |
| k4 | 0.03692312 | 0.03816260 | 0.03654216 | 0.03657518 |
| k5 | 0.03099661 | 0.03101886 | 0.03106109 | 0.03666452 |
| k6 | 0.03174809 | 0.03079558 | 0.02486849 | 0.02488492 |
| k7 | 0.02437169 | 0.02665264 | 0.02538899 | 0.02568506 |
| k8 | 0.02014054 | 0.02046331 | 0.02031996 | 0.02076130 |
| k9 | 0.01995673 | 0.02113800 | 0.02020376 | 0.01991826 |
| k10 | 0.02000044 | 0.02035581 | 0.02055920 | 0.02059065 |

Tabla con la suma de los cuadrados de los residuos (valor a minimizar) calculado para 4 aplicaciones distintas (distintas inicializaciones) del análisis de arquetipos utilizando distinto número de grupos (arquetipos) $k$ variando desde $k = 1$ hasta $k = 10$.

# Apéndice 3: Validación de los resultados

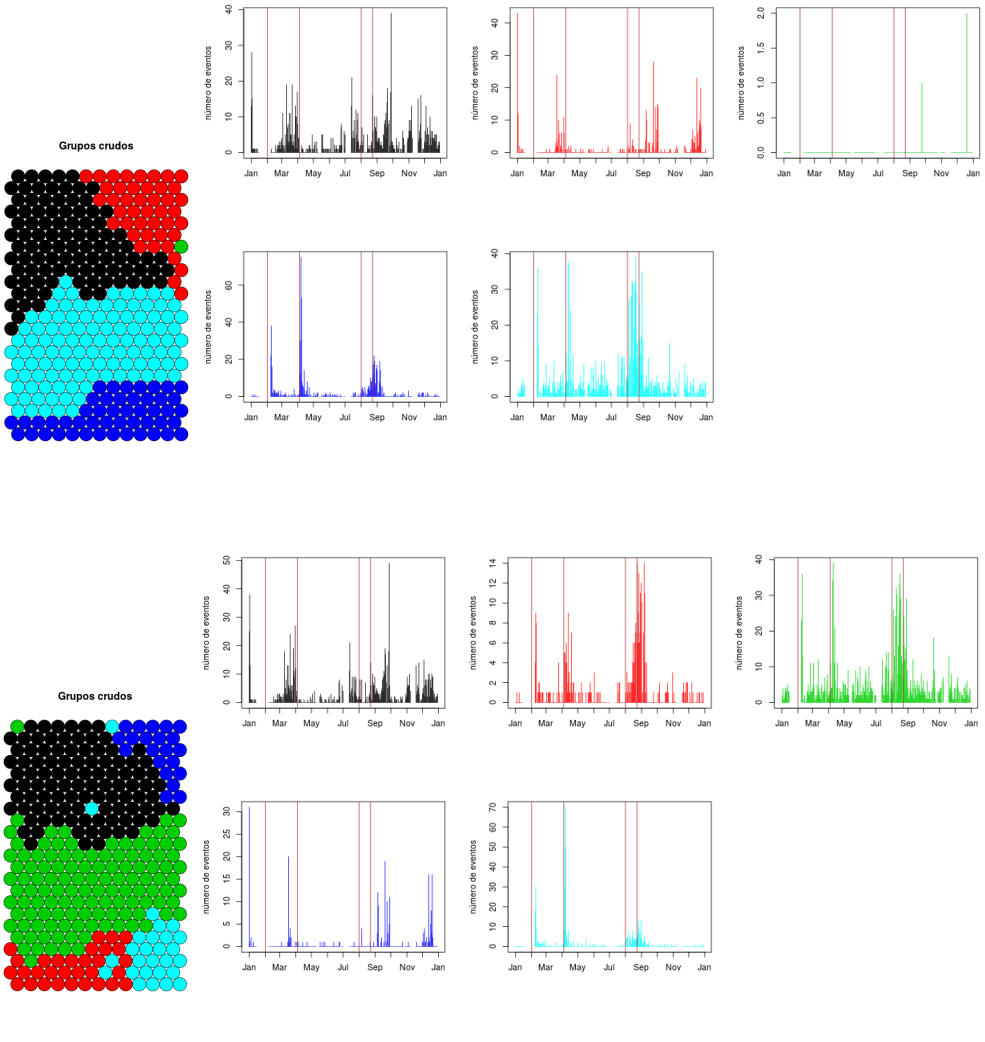
En este apéndice se presentan en sucesión los mapas, la distribución de familias en el tiempo y la tabla comparativa entre los grupos obtenidos por el algoritmo k-means y el análisis de arquetipos para validar la formación de los grupos. La codificación cada grupo por color en los siguientes resultados es: para el grupo 1: negro; grupo 2: rojo; grupo 3: verde; grupo 4: azul; y grupo 5: celeste. En la parte de arriba de cada imagen se presentan el mapa auto-organizado agrupado mediante k-means y la distribución temporal de cada grupo. En la parte de abajo de cada imagen se presenta el mapa auto-organizado y la distribución temporal de los grupos obtenido por medio de análisis de arquetipos. Después la tabla de contraste entre los grupos de los dos algoritmos para cada re-muestreo de los datos. El grado de coincidencia entre los grupos conocidos/bien definidos (ruido y explosiones/emisiones) se toma como el promedio del porcentaje de coincidencias de eventos de grupo en k-means respecto a la asignación de análisis de arquetipos y el porcentaje las coincidencias en el otro sentido (eventos de grupos obtenidos mediante

análsis de arquetipos respecto a grupos obtenidos por medio de k-means).



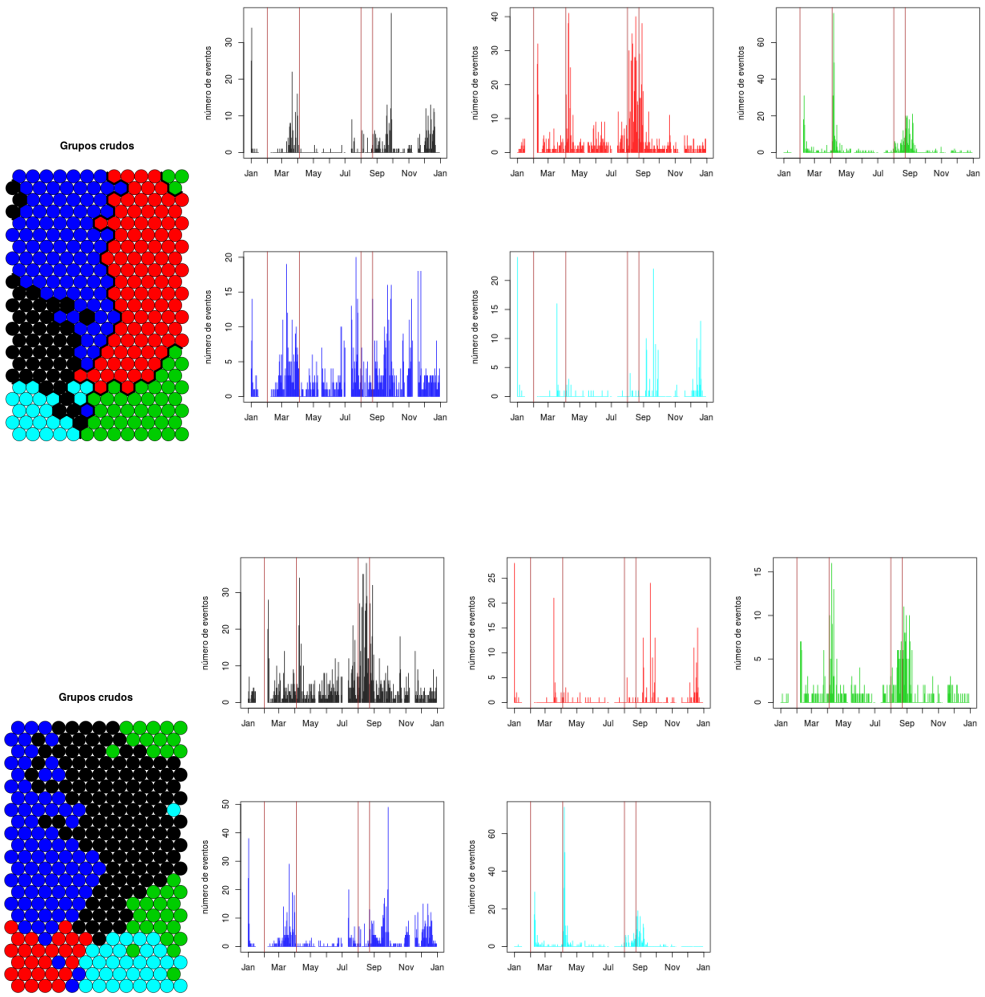

| | arch1 | arch2 | arch3 | arch4 | arch5 |
|---|---|---|---|---|---|
| kmeans1 | 0 | 0 | 0 | 186 | 0 |
| kmeans2 | 0 | 0 | 380 | 91 | 0 |
| kmeans3 | 0 | 0 | 733 | 0 | 288 |
| kmeans4 | 85 | 0 | 0 | 0 | 1152 |
| kmeans5 | 242 | 448 | 0 | 0 | 0 |

Resultados del primer re-muestreo y análisis del 90 % de los datos. Los datos asociados a ruido tuvieron un 83,57 % de coincidencias. Los datos asociados a explosiones/emisiones tuvieron una coincidencia del 94,52 %.



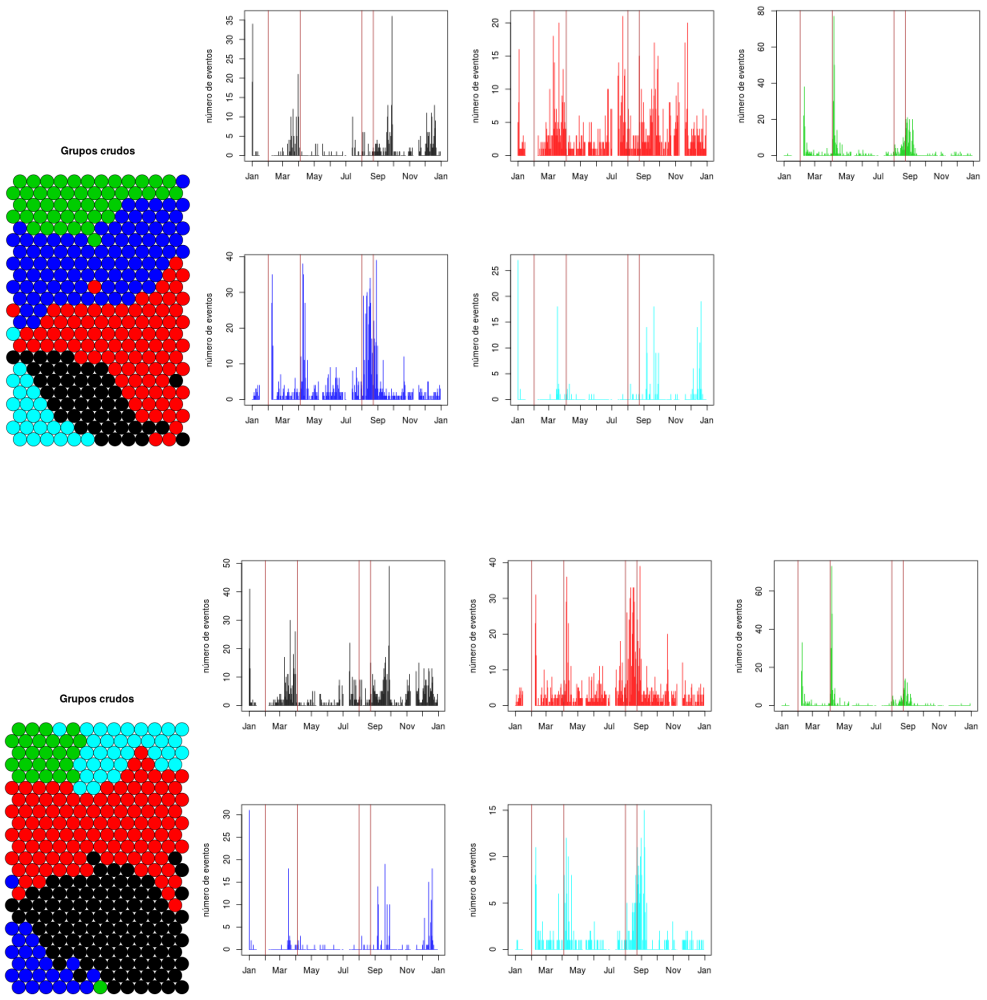

| | arch1 | arch2 | arch3 | arch4 | arch5 |
|---|---|---|---|---|---|
| kmeans1 | 877 | 0 | 186 | 0 | 4 |
| kmeans2 | 192 | 0 | 19 | 227 | 7 |
| kmeans3 | 0 | 0 | 0 | 3 | 0 |
| kmeans4 | 0 | 319 | 0 | 0 | 375 |
| kmeans5 | 0 | 39 | 1306 | 0 | 51 |

Resultados del segundo re-muestreo y análisis del 90 % de los datos. Los datos asociados a ruido tuvieron un 74,86 % de coincidencias. Los datos asociados a explosiones/emisiones tuvieron una coincidencia del 93,65 %.



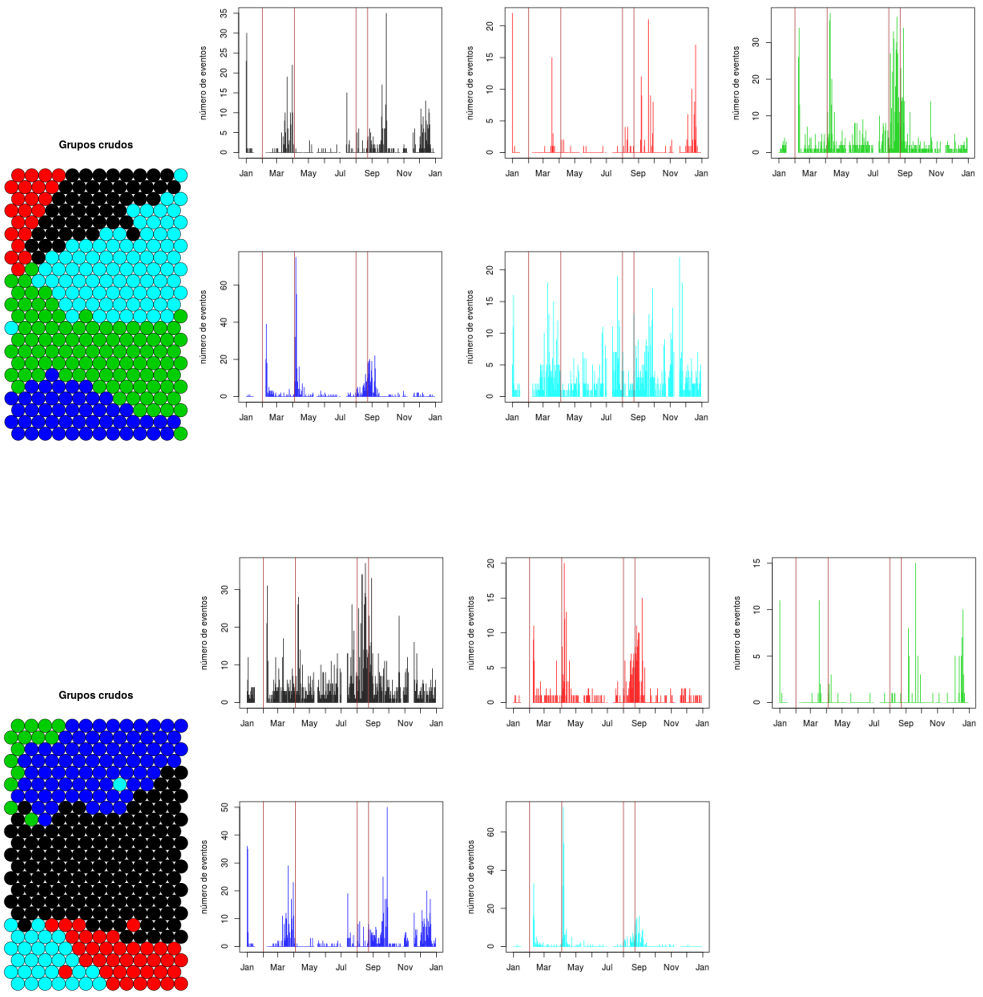

| | arch1 | arch2 | arch3 | arch4 | arch5 |
|---|---|---|---|---|---|
| kmeans1 | 0 | 47 | 0 | 485 | 0 |
| kmeans2 | 973 | 0 | 224 | 0 | 35 |
| kmeans3 | 0 | 0 | 172 | 0 | 474 |
| kmeans4 | 568 | 0 | 0 | 425 | 3 |
| kmeans5 | 0 | 189 | 0 | 10 | 0 |

Resultados del tercer re-muestreo y análisis del 90 % de los datos. Los datos asociados a ruido tuvieron un 87,53 % de coincidencias. Los datos asociados a explosiones/emisiones tuvieron una coincidencia del 85,57 %.



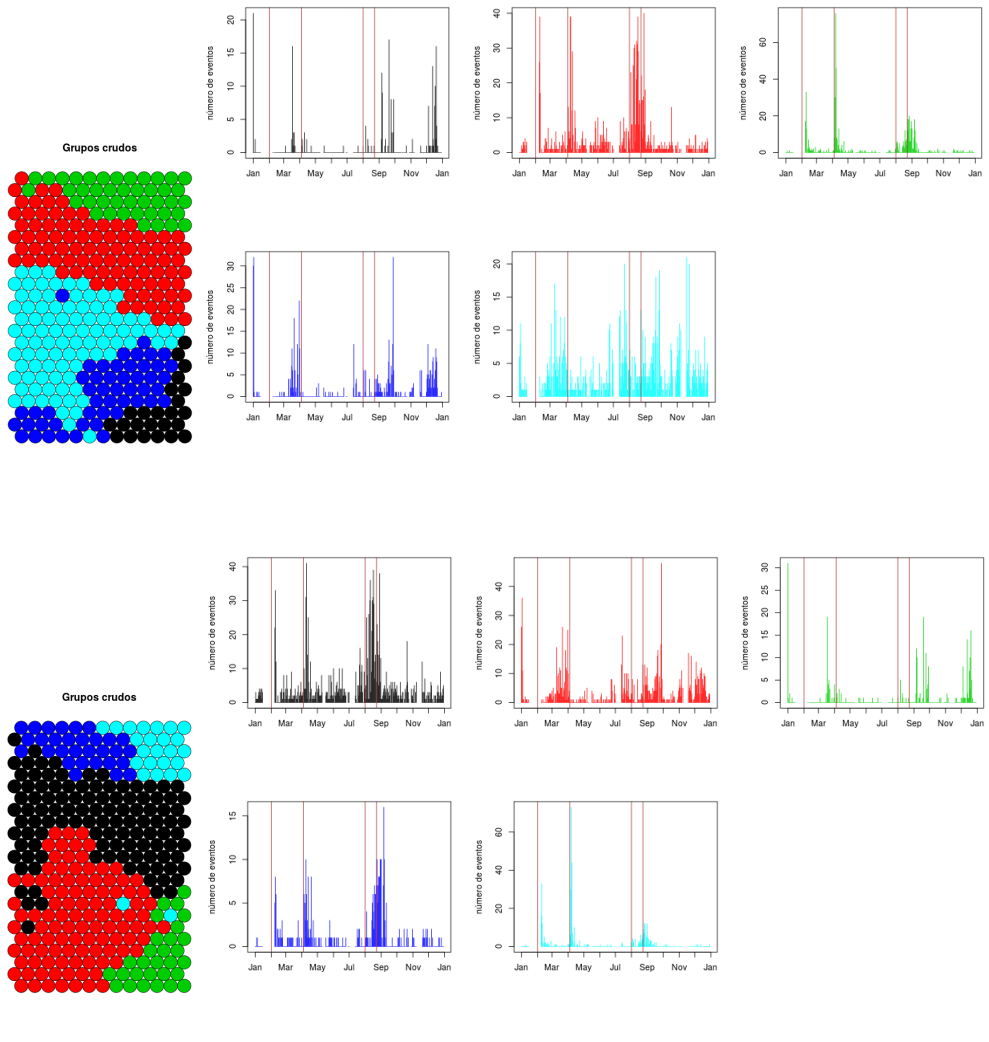

| | arch1 | arch2 | arch3 | arch4 | arch5 |
|---|---|---|---|---|---|
| kmeans1 | 473 | 0 | 6 | 20 | 0 |
| kmeans2 | 636 | 413 | 0 | 0 | 0 |
| kmeans3 | 0 | 0 | 419 | 0 | 264 |
| kmeans4 | 0 | 1034 | 5 | 0 | 111 |
| kmeans5 | 5 | 0 | 0 | 219 | 0 |

Resultados del cuarto re-muestreo y análisis del 90 % de los datos. Los datos asociados a ruido tuvieron un 94,70 % de coincidencias. Los datos asociados a explosiones/emisiones tuvieron una coincidencia del 92,42 %.



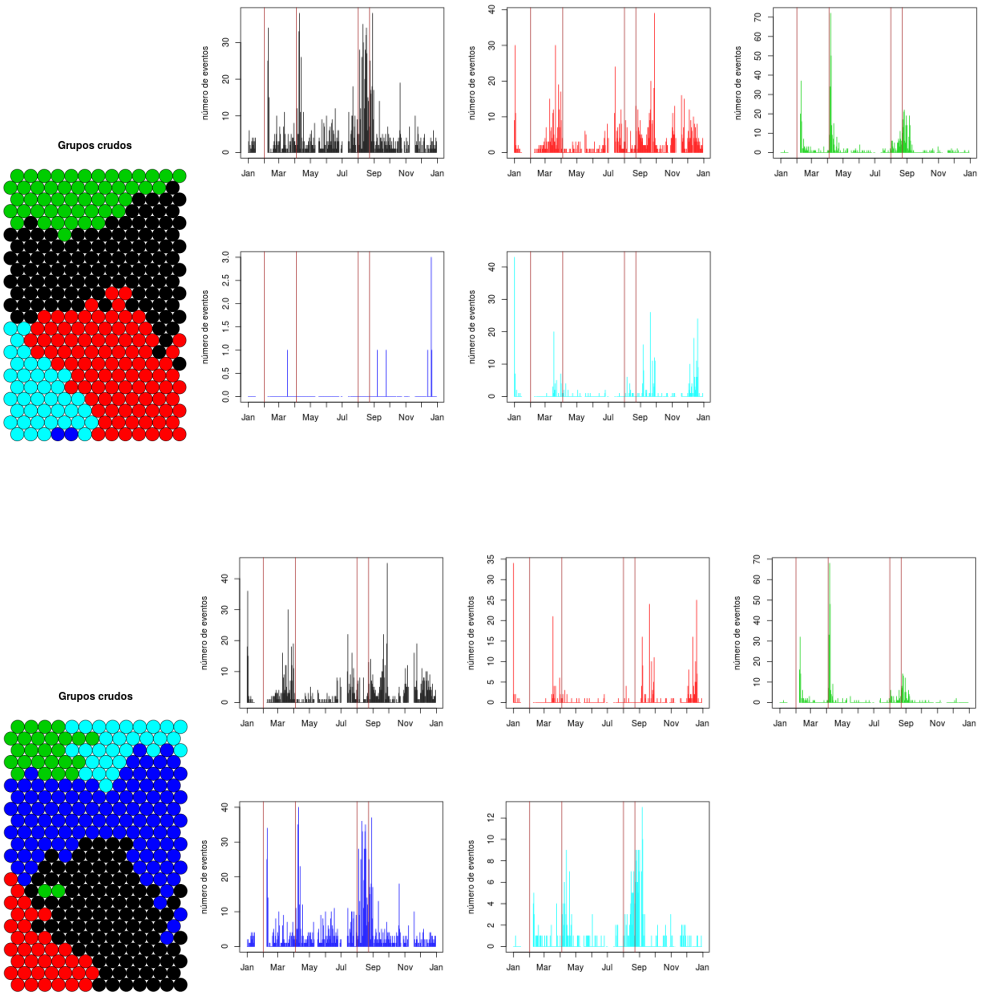

| | arch1 | arch2 | arch3 | arch4 | arch5 |
|---|---|---|---|---|---|
| kmeans1 | 0 | 0 | 0 | 533 | 0 |
| kmeans2 | 10 | 0 | 118 | 76 | 0 |
| kmeans3 | 975 | 152 | 3 | 0 | 27 |
| kmeans4 | 0 | 238 | 0 | 0 | 459 |
| kmeans5 | 760 | 0 | 0 | 250 | 4 |

Resultados del quinto re-muestreo y análisis del 90 % de los datos. Los datos asociados a ruido tuvieron un 77,68 % de coincidencias. Los datos asociados a explosiones/emisiones tuvieron una coincidencia del 89,60 %.



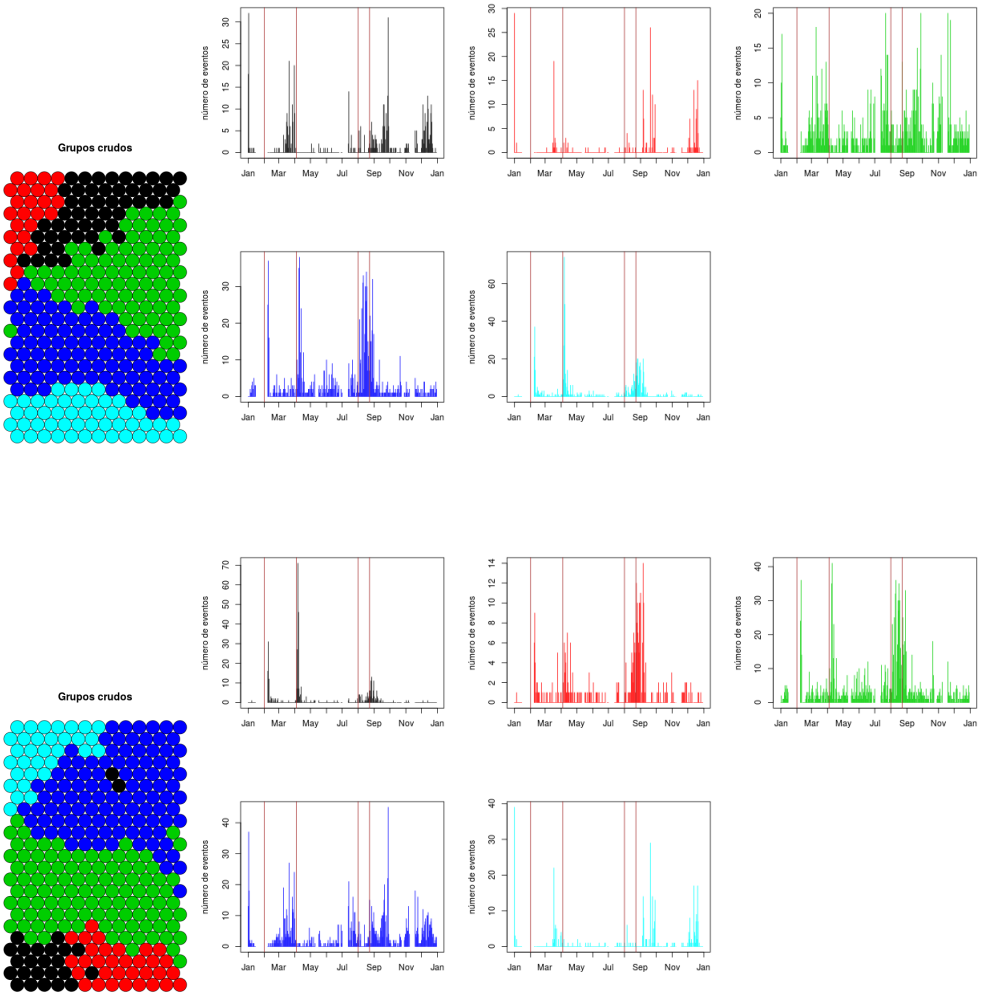

| | arch1 | arch2 | arch3 | arch4 | arch5 |
|---|---|---|---|---|---|
| kmeans1 | 0 | 0 | 210 | 0 | 0 |
| kmeans2 | 1089 | 0 | 0 | 132 | 0 |
| kmeans3 | 0 | 0 | 0 | 219 | 399 |
| kmeans4 | 0 | 451 | 50 | 0 | 17 |
| kmeans5 | 353 | 685 | 0 | 0 | 0 |

Resultados del sexto re-muestreo y análisis del 90 % de los datos. Los datos asociados a ruido tuvieron un 90,38 % de coincidencias. Los datos asociados a explosiones/emisiones tuvieron una coincidencia del 90,29 %.



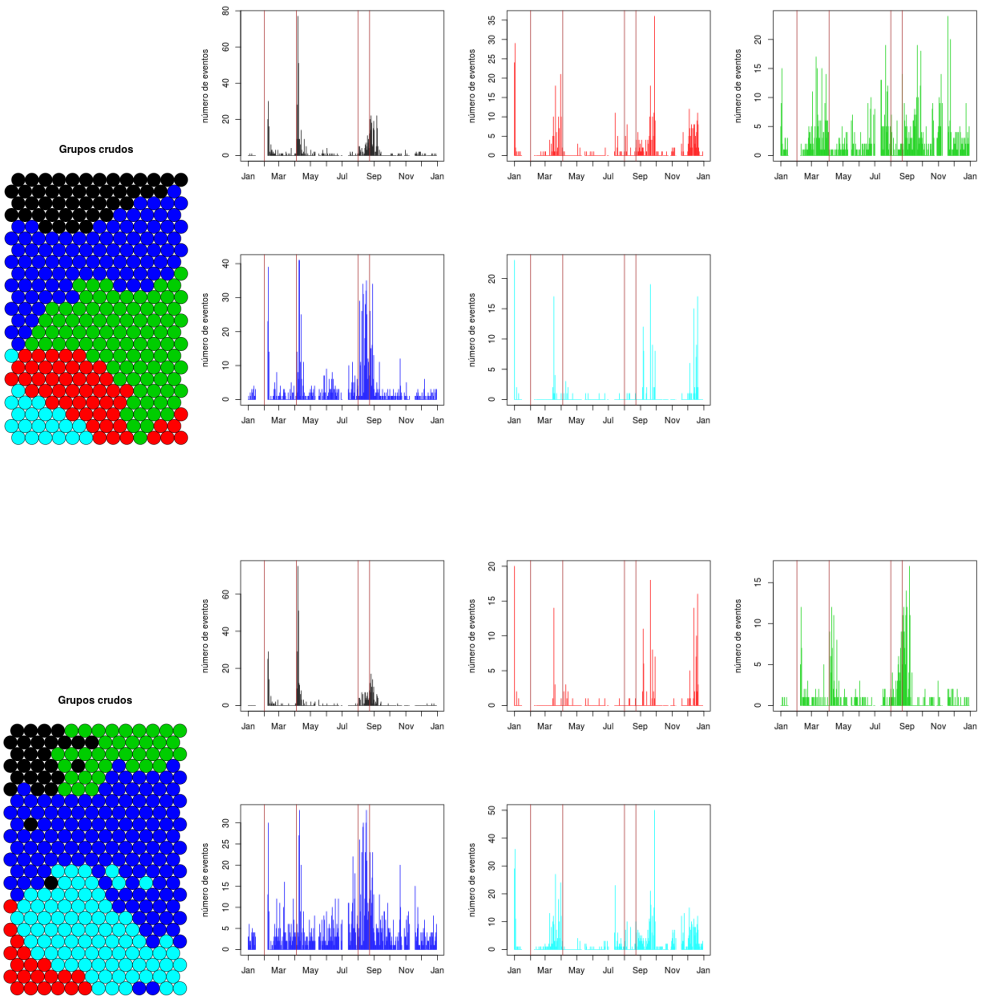

| | arch1 | arch2 | arch3 | arch4 | arch5 |
|---|---|---|---|---|---|
| kmeans1 | 95 | 0 | 0 | 1361 | 37 |
| kmeans2 | 981 | 0 | 16 | 31 | 0 |
| kmeans3 | 0 | 0 | 416 | 15 | 272 |
| kmeans4 | 0 | 8 | 0 | 0 | 0 |
| kmeans5 | 59 | 305 | 0 | 9 | 0 |

Resultados del séptimo re-muestreo y análisis del 90 % de los datos. Los datos asociados a ruido tuvieron un 89,61 % de coincidencias. Los datos asociados a explosiones/emisiones tuvieron una coincidencia del 95,35 %.



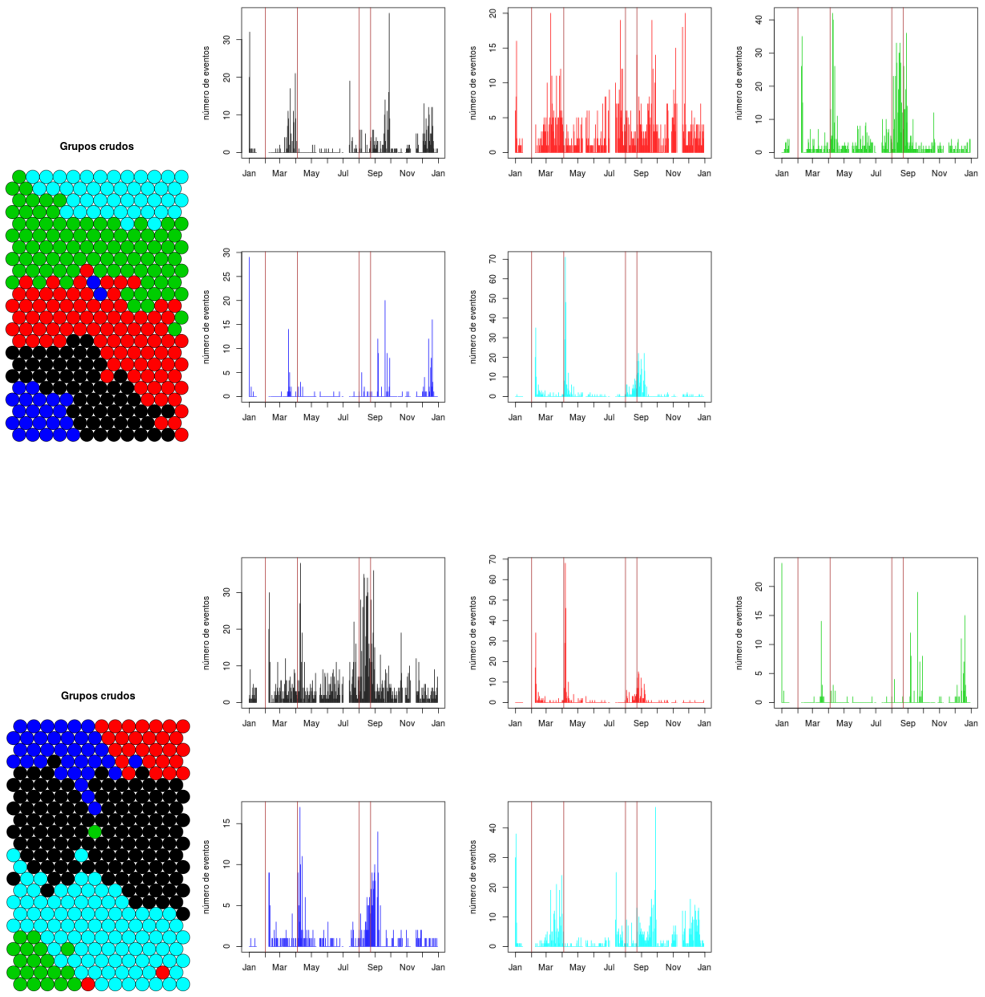

| | arch1 | arch2 | arch3 | arch4 | arch5 |
|---|---|---|---|---|---|
| kmeans1 | 10 | 0 | 0 | 422 | 93 |
| kmeans2 | 0 | 0 | 11 | 0 | 232 |
| kmeans3 | 0 | 0 | 269 | 739 | 0 |
| kmeans4 | 2 | 42 | 1109 | 0 | 0 |
| kmeans5 | 384 | 292 | 0 | 0 | 0 |

Resultados del octavo re-muestreo y análisis del 90 % de los datos. Los datos asociados a ruido tuvieron un 83,43 % de coincidencias. Los datos asociados a explosiones/emisiones tuvieron una coincidencia del 96,30 %.




| | arch1 | arch2 | arch3 | arch4 | arch5 |
|---|---|---|---|---|---|
| kmeans1 | 399 | 0 | 291 | 0 | 0 |
| kmeans2 | 0 | 7 | 0 | 3 | 523 |
| kmeans3 | 5 | 0 | 0 | 686 | 330 |
| kmeans4 | 98 | 0 | 119 | 942 | 0 |
| kmeans5 | 0 | 175 | 0 | 0 | 27 |

Resultados del noveno re-muestreo y análisis del 90 % de los datos. Los datos asociados a ruido tuvieron un 91,39 % de coincidencias. Los datos asociados a explosiones/emisiones tuvieron una coincidencia del 87,83 %.




| | arch1 | arch2 | arch3 | arch4 | arch5 |
|---|---|---|---|---|---|
| kmeans1 | 0 | 7 | 0 | 0 | 561 |
| kmeans2 | 639 | 4 | 0 | 0 | 398 |
| kmeans3 | 979 | 1 | 0 | 178 | 0 |
| kmeans4 | 7 | 0 | 183 | 0 | 25 |
| kmeans5 | 0 | 437 | 0 | 186 | 0 |

Resultados del décimo re-muestreo y análisis del 90 % de los datos. Los datos asociados a ruido tuvieron un 92,56 % de coincidencias. Los datos asociados a explosiones/emisiones tuvieron una coincidencia del 88,31 %.

Las coincidencias entre grupos de ruido puro entre ambos métodos fueron en promedio del 86,57 % con una desviación estándar de 6,54. Las coincidencias entre ambos métodos para los grupos de explosiones/emsiones (un grupo en k-means y dos grupos en análisis de arquetipos) fueron de 91,38 % en promedio con desviación estándar de 3,59. En la mayoría de aplicaciones de los métodos utilizados se obtiene que reaparecen grupos de datos durante épocas de alta actividad volcánica asociados a explosiones y emisiones. En el caso de análisis de arquetipos este grupo vuelve a dividirse en dos subgrupos. Además nuevamente se identifican grupos de datos que constan principalmente de ruido y que tienen altas coincidencias en los agrupamientos mediante k-means y análisis de arquetipos. Hay que mencionar que en el segundo y séptimo remuestreos k-means encontró prácticamente sólo 4 grupos (aunque se definió el número de grupos como 5) mientras que el análisis de arquetipos sí encontró 5 agrupamientos como los descritos anteriormente aunque esto no afectó profundamente las coincidencias de los grupos bien definidos. Estas tendencias se mantuvieron aún con variaciones altas en la topología de los grupos obtenidos en cada mapa.

# Referencias

[1] Mothes P. Actividad volcánica y pueblos precolombinos en el Ecuador, Editorial Abya Yala, 1998.

[2] Hall M., Robin C., Beate B., Mothes P. & Monzier M., Tungurahua Volcano, Ecuador: structure, eruptive history and hazards. Journal of Volcanology and Geothermal Research 91 1-21 (1999) DOI: 10.1016/S0377-0273(99)00047-5

[3] McNut S., Seismic monitoring and Eruption Forecasting of Volcanoes: A Review of the State-of-the-Art and Case Histories, Springer, 1996, pp. 99-146. DOI: 10.1007/978-3-642-80087-0_3

[4] Chouet B., Long-period volcano seismicity: its sources and use in eruption forecasting Nature, vol. 380, no. 6572,p 309-316 (1996) DOI: 10.1038/380309a0

[5] Lahr J., Chouet B., Stephens C., Power J. & Page R., Earthquake classification, location, and error analysis in a volcanic environment: implications for the magmatic system of the 1989-1990 eruptions at Redoubt volcano, Alaska. Journal of Volcanology and Geothermal Research, 62:137-151. (1994) DOI: 10.1016/0377-0273(94)90031-0

[6] Neuberg J., Luckett R., Baptie B. & Olsen K., Models of tremor and low-frequency earthquake swarms on Montserrat. Journal of Volcanology and Geothermal Research, 101:83-104. (2000) DOI: 10.1016/S0377-0273(00)00169-4

[7] Sigurdsson H., Houghton B., McNutt S., Rymer H. & Stix J., The Encyclopedia of Volcanoes. Academic Press, Elsevier, Segunda Edición, (2015) ISBN: 978-0-12-385938-9

[8] Chouet, B. A. & Matoza, R. S., A multi-decadal view of seismic methods for detecting precursors of magma movement and eruption. Journal of Volcanology and Geothermal Research, 252:108-175. (2013) DOI: 10.1016/j.jvolgeores.2012.11.013

[9] Orhnberger M. . Continuous automatic classification of seismic signals of volcanic origin at Mt . Merapi, Java, Indonesia. Tesis Doctoral, Mathematisch-Naturwissenschaftlichen Facultat der Universitat Potsdam, Alemania (2001)

[10] VOLUME Project Consortium, VOLUME Project - VOLcanoes: Understanding subsurface mass moveMEnt, VOLUME Project Consortium (2005) ISBN: 978-1-905254-39-2

[11] Lara R. A., Real-time Volcanic Monitoring Using Wireless Sensor Networks, Tesis Doctoral, Universidad Rey Juan Carlos, España 2015

[12] Bicego M., Acosta-Muñoz C. & Orozco-Alzate M., Classification of seismic volcanic signals using Hidden Markov Model based generative embeddings, IEEE, Transactions on Geoscience and Remote Sensing 51 (6) (2013) 3400-3409 DOI: 10.1109/TGRS.2012.2220370 (2012)

[13] Orozco-Alzate M, García M. E., Duin R. P. & C. G. Castellanos, Dissimilarity-based classification of seismic signals at Nevado del Ruiz Volcano, Earth Sciences Research Journal 10 (2) 57-66 (2006)

[14] Gutiérrez L. Sistema de Detección y Clasificación de señales sísmico-volcánicas utilizando Modelos Ocultos de Markov (HMMs): Aplicación a volcanes activos de Nicaragua e Italia, Tesis Doctoral, Universidad de Granada, España 2013

[15] Boué A., Data mining and volcanic eruption forecasting, Tesis Doctoral, Universidad de Grenoble, Francia 2015

[16] Benítez C., Ramírez J., Segura J. C., Rubio A., Ibáñez J. M., Almendros J. & García-Yeguas A., Continuous HMM-based volcano monitoring at Deception Island, Antarctica, IEEE, Transactions on Geoscience and Remote Sensing 45 (1) (2007) 138-146 DOI: 10.1109/TGRS.2006.882264 (2007)

[17] Langer H. & Falsaperla S. Long-term observation of volcanic tremor on Stromboli volcano (Italy): A synopsis, Pure and Applied Geophysics Volume 147, Issue 1, pp 57-82 DOI: 10.1007/BF00876436 (1996)

[18] Esposito A., Giudicepietro F., Scarpetta S., D'Auria L., Marinaro M. & Martini M., Automatic Discrimination among Landslide, Explosion-Quake, and Microtremor Seismic Signals at Stromboli Volcano Using Neural Networks, Bulletin of the Seismological Society of America, Vol. 96, No. 4A, pp. 1230-1240, DOI: 10.1785/0120050097 (2006)

[19] Esposito A., Giudicepietro F., D'Auria L., Scarpetta S., Martini M., Coltelli M. & Marinaro M. Unsupervised Neural Analysis of Very-Long-Period Events at Stromboli Volcano Using the Self-Organizing Maps, Bulletin of the Seismological Society of America, Vol. 98, No. 5, pp. 2449-2459, DOI: 10.1785/0120070110 (2008)

[20] Messina A. & Langer H., Pattern recognition of volcanic tremor data on Mt. Etna (Italy) with KKAnalysis-A software program for unsupervised classification, Elsevier Computers & Geosciences 37 953-961 DOI: 10.1016/j.cageo.2011.03.015 (2011)

[21] Langer H., Falsaperla S., Masotti M., Campanini R., Spampinato S. & Messina A., Synopsis of supervised and unsupervised pattern classification techniques applied to volcanic tremor data at Mt Etna, Italy, Geophysical Journal International 178, 1132-1144 DOI: 10.1111/j.1365-246X.2009.04179.x (2009)

[22] Esposito A., D'Auria L., Giudicepietro F., Caputo T. & Martini M. Neural analysis of seismic data: applications to the monitoring of Mt. Vesuvius, Annals of Geophysics, 56, 4, S0446, DOI: 10.4401/ag-6452 (2013)

[23] Köhler A., Ohrnberger M. & Scherbaum, F., Unsupervised pattern recognition in continuous seismic wavefield records using Self-Organizing Maps, Geophysycal Journal International, 2010, 182: 1619-1630, DOI: 10.1111/j.1365-246X.2010.04709.x (2010)

[24] Carniel R., Barbui L. & Jolly A. D., Detecting dynamical regimes by Self-Organizing Map (SOM) analysis: an example from the March 2006 phreatic eruption at Raoul Island, New Zealand Kermadec Arc Bollettino di Geofisica Teorica ed Applicata Vol. 54, n. 1, pp. 39-52 DOI: 10.4430/bgta0077 (2013)

[25] Hastie T., Tibshirani R. & Friedman J., The Elements of Statistical Learning Data Mining, Inference, and Prediction Springer, Segunda Edición, Febrero 2009

[26] Beyreuther M., Barsch R., Krischer L., Megies T., Behr Y. & Wassermann J. ObsPy: A Python Toolbox for Seismology Seismological Research Letters, 81 (3), 530-533. (2010) DOI: 10.1785/gssrl.81.3.530

[27] Allen R., Automatic earthquake recognition and timing from single traces. Bulletin of the Seismological Society of America, vol 68, pp. 1521-1532. (1978)

[28] Withers M., Aster R., Young C., Beiriger J., Harris M., Moore S. & Trujillo J. 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 (1998)

[29] Köhler A., Ohrnberger M, Riggelsen C. & Scherbaum F., Unsupervised Feature Selection for Pattern Search in Seismic Time Series, J. Mach. Learn. Res., Workshop and Conference Proceedings 4, 106-121 (2008)

[30] Davies D. & Bouldin D., A Cluster Separation Measure IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-1(2), 224 - 227 (1979) DOI: 10.1109/TPAMI.1979.4766909

[31] Cutler A. & Breiman L., Archetypal Analysis Technometrics, Vol. 36, No. 4 pp. 338-347 (1994)

[32] Eugster M. & Leisch F., From Spider-Man to Hero - Archetypal Analysis in R, Journal of Statistical Software, Vol 30, Issue 8 (2009) DOI: 10.18637/jss.v030.i08

[33] Kohonen T., Self-Organizing Maps, Tercera Edición. Springer. 2001 ISBN: 978-3-540-67921-9